COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales ICADE

Análisis de los viajes del servicio de bikesharing “Bluebikes” en Boston: Un caso de estudio

Clave: 201907760

MADRID | ABRIL 2024

RESUMEN EJECUTIVO

Este caso de estudio presenta un análisis detallado del uso del sistema de bicicletas compartidas en la ciudad de Boston durante el año 2021, para comprender cómo ha impactado el contexto post-pandémico en el uso de este sistema de movilidad urbana. Se ha examinado la información de los viajes registrados por la empresa Blue Bikes, así como la influencia de distintas variables en los patrones de uso de este medio de transporte.

Se ha realizado un análisis exploratorio de datos (EDA) y se realizaron tareas de extracción, tratamiento y limpieza de datos. Además, se emplearon modelos predictivos, como el bosque aleatorio "random forest", para investigar la duración de los viajes y la probabilidad de que un viaje sea corto o largo. Estos resultados ofrecen una visión integral del comportamiento de los usuarios del sistema de bicicletas compartidas tras el COVID-19 y pueden servir como base para futuras estrategias de movilidad urbana en la ciudad.

Palabras clave: bicicletas compartidas, Boston, Blue Bikes, post-pandemia

ABSTRACT

This case study presents a detailed analysis of the use of the bike-sharing system in the city of Boston during the year 2021, in order to understand how the post-pandemic context has impacted the use of this urban mobility system. The information on trips recorded by the company Blue Bikes has been examined, as well as the influence of different variables on the patterns of use of this means of transportation.

An exploratory data analysis (EDA) was carried out and data extraction, processing and cleaning tasks were performed. In addition, predictive models, such as the random forest, were employed to investigate trip length and the probability of a trip being short or long. These results provide a comprehensive view of user behavior of the bike sharing system after COVID-19 and can serve as a basis for future urban mobility strategies in the city.

Key words: bikesharing, Boston, Blue Bikes, post-pandemic

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. Introducción

1.1 Introducción del tema

Actualmente, la sociedad se enfrenta a desafíos ambientales que impulsan la búsqueda de soluciones sostenibles. En este contexto, ha surgido un aumento significativo en la popularidad de los sistemas de bicicletas compartidas.

Como consecuencia del reconocimiento de la significante contribución del sector del transporte al cambio climático, organismos internacionales como Las Naciones Unidas han establecido los Objetivos de Desarrollo Sostenible (ODS) en la Agenda 2030 con el objetivo de paliar los efectos contaminantes del transporte y buscar alternativas más sostenibles. Esta creciente preocupación ha llevado a los diversos organismos a cuestionarse acerca de las posibilidades de hacer más sostenible el transporte actual (Pérez-Morales, Gil-Guirado & Maqueda-Belmonte, 2022).

Cada año, son más los vehículos circulando por las carreteras dando lugar a una cantidad de tráfico significante en las intersecciones e importantes carreteras. Consecuentemente, es mayor el número de gases de efecto invernadero que se emiten a la atmosfera y, por ende, mayor la contaminación. La movilidad como servicio (MaaS) ha ganado popularidad en los últimos años como solución posible a este problema. Esta tendencia se debe a una serie de beneficios que ofrece, incluyendo la reducción del número de accidentes de tráfico, la disminución de la contaminación atmosférica y la descongestión vial. Además, el concepto de Movilidad como Servicio (MaaS) incluye una variedad de medios de transporte que brindan flexibilidad, personalización, precios competitivos y una experiencia de viaje sin interrupciones. La idea del MaaS es que el proveedor del servicio ofrece al individuo una variedad de modos de transporte a través de una única interfaz, adaptándose a sus necesidades y considerando aspectos como su edad (Ayuba, Supangkat & Wibowo, 2023).

Los sistemas de transporte compartido tienen larga historia. El primer sistema de coches compartido empezó en 1.948 en Zurich y el primer sistema de bicicletas compartidas empezó en 1.965 en Amsterdam. Los avances en las Tecnologías de la Información y la Comunicación (TIC) han posibilitado ofrecer flexibilidad y comodidad a los usuarios al poner este tipo de servicio en sus manos, a través de aplicaciones móviles o plataformas digitales. El sistema de MaaS permite a los clientes ahorrar costes y ha

reducido, como he mencionado previamente, la necesidad de adquirir un vehículo propio (Narayanan & Antoniou, 2023).

Cada vez más jóvenes están optando por servicios de transporte compartido en lugar de adquirir un vehículo privado. Entre los ejemplos de servicios de transporte compartido, encontramos servicios de ridesharing como BlaBlaCar, de carsharing como WiBLE o de bike sharing como Blue Bikes (Eckhardt, Aapaoja, Nykänen, Sochor, Karlsson & König, 2018).

1.2 Justificación de interés del problema

En la página oficial de Blue Bikes, se ofrece un análisis detallado del uso de bicicletas compartidas durante el año 2020, con el propósito de analizar los efectos de la pandemia en este medio de transporte. Sin embargo, se ha notado una brecha en la investigación respecto al impacto post-pandemia en los sistemas de bicicletas compartidas.

Dada esta falta de investigación, hemos decidido enfocar este caso de estudio en el análisis del servicio durante el año 2021. Este enfoque nos permite evaluar cómo ha evolucionado la demanda de bicicletas compartidas después del período crítico de la pandemia. Además, mediante este análisis, se podrán identificar posibles tendencias emergentes y los cambios que ha provocado el COVID-19 en los patrones de uso. Esta información proporcionará una comprensión de cómo los usuarios se han adaptado a las nuevas condiciones socioeconómicas y de salud pública después de la pandemia.

La elección de investigar el sistema de bicicletas compartidas en Boston durante el año 2021 se justifica por varias razones. En primer lugar, en esta ciudad, el sistema de bicicletas compartidas ha ganado gran popularidad como una alternativa de transporte sostenible y cómoda, gracias a la implementación de iniciativas destinadas a promover su uso. Además, el año 2021 marcó un punto de inflexión tras el impacto de la pandemia, lo que lo convierte en un período relevante para analizar la recuperación y adaptación de los hábitos de movilidad urbana. El análisis de este período es importante para estudiar cómo los cambios generados por el COVID-19 podrían influir en el futuro de la movilidad urbana. Es una oportunidad para entender mejor los nuevos hábitos de movilidad que surgieron después de la pandemia.

1.3 Objetivos del trabajo & preguntas de investigación

Este Trabajo de Fin de Grado tiene como objetivo principal realizar un análisis exhaustivo del servicio de bikesharing proporcionado por la compañía Bluebikes en Boston durante el año 2021. Se pretende obtener una comprensión detallada de su funcionamiento y desempeño, con el fin de extraer información relevante que contribuya a una mejor comprensión de los patrones de uso del servicio y cómo las diferentes variables se relacionan entre sí.

Las preguntas de investigación han sido las siguientes:

1. ¿Qué bicicletas tienen que ir a mantenimiento?
2. ¿Existe relación entre la distancia entre estaciones de Blue Bikes en Boston y el uso del servicio por parte de los usuarios?
3. ¿Cómo afecta la hora del día, la latitud de origen y la latitud de destino a la duración de los viajes en bicicleta compartida?
4. ¿Cómo influyen la hora del día y la ubicación en la probabilidad de que un viaje en bicicleta compartida sea corto (menos de 29 minutos) o largo (más de 29 minutos)?

2. Marco Teórico

2.1. Historia del Bike Sharing

Hace unos años, la bicicleta no se contemplaba como una opción de transporte público. Sin embargo, con el transcurso del tiempo, este medio de transporte ha demostrado ser muy cómodo para su uso en áreas urbanas. Además, resulta una opción de movilidad atractiva por su bajo coste de adquisición y mantenimiento en comparación con otros medios de transporte (DeMaio, 2003).

Respecto a su historia, podemos establecer que existen cuatro generaciones de sistemas de bicicletas compartidas. La primera generación tuvo su origen en Ámsterdam en 1965 con las denominadas “White Bikes”, unas bicicletas ordinarias blancas que se ofrecieron abiertamente al público (DeMaio, 2003). Sin embargo, este sistema resultó en un fracaso a los pocos días al verse colapsado el sistema como consecuencia del uso inapropiado de las bicicletas por parte de los usuarios, incluyendo actos como arrojarlas a canales o apropiárselas para uso personal (DeMaio, 2009).

Más adelante, en 1995, un segundo intento tuvo lugar en Copenhague con el lanzamiento del programa "City Bike" que permitía a los usuarios el uso público de este vehículo por el casco urbano de la ciudad a cambio del depósito de unas monedas (DeMaio, 2003). Estas bicicletas de segunda generación presentaban mejoras significativas frente a las anteriores, como neumáticos de caucho macizo y ruedas con placas publicitarias, diseñadas para un uso intensivo y práctico. Sin embargo, el robo de estas bicicletas por el anonimato del usuario seguía impidiendo la consolidación de este sistema, lo que dio lugar al lanzamiento de la tercera generación del sistema de bicicletas compartidas (DeMaio, 2009).

El primer programa de esta generación, "Bikeabout", nació en 1996 en la universidad de Portsmouth (Inglaterra). Para usar este vehículo, los universitarios hacían uso de una tarjeta de banda magnética. Estas bicicletas, conocidas como "Smart Bikes", integraron innovaciones tecnológicas para facilitar su monitoreo como sistemas de telecomunicación, tarjetas inteligentes, candados electrónicos o sistemas informáticos (DeMaio, 2009).

En los años posteriores, se lanzaron anualmente uno o dos programas nuevos como Vélo à la Carte en Rennes en 1998 y Call a Bike en Múnich en 2000. Sin embargo, el programa más destacado de esta tercera generación de bicicletas compartidas fue Velo'v, con 1.500 bicicletas en Lyon lanzado por JCDecaux en 2005. Le siguió en 2007 el sistema de bicicletas compartidas Vélib' de París, con unas 7.000 bicicletas, que desde entonces se ha expandido a 23.600 bicicletas. A finales de 2007, había alrededor de 60 programas de tercera generación en todo el mundo (DeMaio, 2009).

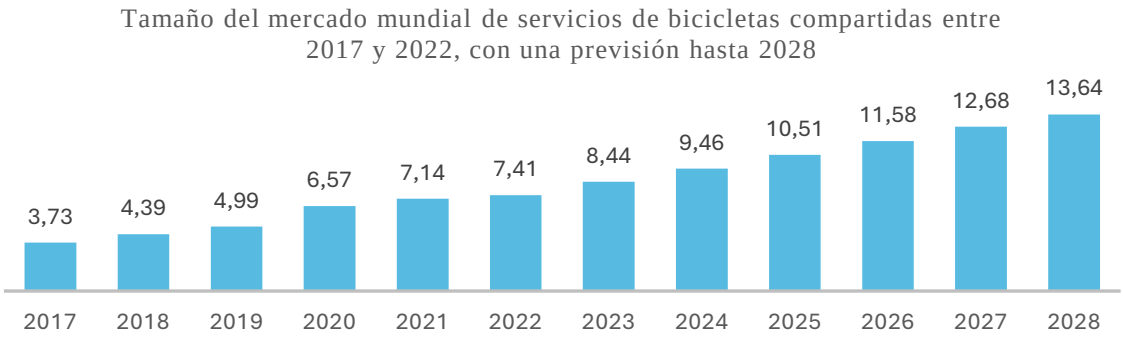
Por último, la cuarta generación de sistemas de uso compartido de bicicletas incluye sistemas inteligentes con el objetivo ser sostenibles, eficientes y proporcionar un servicio de calidad (Eren & Uz, 2020). Estas bicicletas incluyen bloqueo automatizado y sistema de electrificación. Además, en las estaciones de bicicletas se han incluido paneles solares para dar energía a las bicicletas. Por último, algunas empresas de bicicletas compartidas están incorporando un candado automático que permite al usuario asegurar su bicicleta a cualquier objeto fijo al final de su trayecto (Moon, Sharpin, De La Lanza, Khan, Re & Maassen, 2019). A continuación, se presenta una tabla resumen que detalla las características de las cuatro generaciones de bicicletas:

Tabla 1: Generaciones Bike Sharing

	1º Generación (mediados de los 60)	2º Generación (principio de los 90)	3º Generación (finales de los 90)	4º Generación (2010s)
Descripción	· Bicicletas ordinarias blancas · Acceso libre al público	· Bicicletas con neumáticos de caucho macizo · Depósito de monedas para acceso · Estaciones de acoplamiento	· Bicicletas inteligentes · Tarjetas de banda magnética · Sistema IT con estaciones de acoplamiento ·Aplicación Móvil	· Bicicletas inteligentes con bloqueo automatizado y electrificación · Sistema IT sin estaciones de acoplamiento · Aplicación Móvil
Características	· Servicio gratuito · Usuarios anónimos · Bicicletas desbloqueadas	· Servicio Gratuito · Usuarios anónimos · Bicicletas bloqueadas en una estación de acoplamiento específica · Monedas como depósito utilizado para desbloquear las bicicletas y estas se recuperan al devolver las bicicletas	· Método de pago de suscripción o uso esporádico · Verificación del usuario con identidad real · Bicicletas bloqueadas en una estación de acoplamiento específica ·Tecnología	· Pago de cuota para hacerse socio · Verificación de usuarios con identidad real · Cobro en función del tiempo de uso en la app móvil o suscripción estacional · Bicicletas bloqueadas
Línea de tiempo	Origen en 1.965 en Amsterdam, Países Bajos con las denominadas "White Bikes"	Origen en 1.995 en Copenhague, Dinamarca con el lanzamiento del programa "City Bike"	· 1996: Bikeabout en la Universidad de Portsmouth (Inglaterra) · 1998: Vélo à la Carte en Rennes · 2000: Call a Bike en Múnich · 2005: Velo'v en Lyon (1,500 bicicletas) ·2007: Vélib' en París (7,000 bicicletas, ahora 23,600 bicicletas)	Se implantó por primera vez a gran escala en Pekín (China) y posteriormente se extendió a Asia, Europa y América
Principal recurso financiero	Inversión Pública	Inversión Pública	Inversión Pública	Financiación de capital riesgo

Fuente:(Chen, Van Lierop & Ettema, 2020)

Figura 1: Tamaño mercado mundial bike sharing



Fuente: Statista

Como podemos observar en el gráfico, el mercado de bikesharing presenta un crecimiento constante y continuo a lo largo de los años. Se prevé que, en 2028, alcance un valor de 13,64 billones de dólares estadounidenses. Además, destaca que la pandemia no impidió el crecimiento de este servicio entre 2019 y 2020. Respecto a 2021, se puede concluir que el mercado mundial de servicios de bicicletas compartidas superó incluso los niveles prepandémicos.

2.2 Smart Bikes

Las Smart Bikes permiten a sus usuarios satisfacer sus necesidades de desplazamiento de forma sostenible. Para acceder a una *Smart bike*, el cliente tiene que hacer uso de una tarjeta inteligente, de forma que, si el usuario no devuelve la bicicleta después de su uso, se le cargará el coste de sustitución al ser conocida la identidad del usuario. Estas bicicletas se suelen encontrar en estaciones con bastidores de bloqueo electrónico diseñados para ellas específicamente. Existen dos tecnologías de cierre para estas; la primera requiere el uso de la tarjeta inteligente o una tarjeta de banda magnética para desbloquearla; la segunda, exige al usuario el uso de su móvil para la obtención de un código numérico que desbloquee el candado automatizado colocado en la bicicleta (DeMaio 2003, 2004).

La idea es que el usuario lleve la bicicleta hasta su destino y pueda dejarla aparcada en una estación cercana. En cada estación, suele haber una pantalla que indica el área donde deben circular las bicis sin salirse de ella y suele indicarte también estaciones cercanas al destino donde tienes pensado ir. El sistema no se hace responsable de tu seguridad, haciéndote firmar digitalmente un documento antes del uso de sus bicis por el

que accedes a no hacerles responsables de ningún accidente en el que puedas incurrir (DeMaio 2003, 2004).

Las *Smart bikes* están pensadas para viajes cortos por las zonas urbanas de las principales ciudades. Este sistema de transporte funciona bajo demanda y permite acceder a destinos no accesibles por otros medios de transporte. Al contar con una infraestructura necesaria menor para operar, resulta una opción barata para viajar. Las estaciones con *smart bikes* suelen estar ubicadas en zonas de mucho tráfico y cercanas a estaciones de bus, tren u otras alternativas pues este programa está pensado como sistema complementario de otros modos de transporte público, incrementando por ende su uso. Este vehículo tiene beneficios para la salud de los usuarios pues mediante su uso los usuarios hacen ejercicio. Además, tiene un impacto positivo para el medioambiente, pues reduce el tráfico en las ciudades y reduce la contaminación (DeMaio, 2003).

Pese a los beneficios recién mencionados, estas bicicletas cuentan con algunos inconvenientes frente a otros modos de transporte, como pueden ser la inaccesibilidad a estas por parte de personas discapacitadas o su dificultad de uso en determinadas localizaciones. Además, este vehículo requiere que el usuario tenga conocimiento acerca de cómo usarlas, existiendo el riesgo de que sean peligrosas para otros ciclistas o peatones si se conducen incorrectamente. Y, por último, cuentan con algunos límites como su uso exclusivo para distancias de viaje reducidas y solo en situaciones meteorológicas adecuadas (DeMaio, 2004).

Es esencial crear conciencia entre los usuarios sobre el respeto mutuo necesario entre ciclistas, peatones y conductores de automóviles, especialmente en áreas donde el uso de la bicicleta sea menos común. Para asegurar el éxito de este servicio en cualquier ciudad, se requiere la disponibilidad de una extensa y continua red de carriles bici o de carriles sin coches, así como la combinación de una topografía y un clima favorables para su uso (Midgley, 2009).

2.3 Factores que afectan a la demanda de las bicicletas compartidas

La demanda de bicicletas compartidas se ha visto afectada por diversos factores. El consumo de combustibles fósiles causado por el tráfico ha provocado un aumento en las emisiones de gases de efecto invernadero. El uso de vehículos de motor es responsable del 40% de las emisiones totales de gases de efecto invernadero en Europa, y del 20% en Estados Unidos. Este factor ha llevado a un aumento en la demanda de bicicletas

compartidas como medio de transporte sostenible para reducir la contaminación provocada por los vehículos de motor. Además de los beneficios que genera en el medio ambiente, BSPs ('Bike Sharing Programs') contribuye a la mejora de la salud de los usuarios al incrementar su actividad física, así como de suponer un medio de transporte más económico (Eren & Uz, 2020).

El impacto del tiempo en la demanda de las bicis compartidas se ha analizado ampliamente, reflejando una correlación positiva cuando la temperatura ronda entre los 0 y 20 ºC. Sin embargo, cuando la temperatura supera los 30º, los investigadores presentan resultados contradictorios. Cuando se dan condiciones meteorológicas desfavorables, lluvia, nieve o bajas temperaturas, los investigadores están de acuerdo en que la demanda de las bicis compartidas disminuye. La velocidad del viento y la humedad también han demostrado impactar negativamente al uso compartido de bicicletas.

Respecto al entorno construido y uso del suelo, la demanda de las bicis compartidas disminuye en terrenos con pendientes ascendentes. Por otro lado, la proximidad a zonas verdes, locales comerciales, instalaciones deportivas, universidades y otras áreas con gran afluencia de personas, influye positivamente en la demanda de BSPs.

Respecto al transporte público, los BSP pueden formar parte de la red de transporte público como medio complementario o alternativa de transporte en aquellas áreas con sistemas de transporte público deficientes.

Respecto a aspectos demográficos, en la mayoría de los países occidentales, el perfil de usuario mayoritario de las BSP es masculino, joven, educado, trabajador, con ingresos elevados. Sin embargo, en los países orientales, las bicis compartidas son usadas principalmente por jóvenes y personas de edad media con pocos recursos económicos a pesar de tener trabajo. Por otro lado, artículos académicos señalan que el número de viajes en BPS entre semana y fin de semana es similar exceptuando el fin de semana por la mañana donde se reduce el número de viajes. Además, en días festivos el uso de bicicletas compartidas disminuye.

La percepción de seguridad de las bicicletas inteligentes también es un factor determinante. Estudios han demostrado que el ciclismo es más seguro en países donde hay carriles bici, modificaciones especiales de las intersecciones y semáforos prioritarios. Por ello, los sistemas de bicicletas compartidas deben reducir el riesgo lo máximo posible

ofreciendo clases de ciclismo, incentivando a los usuarios a llevar casco, etc. para aumentar la demanda de este servicio (DeMaio & Gifford, 2004).

El coste económico del uso de las bicicletas influye en la demanda de las Smart Bikes. La imposición de una tasa por su uso desincentiva a los usuarios a utilizarlas por lo que los programas de bicis compartidas sin ánimo de lucro tendrán más éxito entre los usuarios, aumentando la demanda del servicio por parte de los usuarios. Los ingresos de estos vendrán de la publicidad de las bicicletas o del mobiliario urbano instalado por el proveedor (DeMaio & Gifford, 2004).

Por otro lado, el vandalismo y el robo reducen la demanda de estos programas de bicicletas compartidas. Por último, el factor de conectividad multimodal afecta al uso de estos sistemas. Las bicicletas tienden a ubicarse en las zonas del centro de la ciudad para facilitar el acceso y la conexión. Además, una gran mayoría de las estaciones de bicicletas suelen posicionarse cerca de estaciones de transporte público para mejorar la movilidad de los usuarios y actuar como un medio de transporte complementario (DeMaio & Gifford, 2004).

Entre los factores más determinantes en el éxito del sistema de bicicletas compartidas, se encuentran la ubicación y densidad de las estaciones. Es importante que las estaciones se encuentren próximas unas de otras con el objetivo de hacer caminar a los usuarios lo menos posible para acceder al sistema. Esto proporcionará mayor comodidad y mejor accesibilidad a los usuarios. Cuanto mayor sea la densidad dentro de los sistemas bicicletas compartidas, mayor será el número de usuarios en el sistema. Vivir cerca de una estación de bicicletas compartidas aumenta su uso, siendo caminar la forma más común en que las personas se conectan a la estación (Conrow, Murray & Fischer, 2018).

Finalmente, se llevó a cabo una encuesta a una muestra de 125 individuos residentes en Madrid, con el objetivo de analizar los factores relevantes relacionados con el sistema de bicicletas compartidas. La muestra abarcó una amplia variedad de edades y el género femenino predominó. Los principales hallazgos de la encuesta son los siguientes:

- El 85,6% de los encuestados indicó no tener experiencia previa utilizando los servicios de bicicletas compartidas.

- La mayoría de los encuestados (87,9%) posee permiso de conducir, y el 81,6% cuenta con acceso a un coche personal.

- Respecto a los medios de transporte utilizados, el 51,5% de los encuestados indicó utilizar mayoritariamente el coche, mientras que el 39,2% utiliza el transporte público.

- En cuanto a las distancias recorridas, el 55,2% de los encuestados se desplaza con frecuencia en distancias cortas, en contraste con el 32,8% que lo hace en distancias largas.

- Un 38,4% de los encuestados considera que el uso de bicicletas compartidas no es importante para su movilidad urbana, frente a un 12,8% que las considera muy importantes.

- Los tres principales factores que motivarían a los encuestados a utilizar los servicios de bicicletas compartidas con mayor frecuencia son: la promoción de un estilo de vida saludable (42,3%), el ahorro económico (39%) y la comodidad (26%).

- Los tres mayores obstáculos que impiden un mayor uso de las bicicletas compartidas, según los encuestados, son: la falta de infraestructura ciclista (56,8%), las preocupaciones por la seguridad (44,8%) y la preferencia por otros medios de transporte (28,8%).

Para acceder a la encuesta completa, se puede consultar el anexo 1 del trabajo.

2.4 Evolución y tendencias de investigación

Se pueden distinguir 4 etapas de tendencias evolutivas en la investigación del sistema de bicis compartidas. A pesar de que la implementación de sistemas de bicicletas compartidas comenzó anteriormente, no fue hasta el año 2010 que surgió investigación relevante sobre el tema. La primera, de 2010 a 2012, se centró en asuntos de seguridad y políticas. La segunda, de 2013 a 2014, se centró en investigar el beneficio, sistema e impacto de los programas de bicis compartidas. La tercera etapa en 2015 se centró en analizar la optimización, comportamiento, entorno construido, diseño e infraestructura. En la última fase, de 2016 a 2018, la investigación fue acerca de la demanda, reequilibrio, redistribución, elección, tiempo, uso, transporte público y actitud (Si, Shi, Wu, Chen & Zhao, 2019).

