Universidad Pontificia de Comillas

ANÁLISIS DE LOS VIAJES DE *BIKE-SHARING* DE CITI BIKE. ¿CÓMO HA AFECTADO LA PANDEMIA A ESTA PLATAFORMA?

Autor: Alfredo González-Izquierdo Antón
Tutor: Carlos Miguel Vallez Fernández
Clave: 201906395

MADRID | junio 2024

RESUMEN

Este estudio analiza el impacto de la pandemia de la COVID-19 en los patrones de uso de Citi Bike, un destacado sistema de bicicletas compartidas en la ciudad de Nueva York. La investigación identifica cambios significativos en el volumen de viajes, las demografías de los usuarios y la distribución geográfica del uso de bicicletas antes y después de la pandemia. Se observó un aumento del 50% en el total de viajes de 2019 a 2022, con notables cambios en la duración de los trayectos y las horas pico de uso, reflejando cambios en los hábitos laborales y un aumento del teletrabajo. El estudio también destaca una reducción en la brecha de género en el uso de bicicletas y un aumento en la adopción de bicicletas eléctricas. A través de un análisis de regresión múltiple, la investigación examina las variables sociodemográficas que influyen en el uso de bicicletas, proporcionando información sobre las tendencias de movilidad urbana y la resiliencia de los sistemas de bicicletas compartidas en medio de crisis sanitarias globales.

Palabras clave: bicicletas compartidas, Citi Bike, pandemia, COVID-19, sociodemográficas, Nueva York.

ABSTRACT

This paper analyzes the impact of the COVID-19 pandemic on the usage patterns of Citi Bike, a prominent *bike-sharing* system in New York City. The research identifies significant changes in the volume of rides, user demographics, and geographical distribution of bike usage pre- and post-pandemic. A 50% increase in total trips from 2019 to 2022 was observed, with notable shifts in trip durations and peak usage times reflecting changes in work habits and a rise in telecommuting. The study also highlights a narrowing gender gap in bike usage and increased adoption of electric bikes. Through multiple regression analysis, the research examines sociodemographic variables influencing bike usage, providing insights into urban mobility trends and the resilience of *bike-sharing* systems amid global health crises.

Key words: *bike-sharing*, Citi Bike, pandemic, COVID-19, sociodemographic, New York.

Índice de contenidos

Índice de tablas

Índice de ilustraciones

Índice de anexos

Capítulo 1: Introducción

1.1.Justificación del proyecto

Los sistemas de micro movilidad hacen referencia a aquella clase de vehículos pequeños y ligeros que circulan a velocidades de hasta 25 km/h, son conducidos por los propios individuos y son ideales para distancias cortas (Bozzi & Aguilera, 2021).

Dentro de este ámbito se pueden incluir tanto vehículos de tracción humana como eléctrica; sin embargo, los vehículos eléctricos no pueden tener motores de combustión interna ni desplazarse a más de 45 km/h para ser considerados micro movilidad (Oeschger et al., 2020). Algunos ejemplos de estos vehículos pueden ser bicicletas, *e-bikes*, patinetes eléctricos *(e-scooters)* o bicicletas compartidas. Además, pueden tener dueños particulares o estar disponibles como vehículos de alquiler, sobre todo en forma de *dockless sharing*[1] (Spherical Insights, 2023).

Este mercado fue valorado en 40 mil millones de dólares en 2021 y según un estudio publicado por Acumen (2023), se espera que alcance los 186,2 mil millones de dólares en 2030, con una tasa de crecimiento anual compuesta (CAGR) del 18,6%. Además, se estima que esta industria alberga más de 8.000 empleos (aproximadamente un empleo por cada 26 vehículos), cifras que demuestran la relevancia de esta industria.

Como se ha mencionado anteriormente, esta industria engloba una gran variedad de diferentes vehículos de transporte; sin embargo, a lo largo de este trabajo nos vamos a centrar en los sistemas de *bike-sharing* (bicicletas compartidas).

Antes de explicar en profundidad el funcionamiento de estos sistemas, es importante resaltar el concepto de la economía compartida *(sharing economy).* Este concepto, a pesar de no tener una definición muy clara, ha transformado significativamente la manera en la que las personas acceden a los bienes y servicios en pleno siglo XXI. Este modelo económico se basa en la idea de compartir recursos infrautilizados a través de plataformas tecnológicas, permitiendo así a los individuos alquilar, prestar o compartir activos de forma temporal (Zhu et al., 2018). Un ejemplo claro de activos que forman parte de esta economía son los sistemas de *bike-sharing*.

[1] Modelo de uso compartido de vehículos sin utilizar estaciones de acoplamiento fijo

Este tipo de sistemas hacen referencia a una forma de transporte público de autoservicio que ofrece el uso de bicicletas a corto plazo (DeMaio, 2009). Además, estos sistemas tienen el objetivo principal de proporcionar a las personas una mayor comodidad y flexibilidad en el acceso a la propiedad de la bicicleta, sin el coste y las responsabilidades asociadas (Faghih-Imani & Eluru, 2015).

Su funcionamiento es muy sencillo, el cliente puede coger una bicicleta en un muelle (plataforma de estacionamiento de las bicicletas) y devolverla en otro muelle del mismo sistema. Gracias a la flexibilidad que aporta y su bajo coste, este sistema ha crecido de manera muy rápida durante los últimos años, especialmente en países desarrollados (Fishman, 2016). De hecho, según la oficina de transporte de Estados Unidos, en el periodo comprendido entre 2015 y 2019, el crecimiento de sistemas de bicicletas compartidas llegó casi a alcanzar un 200%, mostrando así la rápida evolución y la creciente popularidad de dichos sistemas en el país

Como se ha mencionado anteriormente, las tarifas asociadas a este tipo de servicios no suelen ser muy elevadas. Sin embargo, este no es el principal factor que motiva a los usuarios a optar por estos servicios. Entre los principales motivos para utilizar estos vehículos se encuentran la reducción de emisiones y el uso de combustible, la mejora de la fluidez del tráfico y la integración de la actividad física en las rutinas diarias. Estos beneficios medioambientales y de salud son los que realmente impulsan a los usuarios a utilizar estos sistemas (Shaheen et al., 2013).

Para verificar que estos sistemas consiguen reducir las emisiones y el uso de combustible, es útil revisar la literatura existente. En primer lugar, de acuerdo con lo mencionado en el trabajo científico de Zhang & Mi (2018), el uso de estos servicios en Shanghái consiguió reducir en 2016 el consumo de petróleo en 8.358 toneladas, así como una reducción de las emisiones en 25.240 toneladas. Esto demuestra que la implementación de estos servicios ha conseguido reducir significativamente las emisiones.

En segundo lugar, si realizamos el mismo análisis en Nueva York, el uso de estos sistemas de transporte logró que, entre 2014 y 2017, los usuarios de estas plataformas consiguiesen ahorrar 13.000 toneladas de petróleo, equivalente a una reducción de 30.070 toneladas de Co2 (Chen et al., 2022).

Además de los beneficios medioambientales, la reducción de la congestión del tráfico es otro factor crucial. Estudios recientes han demostrado que los sistemas de *bike-sharing* pueden aliviar significativamente el tráfico urbano. Según el estudio de Hamilton (2018), las ciudades que incorporan sistemas de *bike-sharing* son más propensas a reducir el tráfico en un aproximadamente un 4%, al ofrecer una alternativa eficiente y sostenible al uso de vehículos motorizados.

Por otro lado, según el estudio llevado a cabo por Zheng et al. (2022), el metro de Shanghái (cuyo volumen medio diario de personas supera los 10 millones) ha conseguido reducir la intensidad de flujo de pasajeros de 13.200 a 8.700 personas/kilómetro, gracias a la aparición de los sistemas de *bike-sharing* y nuevas formas de transporte público.

Finalmente, los beneficios para la salud son otra motivación importante para el uso del *bike-sharing*. Según la investigación llevada a cabo por Clockston & Rojas Rueda (2021), el uso de los sistemas de *bike-sharing* redujo las muertes prematuras un 4,74% en Estados Unidos y 36 mil millones de dólares de impacto económico relacionados con temas de la salud.

Del mismo modo, según la encuesta realizada por Capital bikeshare, el 31,5% de los encuestados declaró haber reducido el estrés, y alrededor del 30% indicó que había perdido peso como resultado de utilizar estos sistemas a diario (Ricci, 2015). Por último, según el informe publicado por la asociación de *bike-sharing* y *scooter sharing* de América del Norte (2022), un 27% de los usuarios afirmaba que utilizaba este tipo de sistemas porque quería realizar ejercicio.