Hasta el año 2012, la investigación se centró en el problema de redistribución en BSS (Bike Sharing System), asumiendo niveles óptimos de ocupación por estación y buscando minimizar los costos de redistribución. A partir de 2013, la investigación se orientó hacia la predicción de la demanda y optimización de la redistribución de bicicletas. Además, se introdujeron consideraciones sobre la eficiencia operativa, la calidad del servicio y la

sostenibilidad ambiental en la investigación, abriendo nuevas vías para soluciones más avanzadas en el ámbito de bikesharing. En los últimos años, ha surgido un creciente interés por comprender los patrones de uso de estos sistemas (Vallez, Castro & Contreras, 2021).

El desarrollo de programas de bicicletas compartidas a lo largo de los años ha sido extenso. Las ciudades están impulsando el uso de este medio de transporte ecológico para conseguir sus objetivos financieros y medioambientales, así como para reducir el uso de vehículos privados y su consiguiente contaminación. Los BPS se usan principalmente como medio complementario al desplazamiento a pie y otros sistemas de transporte público. Las bicicletas compartidas proporcionan velocidades de viaje más rápidas y costes más bajos contribuyendo a reducir el uso del coche (Zheng & Li, 2020).

Respecto al impacto económico del sistema de bicicletas compartidas, las investigaciones se han centrado en los beneficios indirectos que generan. Durante los primeros años, estos sistemas no son muy rentables al cobrar bajas rentas por el uso de las bicicletas e incurriendo en costes elevados asociados al equipamiento, cerraduras electrónicas y GPS. Las investigaciones han reflejado que el impacto económico más significativo es el ahorro de tiempo al proporcionar conectividad entre ciudades. Ese ahorro de tiempo puede ser aprovechado para generar valor económico. Además, se ha descubierto que el uso de bicicletas compartidas disminuye el número de accidentes de tráfico y su consecuente coste. Al mismo tiempo, contribuye a reducir los costes de sanidad al mejorar la salud de sus usuarios (Zheng & Li, 2020).

A continuación, se mencionan algunas oportunidades para estudiar y mejorar el sistema de bicicletas compartidas. Los datos históricos pueden analizarse para ganar insights acerca de los patrones de comportamientos de los usuarios de las bicicletas, así como para predecir la demanda y posicionar las bicicletas en aquellas áreas más concurridas de la ciudad para satisfacer la creciente demanda. Por otro lado, se podrían aplicar técnicas de machine learning para extrapolar información relevante de bases de datos e incorporar datos adicionales como por ejemplo condiciones meteorológicas (Chiariotti, Pielli, Cenedese, Zanella & Zorzi, 2018).

Por otro lado, es primordial seguir estudiando los efectos de la pandemia en la integración de opciones de transporte más sostenibles. Es importante evaluar como de atractivos resultaron los sistemas de bicicletas compartidas como alternativa al transporte público durante la pandemia. Se necesita estudiar en más detalle los efectos de la

pandemia en los perfiles socioeconómicos y demográficos de los usuarios de los BPS (Teixeira, Silva & Moura, 2023).

2.5 Boston

Durante la última década, Boston, en el estado de Massachusetts, ha realizado un importante esfuerzo por fomentar el ciclismo en su área metropolitana. Este compromiso se ha reflejado en los rankings, que en los primeros años del 2000 la clasificaban como una de las peores ciudades para el ciclismo debido a su consideración como un entorno hostil, pero para el año 2016, fue reconocida como una de las 50 mejores ciudades para esta actividad. La ciudad ha participado en una serie de programas destinado a fomentar el transporte multimodal y seguro por el centro de la ciudad como el programa Complete Streets en 2009 así como Visión Cero y Go Boston 2030 en 2014 (Karpinski, 2021).

En línea con sus objetivos de Visión Cero, la ciudad de Boston lanzó un proyecto llamado Commonwealth Ave Phase 2ª, con el objetivo de transformar la avenida Commonwealth entre el puente BU y Packard's Corner para reducir el riesgo de colisiones y aumentar la seguridad de los ciclistas. Este proyecto incluye carriles de bici e intersecciones protegidas para mejorar la visibilidad de los ciclistas y así evitar los accidentes denominados 'ganchos a la derecha' que ocurren cuando, al girar a la derecha, un coche choca con el ciclista que sigue recto. Además, se incorporarán señales de tráfico dirigidas exclusivamente para los ciclistas que les guiarán a cruzar por la intersección en el momento más seguro (Boston.gov, 2016).

El programa Complete Streets tenía como objetivo lograr que las calles de Boston resultasen verdes, inteligentes y seguras para todos los usuarios de todos los medios de transporte. Esta iniciativa contribuye a mejorar la salud de la comunidad y a luchar contra el cambio climático (Boston.gov, 2020). Por otro lado, su programa de Visión Cero tiene como objetivo para 2030 eliminar todos los accidentes fatales de tráfico, priorizando la seguridad de los individuos (Boston.gov., 2018). Por último, el programa Go Boston 2030 ha establecido la visión de las inversiones en transporte de Boston apostando por opciones de transporte mejores y más equitativas en el futuro (Boston.gov., 2017).

Con el objetivo de alcanzar mayor seguridad en las calles de la ciudad, el Departamento de Transportes de Boston (BTD) anunció la ampliación en 9,4 millas de los carriles para bicicletas. Para satisfacer esta creciente demanda, Boston ha ampliado su sistema de bicicletas compartidas Blue Bikes. Esta red ciclista segura ayudará a cerrar las

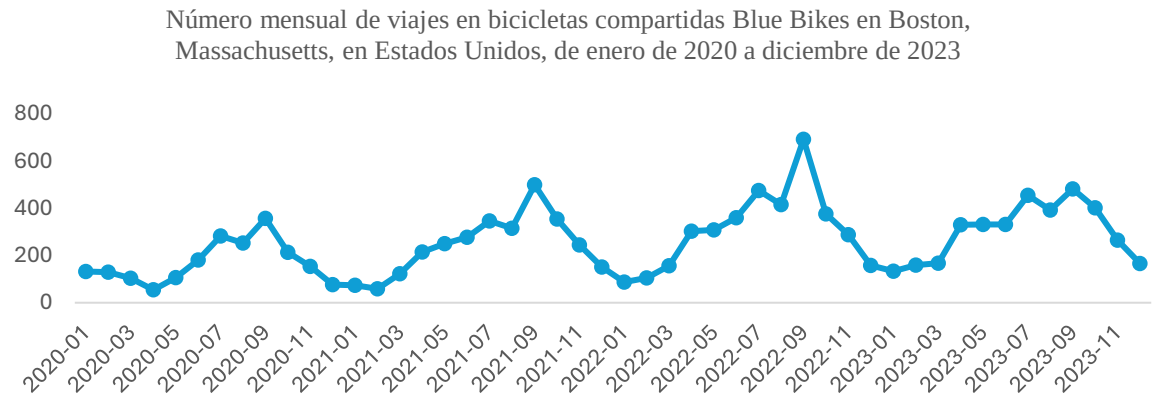
brechas de transporte entre barrios de la ciudad creando conexiones clave. Además, esta iniciativa promoverá el acceso a un medio de transporte económico que contribuye a abordar los problemas de tráfico, medioambientales y de seguridad que existen en la ciudad. Entre las acciones que se llevarán a cabo se incluyen la ampliación de la red ciclista para que el 50% de los residentes se encuentren a 3 minutos a pie de la misma, la ampliación del sistema Blue Bikes en un 40% mediante la inclusión de 100 nuevas estaciones, construir badenes en 30 barrios de la ciudad, la adición de 75 pasos de peatones elevados en parques, bibliotecas y colegios de toda la ciudad y la oferta de talleres gratuitos para aprender a montar en bicicleta dirigidos a mujeres y adultos con diversidad de género (Boston.gov, 2022).

Blue Bikes es el sistema de bicicletas públicas compartidas con más de 450 estaciones y 4.000 bicicletas en 13 municipios del área metropolitana de Boston. Está gestionado conjuntamente por Boston, Brookline, Cambridge, Everett y Somerville. Desde su lanzamiento en 2011, ha registrado más de 22 millones de viajes realizados entre residentes y visitantes (Boston.gov., 2016).

Este sistema de bicicletas compartidas consiste en una flota de bicicletas distribuidas por estaciones de acoplamiento de la red ciclista del área urbana de la ciudad. Este medio de transporte permite coger una bicicleta en una estación y devolverla en cualquier otra estación del sistema. La gente las usa tanto para ir al colegio, universidad o trabajo como para ir a encuentros sociales o hacer recados. Es una opción de movilidad que resulta muy cómoda, eficiente y económica (Bluebikes., n.d.).

En el siguiente gráfico, se puede observar un patrón estacional en el uso de las bicicletas compartidas Blue Bikes en Boston. Durante los meses más cálidos, se observa un aumento significativo en el número de viajes, particularmente en julio, agosto y septiembre, alcanzando su pico durante este último mes. Por otro lado, se registra un descenso en los viajes durante los meses más fríos, desde octubre hasta mayo. Esto refleja una significante influencia del clima en el uso de este medio de transporte por parte de los usuarios.

Figura 2: Viajes mensuales Blue Bikes



Fuente: Statista

3. Metodología

3.1 Adquisición de datos

La base de datos que se ha descargado para llevar a cabo el análisis del sistema de bicicletas compartidas Bluebikes de Boston se he obtenido del sistema de datos de Bluebikes abierto a todo el público. La url de la página web es la siguiente: https://www.bluebikes.com/system-data

3.1.1 Explicación campos de datos

En este sistema de datos, encontramos archivos tipo zip que incluyen ficheros Excel publicados mensualmente desde enero de 2015 hasta el mes actual de 2024. Desde enero de 2015 hasta mayo de 2020, cada fichero publicado cuenta con los siguientes 15 campos de variables:

- **Trip Duration (seconds):** duración del viaje en segundos, de tipo numérico entero.
- **Start Time and Date:** fecha y hora de inicio del viaje de tipo categórica.
- **Stop Time and Date:** fecha y hora de finalización del viaje, de tipo categórica.
- **Start Station ID:** ID de estación de inicio, de tipo numérico entero.
- **Start Station Name:** nombre de la estación de inicio, de tipo categórica.
- **Start Station Latitude:** latitud de la estación de inicio, de tipo numérica decimal.
- **Start Station Longitude:** longitud de la estación de inicio de tipo numérico decimal.

- **Stop Station ID**: ID de estación final de tipo numérico entero.
- **Stop Station Name:** nombre de la estación final de tipo categórica.
- **Stop Station Latitude:** latitud de la estación final**,** de tipo numérica decimal.
- **Stop Station Longitude**: longitud de la estación final, de tipo numérica decimal.
- **Bike ID:** ID de la bicicleta, de tipo numérico entero.
- **User Type:** tipo de usuario, de tipo categórica (casual = usuario de viaje único o pase de un día; miembro = miembro anual o mensual).
- **Birth Year:** año de nacimiento, de tipo numérico entero.
- **Gender**: género, de tipo categórica (Cero=desconocido; 1=hombre; 2=mujer).

Sin embargo, a partir de mayo de 2020, por tema de protección de datos, la empresa dejó de publicar el año de nacimiento y el género de los usuarios. Y al mismo tiempo, comenzó a incluir como nueva variable 'zip code', el código postal (de tipo categórica).

El tipo de cada variable lo hemos podido obtener a través de la función df.*dtypes* en Python :

Figura 3: tipos de variables

```
tripduration               int64
starttime                 object
stoptime                  object
start station id           int64
start station name        object
start station latitude   float64
start station longitude  float64
end station id             int64
end station name          object
end station latitude     float64
end station longitude    float64
bikeid                     int64
usertype                  object
postal code               object
dtype: object
```

Los datos se procesaron por la misma empresa Blue Bikes para eliminar los desplazamientos realizados por el personal durante el mantenimiento y la inspección del sistema, así como los desplazamientos de menos de 60 segundos (posibles salidas incorrectas o usuarios que la activan accidentalmente cuando intentan aparcar su bicicleta y quieren asegurarse de que esté bien sujeta).

En este trabajo de investigación, vamos a realizar un análisis del estado del sistema de bicicletas compartidas Blue Bikes de Boston durante el año posterior al comienzo de

la pandemia, es decir, 2021. Para ello, sacaremos una serie de insights que puedan ayudar a entender mejor el servicio y estudiar los nuevos patrones de uso que surgieron tras la pandemia.

3.1.2 Extracción de datos

Para la **extracción de datos**, vamos a usar dos librerías: *beautiful soup* y *selenium*. *Beautiful soup* es una biblioteca que facilita la extracción de información de páginas web mediante webscrapping. Permite analizar documentos HTML y XML (Uzun, Yerlikaya & Kirat, 2018). Normalmente, esta librería solo puede utilizarse con páginas estáticas. Sin embargo, la página de viajes se crea de forma dinámica con datos en tiempo real de Blue Bikes, siguiendo el estándar de datos recomendado por la North American Bike Share Association (NABSA).

Para poder hacer Web Scraping de manera efectiva sin perder parte de la información que queremos conseguir, vamos a usar la herramienta selenium. Esta biblioteca de código abierto es utilizada para pruebas automatizadas en aplicaciones web (Nyamathulla, Ratnababu & Shaik, 2021). *Selenium* automatiza la interacción con el navegador, permitiendo capturar la página web y acceder al código HTML generado y luego procesarlo con Beautiful Soup para extraer los datos necesarios. Para ello, requiere un controlador, en este caso, Chrome exe, el motor del propio navegador Chrome.

3.1.3 Descarga y descompresión de archivo

Primero, se adjunta la URL a analizar, aquella que da acceso a la página con datos del historial de viajes realizados con Blue Bikes. A continuación, se configura el driver del navegador Chrome indiciando la ruta de Chrome exe en mi equipo. Abrimos la URL con selenium mediante la función *driver.get()* y esperamos que la página cargue correctamente ajustando el tiempo requerido según mi conexión, en este caso deteniendo el programa durante 5 segundos.

Usamos *driver.page_source* para obtener el código fuente HTML de la página web. A continuación, se lo pasamos a Beautiful Soup para analizarlo. Y usamos la función soup.find_all('a') para encontrar todos los enlaces en la página. Después, mediante un bucle for, obtenemos el elemento 'href' de cada enlace de la página obtenido previamente para extraer las url asociadas y así poder descargar los archivos después. Una vez obtenidos los atributos href, se puede hacer click sobre los hipervínculos para descargar directamente la carpeta zip que escojamos.

Se crea la variable download_ dir, con el nombre del directorio donde se van a descargar los archivos. Mediante el filtro *.endswith('.zip')* and *'2021',* verificamos que solo se descargan los enlaces zip del año 2021 que son los que se quiere analizar. Obtenemos el nombre del archivo ZIP y establecemos la ruta del directorio de descarga. Luego, abrimos el archivo en escritura binaria ('wb') para asegurarnos de almacenar los datos tal como los recibimos, y se escribe en el archivo descargado la respuesta HTTP. Posteriormente, se descomprime el archivo descargado y se extrae su contenido en un directorio designado. Finalmente, elimina el archivo ZIP para ahorrar espacio en disco una vez que ha sido descomprimido.

Importamos las librerías *requests, para hacer peticiones HTTP* y así poder extraer información de una página*; os,* para manipular las rutas de los archivos descargados y establecer el directorio de descargas; *zipfile*, para trabajar y descomprimir los archivos zip en el directorio señalado. Por último, mediante os.remove() eliminamos los archivos zip después de descomprimirlos con el objetivo de liberar espacio en disco y eliminarlos permanentemente.

3.1.4 Procesamiento de archivos CSV y creación de un DataFrame

A continuación, creamos la variable '*extracted_dir'* para indicar la ruta del directorio donde se han descomprimido los archivos zip. Posteriormente, una lista es creada con los nombres de los archivos que presentan formato 'csv' y que se encuentran en el directorio de los archivos zip descomprimidos.

Se crea una lista vacía, 'dfs', para almacenar todos los dataframes de los archivos csv. Para ello, mediante un bucle que recorre los nombres de archivo csv seleccionados, se construye la ruta completa del archivo. Posteriormente, se lee su contenido en un DataFrame de Pandas, el cual se agrega a la lista 'dfs'.

Finalmente, se concatenan todos los dataframes en uno solo, al que hemos denominado 'combined_df'.

3.2 Transformación de datos

Las variables "starttime" y "stoptime", que originalmente eran de tipo categórico, se han cambiado a tipo "datetime" con el fin de poder analizar diferentes fechas y extraer más información y campos para su posterior formateo o manipulación.

A continuación, utilizando la función *'.dt.weekday'* de la librería Pandas, se han creado dos nuevas columnas en el dataframe. La primera columna representa el día de la semana correspondiente al tiempo de inicio de cada viaje, mientras que la segunda columna representa el día de la semana correspondiente al tiempo de finalización de cada viaje. Esta función devuelve valores en un rango del 0 al 6, donde 0 representa el lunes y 6 representa el domingo. Se ha empleado la función *pd.Categorical()* para convertir las columnas en variables categóricas ordinales, y se ha establecido el valor `True` para el parámetro 'ordered', indicando que siguen un orden específico.

Utilizando las dos nuevas variables mencionadas, se han creado otras dos variables adicionales que indican con valores binarios (0,1) si es fin de semana o no. Para lograr esto, se estableció que, si el valor es igual o mayor que 5, se asigna 1, indicando que es fin de semana; de lo contrario, se asigna 0, indicando que es entre semana. La primera variable se denomina 'fin_de_semana_start' y corresponde al tiempo de inicio del viaje, mientras que la segunda se denomina 'fin_de_semana_stop' y corresponde al tiempo de finalización del viaje. Se ha definido el tipo de variable como 'int' para obtener valores 0 y 1.

Por otro lado, se ha creado una variable denominada 'usertype_binary', codificando con valores dummy la variable categórica de usertype para facilitar su interpretación y su uso en el modelo posteriormente. Esta variable toma el valor de 1 si el usuario es abonado y 0 si es cliente esporádico. Se pasa al tipo int para que devuelva valores '1' y '0' en lugar de 'True' y 'False'.

Las columnas de latitud y longitud de las estaciones de origen y destino se combinaron para crear una nueva columna con la distancia que hay entre ambos puntos geográficos. Apliqué la fórmula de la distancia Haversine para que, a partir de la latitud y longitud de origen y destino, el algoritmo devolviese la distancia en kilómetros. Esta fórmula es la siguiente:

$$d_H(p,q) = 2 * R * \arcsin \sqrt{\sin^2\left(\frac{lat1 - lat2}{2}\right) + \cos(lat1) * \cos(lat2) * \sin^2\left(\frac{long1 - long2}{2}\right)} * 1000$$

Se ha creado una función en Python llamada 'funct_dist_Haversine' para calcular la distancia Haversine entre las estaciones de inicio y de parada. En esta función, se convierten los valores de latitud y longitud de las estaciones a radianes y se importa la librería 'math' para realizar los cálculos necesarios, incluyendo seno, coseno, arcoseno,

raíz cuadrada, arcotangente y conversión a radianes. Finalmente, se multiplica por 1.000 para pasar el valor a metros.

Luego, se ha aplicado esta función a cada fila del dataFrame, 'combined_df', para calcular la distancia entre las estaciones de inicio y fin de cada viaje. El resultado de este cálculo se ha almacenado en una nueva columna del DataFrame llamada 'distance_haversine'.

Se ha elegido emplear la distancia Haversine en lugar de la distancia Euclidea pues esta es más precisa para calcular distancias en objetos esféricos como la Tierra (Sharmila & Sabarish, 2021). La distancia Haversine se utiliza para calcular la distancia entre dos puntos geográficos representados por latitud y longitud, asumiendo una Tierra esférica con un radio constante. Esta fórmula no tiene en cuenta la topografía de la superficie terrestre, como las diferentes alturas de las colinas o la profundidad de los valles. Por lo tanto, simplifica los cálculos al ignorar el efecto elipsoidal (Maria, Budiman & Taruk, 2020).

3.3 Limpieza de datos

En esta sección, se realiza el proceso de limpieza de los datos mediante la identificación y posterior eliminación de errores y la validación de los datos (Requiz, Silva, Silva, Enriquez & Orbegoso, 2023).

Primero, se analiza si hay valores duplicados en la base de datos. Para ello, empleo la función *.duplicated().sum().* Esta me da como resultado valor cero, indicando que hay ausencia de duplicaciones.

A continuación, se estudia si hay valores nulos en la base de datos aplicando la función *.isnull().sum().* Esta fórmula me devuelve valor cero en todas las variables de la base de datos exceptuando en la de código postal que resulta presentar 222.076 valores nulos. Estos datos en la variable de código postal pueden ser el resultado de varios factores. Por un lado, podrían deberse a la falta de registro del código postal por parte de los usuarios. También es posible que se produzcan debido a errores durante el proceso de registro en la aplicación, ya sea por parte del usuario o por fallos técnicos del sistema. Además, si el código postal se registra automáticamente mediante un sistema de geolocalización, este puede resultar inexacto identificando la ubicación del usuario.

Para mejorar la precisión del análisis, se ha optado por excluir la columna de código postal, aplicando la función *.drop().* Esta variable no será utilizada en el modelo

de análisis debido a su cantidad significativa de valores nulos. Además, dado que contamos con las variables de latitud y longitud, podemos emplearlas para inferir la ubicación de las estaciones sin depender del código postal.

3.4 EDA

3.4.1 Descripción general del conjunto de datos y librerías empleadas

A continuación, se muestra una tabla con las versiones de las librerías empleadas:

Tabla 2: Librerías Pyhton

Librerías	Descripción
Pandas	Es una biblioteca de software de código abierto destinada principalmente a la manipulación y el análisis de datos en el lenguaje Python. Es potente, adaptable y simple de usar (*Pandas: La biblioteca de Python dedicada a la Data Science*, 2022).
Matplotlib	Librería open source enfocada en la creación de visualizaciones de datos (gráficos en dos dimensiones) con la ayuda de pandas y numpy (Daniel, 2022).
Seaborn	Librería creada sobre matplotlib y al igual que ésta sirve para crear gráficos con pocas líneas de código.
Numpy	Librería usada principalmente para realizar cálculos matemáticos y operaciones estadísticas. Otras librerías dependen de los arrays NumPy que usan como datos de entrada y salida (*NumPy: La biblioteca de Python más utilizada en Data Science,* 2023).
Stats (from scipy	Este módulo incluye muchas funciones estadísticas como distribuciones de probabilidad, estadísticas de resumen y frecuencia, estadísticas, funciones de correlación (*Statistical functions (scipy.stats) — SciPy v1.12.0 Manual*, s/f).
Sklearn	Esta librería, basada en numpy, scipy y matplotlib, es usada principalmente para el análisis predictivo de los datos. Es una herramienta para realizar técnicas de machine learning en Python (*Scikit-learn*, s/f).

Elaboración Propia

Para ver el código de Python, acceder a: https://github.com/itscarmengo/An-lisisdedatos_BlueBikes_Boston2021/blob/739a6f724dc44ae68e79440ed99aac311477e8e1/An%C3%A1lisisDatos_BlueBikes_Boston2021.ipynb

Como se mencionó previamente, se han concatenado los ficheros de enero a diciembre de 2021 para obtener la información anual mediante la función *pd.concat().*

Para comenzar con el análisis, se han calculado las dimensiones del dataset con la función *shape* y se ha comprobado que éste cuenta con 2.934.378 registros y 19 columnas. Esto incluye las seis nuevas columnas añadidas: weekday_start, weekday_stop, fin_de_semana_start, fin_de_semana_stop, usertype_binary, distance_haversine. Así mismo, se excluye la columna original de código postal.

Además, se ha empleado la función *.dtypes* para verificar la modificación correcta del tipo de variable de 'starttime' y 'stoptime', de categórica a tipo fecha. También hemos analizado el tipo de las nuevas variables incorporadas al conjunto de datos, todas ellas de tipo numérico entero, excepto la variable de distancia haversine, que es de tipo numérico decimal.

A continuación, usando la función *head* el programa muestra por defecto las 5 primeras observaciones del dataset para tener una visión general de cómo se estructura éste. Mediante la función *columns,* se pueden obtener también los nombres de las columnas del conjunto de datos

Figura 4: descripción datos

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	weekday_start	weekday_stop	fin_de_semana_start	fin_de_semana_stop	usertype_binary	distance_haversine
0	914	2021-01-01 00:00:04.590	2021-01-01 00:15:19.168	91	One Kendall Square at Hampshire St / Portland St	42.366277	-71.091690	370	Dartmouth St at Newbury St	42.350961	-71.077828	5316	Customer	4	4	0	0	0	2049.432144
1	1085	2021-01-01 00:00:21.803	2021-01-01 00:18:27.464	370	Dartmouth St at Newbury St	42.350961	-71.077828	169	Edwards Playground - Main St at Eden St	42.378965	-71.068607	4917	Subscriber	4	4	0	0	1	3205.694222
2	946	2021-01-01 00:00:26.009	2021-01-01 00:16:12.090	46	Christian Science Plaza - Massachusetts Ave at...	42.343666	-71.085824	21	Prudential Center - 101 Huntington Ave	42.346520	-71.080658	2881	Customer	4	4	0	0	0	530.243506
3	355	2021-01-01 00:00:30.921	2021-01-01 00:06:26.600	178	MIT Pacific St at Purrington St	42.359573	-71.101295	107	Ames St at Main St	42.362500	-71.088220	4792	Subscriber	4	4	0	0	1	1122.833764
4	511	2021-01-01 00:01:11.227	2021-01-01 00:09:43.195	386	Sennott Park Broadway at Norfolk Street	42.368605	-71.099302	413	Kennedy-Longfellow School 158 Spring St	42.369553	-71.085790	6062	Subscriber	4	4	0	0	1	1115.367617

Fuente: elaboración propia

3.4.2 Análisis de las variables numéricas

Para describir las variables numéricas, se ha aplicado la función *.describe(),* la cual nos devuelve el count, media, mínimo, máximo, percentiles (24%,50%,75%) y la desviación típica de las variables numéricas. Se han excluido variables numéricas que carecen de información estadística relevante.