Tras revisar la literatura existente, hemos podido corroborar la veracidad de las ventajas de utilizar este tipo de servicio. Dentro de este contexto, voy a desarrollar un análisis que demuestre si la pandemia ha tenido efecto en los hábitos y patrones de uso en los sistemas de *bike-sharing*.

Para poder acotar este trabajo, me voy a centrar en la ciudad de Nueva York y especialmente en la empresa de Citi Bike, que cuenta con el sistema de bicicletas más grande de Estados Unidos, con más de 25.000 bicicletas y 1.500 estaciones distribuidas por las ciudades de Manhattan, Brooklyn, Queens, el Bronx, Jersey City y Hoboken (Citi Bike, s.f.).

Además, para poder analizar si la pandemia ha tenido impacto o no en los sistemas de *bike-sharing* de esta compañía, voy a seleccionar tres fechas significativas con las que trabajaremos a lo largo del trabajo.

En primer lugar, utilizaré los datos de 2019 para representar la situación prepandemia ya que, la pandemia no comenzó hasta el año siguiente. Para analizar los efectos de la COVID-19, seleccionaré los datos de 2020. Finalmente, utilizaré los datos de 2022 para representar la situación post pandemia, asumiendo que se ha vuelto a la normalidad y que ya no existen las restricciones que estuvieron vigentes durante la pandemia, como confinamientos o limitaciones al desplazamiento.

1.2.Objetivos

El propósito principal de este trabajo consiste en analizar la evolución en el uso de los sistemas de *bike-sharing* de Citi Bike en Nueva York, para ver si la pandemia ha supuesto un aumento o disminución del uso de estos servicios o si los hábitos de uso de los usuarios se han visto modificados. Para ello, este trabajo tendrá que completar los siguientes subobjetivos:

I. Extracción y tratamiento de los datos para que sean lo más precisos posibles y así evitar incurrir en errores a lo largo del análisis.

II. Realizar un análisis exploratorio de los datos que incluya la identificación de tendencias, patrones y anomalías en los datos utilizando herramientas estadísticas.

III. Cruzar los datos de Citi Bike con conjuntos de datos que incluyan variables sociodemográficas de Nueva York, para investigar cómo estos factores afectan los patrones de uso del servicio del *bike-sharing*.

IV. Visualización gráfica de los datos para facilitar la extracción de conclusiones.

1.3.Metodología

Para poder alcanzar los subobjetivos mencionados previamente, se realizará una investigación de acuerdo con la estructura ETL *(Extraction, Transformation and Loading)*, metodología clásica de un ciclo de vida de datos comprendida en las siguientes partes:

En primer lugar, se extraerán los datos que recogen los viajes efectuados por los usuarios de Citi Bike mediante las técnicas de *web scrapping* adecuadas en el lenguaje de programación de Python. Además, también se buscarán y se descargarán diferentes bases de datos que recojan las características sociodemográficas por código postal de Nueva York, como población, ingreso medio o género, para poder posteriormente comparar dicha información.

En segundo lugar, se procederá al tratamiento de ambos datasets, eliminando valores nulos, ajustando posibles valores atípicos o transformando ciertas variables que puedan interferir en nuestro análisis y así mejorar la calidad de nuestros datos. Para ello, se utilizará la herramienta de R Studio.

Una vez analizada la evolución del uso de estos sistemas, se procederá a comparar los datasets de Citi Bike con las variables sociodemográficas de Nueva York, para evaluar si el uso de las bicicletas está sujeto a alguna variable sociodemográfica o si existe algún patrón que pueda demostrar el mayor o menor uso de los sistemas de bicicletas compartidas después de la pandemia. Finalmente, se utilizará la herramienta de Power BI para la visualización de nuestros resultados y así poder respaldar las conclusiones de este trabajo de una manera más visual y rápida.

1.4.Estructura del trabajo

El presente trabajo está estructurado de la siguiente manera. En primer lugar, se explicará en detalle la industria de los sistemas del *bike-sharing*, en donde se explicará el contexto histórico y la evolución que han ido tomando durante todos estos años. En segundo lugar, se explicarán las herramientas utilizadas para la extracción de nuestros datasets mediante las técnicas de *web scrapping* que hemos aprendido a lo largo de la carrera de Business Analytics.

En tercer lugar, se mostrará un análisis más profundo de cada variable de nuestra base de datos, explicando el significado de cada una de ellas, y mostrando los tipos de datos que recogen las mismas, así como de sus estadísticas principales (media, mediana, primer y tercer cuartil, etc.). En cuarto lugar, se explicarán los métodos utilizados para el tratamiento de los datasets y la razón para transformar algunas variables.

A partir de aquí, se procederá a analizar si ha habido una evolución en el uso de los sistemas después de la pandemia, y estudiar posibles relaciones con los indicadores sociodemográficos que nos ayuden a comprender un mayor o menor uso en los diferentes distritos.

Capítulo 2: Análisis de la industria de bike-sharing

Como se mencionaba anteriormente, la industria del *bike-sharing* gira en torno al ámbito de la micro movilidad y la economía compartida. Además, durante esta última década, esta industria ha experimentado un notable crecimiento y cada vez son más las ciudades que integran este tipo de sistemas (Galatoulas et al., 2020).

Para poder explicar en profundidad todo lo que rodea a esta la industria, dividiremos este apartado en cuatro secciones. En primer lugar, se explicará el contexto histórico del *bike-sharing* para así conocer sus orígenes. En segundo lugar, se expondrán las principales cifras que rodean a esta industria. En tercer lugar, nos centraremos en la historia que hay detrás de Citi Bike, y finalmente se identificarán los competidores que tiene esta empresa en Nueva York.

2.1. Contexto histórico de los sistemas de *bike-sharing*

El primer sistema de *bike-sharing* apareció en Ámsterdam a finales de 1960 (Schimmelpennick, 2009). Ofrecía bicicletas pintadas de blanco para uso público, que cualquiera podía usar, siempre y cuando se devolviesen de vuelta. Sin embargo, la escasa regulación y la falta de control aumentaron el vandalismo y el uso indebido de estas bicicletas, acabando muchas de ellas en los canales de la ciudad. Estos factores hicieron que el sistema colapsara en pocos días.

Este primer modelo fue un fracaso; sin embargo, en 1991 apareció la segunda generación del *bike-sharing* en Farsø y Grenå, dos pueblos al norte de Dinamarca (Nielse, 1993), pero finalmente acabó implementándose en Nakskov en 1993. Este sistema era pequeño, con solo veintiséis bicicletas y cuatro estaciones. Dos años más tarde, se lanzó en Copenhague bajo el nombre de Bycyklen, presentando mejoras respecto al modelo anterior. La principal diferencia de este sistema era que, para usar las bicicletas, se tenía que introducir una moneda en un depósito localizado en las estaciones que se devolvía al devolver la bicicleta. Aunque había algo más de control, el robo siguió ocurriendo debido al anonimato del usuario, lo que acabó colapsando el sistema.

La tercera generación no tardó en aparecer, y en 1996 la Universidad de Portsmouth, en Inglaterra, comenzó a incorporar estos sistemas. Estos ya incluían nuevas medidas tecnológicas, como tarjetas de acceso inteligentes para desbloquear las bicicletas. Además, ofrecían treinta minutos gratis de cortesía y permitían identificar en todo momento a los usuarios que habían utilizado el servicio ya que, la información se guardaba en una base de datos.

Desde entonces, este sistema ha evolucionado hasta llegar a su versión actual, la cuarta generación del *bike-sharing*. Esta incluye bicicletas "inteligentes" equipadas con tecnología avanzada, como sensores, geolocalización y acceso mediante una aplicación dedicada, junto con estaciones flexibles (Chen et al., 2018). Estos avances han facilitado la integración de estos vehículos en la rutina diaria de la población, utilizándolos para desplazarse al trabajo y a otros lugares de la ciudad.

2.2. Principales cifras de esta industria

Una vez explicado los orígenes de los sistemas del *bike-sharing*, vamos a exponer las principales cifras que rodean a esta industria. Como la empresa que estamos analizando opera principalmente en Nueva York, en este apartado nos centraremos especialmente en las cifras de Estados Unidos.