Este es el caso de las variables binarias como usertype_binary, fin_de_semana_start o fin_de_semana_stop porque estas columnas solo representan la presencia o ausencia de una categoría. Estas variables no aportan información estadística relevante cuando se

calculan medidas como la media, el mínimo, el máximo, etc., ya que su interpretación se basa únicamente en la frecuencia de ocurrencia de cada categoría.

Además, excluimos las variables de identificación y tiempo porque generalmente estas columnas no tienen sentido estadístico en sí mismas. Por ejemplo, las variables de identificación como ID de estación o de bicicleta no tienen un valor numérico intrínseco que pueda ser interpretado en el contexto de un análisis estadístico. Del mismo modo, las variables de tiempo, como la fecha de inicio del viaje o la de parada del viaje, no se interpretan de manera útil cuando se calculan medidas de resumen como la media o la desviación estándar.

Figura 5: resumen estadístico variables numéricas

	tripduration	start station latitude	start station longitude	end station latitude	end station longitude	distance_haversine
count	2.934378e+06	2.934378e+06	2.934378e+06	2.934378e+06	2.934378e+06	2.934378e+06
mean	1.785841e+03	4.235761e+01	-7.109009e+01	4.235753e+01	-7.108996e+01	1.899958e+03
std	3.259313e+04	1.781573e-02	2.738765e-02	1.788794e-02	2.743911e-02	1.385917e+03
min	6.100000e+01	4.225560e+01	-7.124776e+01	4.225560e+01	-7.124776e+01	0.000000e+00
25%	4.380000e+02	4.234810e+01	-7.110650e+01	4.234807e+01	-7.110650e+01	9.330679e+02
50%	7.420000e+02	4.235810e+01	-7.109039e+01	4.235810e+01	-7.109018e+01	1.562819e+03
75%	1.272000e+03	4.236628e+01	-7.107119e+01	4.236628e+01	-7.107116e+01	2.545343e+03
max	8.277198e+06	4.253467e+01	-7.087021e+01	4.253467e+01	-7.087021e+01	3.739266e+04

Fuente: elaboración propia

La variable "*tripduration*" muestra un valor máximo de 8,28 millones de segundos, que podría deberse a la presencia de valores atípicos ya que es un valor muy poco probable que equivale a aproximadamente 2,299 horas, lo cual es poco realista para la duración de un viaje en bicicleta; un valor mínimo de 61 segundos, que confirma que la empresa descartó los viajes de menos de 1 minuto; y una media de 1.785,84 segundos, que indica que los viajes suelen durar alrededor de 29 minutos. Sin embargo, la desviación típica de esta variable es significativamente mayor que la media, 32.593,13, lo que indica una variabilidad significativa en los tiempos de duración de los viajes. Esto podría ser el resultado de una distribución sesgada a la derecha, como se observará con posterioridad en la matriz de dispersión, o de outliers en el extremo derecho.

Finalmente, en cuanto a los percentiles, el de 25 tiene un valor de 438 segundos, lo que indica que el 25 % de los viajes tienen una duración de alrededor de 7,3 minutos o menos; la mediana tiene un valor de 742 segundos, lo que indica que el 50 % de los viajes tienen una duración de alrededor de 12,4 minutos o menos; y el tercer cuartil tiene un

valor de 1.272 segundos, lo que indica que el 75 % de los viajes tienen una duración de alrededor de 21,2 minutos o menos.

Respecto a la variable *"start station latitude",* los valores oscilan entre 42.25 y 42.53. Esto muestra el rango total de latitudes de las estaciones de inicio en el conjunto de datos. La media de 42.36 indica que, en promedio, las estaciones de inicio se encuentran alrededor de esa latitud. La desviación típica de 0.0178 es pequeña, lo que indica una dispersión moderada alrededor de la ubicación central. Esto indica que las latitudes de las estaciones de inicio son bastante consistentes donde se encuentran. El primer cuartil con un valor de 42.35 refleja que un cuarto de las estaciones tiene latitudes relativamente bajas. La mediana de 42.36 muestra que el 50% de las estaciones tienen latitudes de 42.36 o menos. Y, por último, el tercer cuartil muestra que el 75% de las estaciones tienen latitudes de 42,37 o menos. En conclusión, las estaciones de inicio se encuentran ubicadas en áreas geográficas bastantes cercanas entre sí.

Respecto a la variable *"start station longitude",* los valores oscilan entre -71,25 y -70,87. Esto muestra el rango total de longitudes de las estaciones de inicio en el conjunto de datos. La media de -71,09 indica que, en promedio, las estaciones de inicio se encuentran alrededor de esa longitud. La desviación típica de 0.027 es pequeña, lo que indica una dispersión moderada. Esto indica que las longitudes de las estaciones de inicio son bastante consistentes donde se encuentran. El primer cuartil con un valor de -71,11 refleja que un cuarto de las estaciones tiene latitudes relativamente bajas. La mediana de -71,09 muestra que el 50% de las estaciones tienen latitudes de -71,09 o menos. Y, por último, el tercer cuartil muestra que el 75% de las estaciones tienen latitudes de -71,07 o menos.

Respecto a la variable *"end station latitude",* los valores oscilan entre 42,26 y 42,53. Esto muestra el rango total de latitudes de las estaciones de fin en el conjunto de datos. La media de 42,36 indica que, en promedio, las estaciones de fin se encuentran alrededor de esa latitud. La desviación típica de 0,0179 es pequeña, lo que indica una dispersión moderada. Esto indica que las latitudes de las estaciones de fin son bastante consistentes donde se encuentran. El primer cuartil con un valor de 42,35 refleja que un cuarto de las estaciones tiene latitudes relativamente bajas. La mediana de 42,36 muestra que el 50% de las estaciones tienen latitudes de 42,36 o menos. Y, por último, el tercer cuartil muestra que el 75% de las estaciones tienen latitudes de 42,37 o menos. En

conclusión, las estaciones de fin se encuentran ubicadas en áreas geográficas bastantes cercanas entre sí.

Respecto a la variable *"end station longitude",* los valores oscilan entre -71,25 y -70,87. Esto muestra el rango total de longitudes de las estaciones de inicio en el conjunto de datos. La media de -71,09 indica que, en promedio, las estaciones de inicio se encuentran alrededor de esa longitud. La desviación típica de 0.0274 es pequeña, lo que indica una dispersión moderada. Esto indica que las longitudes de las estaciones de inicio son bastante consistentes donde se encuentran. El primer cuartil con un valor de -71,11 refleja que un cuarto de las estaciones tiene latitudes relativamente bajas. La mediana de -71,09 muestra que el 50% de las estaciones tienen latitudes de -71,09 o menos. Y, por último, el tercer cuartil muestra que el 75% de las estaciones tienen latitudes de -71,07 o menos.

Respecto a la variable *"distance_haversine",* los valores oscilan entre 0 y 37.392,66 metros. Esto muestra el rango total de distancias que existen entre las estaciones de origen y las estaciones de fin en el conjunto de datos. La media de 1.562,81 metros indica que, en promedio, las distancias Haversine se encuentran alrededor de esa latitud. La desviación típica de 1.385,92 es grande, lo que indica una dispersión considerable en los datos. Esto indica que hay una amplia variabilidad en las distancias Haversine. El primer cuartil con un valor de 933,06 metros refleja que un cuarto de las distancias son relativamente cortas. La mediana de 1.562,82 muestra que el 50% de las estaciones son de 1.562,82 metros o menos. Y, por último, el tercer cuartil muestra que el 75% de las estaciones son de 2.545,34 metros o menos. En conclusión, aunque las estaciones de origen y fin estén cercanas entre sí, la red de estaciones abarca una amplia área geográfica, lo que resulta en una variedad significativa de distancias recorridas por los usuarios.

A continuación, se ha empleado la función *.corr()* para obtener las correlaciones entre las variables numéricas.

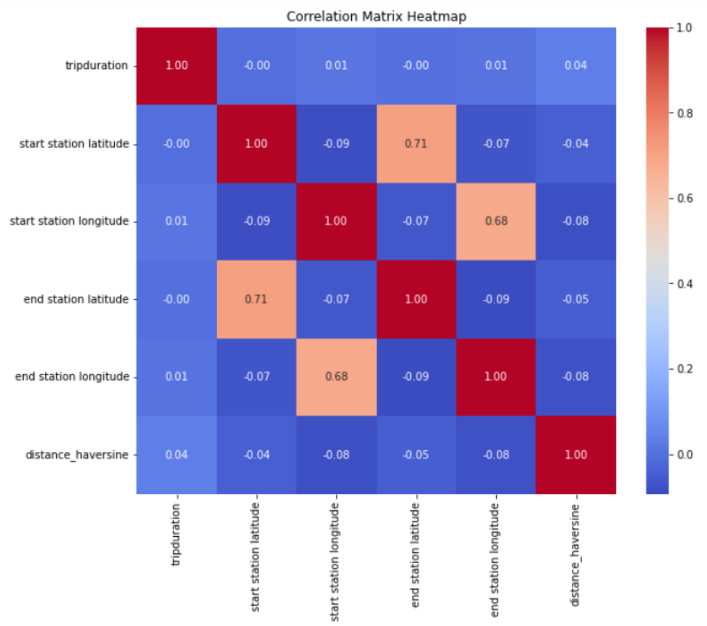
Tabla 3: correlaciones

	trip duration	start station latitude	start station longitude	end station latitude	end station longitude	distance_haversine
trip duration	1.000.000	-0.003988	0.011746	-0.000196	0.005270	0.037319
start station latitude	-0.003988	1.000.000	-0.086127	0.706254	-0.069476	-0.044259
start station longitude	0.011746	-0.086127	1.000.000	-0.067214	0.683482	-0.084243
end station latitude	-0.000196	0.706254	-0.067214	1.000.000	-0.093214	-0.047262
end station longitude	0.005270	-0.069476	0.683482	-0.093214	1.000.000	-0.078333
distance_haversine	0.037319	-0.044259	-0.084243	-0.047262	-0.078333	1.000.000

Fuente: elaboración propia

Además, se ha representado la matriz de correlación utilizando un mapa de calor para obtener una visualización más clara de los resultados.

Figura 6: mapa de calor



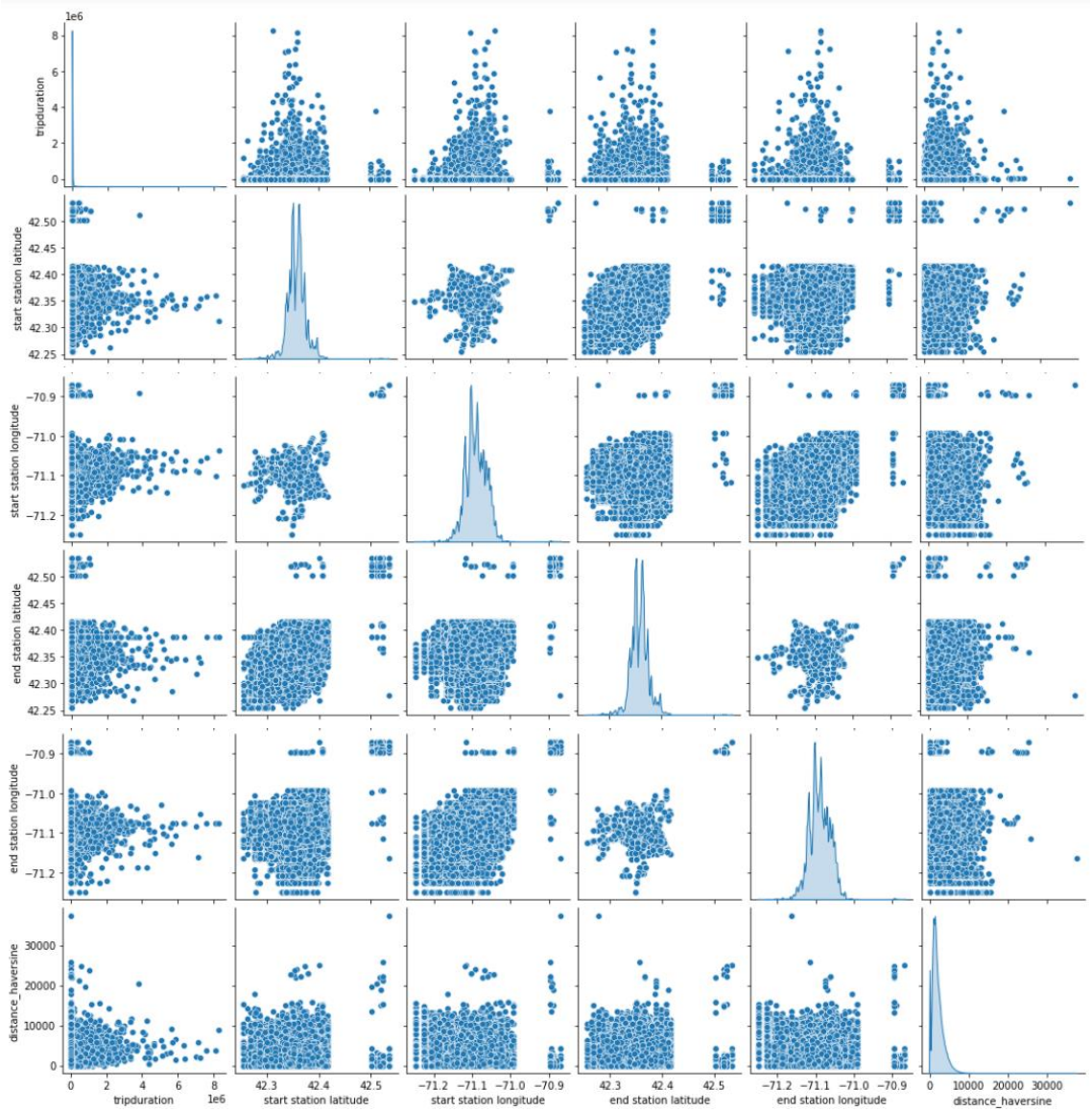
Fuente: elaboración propia

Podemos observar una correlación positiva débil entre las variables de tripduration y distance_haversine lo que indica que los viajes más largos cubren distancias un poco mayores. Es decir, a medida que aumenta la distancia recorrida, también tiende a incrementarse la duración del viaje. Sin embargo, esta relación no es un factor determinante pues otros factores pueden influir.

Hay una correlación positiva moderada entre las latitudes y longitudes de las estaciones de inicio y fin lo que puede indicar que estas se ubican en áreas geográficas similares o cercanas.

Para visualizar la correlación lineal existente entre las variables, se ha creado una matriz de dispersión (Millán, 2020). Para ello, se ha aplicado la función *sns.pairplot()*.

Figura 7: matriz de dispersión



Fuente: elaboración propia

Para realizar el gráfico, he seleccionado "kde" para la diagonal con el objetivo de visualizar estimaciones de densidad tipo núcleo (Coder, 2021).

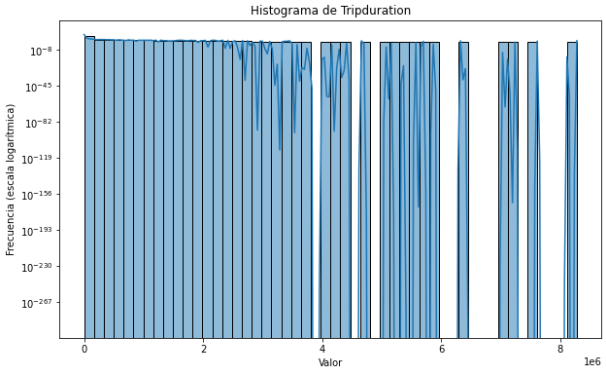
En el gráfico de dispersión, podemos observar variables numéricas tanto en el eje x como en el eje y. Mediante histogramas a lo largo de la diagonal, podemos ver la

distribución de cada una de las variables. Y fuera de la diagonal, podemos ver gráficos de dispersión que muestran las relaciones entre pares de variables del conjunto de datos.

De los histogramas de la diagonal, podemos deducir que la variable de *tripduration* presenta valores atípicos pues vemos que la mayoría de los datos se concentran en valores cercanos al 0 y son pocos los valores grandes.

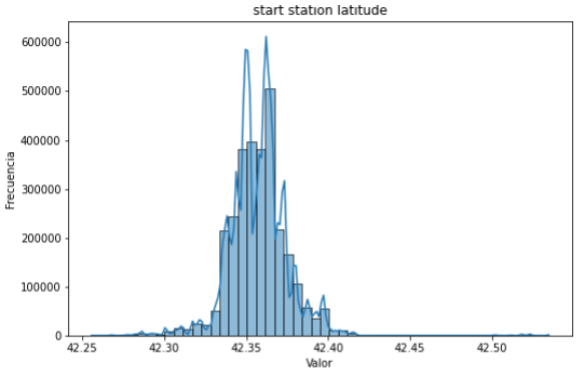
Con el objetivo de analizar más en profundidad el comportamiento de las variables numéricas en el conjunto de datos, hemos decidido realizar histogramas para cada una de ellas mediante la función *sns.histplot()*.

Figura 8: histograma tripduration



Fuente: elaboración propia

Figura 9: histograma start station latitude



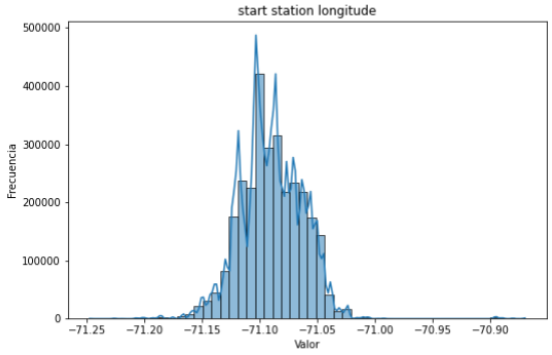
Fuente: elaboración propia

En el histograma de "tripduration", hemos aplicado una escala logarítmica en el eje 'y' debido a la significante variabilidad en los valores de esta variable. Esta presenta un amplio rango que abarca desde un mínimo de 61 hasta un máximo que supera los 8 millones. Además, su distribución está notablemente sesgada hacia valores más pequeños. Esta adaptación nos ha permitido lograr una visualización más clara y efectiva de la frecuencia de ocurrencia de cada valor.

En el gráfico de "tripduration", podemos observar claramente la distribución sesgada a la derecha de los datos. Predominan los valores pequeños, lo que sugiere que la mayoría de los viajes presentan una duración muy corta, mientras que son pocos los que presentan valores grandes.

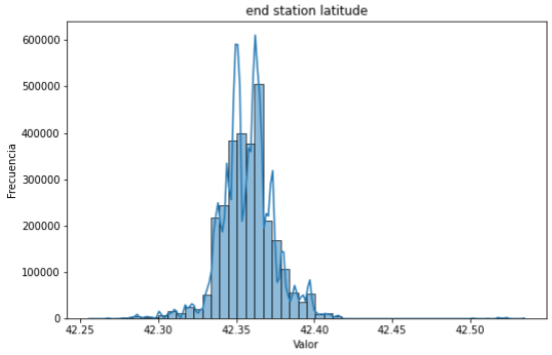
En los histogramas de las variables 'start station latitude' y 'start station longitude', podemos ver que los datos presentan una distribución normal. En ambos casos, la distribución de los datos está un poco sesgada a la izquierda pues hay menos datos con valores inferiores, y los registros presentan un rango pequeño de valores lo que indica poca variabilidad.

Figura 10: histograma start station longitude



Fuente: elaboración propia

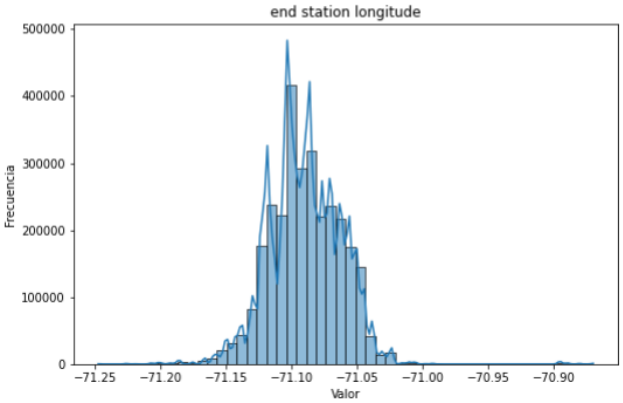
Figura 11: histograma end station latitude



Fuente: elaboración propia

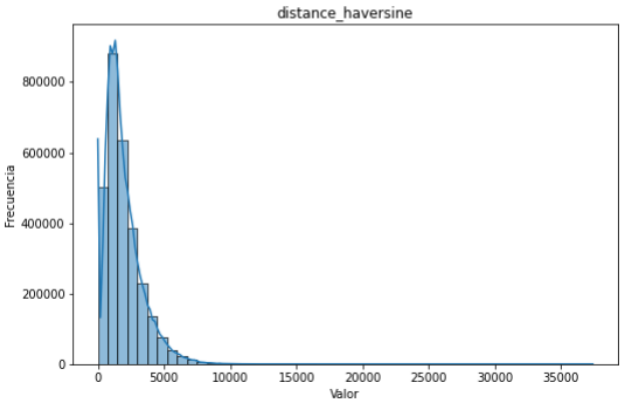
En los histogramas de "*end station latitude*" y "*end station longitude*", podemos ver que los datos siguen una distribución medianamente normal y un poco sesgada a la izquierda. Por lo que podemos deducir que las estaciones finales presentan sobre todo identificadores con valores cercanos a la media y que son más las estaciones que toman valores de latitud y longitud más altos que las que presentan una latitud y longitud más baja.

Figura 12: histograma end station longitude



Fuente: elaboración propia

Figura 13: histograma distance_haversine



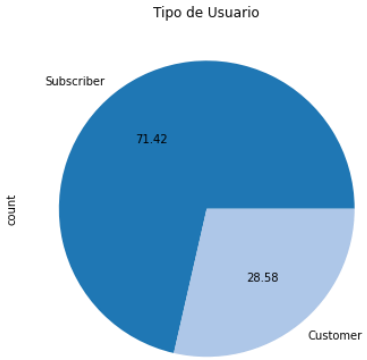
Fuente: elaboración propia

El histograma de *"distance_haversine"* refleja una distribución sesgada a la derecha concentrándose la mayoría de los valores entre el rango de 0-5.000 y pocos valores entre los rangos 5.000-35.000.

3.4.2 Análisis de las variables categóricas

Para el análisis de la variable *usertype*, se ha elaborado un gráfico de tarta. En este podemos observar que la mayoría de los usuarios, un 71,42%, son abonados y tan solo, un 28,58% son clientes esporádicos. Este hallazgo sugiere que el sistema de bicicletas compartidas Blue Bikes tiene una alta retención de clientes, ya que la mayoría de ellos optan por suscribirse al programa después de utilizar el servicio.

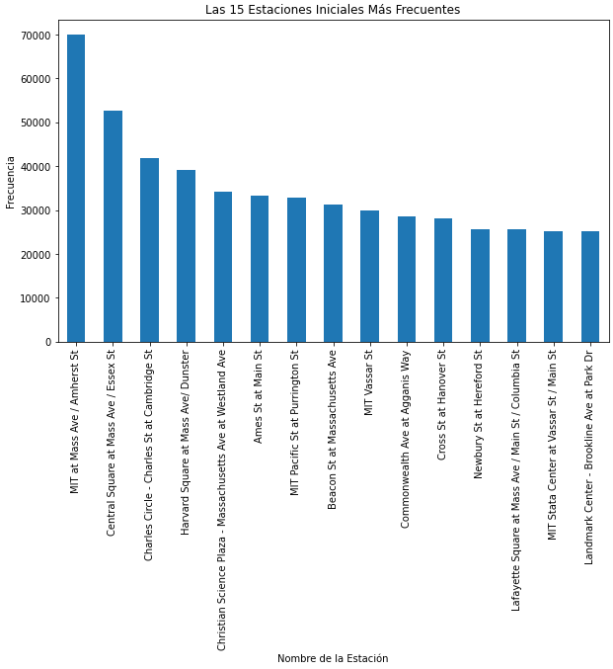
Figura 14: Gráfico de tarta tipo de usuario



Fuente: Elaboración propia

Se ha generado un gráfico de barras para analizar el recuento de viajes desde las 15 estaciones iniciales más frecuentes. Esta visualización resulta útil dado el gran número de estaciones, lo que dificulta la interpretación en un solo gráfico. En el gráfico, se puede observar que la estación de inicio "MIT at Mass Ave/Amherst St" encabeza la lista en cuanto a la cantidad de viajes iniciados. Le siguen "Central Square at Mass Ave /Essex St" y "Charles Circle – Charles St at Cambridge St". Las dos primeras estaciones se ubican en áreas de alto tránsito debido a su ubicación en una importante vía de la ciudad. Por otro lado, la tercera posición la ocupa una estación ubicada en Beacon Hill, un prestigioso y concurrido barrio residencial de la ciudad.