El primer sistema de *bike-sharing* apareció en Estados Unidos en 2008 a manos de la empresa SmartBike DC, que fue luego reemplazado por el sistema de Capital Bikeshare en 2010. Desde entonces, esta industria ha estado creciendo a un ritmo exponencial, estando presente en más de 50 de sus estados (Pan American Health Organization, 2020).

A lo largo de este apartado analizaremos la evolución de los sistemas de *bike-sharing*, enfocándonos principalmente en la cantidad de sistemas que han ido apareciendo durante estos últimos años, así como la evolución de la duración, precio y distancia media de los viajes realizados por sus usuarios.

Número de sistemas de bike-sharing en EE.UU.

Tabla 1. Número de sistemas de bike-sharing desde 2015 hasta 2023

	2015	2016	2017	2018	2019	2020	2021	2022	2023	Var. (15-23)
Sistemas Existentes	43	65	78	99	104	66	64	58	52	21%
Nuevos Sistemas	23	20	27	13	5	1	5	4	4	(82%)
Total	**66**	**85**	**105**	**112**	**109**	**67**	**69**	**62**	**56**	**(15%)**

Fuente: elaboración propia en base a *Bureau of Transportaion Statistics 2023*

Siete años después de la introducción de los servicios de *bike-sharing* en Estados Unidos, el país contaba con 66 sistemas diferentes distribuidos en todos los estados, incluyendo 43 sistemas ya existentes y 23 creados ese mismo año. Entre 2015 y 2019, todos estos sistemas experimentaron un crecimiento cercano al 200%, reflejando la rápida evolución y la creciente popularidad de los servicios de *bike-sharing*.

Sin embargo, a partir de 2017 la creación de nuevos sistemas empezó a reducirse pasando de 27 nuevos sistemas creados en ese año a la creación de tan solo 5 nuevos sistemas en 2019, una reducción del 80% en tan solo dos años. Además, también podemos apreciar que la cantidad de sistemas existentes comenzaba a decrecer significativamente a partir de 2019, momento en el que aparecía la COVID-19, llegando a pasar de 109 sistemas en ese año a 56 en 2023, una reducción del 40%.

Evolución de los viajes por hora (2019-2020)

Ilustración 1. Evolución de los viajes por hora (2019-2020)

Fuente: externa (NACTO, 2022)

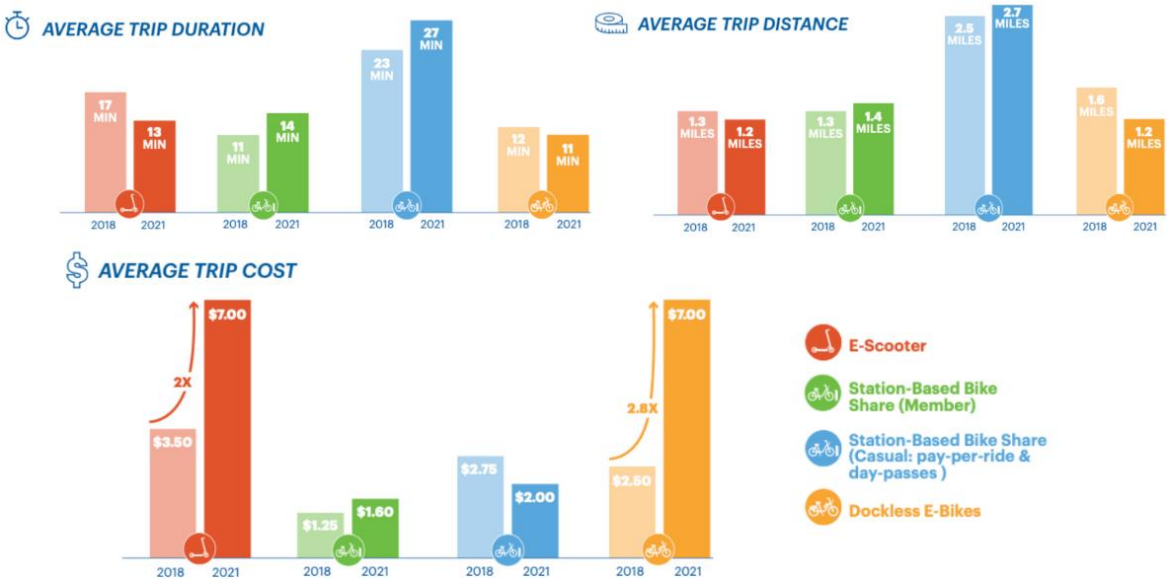
Una vez analizado la evolución de los sistemas de *bike-sharing* en Estados Unidos, vamos a comparar la frecuencia de viajes por franja horaria entre 2019 y 2020.