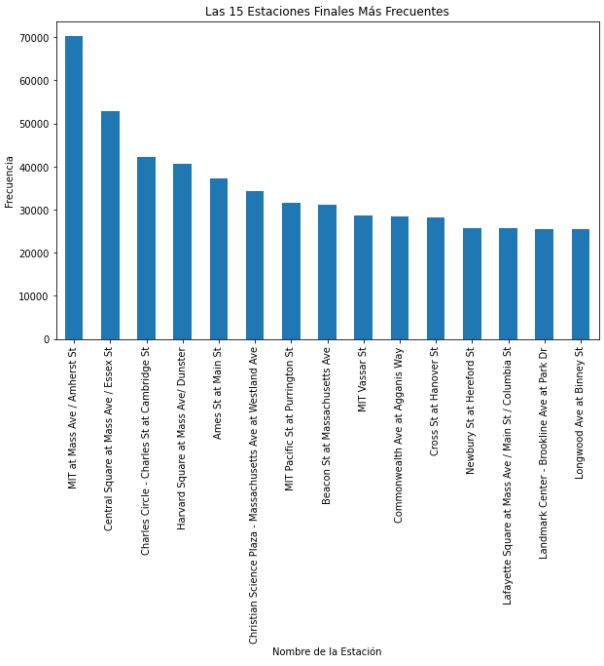
Figura 15: Gráfico de Barras Estaciones Iniciales



Fuente: Elaboración propia

El mismo tipo de gráfico se ha utilizado para analizar las quince estaciones finales más frecuentes. Curiosamente, las tres estaciones finales más frecuentes coinciden con las mismas que las estaciones iniciales más populares. Esto podría indicar que ciertas estaciones son puntos de inicio y fin preferidos para los usuarios, quizás debido a su conveniente ubicación.

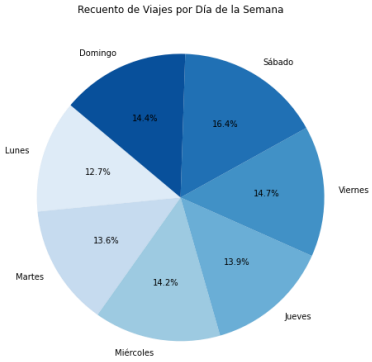
Figura 16: Gráfico de Barras Estaciones Finales



Fuente: Elaboración propia

Para analizar las variables 'weekday_start' y 'weekday_stop', se ha creado un gráfico de tarta para visualizar el número de viajes por día de la semana. Se ha optado por un solo gráfico debido a que los viajes comienzan y terminan en el mismo día. El análisis reveló que el sábado es el día con mayor número de viajes, representando el 16.4% del total, seguido por el viernes con un 14.7% y el domingo con un 14.4%. Por otro lado, el lunes registra el menor porcentaje de viajes, con un 12.7%. Este patrón sugiere un aumento en la actividad de viaje durante los fines de semana, probablemente debido a turistas y personas que disfrutan de actividades de ocio.

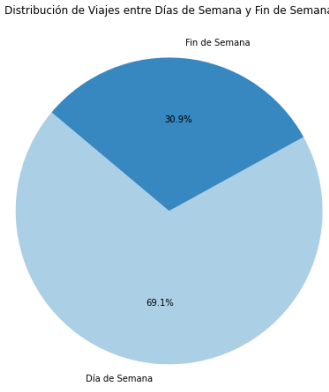
Figura 17: Gráfico de tarta Viajes por Día de la Semana



Fuente: Elaboración Propia

El análisis revela que la mayoría de los viajes se realizan durante la semana. Sin embargo, como hemos mencionado previamente, al observar los datos individualmente por día, se evidencia que los días del fin de semana tienen un mayor número de viajes en comparación con los días laborales.

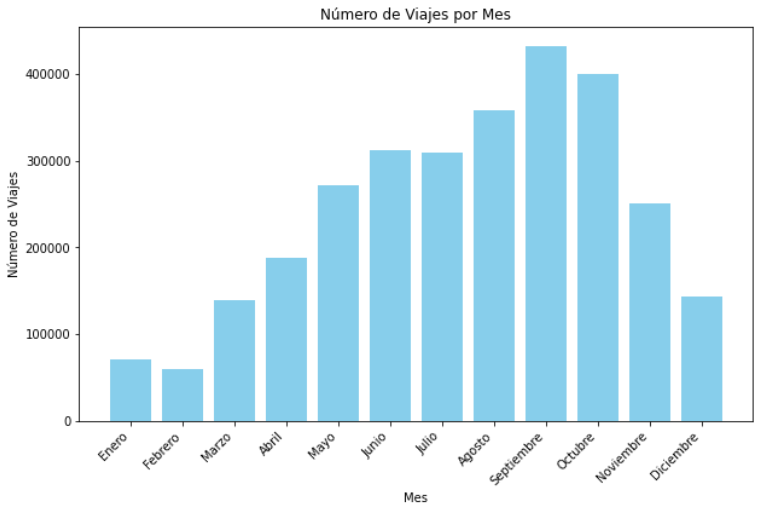
Figura 18: Gráfico de tarta Viajes Fin de Semana



Fuente: Elaboración Propia

Para analizar las variables starttime y stoptime, se ha creado un histograma que ilustra el número de viajes por mes. Los datos revelan que septiembre es el mes con mayor cantidad de viajes, seguido por octubre y agosto en orden de frecuencia.

Figura 19: Número de viajes por mes



Fuente: Elaboración Propia

Enero y febrero destacan como los meses con menor registro de viajes. Esta tendencia puede atribuirse a las bajas temperaturas que caracterizan a estos meses, así como a una mayor incidencia de precipitaciones. Las condiciones climáticas adversas, que incluyen frío y lluvias, pueden desalentar la movilidad y la participación en actividades al aire libre, lo que resulta en una disminución en la cantidad de viajes realizados. Además, es posible que, durante estos meses, las personas opten por alternativas de transporte más cálidas y resguardadas, como el transporte público o los vehículos privados, en lugar de la bicicleta.

Tabla 4: Clima Boston

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Temperatura media (ºC)	-3.1	-1.8	2	8.4	14	19.2	22.8	22.1	18.4	12	6.1	0.6
Temperatura min. (ºC)	-6.8	-5.7	-2	4.1	9.8	15.1	18.8	18.3	14.9	8.7	3	-2.6
Temperatura máx. (ºC)	1.6	3.2	7	13.8	19.2	24.2	27.8	26.9	23	16.3	10.2	4.7
Precipitación (mm)	99	87	117	103	88	90	76	90	90	110	95	117
Humedad (%)	67%	63%	66%	65%	70%	72%	69%	70%	73%	72%	73%	71%
Días lluviosos (días)	8	7	8	8	8	7	7	7	6	7	7	8
Horas de sol (horas)	5.9	6.7	7.2	8.3	8.6	9.7	10.6	9.7	8.0	6.6	5.9	5.4

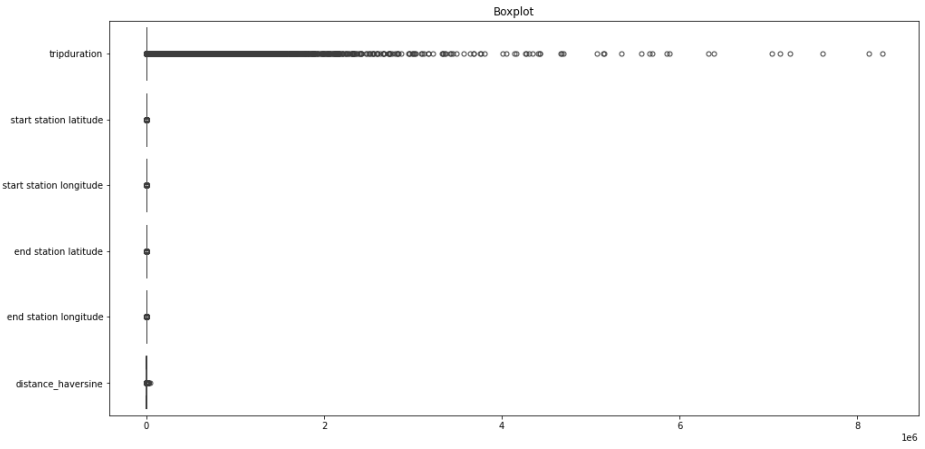
Fuente: (Clima Boston: Temperatura, Climograma y Temperatura del agua de Boston, s/f)

El incremento en la cantidad de viajes de agosto a septiembre, alcanzando su punto máximo en este último mes, puede atribuirse al inicio del período académico en las universidades. Con la vuelta a la ciudad de muchos estudiantes, aumenta la demanda de movilidad y transporte dentro de la misma. Este fenómeno refleja la influencia que tienen los calendarios académicos en los patrones de movilidad urbana, donde el regreso a clases genera un incremento notable en la actividad de transporte público y otros medios de movilidad, incluyendo el uso de bicicletas.

3.4.4 Outliers

Para realizar la búsqueda de outliers, se han realizado boxplots de las variables con el objetivo de visualizar los valores atípicos.

Figura 20: Boxplot



Fuente: Elaboración Propia

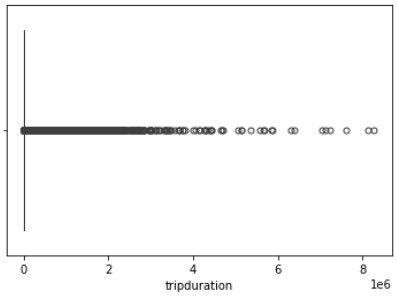
En el gráfico podemos ver que efectivamente la variable *"tripduration"* presenta una cantidad significativa de valores extremos. Por ende, a continuación, procederemos a un análisis más detallado de la presencia de éstos enfocándonos exclusivamente en esta columna de duración del viaje.

Para calcular los outliers, hemos establecido un umbral con valor 3. Esto se debe a la regla empírica, según la cual los datos dentro de 3 veces la desviación estándar respecto a la media representan el 99.7% de los datos de la distribución. Sabiendo esto, podemos concluir con bastante seguridad que los datos que caen más allá de este umbral son atípicos, pues son distintos al 99.7% de los datos. El resultado ha sido de 2.853 outliers (Zulmuthi, 2022).

Para tratar con los outliers, se ha empleado el método de winsorización. Esta es una técnica que reemplaza los valores atípicos por el valor más cercano que no se considera un outlier según ciertos criterios (Zulmuthi, 2022).

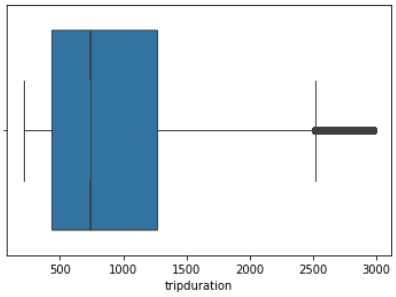
A continuación, se muestran los boxplot antes y después de aplicar el método de winsorización.

Figura 21: Boxplot Antes de Winsorización



Fuente: Elaboración Propia

Figura 22: Boxplot Después de Winsorización



Fuente: Elaboración Propia

4.Preguntas de investigación

4.1 Pregunta 1: ¿Qué bicicletas tienen que ir a mantenimiento?

Para comenzar, utilizando la función *.value_counts(),* calculamos el número de viajes realizados por cada bicicleta y guardamos el resultado en la variable *bike_counts*. Es importante destacar que pandas nos presenta los resultados ordenados de mayor a menor por defecto. Observamos que la bicicleta con ID 3489 encabeza la lista con la mayor cantidad de viajes, con un total de 1292. Le sigue la bicicleta con ID 5615, con 1288 viajes, en segunda posición. En tercer lugar, encontramos la bicicleta con ID 6678, con un total de 1273 viajes. Por otro lado, las bicicletas con ID 7622, 5226 y 2411 son las que menos viajes han realizado, con un único viaje cada una.

A continuación, creo una variable, 'top_bicicletas', para ver las diez bicicletas que más viajes realizan. Para ello, aplico la función *.head(10).*

Tabla 5: Bicicletas más usadas

Bike ID	3489	5615	6678	6304	4507	6241	6564	6858	5666	5712
Nº Viajes	1292	1288	1273	1258	1255	1254	1253	1244	1212	1201

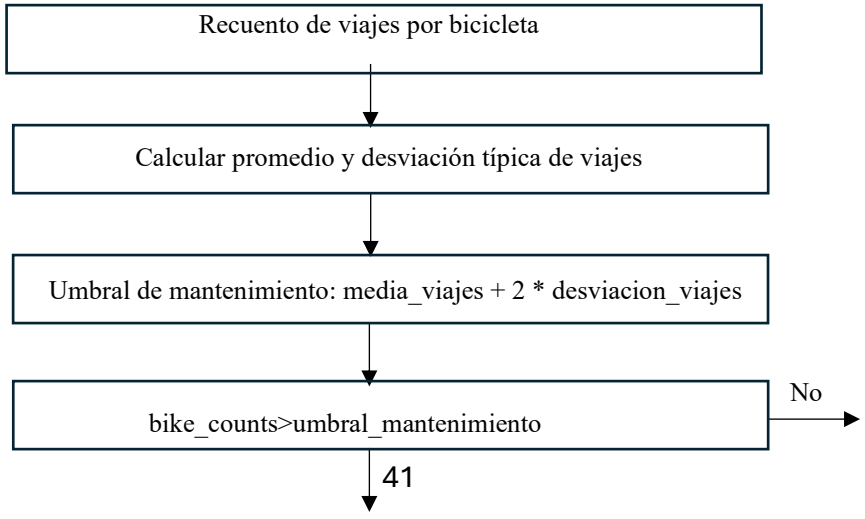
Fuente: Elaboración propia

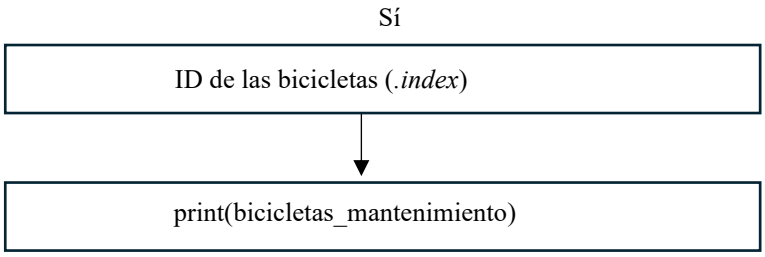
Se ha creado una nueva variable llamada 'viajes_top_bicicletas' para almacenar la información de los viajes realizados con las bicicletas más frecuentes. Este paso implica filtrar el DataFrame para incluir solo las filas en las que el ID de la bicicleta se encuentra dentro del índice de la variable 'top_bicicletas', que previamente ha sido creada y contiene las diez bicicletas que han realizado más viajes.

Para determinar qué bicicletas requieren mantenimiento, primero calculamos el promedio y la desviación estándar de los viajes mediante la función *bike_counts.mean()* y *bike_counts.std(),* obteniendo valores de 652 y 281, respectivamente. Definimos el umbral como la suma de la media de los viajes más dos desviaciones estándar. Esta elección se basa en el hecho de que, en una distribución normal, aproximadamente el 95% de los valores se encuentran dentro de dos desviaciones estándar de la media. Por lo tanto, al establecer este umbral, estamos seleccionando el 5% de los valores que están más allá de este rango, indicando desviaciones significativas de la tendencia central de la distribución (Juan, Sedano & Vila, 2006). El umbral de mantenimiento tiene un valor de 1.214. Las bicicletas con un número de viajes mayor a este umbral serán las que necesiten mantenimiento, pues indican un uso alto y mucho mayor que el del resto.

Se ha creado una variable, 'bicicletas_mantenimiento, que contiene los IDs de las bicicletas con un número de viajes mayor al umbral de viajes establecido. Primero, se realizó una comparativa de cada valor de la variable bike_counts con el umbral de mantenimiento mediante la función *bike_counts > umbral_mantenimiento.* Esta expresión devuelve 'True' si el valor supera el umbral y 'False' en caso contrario. Luego, estos resultados booleanos se usan para filtrar la varia bike_counts, dejando solo las bicicletas que han superado el umbral. Finalmente, .index se utiliza para obtener los IDs de las bicicletas resultantes.

Figura 23: Flujos de caja pregunta 1





Fuente: Elaboración propia

Se ha obtenido que las bicicletas que necesitan mantenimiento son las siguientes:

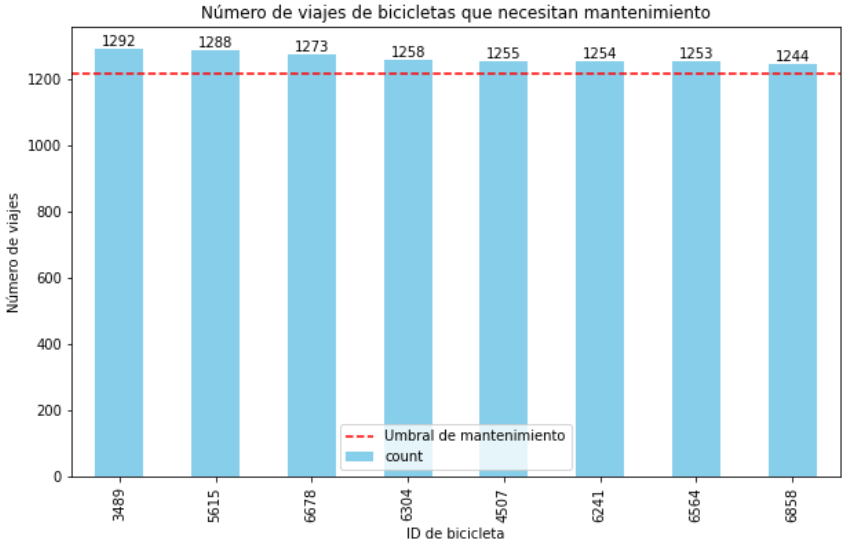
Tabla 6: IDs bicicletas mantenimiento

Bike ID	3489	5615	6678	6304	4507	6241	6564	6858

Fuente: elaboración propia

En el siguiente gráfico, se muestra el número de viajes por bicicleta que necesita mantenimiento y que supera el umbral establecido.

Figura 24 Viajes bicicletas mantenimiento



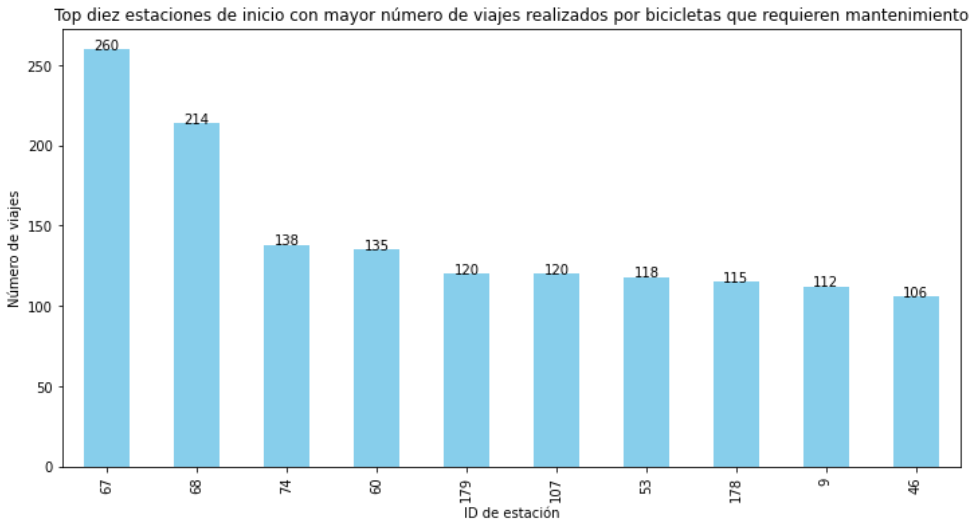
Fuente: Elaboración propia

A continuación, se presenta un gráfico que muestra las diez estaciones de inicio desde donde se han emitido más viajes con bicicletas que requieren mantenimiento.

Primero, se recopilan las estaciones de inicio únicas de las bicicletas que necesitan mantenimiento en la variable 'estacionesinicio_bibicletasmantenimiento'. Luego, se agrupan los viajes de las bicicletas que necesitan mantenimiento por ID de la estación de inicio mediante la función *.groupby*() y se calcula el número de viajes desde cada estación de inicio utilizando la función *.size(),* almacenando este resultado en la variable 'viajes_por_estacion_inicio'. Posteriormente, estas estaciones se ordenan en orden descendente según el número de viajes y se seleccionan las diez estaciones principales

mediante la función *.head(10),* almacenando los resultados en la variable 'top_estacionesinicio_bicicletasmantenimiento'. Finalmente, estos datos se visualizan en un gráfico de barras para identificar claramente las estaciones de inicio con mayor número de viajes de bicicletas de mantenimiento, lo que proporciona información valiosa para la gestión y mantenimiento del sistema de bicicletas compartidas.

Figura 25 Top 10 estaciones de inicio de bicicletas mantenimiento



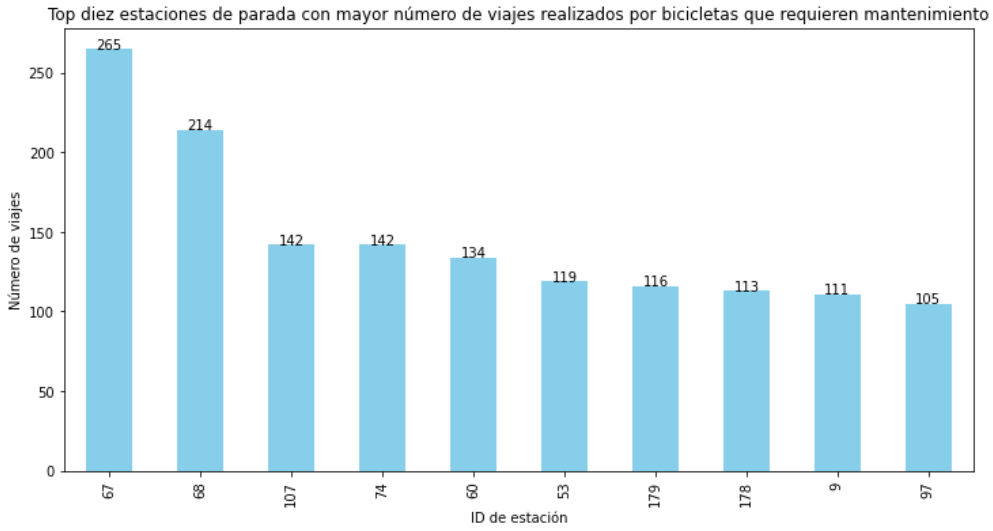
Fuente: Elaboración propia

Este gráfico representa las diez estaciones de inicio con mayor número de viajes para las bicicletas que requieren mantenimiento, es decir, aquellas cuyo número de viajes ha superado el umbral establecido para mantenimiento. Es importante tener en cuenta que este recuento corresponde exclusivamente a los viajes realizados por estas bicicletas que han superado el umbral definido para el mantenimiento, no refleja el número total de viajes registrados en cada estación.

Podemos observar que la estación de inicio con ID 67 ha registrado el mayor número de viajes por parte de las bicicletas que necesitan mantenimiento, totalizando 260 viajes. Le sigue de cerca la estación número 68 con un total de 214 viajes.

A continuación, se repitió el proceso para identificar las estaciones de destino más comunes.

Figura 26: Top 10 estaciones de parada de bicicletas de mantenimiento



Fuente: Elaboración propia

Este gráfico revela un patrón destacado en los viajes de las bicicletas que necesitan mantenimiento: las estaciones identificadas con los ID 67 y 68 no solo son las principales estaciones de inicio para estas bicicletas, sino que también son las estaciones de destino más frecuentes. Este análisis revela que las bicicletas están siendo utilizadas repetidamente entre un conjunto limitado de estaciones, donde los usuarios las retiran de una estación, las utilizan y luego las devuelven al mismo lugar de origen. Este descubrimiento destaca la importancia de estas estaciones como puntos clave de actividad y mantenimiento en la red de bicicletas compartidas de Boston.