En base a la finalidad que tiene cada usuario para usar los sistemas, la demanda de estos puede variar. Si analizamos las horas del día en las que se registra un mayor uso de estos vehículos, no sería extraño que las más solicitadas fuesen las horas de entrada y salida del trabajo. De acuerdo con el gráfico extraído de un estudio publicado por NACTO (2022), en 2019 se veía claramente el patrón previamente mencionado, de hecho, eran las bicicletas los vehículos de transporte cuya demanda era mayor a lo largo del día con un 10% de viajes a las 08:00 y un 12% a las 20:00.

Sin embargo, en 2020 se ve claramente que la demanda es totalmente diferente ya que, durante las primeras horas de la mañana la demanda de las bicicletas compartidas no era tan grande. Además, había un reparto más equitativo a lo largo del día en comparación a las cifras de 2019. Una explicación que podría explicar esta diferencia, podrían ser los grandes cambios que ha introducido la pandemia en el sector laboral como el teletrabajo.

Evolución de la duración, distancia y precio medio de los viajes (2018-2021)

Ilustración 2. Evolución de la duración, distancia y precio medio de los viajes (2018-2021)



Fuente: externa (NACTO, 2022)

Gracias a las cifras previamente comentadas, hemos identificado que la cantidad de viajes por hora se han visto modificados entre 2019 y 2020. A continuación, vamos a analizar si los precios y las duraciones de los trayectos también se han visto modificados o mantienen las cifras previas a la pandemia. Para ello, volveremos a utilizar la información recogida por NACTO (2022), con la que compararemos la evolución en la duración y el precio de los viajes desde 2018 hasta 2021.

Si analizamos la evolución de la duración y la distancia de los viajes, observamos una estrecha relación entre ambas variables. Desde 2018 hasta 2021, la mayoría de los sistemas de *bike-sharing* experimentaron un aumento en estas dos métricas, excepto aquellos que ofrecían exclusivamente bicicletas eléctricas. Además, el costo promedio por viaje también experimentó cambios, especialmente en el caso de las bicicletas eléctricas, que suelen ser más caras que las tradicionales (Campbell et al., 2016).

Tras explicar la evolución y las principales cifras de la industria del *bike-sharing*, podemos concluir que la pandemia ha tenido un impacto significativo en el uso de estos

servicios. El número total de viajes realizados se vio reducido, así como el número de sistemas existentes en Estados Unidos. Además, tanto la duración, distancia y precio medio de cada viaje se vieron modificados, lo que nos ayuda a concluir que la pandemia puede estar detrás de estos comportamientos.

2.3.Citi Bike

Una vez expuesto las principales cifras de estos sistemas en Estados Unidos, vamos a explicar la historia que rodea a la empresa utilizada en este trabajo.

Citi Bike, lanzada el 27 de mayo de 2013, se propuso establecer uno de los mayores sistemas de bicicletas compartidas en Estados Unidos, inicialmente con 332 estaciones en Manhattan, Brooklyn, Queens y la ciudad de Jersey, y alrededor de 6.000 bicicletas. Desde entonces, ha sido fundamental para el transporte urbano, fomentando la salud y la sostenibilidad (Citi Bike, s.f.).

Durante su primer año, atrajo a más de 100.000 usuarios y generó más de 23,7 millones de kilómetros de viaje (Chaban, 2014), alcanzando un récord en Norteamérica el 6 de agosto de 2013, con 42.010 viajes en un solo día (Citi Bike, s.f.). Sin embargo, se enfrentó a desafíos técnicos como anclajes defectuosos y estaciones inoperativas que obligaron a los usuarios a buscar otras alternativas (Flegenheimer, 2013).

La repentina popularidad de los sistemas de *bike-sharing* también trajo consigo problemas financieros. En enero de 2014, la empresa responsable de los diseños, Public Bike System Company (PBSC), se declaró en bancarrota (Bloomberg, 2014). Los retrasos en las expansiones planificadas llevaron a la renuncia del gerente general en marzo de 2014 (Flegenheimer, 2014). En octubre de 2014, se formó Motivate con el objetivo de reorganizar la empresa y prometió mejoras y expansión mediante grandes inversiones. A pesar de estos esfuerzos, los problemas técnicos persistieron, lo que resultó en un cierre temporal en marzo de 2015 para realizar mejoras (Fitzsimmons, 2015).

Después de corregir errores, Citi Bike continuó su expansión con más estaciones en Queens y Brooklyn en 2015, seguidas de nuevas estaciones en Manhattan y Brooklyn en 2016. En julio de 2018, Lyft adquirió Motivate, asumiendo el control de Citi Bike

(Hardawar, 2018) e introduciendo bicicletas eléctricas de pedaleo asistido al mes siguiente, las cuales fueron muy bien recibidas y rápidamente se popularizaron (Berguer, 2018). Se anunció una expansión a cinco años con una inversión de 100 millones de dólares para ampliar el área de servicio a 91 kilómetros cuadrados y aumentar la flota de bicicletas de 12.000 a 40.000, convirtiéndose en el mayor sistema de *bike-sharing* del mundo (Hawkins, 2018).

Desde entonces, ha introducido una mayor flota de bicicletas eléctricas y en 2020 firmó un acuerdo para operar en Jersey City y Hoboken durante al menos cinco años (Gannon, 2020). En 2021, durante la pandemia de 2019, Citi Bike superó su récord anterior de viajes en un día y alcanzó su viaje número 100 millones (Citi Bike, s.f.). Finalmente, en 2023, Lyft anunció planes para duplicar la cantidad de bicicletas eléctricas en los próximos años (Lee, 2023).

2.4.Competidores

Como se ha comentado previamente, los servicios de *bike-sharing* en Estados Unidos contaban con más de 60 sistemas en 2023, lo que demuestra el alto grado de competencia de esta industria. Es por eso, que las empresas tratan de diferenciarse de la competencia ya sea mediante diferentes promociones o precios. A continuación, vamos a explicar los principales competidores que tiene Citi Bike en la región de Nueva York.

- **CycleHop:** Fundada en 1997 por Josh Squire, ofrece un sistema de bicicletas en Chicago y otras ciudades de Estados Unidos y Canadá, incluyendo Nueva York. Su flota incluye bicicletas y patinetes eléctricos (CycleHop, s.f.).
- **Spinlister:** Fundada en 2011 por Will Dennis y Jeff Noh, permitía a los usuarios alquilar bicicletas por períodos prolongados, inicialmente en áreas como Nueva York y San Francisco, y recientemente también en Europa. A pesar de varios desafíos financieros, consiguieron diversificar su negocio incluyendo el alquiler de tablas de surf y snowboard (Spinlister, s.f.).
- **JOCO:** Fundada en Nueva York en 2020, ofrece una plataforma integral para repartidores y empresas que utilizan vehículos eléctricos. Recientemente lanzaron

su propio sistema de *bike-sharing* en la ciudad de Nueva York, con estaciones ubicadas en parcelas privadas en lugar de la vía pública, lo que ayudaba a minimizar el impacto en el espacio público (JOCO, s.f.).

A pesar de que Citi Bike tiene ciertos competidores repartidos por esta ciudad, sigue siendo el patrocinador oficial del *bike-sharing* en esta región, lo que le permite mantener una cuota de mercado significativamente superior a la de la competencia.

Capítulo 3: Metodología de extracción, exploración, limpieza y creación de variables

Una vez explicado el contexto que rodea a la industria del *bike-sharing* y a los servicios de Citi Bike, es hora de descargar, limpiar y analizar los datos que vamos a utilizar para este análisis. A continuación, explicamos mediante el siguiente flujograma las acciones que se han ido tomando para asegurarnos que trabajamos con unos datos libres de errores y de ruido que pueda interferir en nuestro análisis.

Ilustración 3. Flujograma con la metodología de nuestro análisis

Fuente: elaboración propia

Dividiremos este apartado en las siguientes secciones: en primer lugar, explicaremos cómo se ha llevado a cabo la extracción de los datos para el análisis. En segundo lugar, analizaremos y definiremos las variables contenidas en estos datasets. En tercer lugar, aplicaremos diferentes técnicas para el tratamiento de nuestros datos y, finalmente, explicaremos el proceso de creación de nuevas variables en nuestros datasets.