La estación más concurrida, identificada como la número 67, es conocida como 'MIT at Mass Ave / Amherst St'. La segunda estación más transitada, con el ID 68, es la ubicada en 'Central Square at Mass Ave / Essex St'.

A continuación, se han obtenido las diez rutas más comunes realizadas por bicicletas que requieren mantenimiento. Primero, se han agrupado los viajes de las bicicletas que necesitan mantenimiento por las columnas ‘start station id’ y ‘end station id’ mediante la función *.groupby()*. Luego, mediante la función *.size()*, se ha calculado el número de viajes de cada ruta realizada desde la estación de inicio hasta la de fin. Por último, se han seleccionado las diez rutas con mayor número de viajes registrados utilizando el método *.nlargest(10)* y se almacenaron en la variable 'top_rutas’.

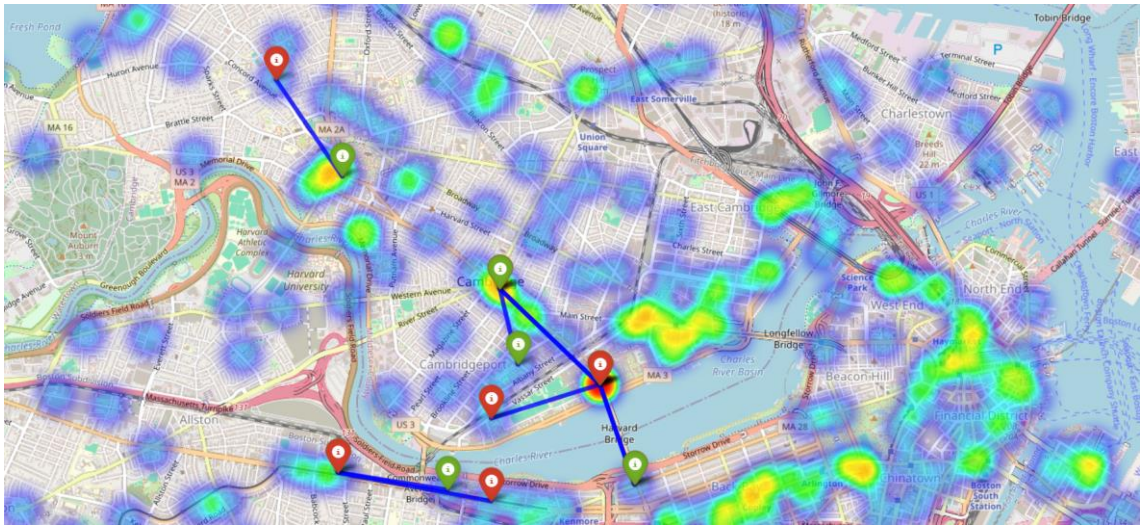
Con el objetivo de obtener los nombres de dichas estaciones de inicio y fin de esas rutas, se utilizó un bucle 'for' que itera sobre los elementos de 'top_rutas' que contiene el par de identificadores de las estaciones de inicio y fin, así como el número de viajes registrados para cada ruta. En cada iteración, se filtra el dataFrame de viajes de las bicicletas que necesitan mantenimiento para obtener las filas que corresponden a la estación de inicio y fin de la ruta actual. Luego, se extraen los nombres de estas estaciones. Por último, se imprimen junto con los nombres de las estaciones de inicio y fin, el número de viajes registrados para cada ruta.

El resultado obtenido ha sido el siguiente:

1. Ruta: MIT at Mass Ave / Amherst St -> Beacon St at Massachusetts Ave, con un total de 20 viajes registrados.
2. Ruta: Commonwealth Ave at Agganis Way -> 700 Commonwealth Ave., con un total de 19 viajes registrados.
3. Ruta: Beacon St at Massachusetts Ave -> MIT at Mass Ave / Amherst St, con un total de 19 viajes registrados.
4. Ruta: MIT Pacific St at Purrington St -> Central Square at Mass Ave / Essex St, con un total de 17 viajes registrados.
5. Ruta: Murphy Skating Rink - 1880 Day Blvd -> Murphy Skating Rink - 1880 Day Blvd, con un total de 15 viajes registrados.
6. Ruta: MIT at Mass Ave / Amherst St -> Central Square at Mass Ave / Essex St, con un total de 14 viajes registrados.
7. Ruta: MIT at Mass Ave / Amherst St -> MIT Vassar St, Viajes registrados: 14
8. Ruta: B.U. Central - 725 Comm. Ave. -> Commonwealth Ave at Agganis Way, con un total de 13 viajes registrados.
9. Ruta: Central Square at Mass Ave / Essex St -> MIT at Mass Ave / Amherst St, con un total de 13 viajes registrados.
10. Ruta: Harvard Square at Mass Ave/ Dunster -> Harvard University Radcliffe Quadrangle at Shepard St / Garden St, con un total de 13 viajes registrados.

A continuación, hemos representado las rutas más frecuentadas en un mapa de Boston con capas de calor para identificar las estaciones que presentan mayor número de viajes.

Figura 27: Mapa rutas



Fuente: elaboración propia

Como se puede observar en el gráfico, las estaciones más populares se encuentran bastantes cercanas unas de otras lo que confirma la relevancia del factor de densidad de estaciones para el éxito del sistema de bicicletas compartidas.

Para elaborar este gráfico, se ha utilizado la librería Folium, especializada en la visualización geoespacial de datos (Tiempo & Data, 2022). Dentro de esta librería, se han importado también las funciones de Heatmap, para agregar capas de calor; PolyLine, para agregar polilíneas; CircleMarker, para añadir marcadores circulares. Se han identificado las 10 rutas más realizadas mediante el análisis de los datos de los viajes de bicicletas que requieren mantenimiento, agrupando los registros por las estaciones de inicio y fin.

A continuación, mediante la función folium.Map(), se ha creado un mapa de Boston asignando como valores de entrada su latitud, 42.3601, y longitud, -71.0589. Posteriormente, se ha agregado una capa de calor mediante la función HeatMap() para mostrar la densidad de viajes de bicicletas de mantenimiento. Finalmente, se ha iterado sobre las 10 rutas más realizadas, creando polilíneas que conecta las estaciones de inicio y fin de cada ruta. Además, se han agregado también marcadores circulares en las ubicaciones de inicio y fin de cada ruta, con colores verdes para la estación de inicio y rojos para la estación de fin.

4.2 Pregunta 2: ¿Existe relación entre distancia entre estaciones de Blue Bikes en Boston y uso del servicio por parte de los usuarios?

Antes de empezar a estudiar la relación entre la distancia entre estaciones de Blue Bikes en Boston y uso del servicio por parte de los usuarios, vamos a analizar cómo se comporta y distribuye la variable de 'distance_haversine'. Primero, usamos la función *.describe()* para ver un resumen estadístico de la misma.

Tabla 7: estadísticos distancia Haversine

Resumen estadístico	
Recuento	2.934.378
Media	1.900
Desviación típica	1.386
Mínimo	0
25%	933
50%	1.563
75%	2.545
Máximo	37.393

Fuente: elaboración propia

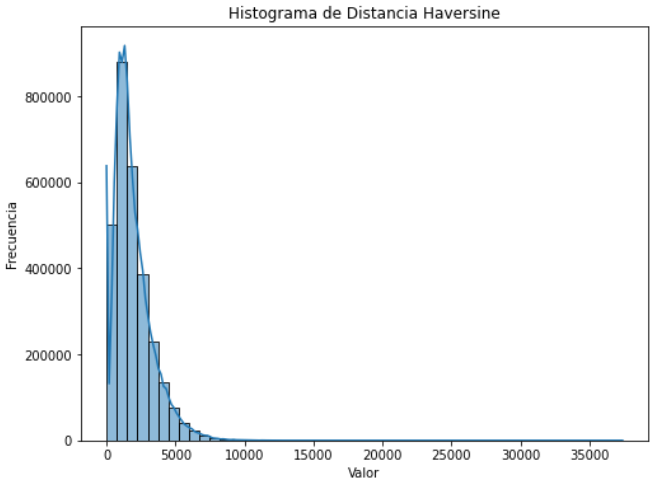
De acuerdo con los datos analizados, se puede concluir que, en promedio, las estaciones de Blue Bikes están separadas por una distancia aproximada de 1.900 metros. Sin embargo, la desviación estándar de 1.386 indica una dispersión considerable alrededor de esta media, lo que sugiere una variabilidad significativa en las distancias entre estaciones.

Es importante destacar que la presencia de un valor mínimo de 0 metros y un máximo de 37.393 metros plantea interrogantes sobre la calidad de los datos. Es posible que el valor mínimo se deba a registros erróneos, como bicicletas que fueron sacadas o devueltas en la misma estación sin realizar un viaje real. Por otro lado, el valor máximo podría representar un outlier, ya que parece irrealista en comparación con la distribución general de los datos.

Al observar los percentiles, se observa que el 25% de los viajes tienen una distancia de 933 metros o menos, el 50% tienen una distancia de 1.563 metros o menos y el 75% tienen una distancia de 2545 metros o menos. Estos valores proporcionan una idea de la distribución de las distancias entre estaciones y refuerzan la posible presencia de outliers, especialmente en el extremo superior de la distribución.

A continuación, vamos a utilizar un histograma para visualizar la distribución de los valores de la variable "distance_haversine". Esto nos permitirá identificar posibles outliers de manera más clara y precisa.

Figura 28: histograma distancia Haversine



Fuente: elaboración propia

Efectivamente, podemos observar una distribución sesgada a la derecha, donde la mayoría de los valores se concentran en el rango entre 0 y 5.000 metros. Esto sugiere que la mayoría de los viajes realizados con Blue Bikes presentan distancias cortas. Sin embargo, hay pocos valores en la cola de la derecha, a partir del valor 5.000 metros hasta alrededor de 35.000 metros, lo que indica la presencia de viajes excepcionalmente largos, posiblemente outliers.

Para analizar y tratar con los outliers, hemos usado la función *zscore* de la biblioteca *scipy.stats* para calcular la puntuación z de cada dato de la variable distance_haversine. Esto nos indicará a cuántas desviaciones estándar por encima o por debajo de la media del conjunto de datos se encuentra cada punto de datos específico. A continuación, se ha establecido un umbral de 3 y se han identificado como outliers todos aquellos datos cuya puntuación Z absoluta sea superior a este valor (Zulmuthi, 2022). Este umbral ha sido elegido debido a que en una distribución normal aproximadamente el 99.7% de los datos están dentro de más o menos 3 desviaciones estándar de la media. Como resultado, hemos obtenido que existen 43.362 outliers en la base de datos. Para mitigar el impacto de estos valores atípicos, hemos aplicado el método de winsorización. Esto implica ajustar los valores atípicos por el valor correspondiente al percentil 99.7%, estableciendo los límites inferiores y superiores en 0.25% (0.0025).

Figura 29:Boxplot Antes de Wins. distancia

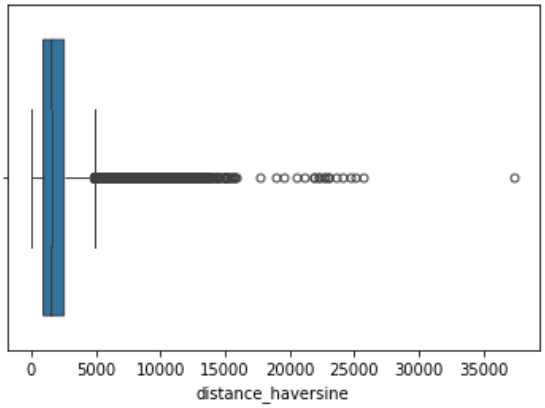
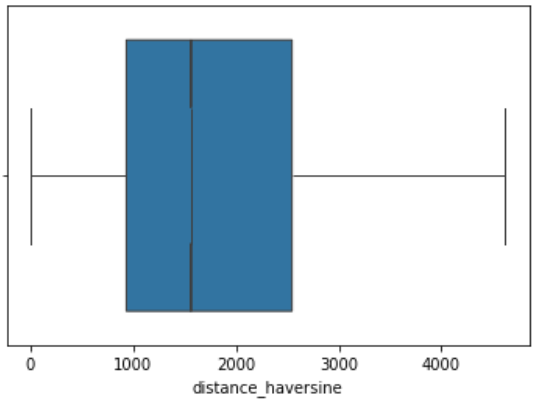
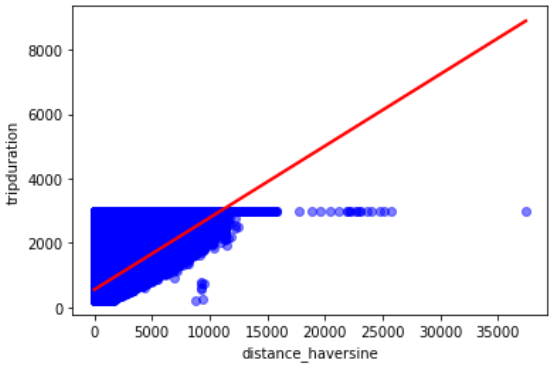


Figura 30:Boxplot Después de Wins. distancia



A continuación, hemos elaborado un gráfico de dispersión con la variable 'tripduration' en el eje 'y' y la variable 'distance_haversine' en el eje 'x' con el objetivo de analizar la relación entre estas dos variables. Al emplear la función *sns.regplot* de la biblioteca Seaborn, hemos trazado además una línea de regresión para modelar la tendencia general que siguen los datos en el gráfico de dispersión. Este enfoque nos permite visualizar cómo la duración del viaje varía en función de la distancia existente entre las estaciones de inicio y fin, proporcionando información relevante sobre los patrones de uso de los usuarios de Blue Bikes en Boston.

Figura 31: Matriz de dispersión



Fuente: elaboración propia

El gráfico de dispersión con la línea roja diagonal sugiere una relación positiva entre las variables 'distance_haversine' y 'tripduration' lo que significa que a medida que aumenta la distancia entre las estaciones de inicio y fin, también se incrementa la duración del viaje. La concentración de puntos azules en la esquina inferior izquierda indica que hay muchos viajes con distancias pequeñas y, por lo tanto, duraciones cortas. Sin embargo, algunos puntos se extienden hacia la derecha del gráfico, indicando que hay

algunos viajes con distancias más largas y, por lo tanto, duraciones más largas, pero no tantos como los viajes más cortos.

Con el objetivo de representar y analizar mejor la relación entre la distancia haversine entre estaciones y el uso del servicio bikesharing, se han clasificado los valores de esta variable en tres grupos: 'Distancia Corta', entre 0 y 1.000 metros; 'Distancia Media', entre 1.000 y 2.000 metros y 'Distancia Larga', entre 2.000 y 40.000 metros. Se ha agregado una nueva columna en el dataframe con esta información, denominada 'Grupo_distancia'. Este proceso se llevó a cabo utilizando la función *pd.cut* de pandas, que permite dividir una columna numérica en intervalos y asignar etiquetas a cada intervalo. De esta manera, hemos creado una variable categórica que clasifica las distancias entre estaciones.

A continuación, se muestran los cinco primeros valores del dataframe con esta nueva columna incorporada.

Tabla 8: grupos distancia

	ID Estación inicio	Distancia Haversine	Grupo_distancia
0	91	2049	Larga
1	370	3206	Larga
2	46	530	Corta
3	178	1123	Media
4	386	1115	Media

Fuente: elaboración propia

A continuación, se ha calculado tanto la duración promedio de los viajes como el número total de viajes realizados por cada grupo.

Tabla 9: viajes grupos distancia

Grupo Distancia	Duración Promedio de viajes	Número de viajes
Distancia Corta	1.036	802.020
Distancia Media	815	1.050.050
Distancia Larga	1.524	1.082.308

Fuente: elaboración propia

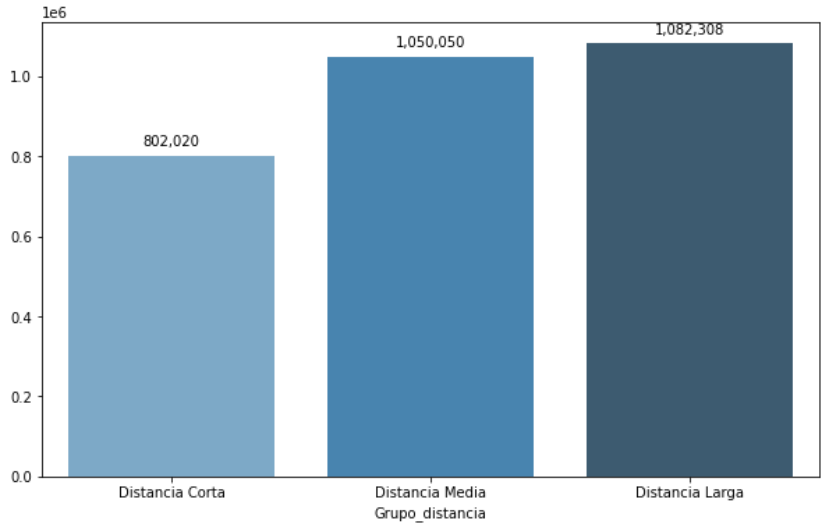
Los viajes de distancia corta tienen una duración promedio de 1.036 segundos, lo cual es mayor que la duración promedio de los viajes de distancias medias. Esto sugiere que los usuarios suelen tomarse más tiempo para disfrutar del paseo en distancias cortas, a diferencia de los viajes de distancias medias, donde es probable que pedaleen más rápido o hagan menos paradas intermedias. Por otro lado, los viajes de distancias largas

presentan el promedio de duración de los viajes más alto, con 1.524 segundos, lo que sugiere que los usuarios tardan más tiempo en recorrer esas distancias mayores y quizás van más relajados.

En cuanto al número total de viajes realizados, el grupo de distancia corta tiene el menor número de viajes, con 802.020 viajes. Esto puede indicar que, aunque los viajes cortos son comunes, pueden no ser tan frecuentes como los de distancias medias o largas. Por otro lado, el grupo de distancia larga tiene el mayor número de viajes, con 1.082.308 en total. Esto sugiere que los viajes largos, que van desde 2.000 hasta 40.000 metros, son los más comunes para el uso de bicicletas compartidas.

En el gráfico siguiente, se pueden visualizar los viajes realizados por cada grupo de distancia:

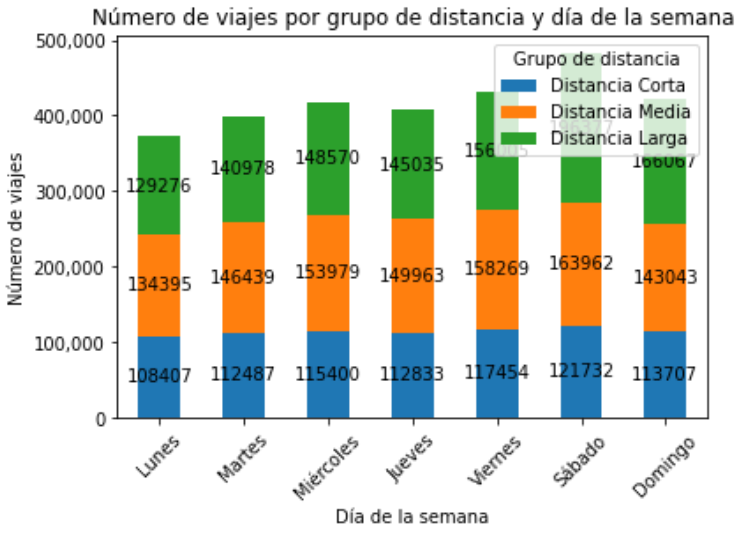
Figura 32: viajes por distancia



Fuente: elaboración propia

A continuación, se ha realizado un análisis del número de viajes según el día de la semana y el grupo de distancia. Se ha observado que el sábado es el día con mayor actividad de viajes, registrando la mayor cantidad tanto de viajes de corta, media como larga distancia. En contraste, el lunes presenta el menor número de viajes en general.

Figura 33: viajes por grupo de distancia y día de la semana

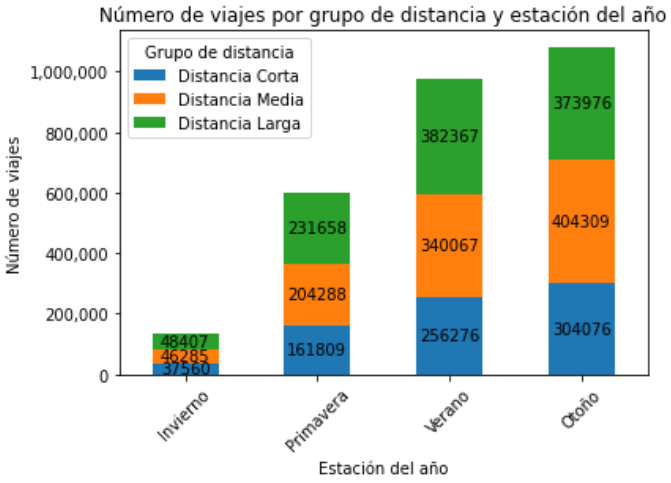


Fuente: elaboración propia

En cuanto a la distribución de los viajes por distancia durante la semana, se destaca que los viajes de corta distancia son predominantes los sábados, seguidos por los viernes y miércoles en orden. Este patrón se repite para los viajes de distancia media. Es estos dos grupos, el miércoles presenta una actividad considerable, aunque en menor medida que los viernes y sábados lo que podría deberse a factores, como compromisos laborales, actividades de ocio o recados diarios que requieren desplazamientos dentro de la ciudad. Por otro lado, los viajes de larga distancia son más comunes los sábados, seguidos por los domingos y viernes. Este aumento de los viajes de larga distancia durante los fines de semana sugiere que las personas podrían estar aprovechando este período para realizar planes o actividades que requieren desplazamientos más largos como rutas turísticas.

Por otro lado, hemos querido analizar el número de viajes por grupo de distancia y estación del año. En el gráfico, se puede observar que el otoño es la estación en la que más viajes se realizan, seguido por el verano. Este hallazgo coincide con los resultados previamente obtenidos en el trabajo, donde se destacó que septiembre y octubre son los meses con mayor actividad de viajes.

Figura 34: viajes por grupo de distancia y estación del año



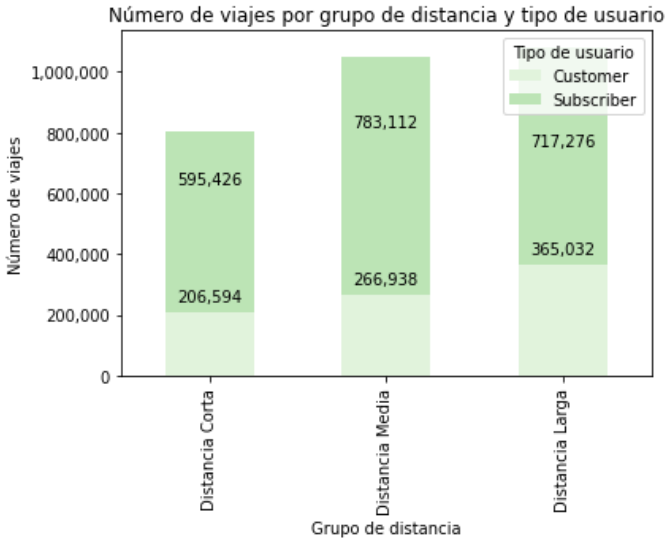
Fuente: elaboración propia

Este aumento de la actividad podría atribuirse a varios factores. Por un lado, en otoño las temperaturas son más moderadas en comparación con el verano, lo que hace que sea más cómodo para las personas realizar actividades al aire libre, como viajar en bicicleta. Además, el comienzo del otoño coincide con el regreso a clases y a la rutina laboral después de las vacaciones de verano, lo que puede aumentar la necesidad de movilidad dentro de la ciudad. En resumen, el otoño se presenta como una época favorable para los desplazamientos en bicicleta, impulsado por el buen clima pues sigue haciendo bueno, pero no hacen temperaturas extremas de calor y la vuelta a las actividades cotidianas.

Por último, se han analizado el número de viajes según la distancia recorrida y el tipo de usuario. En los tres grupos de distancia, los suscriptores son el tipo de usuario predominante. Además, se observa que los usuarios esporádicos tienden a realizar principalmente viajes de distancia larga, seguidos por distancias medias y, en menor medida, cortas.

Por otro lado, los suscriptores tienen una distribución diferente: realizan principalmente viajes de distancia media, seguidos por distancias largas y, por último, cortas. Esta diferencia puede atribuirse a los distintos comportamientos de cada grupo de usuarios. Los usuarios esporádicos, posiblemente visitantes o turistas, pueden optar por viajes más largos para explorar la ciudad o realizar desplazamientos entre puntos de interés. Sin embargo, los suscriptores pueden utilizar el servicio para trayectos diarios más comunes, como ir al trabajo o a actividades regulares, que suelen tener una distancia media.

Figura 35: viajes por grupo de distancia y tipo de usuario



Fuente: elaboración propia

4.3 Pregunta 3. ¿Cómo afecta la hora del día, la latitud de origen y la latitud de destino a la duración de los viajes en bicicleta compartida?

La siguiente pregunta se ha resuelto mediante la aplicación del algoritmo de aprendizaje automático conocido como random forest. Esta elección se fundamenta en la serie de beneficios que proporciona este método. En primer lugar, el random forest ayuda a mitigar el riesgo de sobreajuste gracias a la combinación de múltiples árboles de decisión no correlacionados. Esta diversificación contribuye a reducir la varianza general y el error de predicción del modelo. Además, el algoritmo random forest puede ser utilizado tanto para problemas de regresión como de clasificación. Por último, otra ventaja significativa radica en su capacidad para evaluar la importancia relativa de cada variable en el modelo (What is random forest?, s/f).

En la siguiente tabla, se describen las funciones que se han utilizado para poder realizar el modelo de bosque aleatorio (“random forest”):

Tabla 10: librerías modelo random forest

Funciones	Descripción
sklearn.model_selection.train_test_split	Para dividir el dataframe en datos de entrenamiento y prueba
sklearn.ensemble.RandomForestRegressor	Aplica el algoritmo de Random Forest para problemas de regresión
sklearn.preprocessing.StandardScaler	Para estandarizar las variables eliminando la media y escalando a la varianza unitaria
sklearn.metrics.mean_squared_error	Calcula el error cuadrático medio entre las etiquetas verdaderas y las predicciones
sklearn.ensemble.RandomForestClassifier	Aplica el algoritmo de Random Forest para problemas de clasificación
sklearn.metrics.roc_curve	Calcula la curva ROC (Receiver Operating Characteristic) para un modelo de clasificación
sklearn.metrics.roc_auc_score	Calcula el área bajo la curva ROC (AUC) para un modelo de clasificación

Fuente: elaboración propia

Se pretende ajustar un modelo de regresión que permita predecir la duración de un viaje en función de las variables: 'start station latitude', 'end station latitude', 'HoradeldíaInicio','HoradeldíaParada','distance_haversine'.

Las variables 'HoradeldíaInicio' y 'HoradeldíaParada' se han generado a partir de las variables 'starttime' y 'stoptime', respectivamente, utilizando el método *dt.hour()* de la librería pandas datetime.

La elección de estas variables dependientes se basa en la idea de que la distancia entre las estaciones, la ubicación de las mismas y la hora del día pueden tener un impacto importante en cuánto tiempo lleva realizar un viaje en bicicleta compartida. Por lo tanto, al considerar estas variables, esperamos obtener un modelo más preciso y útil para predecir la duración de los viajes.

Con el objetivo de evaluar si el modelo funciona y predice correctamente nuevas observaciones, se han dividido los datos en dos partes: de entrenamiento y datos de prueba o test (de los Santos, 2022).

Se ha reducido el tamaño del conjunto de datos de entrenamiento a 100.000 observaciones para acelerar el proceso de carga, mientras que al conjunto de prueba se le ha asignado el 20% de los datos, lo que equivale a 586.876 observaciones.

A continuación, se ha empleado la función *StandardScaler()* para estandarizar las variables y asegurarnos que todas tienen la misma escala con el objetivo de incrementar la precisión del modelo.

El modelo de bosque aleatorio ha sido creado con la función *RandomForestRegressor()*. Este modelo se ha ajustado con diez árboles y se ha empleado el criterio de error cuadrático medio para evaluar la calidad de las divisiones. No se ha establecido una profundidad máxima para los árboles, lo que permite que se expandan hasta alcanzar el límite de las muestras disponibles. Se ha especificado que solo se considere un predictor en cada división. Además, se han utilizado todos los núcleos disponibles en el sistema para el entrenamiento del modelo y se ha fijado una semilla aleatoria de 123 para garantizar la reproducibilidad de los resultados.

Para entrenar el modelo se ha aplicado la función *.fit()*. Después, se ha evaluado la capacidad predictiva del modelo con el conjunto de test mediante la función *.predict()*. Por último, hemos calculado el error cuadrático medio (MSE) entre las respuestas reales

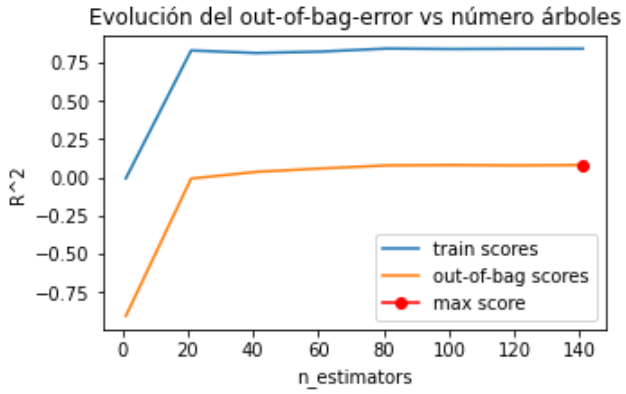
del conjunto de prueba (y_test) y las predicciones realizadas por el modelo (y_pred) para poder analizar la capacidad de predecir nuevas instancias del modelo. El error cuadrático medio ha sido de 1025578106.078.

Esta cifra es muy alta lo que refleja que las predicciones obtenidas por el modelo están muy alejadas de los valores reales, por lo que requiere ajustes para mejorar su rendimiento. Para ello, a continuación, se ha realizado una búsqueda de los hiperparámetros más óptimos mediante estrategias de validación.

Inicialmente, hemos investigado el número óptimo de árboles de decisión. Aunque este hiperparámetro no es crítico, ya que agregar más árboles solo mejora el resultado, su ajuste puede llevar a una pérdida innecesaria de recursos computacionales si se generan más árboles una vez que el resultado se ha estabilizado (Random Forest python, s/f).

Se ha definido un rango de valores para el número de árboles (n_estimators) desde 1 hasta 150 en incrementos de 20. A continuación, se ha generado un bucle para entrenar un modelo con cada valor del rango y extraer su error de entrenamiento y de Out-of-Bag. Cada vez que se entrena un modelo con una cantidad diferente de árboles, se evalúa cómo de bien se ajusta a los datos de entrenamiento (puntaje de entrenamiento) mediante la función *.score()* y cómo de bien podría predecir nuevas observaciones (puntaje OOB). Luego, estos puntajes se comparan para encontrar la cantidad óptima de árboles que debería tener el Bosque Aleatorio ("random forest") para mejorar su rendimiento.

Figura 36: error out-of-bag vs número árboles



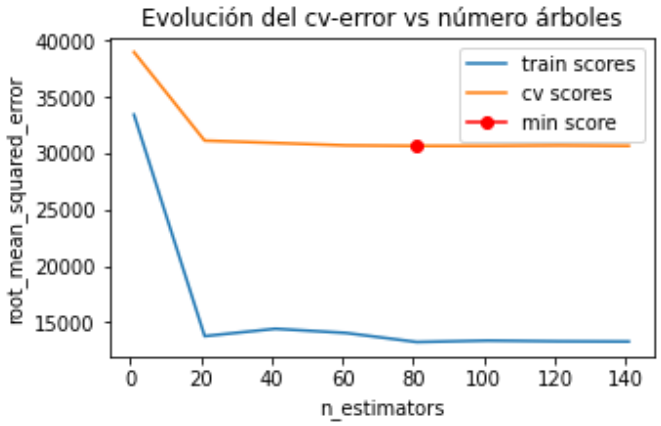
Fuente: elaboración propia

En este gráfico, se muestra la evolución de los errores en nuestro modelo a medida que aumentamos el número de árboles. El valor óptimo de número de árboles que se ha obtenido es de 141, reflejado con el punto rojo del gráfico. La mejora del modelo se estabiliza a partir de los 20 árboles de decisión. La línea azul (train scores) indica cómo

se comporta nuestro modelo cuando se entrena con los datos de entrenamiento en comparación con la línea naranja que muestra los errores *out-of-bag* en los que incurre nuestro modelo al predecir nuevas observaciones. El error out-of-bag sirve como estimación del error de test. La línea azul está significativamente por encima de la línea naranja lo que sugiere que el modelo está "sobreajustando" los datos de entrenamiento, es decir, está aprendiendo demasiado de esos datos específicos y puede tener dificultades para generalizar bien con nuevos datos. Queremos que las dos líneas estén lo más cerca posible para tener un buen equilibrio entre precisión y capacidad de generalización (Grid search de modelos Random Forest con out-of-bag error y early stopping, s/f).

En segundo lugar, se ha creado un bucle para entrenar un modelo con cada valor de n_estimators, cuyo rango de valores va de 1 a 150 en incrementos de 20, y se ha extraído su error de entrenamiento y de k-cross-validation (Random Forest python, s/f). Para ello, se crearon dos listas vacías: 'train_scores', para los errores de entrenamiento y 'cv_scores' para los errores de validación cruzada. Para evaluar los errores de entrenamiento, se empleó el método del error cuadrático medio que calcula la distancia al cuadrado entre los valores reales y predichos por lo que cuanto más cercano a cero mejor. Por otro lado, en el caso de los errores de validación cruzada, se utilizó la métrica de error negativo de la raíz cuadrada media y se dividió el conjunto de entrenamiento en 5 partes. El método de la raíz del error cuadrático medio sirve para identificar las deviaciones de los valores predichos respecto de los reales y poder analizar el impacto de los errores. Cuando más cercano a 0, mejor será el ajuste (Madrigal, 2022).

Figura 37: cv-error vs número árboles

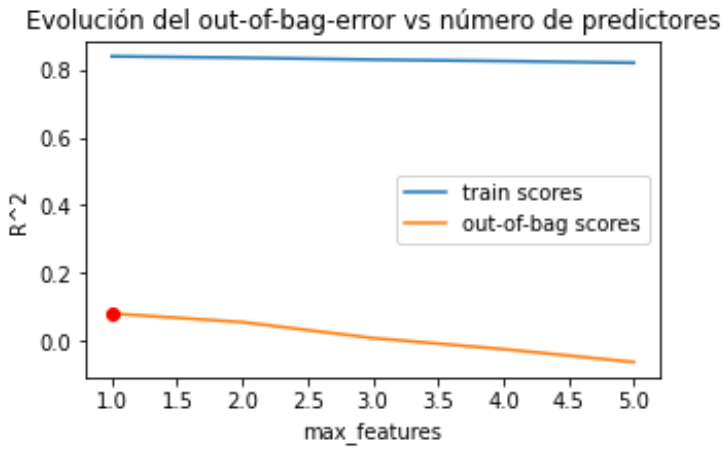


Fuente: elaboración propia

El punto rojo en el gráfico indica el número óptimo de árboles para el cual el modelo alcanza el mínimo error de la raíz cuadrada de la media, situado en 81 árboles. En el gráfico, se muestra la evolución del RMSE (Root Mean Square Error) a medida que se van añadiendo más números de árboles. La línea azul representa los errores de entrenamiento (train scores), mientras que la línea naranja muestra los errores de validación cruzada (cv scores). Ambos errores disminuyen a medida que aumenta el número de árboles. No obstante, el error de entrenamiento es más bajo que el error de validación cruzada, lo que sugiere que el modelo puede estar sobre ajustando los datos de entrenamiento. También, se puede destacar que, a partir de 20 árboles de decisión, ambos errores parecen estabilizarse, reflejando que la adición de árboles no implica una mejora significativa en el rendimiento del modelo, sino que puede llevar a pérdida de recursos computacionales.

En tercer lugar, se ha ajustado el hiperparámetro de número de predictores (máx_features) que índice cuanto están de correlacionados los árboles entre sí (Random Forest python, s/f). Para ello, se ha creado un bucle para entrenar un modelo con cada valor del rango de número de predictores y extraer su error de entrenamiento y de Out-of-Bag. Para ello, se ha empleado el método de error cuadrático que hemos explicado previamente.

Figura 38: out-of-bag error vs número de predictores



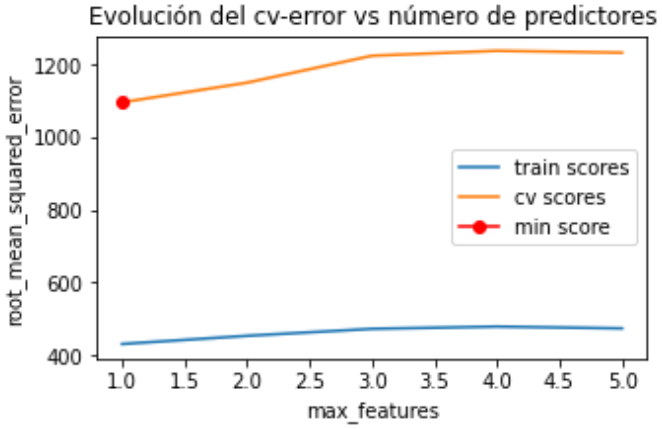
Fuente: elaboración propia

En el gráfico, se observa que la línea azul con los puntajes de entrenamiento se encuentra significativamente más por encima que la línea naranja, que representa los puntajes "out-of-bag". Esto indica que el modelo está sobreajustando los datos de entrenamiento, ya que logra un mejor rendimiento en estos datos que en los datos nuevos

(out-of-bag). La línea naranja disminuye a medida que aumenta el número de predictores lo que sugiere que aumentar el número de predictores puede aumentar el sobreajuste del modelo, lo que resulta en un peor rendimiento en la predicción de las nuevas instancias. Por lo tanto, nos interesa más seleccionar un valor más bajo de número de predictores. En el gráfico, se puede observar que el valor óptimo de número de predictores es 1.

En cuarto lugar, se ha evaluado la evolución de los errores de k-cross-validation al aumentar el número de predictores. Al igual que en el caso anterior, se ha obtenido que el número óptimo de predictores es 1, donde se obtiene el RMSE mínimo en la validación cruzada.

Figura 39: cv-error vs número de predictores



Fuente: elaboración propia

Parece que ambas líneas, tanto la del conjunto de entrenamiento (azul) como la de validación cruzada (naranja), aumentan a medida que aumenta el número de predictores (max_features). Esto sugiere que utilizar más predictores no mejora el rendimiento del modelo y puede conducir a un sobreajuste.

Este análisis individual que se ha realizado de los hiperparámetros ha ayudado a interpretar mejor el impacto de cada uno en el modelo. No obstante, es importante destacar que esta búsqueda de hiperparámetros óptimos se debe realizar de forma conjunta pues hay que tener en cuenta que cada hiperparámetro interacciona con los demás. Por ello, vamos a analizar varias combinaciones de hiperparámetros mediante los métodos grid search y bayesian search (Random Forest python, s/f).

En el método de grid search, mediante la función *ParameterGrid()* hemos definido el siguiente conjunto de hiperparámetros con diferentes combinaciones de valores: 145 árboles de decisión, un número de predictores que varía en cada división del árbol (1, 5 y 7) y diferentes profundidades máximas para los árboles (None, 10 y 20). Se ha creado un

bucle para ajustar el modelo con cada combinación de valores. En este primer caso, hemos realizado el grid search basado en el out-of-bag error.

El resultado ha sido el siguiente:

Tabla 11: grid search out-of.bag-error

	oob_r2	max_depth	max_features	n_estimators
0	0.081453	NaN	1.0	145.0
3	0.074555	10.0	1.0	145.0
6	-0.073903	20.0	1.0	145.0
7	-0.042257	20.0	5.0	145.0

Fuente: elaboración propia

Como podemos observar, la mejor combinación de parámetros es la que tiene 1 predictor, 145 árboles de decisión, ningún límite de profundidad máxima ya que es la que mayor valor de coeficiente de determinación (Out-of-Bag R-squared) presenta siendo este de 0.081453. El coeficiente de determinación es la métrica de evaluación utilizada para determinar la calidad del modelo pues indica la capacidad de generalización de este. Cuanto más cercano a 1 esté el valor, mejor ajuste del modelo habrá.

En el segundo caso, hemos llevado a cabo un grid search basado en validación cruzada para poder realizar la búsqueda de hiperpárametros considerando los datos no observados en el modelo. Así, se podrá evitar el sobreajuste del modelo a los datos de entrenamiento que ocurriría si se emplean los mismos datos para ajustar el modelo y para evaluarlo (Random Forest python, s/f).

Para ello, aplicamos la validación cruzada con los datos de entrenamiento, dividiendo el conjunto en 5 grupos y ajustando el modelo 3 veces, con un grupo diferente como conjunto de validación en cada iteración. Este enfoque nos reduce el riesgo de sobreajuste y proporciona una estimación más precisa del rendimiento del modelo en nuevos datos (Random Forest python, s/f). En este caso, también se ha obtenido que la mejor combinación de hiperparámetros es la que tiene 1 predictor, 145 árboles de decisión y ningún límite de profundidad máxima ya que tiene el menor valor de error cuadrático medio. Esto significa que este conjunto de hiperparámetros produce el modelo con la mejor capacidad predictiva en el conjunto de datos de prueba, según la métrica de error cuadrático medio.

Tabla 12: grid search validación cruzada

param_max_depth	param_max_features	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score

0	None	1	145	-28789.17	14076.73	-11909.96	1303.12
6	10	1	145	-28877.89	14158.80	-15088.90	1920.73
3	3	1	145	-29750.02	14560.72	-32029.03	3539.06
4	3	5	145	-31670.38	14363.90	-28217.99	3698.84

Fuente: elaboración propia

Por último, se ha realizado la optimización bayesiana de hiperparámetros en el que la métrica de validación del modelo (RMSE, AUC, precisión...) es la función objetivo del modelo con el objetivo de enfocar la búsqueda en cada iteración a las regiones de mayor interés (Random Forest python, s/f). Hemos obtenido como mejores hiperparámetros la combinación de 500 árboles de decisión, profundidad máxima de 8, un número mínimo de muestras para la división del nodo de 19, un mínimo de muestras para estar en un nodo hoja de 75, 0.57 predictores, y un parámetro de complejidad de 0.4. El mejor valor que nos han proporcionado estos ha sido de -29776.07. Y se ha obtenido un RMSE de 32538.24.

4.4 Pregunta 4. ¿Cómo influyen la hora del día y la ubicación en la probabilidad de que un viaje en bicicleta compartida sea corto (menos de 29 minutos) o largo (más de 29 minutos)?

La siguiente pregunta se ha resuelto mediante la creación de un modelo de regresión logística. Primero, se calculó el promedio de la duración de los viajes en bicicleta, obteniendo un valor de 1785.84 segundos (aproximadamente 29.8 minutos), que se utilizó como umbral para clasificar los viajes en cortos y largos. La elección de este umbral se justifica debido a que la media es una medida representativa del tiempo que la mayoría de los usuarios de bicicletas comparten. Además, alrededor de los 30 minutos es un tiempo comúnmente utilizado en sistemas de bicicletas compartidas como límite de tiempo para evitar cargos adicionales por exceder el tiempo incluido en la tarifa básica. Por ejemplo, en el caso del sistema de bicicletas compartidas BiciMAD, si un viaje excede los 30 minutos, se aplicará un cargo adicional de 0.50 euros por cada media hora adicional (Novedades en bicimad: llega la tarifa plana y la gratuidad se extiende todo el mes de enero, s/f).

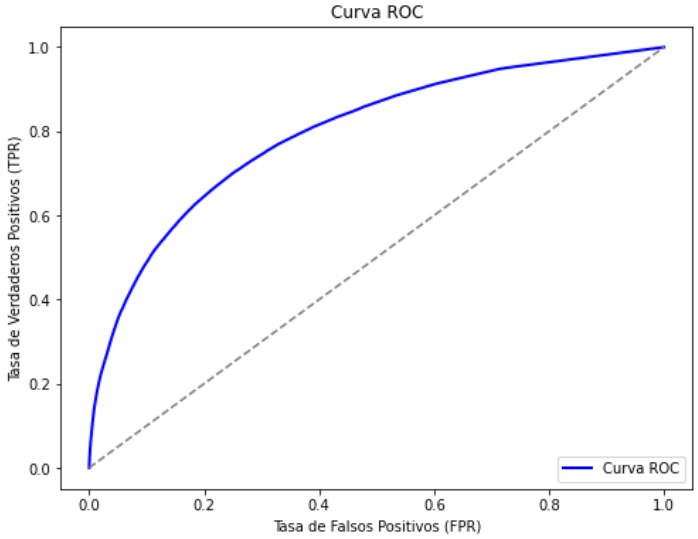
Posteriormente, se creó una variable binaria llamada 'Duración Larga', donde los viajes con una duración mayor que el umbral se clasificaron como '1', indicando duración

larga, mientras que los que no superaron el umbral se clasificaron como '0', indicando duración corta.

El modelo de clasificación utilizado fue un bosque aleatorio (RandomForestClassifier), con el objetivo de predecir si un viaje tiene una duración larga o no, considerando las variables independientes 'start station latitude', 'end station latitude', 'HoradeldíaInicio', y 'distance_haversine'. Para agilizar la carga y el proceso de entrenamiento, se redujo el tamaño del conjunto de entrenamiento a 100,000 datos, mientras que se retuvieron 586,876 datos para el conjunto de prueba.

Posteriormente, se evaluó el modelo calculando el área bajo la curva ROC (AUC), obteniendo un valor de 0.7964. Este valor indica la capacidad del modelo para distinguir entre clases positivas y negativas, donde un valor más cercano a 1 indica un mejor rendimiento del modelo en términos de clasificación.

Figura 40: Curva ROC



Fuente: elaboración propia

Por último, hemos realizado una matriz de confusión para visualizar de forma detallada el rendimiento del modelo de clasificación. En este caso, tenemos cuatro valores:

- Verdaderos Negativos (TN): El modelo predijo en 484,091 casos que un viaje tenía una duración corta (0), y el viaje realmente tenía una duración corta.

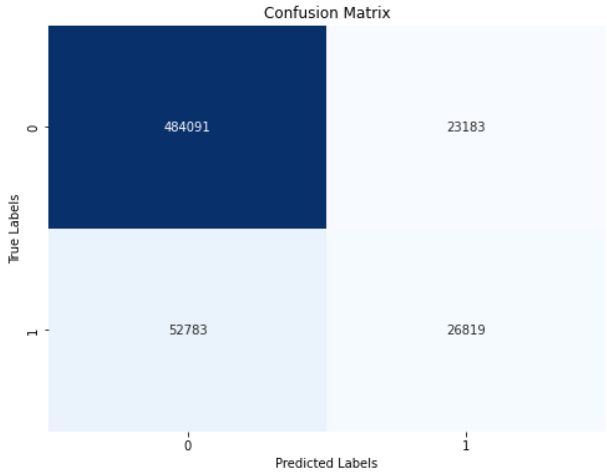
- Falsos Negativos (FN): El modelo predijo en 52,783 que un viaje tenía una duración corta (0), pero en realidad tenía una duración larga (1).

- Falsos Positivos (FP): El modelo predijo en 23,183 casos que un viaje tenía una duración larga (1), pero en realidad tenía una duración corta (0).

- Verdaderos Positivos (TP): El modelo predijo en 26,819 casos que un viaje tenía una duración larga (1), y el viaje realmente tenía una duración larga.

En este análisis, observamos un número relativamente alto de falsos negativos, lo que indica que el modelo tiende a subestimar la duración de los viajes largos, clasificándolos incorrectamente como viajes cortos. Por otro lado, el número de falsos positivos es significativamente menor, lo que sugiere que el modelo tiene una tendencia relativamente menor a sobreestimar la duración de los viajes cortos, clasificándolos incorrectamente como viajes largos. Por último, el modelo clasifica muy bien los viajes que tienen duración corta.

Figura 41: matriz de confusión



Fuente: elaboración propia

5. Consultando datos mediante modelo LLM

Los modelos lingüísticos de gran tamaño (LLM) tienen la capacidad de procesar y analizar grandes volúmenes de datos en un periodo de tiempo reducido. Esta velocidad y eficiencia en la ejecución resultan muy útiles en situaciones que requieren la evaluación de varios criterios en bases de datos extensas o en el procesamiento de varios documentos extensos (Ferrer-Benítez, 2022).

Dentro de estos modelos, destaca la herramienta ChatGPT, lanzada en 2022, la cual representó un avance disruptivo en el ámbito de la inteligencia artificial. Esta herramienta se fundamenta en una serie de modelos GPT y ha sido especialmente adaptada para el diálogo, lo que facilita la interacción con sus usuarios. Los modelos LLM poseen la capacidad de aprender en contexto, lo que les permite predecir el siguiente token en una secuencia basándose en los tokens previos. Esta capacidad les permite analizar la información en detalle y comprenderla, lo que a su vez les capacita para generar texto relevante y coherente, adaptado al contexto específico (Leiva, Vigneau & Sepúlveda, ,2023).

En este trabajo de investigación, se ha incorporado una sección dedicada a los modelos LLM utilizando la librería PandasAI de Python. Esta librería emplea modelos generativos de inteligencia artificial que simplifican el análisis de los datos y permiten formular preguntas a través de un lenguaje natural (Datacamp.com, s/f).

La librería PandasAI ofrece una variedad de funcionalidades, entre las que se incluyen la capacidad de realizar consultas en lenguaje natural, la visualización de datos mediante la generación de gráficos y diagramas, la limpieza de datos para abordar valores faltantes, la mejora de la calidad de los datos a través de la ingeniería de características, y la conexión a diversas fuentes de datos como archivos CSV o bases de datos MySQL, entre otras opciones disponibles (Venturi, s/f). En definitiva, esta librería combina la potencia de la librería de manipulación de datos, Pandas, con modelos de lenguaje (LLM) de última generación como GPT-3.5 de OpenA, permitiendo a los usuarios realizar análisis de datos de forma conversacional (Atiq, 2023).

Para comenzar, instalamos la librería utilizando la función *pip install pandasai*. Luego, obtenemos una clave de API de OpenAI y la almacenamos como una variable de entorno mediante la función *os.environ[]*, en formato de cadena. Para obtener la clave API gratuita, nos registramos en el sitio web https://pandabi.ai. Posteriormente, importamos las siguientes librerías: `os`, que nos permite interactuar con el sistema operativo; `pandas`, para el manejo de datos; y `agent`, un objeto de PandasAI que nos permite recibir una función de ejemplo y ejecutarla como una decisión del agente.

Después, creamos una variable llamada 'agent', la cual resulta de llamar al proceso `agent` y pasarle directamente el DataFrame `combined_df`. A continuación, utilizando la función `agent.chat()`, formulamos una serie de preguntas de los datos directamente.

A continuación, se muestran una serie de preguntas realizadas directamente al modelo LLM. En primer lugar, se consultó por la fecha del año 2021 que registró el mayor número de viajes y la cantidad exacta de estos. La respuesta obtenida fue el 25 de septiembre de 2021, con un total de 18,178 viajes. Posteriormente, se indagó sobre el día de la semana que contaba con la mayor cantidad de viajes, obteniendo como respuesta el número 5, correspondiente al sábado, dado que Python considera los días de la semana del 0 al 6. Además, se interrogó sobre la bicicleta que realizó más viajes, siendo identificada como aquella con el ID 67, corroborando previamente los hallazgos de la investigación. Asimismo, se consultó sobre el tipo de usuario que efectuaba más viajes, siendo la respuesta los suscriptores, lo que confirma también el análisis realizado previamente.

Figura 42: Ejemplo PandasAI

```
agent.chat('Which date of the year has recorded the highest number of trips and how many were there?')
```

```
'The date with the highest number of trips is 2021-09-25 with 18178 trips.'
```

```
agent.chat('What day of the week are most trips made?')
```

```
'The day of the week with the most trips is: 5'
```

```
agent.chat('Which are the bike that have done more trips?')
```

```
'The bike that has done the most trips is 3489.'
```

```
agent.chat('What are the IDs of the start stations that have issued the most trips?')
```

```
'The start station IDs that have issued the most trips are: 67'
```

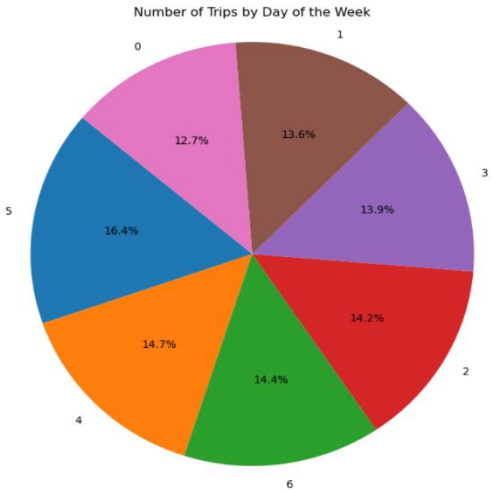
```
agent.chat('What type of user does more trips?')
```

```
'The user type that does the most trips is: Subscriber'
```

Fuente: elaboración propia

Finalmente, importamos la función *SmartDataframe* con el objetivo de añadir características conversacionales al conjunto de datos (Atiq, 2023). Posteriormente, utilizando la función *.chat(),* solicitamos que nos genere un gráfico de tarta que muestre el número de viajes realizados por día de la semana. El programa nos devuelve como resultado el siguiente gráfico:

Figura 43: Gráfico de tarta PandasAI



Fuente: elaboración propia

Se puede acceder al código de esta sección mediante el siguiente link: https://github.com/itscarmengo/PandasAI_BlueBikes_Boston2021/blob/fad1864affc931dd29d4eb0a264e53b483aa475d/PandasAI_Boston2021.ipynb

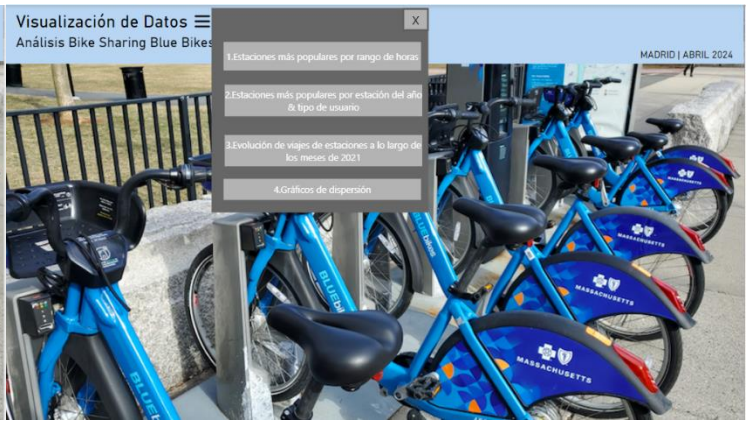
6. Visualización de datos mediante PowerBI

En este apartado, se ha profundizado en el análisis de los datos vinculados a las estaciones de inicio y parada, utilizando la herramienta Power BI para mejorar la comprensión de los mismos. Esta plataforma posibilita la creación de informes interactivos y la visualización de datos de manera intuitiva. Gracias a estas capacidades, se logra obtener una representación visual clara de las tendencias y patrones asociados con las estaciones de inicio y parada, lo que facilita la identificación de insights relevantes. Primero, se ha realizo una portada con un índice reducido del contenido:

Figura 44: Portada PoweBI

Fuente: elaboración propia

Figura 45: Índice PoweBI



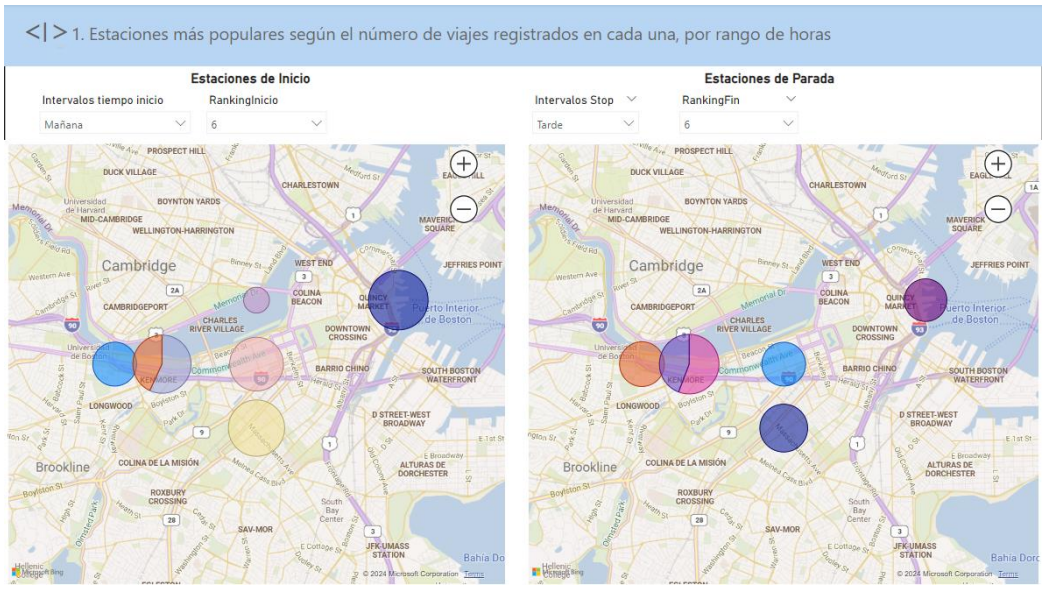
Fuente: elaboración propia

En primer lugar, se ha realizado un análisis de las estaciones más populares en función del número de viajes emitidos, para las estaciones de inicio, o recibidos, en el caso de las estaciones de parada. Este análisis se ha llevado a cabo considerando diferentes intervalos de tiempo: madrugada (de 12 a.m. a 6 a.m.), mañana (de 6 a.m. a 12 p.m.), tarde (de 12 p.m. a 6 p.m.) y noche (de 6 p.m. a 12 a.m.).

Por otro lado, con el propósito de simplificar la visualización del gráfico debido a la gran cantidad de estaciones en el conjunto de datos, se han creado dos variables: una para las estaciones de inicio y otra para las de parada. Estas variables asignan un ranking a las estaciones en función del número de viajes realizados. En este ranking, el número 1 representa las estaciones con más viajes realizados, mientras que el número 10 representa las estaciones con menos viajes.

Los rangos de asignación de ranking se han establecido de la siguiente manera: entre 2 y 7,000 viajes se asigna un 10; entre 7,000 y 14,000, un 9; entre 14,000 y 21,000, un 8; entre 21,000 y 28,000, un 7; entre 28,000 y 35,000, un 6; entre 35,000 y 42,000, un 5; entre 42,000 y 49,000, un 4; entre 49,000 y 56,000, un 3; entre 56,000 y 63,000, un 2; y por último, entre 63,000 y 70,000 viajes, un 1.

Figura 46: Visualización 1 Dashboard

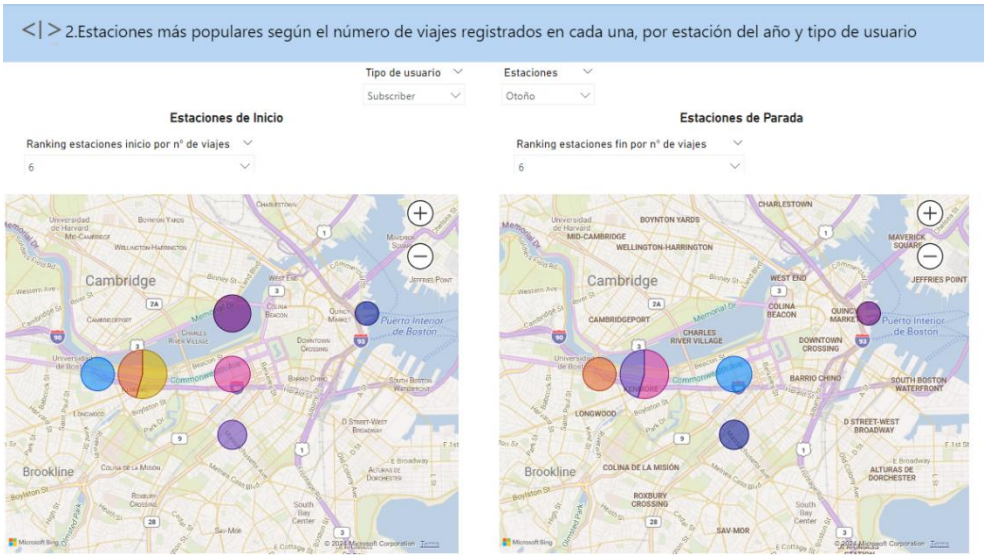


Fuente: elaboración propia

En segundo lugar, se han representado las estaciones más populares, tanto de inicio como de parada, según el número de viajes registrados en cada una. Se aplicaron dos filtros generales para visualizar el análisis en función del tipo de usuario y de la estación del año. Además, se aplicó un filtro adicional basado en el ranking de las

estaciones mencionado previamente para simplificar los gráficos, uno para las estaciones de inicio y otro para las de parada. De esta manera, se permite un análisis dinámico de las estaciones más populares según el número de viajes, considerando la estación del año y el tipo de usuario.

Figura 47: Visualización 2 Dashboard



Fuente: elaboración propia

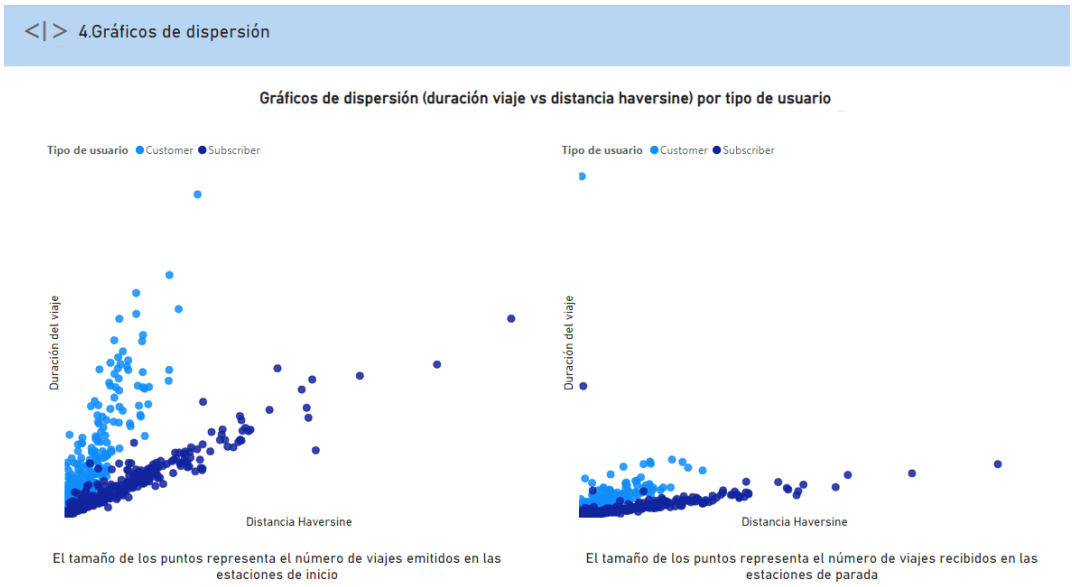
Por último, se llevó a cabo un análisis de la evolución de los viajes emitidos o recibidos en las estaciones a lo largo de los meses de 2021. El primer gráfico muestra la evolución de las estaciones de inicio, aplicando nuevamente el filtro del ranking. El gráfico inferior corresponde a las estaciones de parada, también con su correspondiente filtro. En ambos casos, y como bien hemos confirmado anteriormente en nuestra investigación, septiembre es el mes en el que se realiza el mayor número de viajes.

Figura 48: Visualización 3 Dashboard

Fuente: elaboración propia

Por último, hemos elaborado dos gráficos de dispersión que analizan la relación entre las variables de duración del viaje y distancia haversine según el tipo de usuario (cliente esporádico o suscriptor). En el gráfico de la izquierda, el tamaño de los puntos representa el número de viajes emitidos desde las estaciones de inicio. En el gráfico de la derecha, el tamaño de los puntos representa el número de viajes recibidos en las estaciones de destino.

Figura 49: Visualización 4 Dashboard



Fuente: elaboración propia

Como podemos observar, en el gráfico de la izquierda se muestra una correlación positiva entre ambas variables, lo que indica que cuando aumenta la distancia, también lo hace la duración de los viajes emitidos desde las estaciones de inicio. Los puntos de color azul claro, que representan a los clientes esporádicos, se encuentran por encima de los puntos de color azul oscuro, lo que sugiere que, en promedio, los clientes esporádicos realizan viajes de mayor duración en comparación con los suscriptores. Puede ser que los suscriptores estén más familiarizados con el sistema de bicicletas compartidas y tiendan a utilizarlo para trayectos más cortos y regulares mientras que los clientes esporádicos realizan viajes menos frecuentes, pero de mayor duración para realizar actividades de turismo o para acudir a algún evento ocasional.

Con el siguiente **link**, se puede acceder al reporte completo de PowerBi:

https://app.powerbi.com/view?r=eyJrIjoiZDY0OWQxNWUtODk3Yy00ODA2LWI2ZmMtMWUwMjY3YTNkMmZiIiwidCI6ImJjZDI3MDFjLWFhOWItNGQxMi1iYTIwLWYzZTNiODMwNzBjMSIsImMiOjh9

7. Conclusiones

Este caso de estudio proporciona un análisis del uso del sistema de bicicletas compartidas en la ciudad de Boston durante el año 2021, centrándose en el período postpandemia. Se incluyó un marco teórico que aborda la historia del sistema de bicicletas compartidas y luego se centra en el contexto de Boston. Además, se realizó la adquisición, transformación y limpieza de datos, así como un análisis exploratorio de datos (EDA). A continuación, se presentan las conclusiones derivadas de las preguntas de investigación.

En la primera pregunta de investigación, se identificaron aquellas bicicletas que requieren mantenimiento. Mediante el análisis de los datos de los viajes realizados, se identificaron las bicicletas con mayor número de viajes al mismo tiempo que las estaciones de inicio y fin más frecuentes para estas bicicletas. Los resultados revelaron que las bicicletas con ID 3489, 5615 y 6678 encabezaron la lista de las más utilizadas, lo que sugiere una necesidad de mantenimiento debido a su alto nivel de uso. Por otro lado, las dos estaciones de inicio y fin más frecuentes para estas bicis resultaron coincidir lo que refleja que esas bicicletas están siendo utilizadas frecuentemente entre un conjunto limitado de estaciones. Estas fueron 'MIT at Mass Ave / Amherst St' y 'Central Square at Mass Ave / Essex St', destacando la importancia de estas ubicaciones como puntos clave de actividad en la red de bicicletas compartidas de Boston.

En la segunda pregunta, se analizó la relación entre la distancia entre estaciones de Blue Bikes en Boston y el uso del servicio por parte de los usuarios. En primer lugar, se observó una distribución sesgada a la derecha en las distancias entre estaciones. Por lo que, se realizó un tratamiento de outliers mediante el método de winsorización. Además, se encontró una relación positiva entre la distancia entre estaciones y la duración del viaje, lo que indica que a medida que aumenta la distancia entre estaciones también aumenta la duración del viaje. La clasificación de los viajes en grupos de distancia mostró que los viajes de larga distancia son los más comunes, seguidos por los de distancia media y corta. Además, se identificaron patrones interesantes en la actividad de viajes según el día de la semana y la estación del año, destacando la mayor actividad durante los fines de semana

y en la temporada de otoño. Y, por último, se descubrió que los suscriptores son los usuarios que más viajes realizan para todo tipo de distancia.

En la tercera pregunta de investigación, para comprender cómo la hora del día y la ubicación influyen en la duración de los viajes en bicicleta compartida, se utilizó un algoritmo de aprendizaje automático llamado Random Forest. Se consideraron variables como la latitud de origen y destino, la hora de inicio y parada del viaje, y la distancia entre estaciones. Se dividió el conjunto de datos en entrenamiento y prueba y se ajustó el modelo, pero se observó un alto error de predicción. Se realizaron ajustes en los hiperparámetros del modelo, como el número de árboles y predictores, mediante métodos de validación cruzada y búsqueda grid y bayesiana. Los resultados de la búsqueda bayesiana revelaron que una combinación de 500 árboles de decisión con una profundidad máxima de 8, junto con otros parámetros optimizados, ofreció el mejor rendimiento. Por otro lado, en la búsqueda grid, se identificó una combinación óptima de hiperparámetros que consistía en 1 predictor, 145 árboles de decisión y ninguna limitación en la profundidad máxima.

Por último, en la cuarta pregunta, se analizó cómo la hora del día y la ubicación determinan si un viaje en bicicleta compartida es corto o largo. Para ello, se utilizó un modelo de regresión logística. Se clasificaron los viajes como cortos o largos según si duraban menos o más de 29 minutos. El modelo predijo con éxito si un viaje sería corto o largo con un área bajo la curva ROC de 0.7964. Sin embargo, tiende a subestimar la duración de los viajes largos, lo que se refleja en un número relativamente alto de falsos negativos. Aunque tiene una tendencia menor a sobreestimar los viajes cortos, clasificándolos incorrectamente como largos, clasifica bien los viajes cortos en general.

Cabe destacar que este trabajo se presenta como un proyecto "End to end", ya que abarca los diversos pasos del ciclo de análisis de datos, desde la captura y transformación de los datos, pasando por el análisis estadístico y el uso de técnicas de machine learning, hasta la visualización de los resultados. Además, incluye un apartado de aplicación de modelo LLM mediante la librería PandasAI. Esta integración de diferentes etapas del proceso de análisis de datos permite obtener una visión completa y holística de la problemática abordada.

Como futuras líneas de investigación, sería interesante examinar cómo la infraestructura ciclista influye en la adopción del sistema de bicicletas compartidas y en los patrones de movilidad urbana. Además, explorar la segmentación de usuarios podría

proporcionar una comprensión más profunda de sus necesidades y preferencias. También sería relevante analizar el impacto de políticas gubernamentales y programas de movilidad sostenible en la utilización del sistema. Desarrollar modelos predictivos más precisos y estudiar la equidad en el acceso al sistema son áreas clave que podrían mejorar nuestra comprensión y el funcionamiento del sistema de bicicletas compartidas en Boston. Por último, sería interesante realizar una comparativa entre los resultados obtenidos y los de otros sistemas de bicicletas compartidas en diferentes ciudades.

8. Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Carmen Gómez García-Atance, estudiante de ADE & Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis de los viajes del servicio de bikesharing "Bluebikes" en Boston: Un caso de estudio" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación**: Utilizado para idear y esbozar posibles áreas de investigación.

2. **Metodólogo**: Para descubrir métodos aplicables a problemas específicos de investigación.

3. **Interpretador de código:** Para realizar análisis de datos preliminares.

4. **Corrector de estilo literario y de lenguaje**: Para mejorar la calidad lingüística y estilística del texto.

5. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.

6. **Generador de datos sintéticos de prueba**: Para la creación de conjuntos de datos ficticios.

7. **Generador de problemas de ejemplo**: Para ilustrar conceptos y técnicas.

8. **Generador de encuestas**: Para diseñar cuestionarios preliminares.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 21 de abril de 2024

Firma: Carmen Gómez García-Atance

9. Bibliografía

Ayuba, Y., Supangkat, S. H., & Wibowo, S. S. (2023, September). Mobility as a Service: A Problems and Research Opportunities. In *2023 10th International Conference on ICT for Smart Society (ICISS)* (pp. 1-7). IEEE. https://ieeexplore.ieee.org/abstract/document/10292068

Atiq, A. (2023, octubre 6). *PandasAI: Making data analysis conversational and fun.* Medium. https://medium.com/@amadatiq/pandasai-making-data-analysis-conversational-and-fun-3acc76584cb3

Bluebikes. Boston.gov. (2016a, July 25). https://www.boston.gov/bluebikes

Bluebikes. (n.d.). https://bluebikes.com/about

Boston Complete Streets. Boston.gov. (2020, September 28). https://www.boston.gov/departments/transportation/boston-complete-streets

Boston's Bike Network and safer streets expanding. Boston.gov. (2022, September 6). https://www.boston.gov/news/bostons-bike-network-and-safer-streets-expanding

Chen, Z., Van Lierop, D., & Ettema, D. (2020). Dockless bike-sharing systems: what are the implications?. *Transport reviews*, *40*(3), 333-353. https://www.tandfonline.com/doi/full/10.1080/01441647.2019.1710306

Chiariotti, F., Pielli, C., Cenedese, A., Zanella, A., & Zorzi, M. (2018, May). Bike sharing as a key smart city service: State of the art and future developments. In *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)* (pp. 1-6). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8376628

Clima Boston: Temperatura, Climograma y Temperatura del agua de Boston. (s/f). Climate-data.org.https://es.climate-data.org/america-del-norte/estados-unidos-de-america/massachusetts/boston-1722/

Coder, R. (2021, octubre 13). *Pairs plot (gráfico por pares) en seaborn con la función pairplot*. PYTHON CHARTS | Visualización de datos con Python; R CODER. https://python-charts.com/es/correlacion/pairplot-seaborn/?utm_content=cmp-true

Commonwealth Ave Phase 2A. Boston.gov. (2016, December 8). https://www.boston.gov/departments/transportation/commonwealth-ave-phase-2a

Conrow, L., Murray, A. T., & Fischer, H. A. (2018). An optimization approach for equitable bicycle share station siting. *Journal of transport geography*, *69*, 163-170.

https://www.sciencedirect.com/science/article/pii/S0966692317303563

Daniel. (2022, abril 25). *Matplotlib: todo lo que tienes que saber sobre la librería Python de Dataviz*. Formación en ciencia de datos | Datascientest.com; DataScientest. https://datascientest.com/es/todo-sobre-matplotlib

Datacamp.com. (s/f). https://www.datacamp.com/es/blog/an-introduction-to-pandas-ai

de los Santos, P. R. (2022, enero 24). Datos de entrenamiento vs datos de test. Telefónica Tech. https://telefonicatech.com/blog/datos-entrenamiento-vs-datos-de-test

DeMaio, P. J. (2003). Smart bikes: Public transportation for the 21st century. *Transportation Quarterly*, *57*(1), 9-11.

https://www.metrobike.net/wp-content/uploads/2013/10/Smart-Bikes.pdf

DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, *12*(4), 41-56.

https://www.sciencedirect.com/science/article/pii/S1077291X22002600

Eckhardt, J., Aapaoja, A., Nykänen, L., Sochor, J., Karlsson, M., & König, D. (2018). The European roadmap 2025 for mobility as a service. *Proceedings of the 7th Transport Research Arena TRA*. https://www.researchgate.net/profile/Aki-Aapaoja/publication/321586613_The_European_Roadmap_2025_for_Mobility_as_a_Service/links/5b28e60a0f7e9b1d0034ab4d/The-European-Roadmap-2025-for-Mobility-as-a-Service.pdf

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable cities and society*, *54*, 101882.

https://www.sciencedirect.com/science/article/pii/S2210670719312387

Ferrer-Benítez, M. (2022). Online dispute resolution: can we leave the initial decision to Large Language Models (LLM)?. *Metaverse Basic and Applied Research*, *1*, 23-23. https://mr.saludcyt.ar/index.php/mr/article/view/23

Go Boston 2030 revisioned. Boston.gov. (2017, February 24). https://www.boston.gov/departments/transportation/go-boston-2030

Grid search de modelos Random Forest con out-of-bag error y early stopping. (s/f). Cienciadedatos.net. https://cienciadedatos.net/documentos/py36-grid-search-random-forest-out-of-bag-error-early-stopping

Juan, Á. A., Sedano, M., & Vila, A. (2006). La distribución normal. *Universitat Oberta de Catalunya*.

https://d1wqtxts1xzle7.cloudfront.net/43169354/Distrib_Normal-libre.pdf?1456688138=&response-content-disposition=inline%3B+filename%3DDistrib_Normal.pdf&Expires=1712485052&Signature=K~1Ad18DK4nOG22APiIcTARqTymZ4Ss~Et91EOutiztqRZDfkjNzPz3GLPHaFqeMQifyS1s-JSeZ3sjTdO6fWlC5EyG3XzGAxHplG4bRh8WMmE60piRLUVdxYwBUXrUj6NMJ38rF1OT2M4mCSPH8lBK0jajip-wq09j3-WXtM-B1gjVFaqzW90IEZ~tn4urR1wf79z0455nRuCdyDvCPBYINruoYZZeghRI6JrxnwyAfWf5J3R2YERPKAi00jj9Ul2aRMNklvlxioeLap4KUzLvhp09KNJr~PYz3c42Z2H5M43lkAzowd-SNuptjSJs9YnPZZdZtBzPovtG415dCIQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Karpinski, E. (2021). Estimating the effect of protected bike lanes on bike-share ridership in Boston: A case study on Commonwealth Avenue. *Case studies on transport policy*, *9*(3), 1313-1323.

https://www.sciencedirect.com/science/article/pii/S2213624X21001097

Leiva, G. O., Vigneau, G. H., & Sepúlveda, D. Y. (2023, December). Imitating Teaching: An Automated Approach Using Large Language Model. In *2023 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)* (pp. 1-5). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10418678

Madrigal, E. (2022, octubre 28). Conoce las métricas de precisión más comunes para Modelos de Regresión. Grow Up. https://www.growupcr.com/post/metricas-precision

Maria, E., Budiman, E., & Taruk, M. (2020, February). Measure distance locating nearest public facilities using Haversine and Euclidean Methods. In *Journal of Physics: Conference Series* (Vol. 1450, No. 1, p. 012080). IOP Publishing.

https://iopscience.iop.org/article/10.1088/1742-6596/1450/1/012080/meta

Millán, J. (2020, mayo 15). *Covid-19*. Covid-19 | Inteligencia Artificial, Machine Learning, Python y Esas Cosas. https://medium.com/previsi%C3%B3n-casos-covid-19-espa%C3%B1a-regresi%C3%B3n/estudio-de-factores-que-influyen-en-la-propagaci%C3%B3n-del-covid-19-scatter-matrix-y-correlaciones-3b73fa9c62e2

Midgley, P. (2009). The role of smart bike-sharing systems in urban mobility. Journeys, 2(1), 23-31. https://d1wqtxts1xzle7.cloudfront.net/79547632/The-Role-of-Smart-Bike-sharing-Systems-libre.pdf?1643173247=&response-content-disposition=inline%3B+filename%3DThe_role_of_smart_bike_sharing_systems_i.pdf&Expires=1713385931&Signature=exeJOLshE7zBPtOs8knjmo8OBu1wonitJk7UgGSNXx4x3TPQBbFZ9SLR5OQBjYo8ygtI~S13huI9Q2chOMKt2iigKCnkYIrKz8nkNkVzkVIuVRIMcjK8~lFC64U1C34EuKKlkzNe5bdTFdgIoy0QOaT79F8ffCUSTq1HFtvDcQgC2AW1GpyF2D1EvGwYeU~sBL0eUydscubMV6IpVq-CwEM91ZvRpCB5JsQviWyVvIBaE5aVuS3cYFX7vj82i8PdvhptkOO3Vhoy4UjdqhfF7X8Xe5cAOXoubt-0PRTqtqjC8hZdtIREovDsH07WBeaI~MSgz7JgxQkjHtUEpu6yWg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Moon, C., Sharpin, A. B., De La Lanza, I., Khan, A., Re, L. L., & Maassen, A. (2019). The evolution of bike sharing: 10 questions on the emergence of new technologies, opportunities, and risks.

https://apo.org.au/sites/default/files/resource-files/2019-01/apo-nid229926.pdf

Narayanan, S., & Antoniou, C. (2023). Shared mobility services towards Mobility as a Service (MaaS): What, who and when?. *Transportation research part A: policy and practice*, *168*, 103581.

https://www.sciencedirect.com/science/article/pii/S0965856423000010

Novedades en bicimad: llega la tarifa plana y la gratuidad se extiende todo el mes de enero. (s/f). Bicimad.com. https://www.bicimad.com/noticias/novedades-en-bicimad-llega-la-tarifa-plana-y-la-gratuidad-se-extiende-todo-el-mes-de-enero

NumPy: La biblioteca de Python más utilizada en Data Science. (2023, enero 18). Formación en ciencia de datos | Datascientest.com; DataScientest. https://datascientest.com/es/numpy-la-biblioteca-python

Nyamathulla, S., Ratnababu, P., & Shaik, N. S. (2021). A review on selenium web driver with python. *Annals of the Romanian Society for Cell Biology*, 16760-16768.

https://annalsofrscb.ro/index.php/journal/article/view/7087

Pandas: La biblioteca de Python dedicada a la Data Science. (2022, diciembre 19). Formación en ciencia de datos | Datascientest.com; DataScientest. https://datascientest.com/es/pandas-python

Pérez-Morales, A., Gil-Guirado, S., & Maqueda-Belmonte, F. (2022). Movilidad sostenible: interdisciplinariedad, articulación conceptual y frentes de investigación. *Documents d'anàlisi geogràfica*, *68*(2), 0393-422.

https://ddd.uab.cat/pub/dag/dag_a2022v68n2/dag_a2022v68n2p393.pdf

Random Forest python. (s/f). Cienciadedatos.net. https://cienciadedatos.net/documentos/py08_random_forest_python

Requiz, B. J. C., Silva, J. D. T., Silva, C. E. T., Enriquez, C. H., & Orbegoso, F. A. C. (2023). Automatización del análisis exploratorio de datos y procesamiento geoquímico univariado empleando Python. *Revista del Instituto de investigación de la Facultad de minas, metalurgia y ciencias geográficas*, *26*(51), e24493-e24493.

https://www.researchgate.net/profile/Christian-Hurtado-Enriquez/publication/372077059_Automatizacion_del_analisis_exploratorio_de_datos_y_procesamiento_geoquimico_univariado_empleando_Python/links/64a3ae7eb9ed6874a5f4e9cc/Automatizacion-del-analisis-exploratorio-de-datos-y-procesamiento-geoquimico-univariado-empleando-Python.pdf

Scikit-learn. (s/f). Scikit-learn.org. Recuperado el 30 de marzo de 2024, de https://scikit-learn.org/stable/

Sharmila, S., & Sabarish, B. A. (2021, February). Analysis of distance measures in spatial trajectory data clustering. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1085, No. 1, p. 012021). IOP Publishing.

https://iopscience.iop.org/article/10.1088/1757-899X/1085/1/012021/meta

Si, H., Shi, J. G., Wu, G., Chen, J., & Zhao, X. (2019). Mapping the bike sharing research published from 2010 to 2018: A scientometric review. *Journal of cleaner production*, *213*, 415-427.

https://www.sciencedirect.com/science/article/pii/S0959652618338678?casa_token=icwlxSqS8LQAAAAA:3ETyjwICPuV_F4xuOskLKZ7IPqd4VfuLkOG5bzHIr5CIeMXqukZIZzOOO82lNFWyno0Mj2n4VNI

Statistical functions (scipy.stats) — SciPy v1.12.0 Manual. (s/f). Scipy.org. Recuperado el 30 de marzo de 2024, de https://docs.scipy.org/doc/scipy/reference/stats.html

Teixeira, J. F., Silva, C., & Moura e Sá, F. (2023). Potential of bike sharing during disruptive public health crises: A review of COVID-19 impacts. *Transportation Research Record*, 03611981231160537.

https://journals.sagepub.com/doi/full/10.1177/03611981231160537

Tiempo, A. M., & Data, B. (2022, octubre 28). ¿Qué es Folium? - Al mal tiempo, buena data. Medium. https://lauralpezb.medium.com/qu%C3%A9-es-folium-b16d39797692

Uzun, E., Yerlikaya, T., & Kirat, O. Ğ. U. Z. (2018). Comparison of python libraries used for web data extraction. *Journal of the Technical University-Sofia Plovdiv Branch, Bulgaria*, *24*, 87-92.

https://erdincuzun.com/wp-content/uploads/download/plovdiv_2018_01.pdf

Vallez, C. M., Castro, M., & Contreras, D. (2021). Challenges and opportunities in dock-based bike-sharing rebalancing: A systematic review. *Sustainability*, *13*(4), 1829.

file:///C:/Users/hp/Downloads/sustainability-13-01829-v2.pdf

Venturi, G. (s/f). PandasAI. Pandas-ai.com. Recuperado el 20 de abril de 2024, de https://docs.pandas-ai.com/en/latest/

Vision zero. Boston.gov. (2018, March 1). https://www.boston.gov/transportation/vision-zero

What is random forest? (s/f). Ibm.com. https://www.ibm.com/topics/random-forest

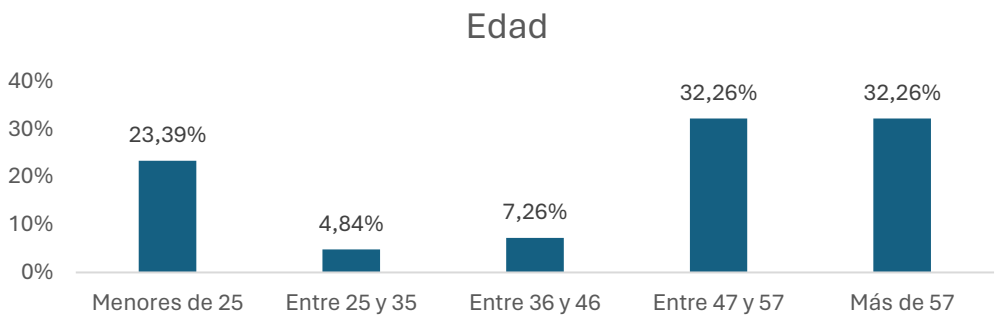
Zheng, L., & Li, Y. (2020). The Development, Characteristics and Impact of Bike Sharing Systems: A Literature Review. *International review for spatial planning and sustainable development*, *8*(2), 37-52.

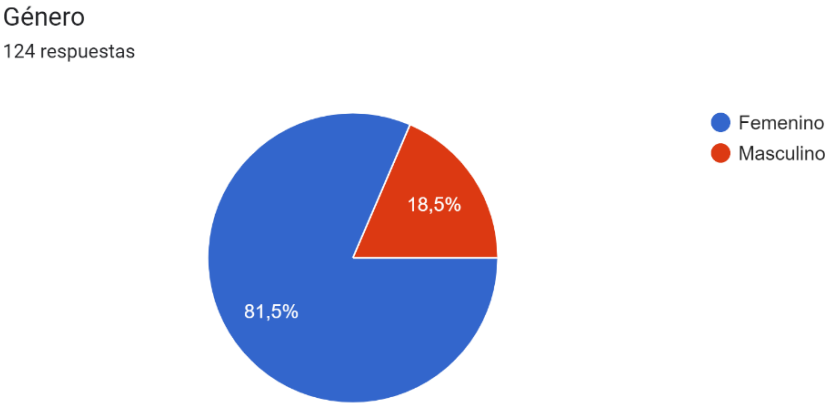
https://www.jstage.jst.go.jp/article/irspsd/8/2/8_37/_pdf/-char/ja

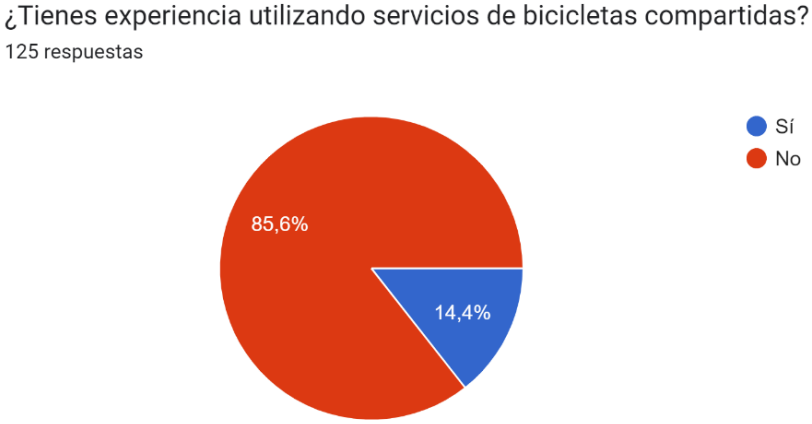
Zulmuthi, H. (2022, mayo 12). *Out with the outliers - hanis zulmuthi*. Medium. https://medium.com/@haniszulaikha/out-with-the-outliers-fc39c2bcacd7

10. Anexos

Anexo 10.1 Encuesta

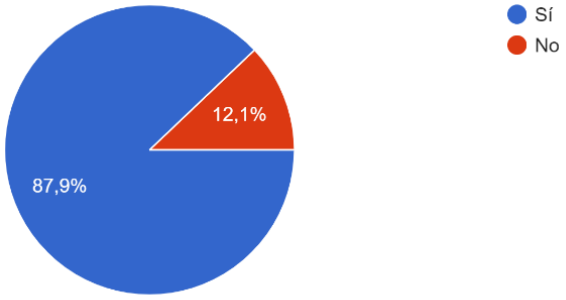






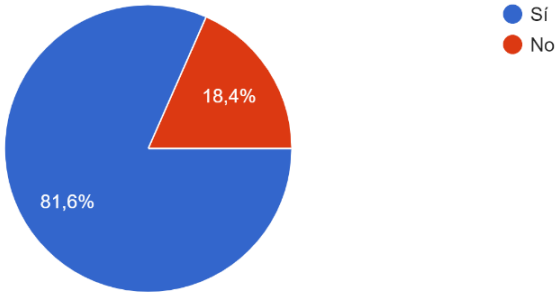
¿Posees permiso de conducir?

124 respuestas



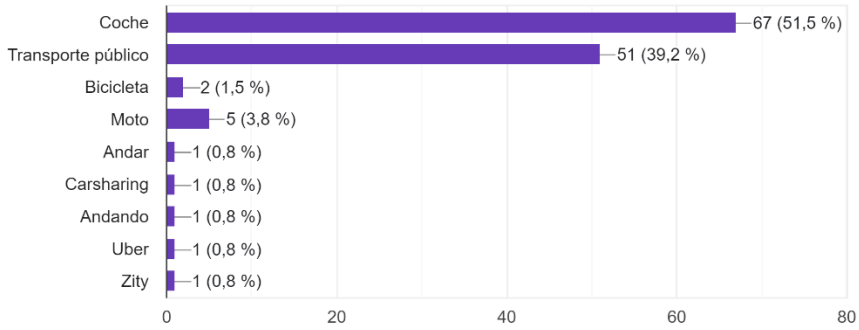
¿Tienes acceso a un coche personal?

125 respuestas



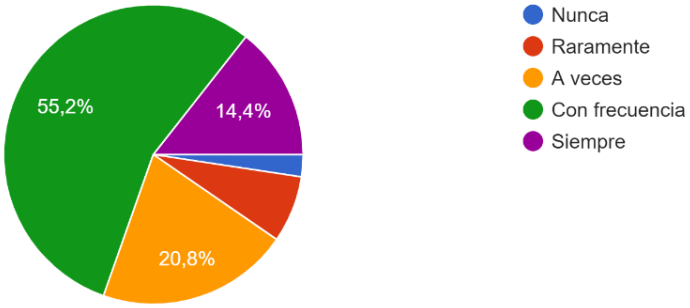
¿Qué medio de transporte sueles utilizar para desplazarte?

130 respuestas



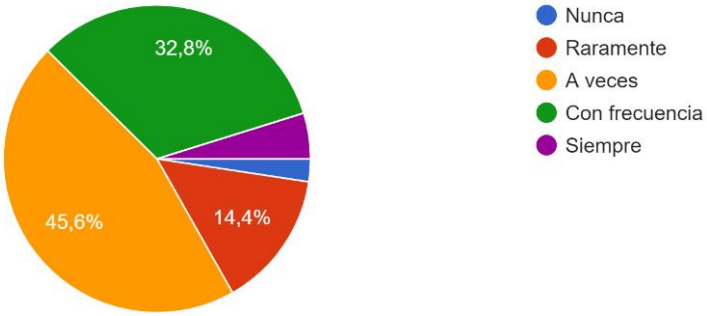
¿Cuánto sueles desplazarte en distancias cortas? (menos de 5 km)

125 respuestas



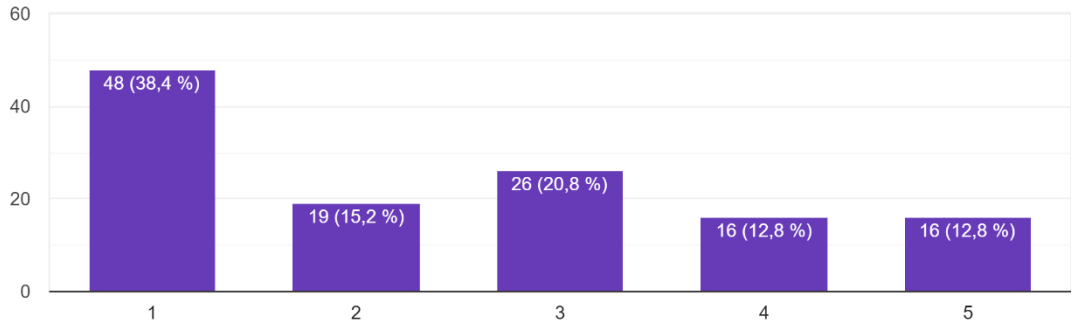
¿Con qué frecuencia te desplazas en distancias largas? (más de 25 km)

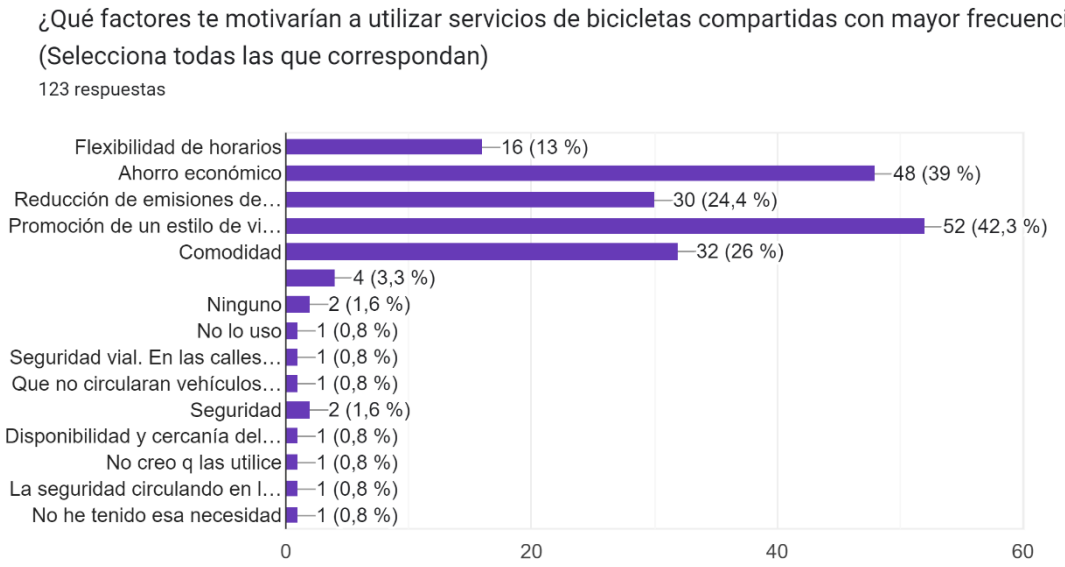
125 respuestas

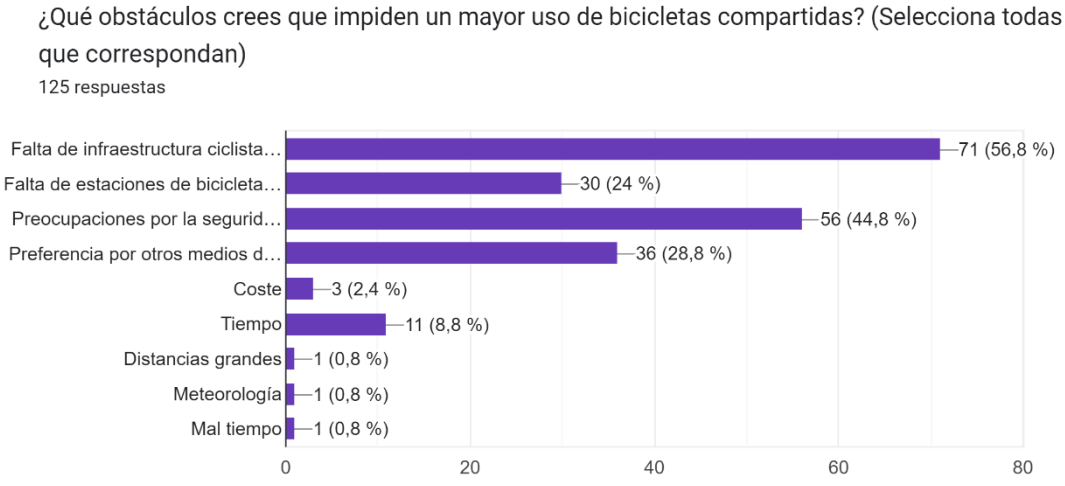


En una escala del 1 al 5, ¿qué tan importante consideras el uso de bicicletas compartidas para tu movilidad urbana? (1 = Nada importante, 5 = Muy importante)

125 respuestas







¿Cómo crees que se podría incrementar el uso de bicicletas compartidas?

- Ampliando y mejorando la infraestructura de carriles bici para garantizar la seguridad de los ciclistas.

- Incrementando el número de estaciones de bicicletas compartidas para mejorar la accesibilidad y conveniencia del servicio.

- Implementando aplicaciones que ofrezcan rutas más rápidas y seguras para los ciclistas, similar a un "Waze para bicicletas".

- Realizando campañas de promoción y concientización sobre el uso de la bicicleta como medio de transporte.

- Reduciendo los costos del servicio y haciendo que sea más económico y accesible para los usuarios.

- Incluyendo el servicio de bicicletas compartidas como parte del abono transporte o tarifas de transporte público.

- Mejorando la seguridad vial en general para los ciclistas, incluyendo la creación de carriles bici segregados de la circulación de vehículos motorizados.

- Proporcionando garantías de seguridad en el tráfico, como la regulación del flujo vehicular y la educación vial.