3.1. Extracción de los datos

Para poder llevar a cabo este análisis, primero hay que identificar los datos con los que vamos a trabajar y así poder descargarlos. Vamos a necesitar las bases de datos de Citi Bike que recojan todos los viajes realizados en 2019, 2020 y 2022. Además, necesitaremos descargar las características sociodemográficas de las zonas de Nueva York para poder estudiar posibles patrones o conductas de comportamiento entre los usuarios de este servicio.

Bases de datos de Citi Bike

Para descargar las bases de datos de Citi Bike, que contienen los viajes de los usuarios, hay que acceder a su página web, donde están disponibles como Open Data para el público general.

En este apartado se encuentran los archivos con los viajes realizados desde 2013 hasta la fecha, divididos en dos regiones: Nueva Jersey y Nueva York, las principales áreas donde opera la compañía. Además, en la propia página web, también viene una breve explicación de las diferentes variables que componen estos archivos, que se explicarán en detalle más adelante.

Como se ha mencionado previamente, queremos descargarnos todos los ficheros que contengan los viajes de los usuarios de 2019, 2020 y 2022 para así poder estudiar si ha habido una evolución de los viajes desde la COVID-19. Sin embargo, estos archivos son mensuales, es decir, que tendríamos que descargarnos 36 ficheros para poder tener los viajes comprendidos en esos años.

Dado el gran número de archivos, se ha procedido a utilizar la herramienta Python para automatizar la descarga. Dentro de esta herramienta, hemos desarrollado un código que se conecta a la página web de Citi Bike y descarga todos los archivos de la plataforma comprendidos entre un rango temporal que nosotros le especificamos, en este caso 2019, 2021 y 2022.

Por defecto, los archivos se descargan en formato ZIP. Luego, el código los descomprime y extrae los archivos en formato CSV. También hemos agregado líneas de código para combinar los archivos mensuales de cada año en uno solo, facilitando así la organización de los viajes de 2019, 2020 y 2022 en tres archivos diferentes.

Por otro lado, dado que cada base de datos contiene más de 30 millones de observaciones, hemos descargado una muestra aleatoria representativa del 10% de cada conjunto de datos para facilitar el manejo de datos en nuestro análisis.

Datos sociodemográficos

Una vez se han descargado los datos con los viajes de los usuarios de Citi Bike de 2019, 2020 y 2022, procederemos a buscar las diferentes características sociodemográficas de la población de Nueva York. Esta ciudad es la más poblada de Estados Unidos y una de las más pobladas del mundo. A fecha de 2024, este estado cuenta con más de 19,84 millones de habitantes, 3 veces más grande que la población de la Comunidad de Madrid.

Para hacer más preciso el análisis, se procederá a descargar las características sociodemográficas de esta ciudad agrupando por código postal, razón que se explicará en detalle más adelante.

Al revisar diversas fuentes oficiales para obtener esta información, he decidido extraer los datos de la página web *Simple Maps*. Esta plataforma recopila múltiples variables sociales, demográficas y económicas por código postal de los diferentes estados de Estados Unidos (Nueva York entre ellos). Los datos se extraen del censo oficial y la última actualización fue realizada en enero de 2024, lo que garantiza que contamos con información actualizada y relevante para nuestro análisis.

Al tratarse de una página web externa al censo, había bastantes problemas para poder descargar los datos. De hecho, o pagabas una cantidad de dinero o no te dejaba exportar

los ficheros en formato Excel o CSV. Como alternativa, pensé en descargarme estos datos utilizando la técnica de *web scrapping*, que permite automatizar la recopilación de datos de sitios web de manera eficiente. Sin embargo, debido a que el dueño de la página web tenía bloqueado este tipo de técnicas en su plataforma, no tuve ningún otro remedio que descargarme la información manualmente.

3.2. Análisis de los datos descargados

Una vez explicado la metodología que hemos empleado para la extracción de nuestros datasets, es hora de explicar qué información contiene cada uno de ellos. Para ello, dividiremos este apartado en dos secciones, la descripción de las variables y un análisis exploratorio de la información contenida en los datos del análisis.

Descripción de las variables

Una vez hemos descargado la información necesaria para poder llevar a cabo nuestro análisis, procederemos a visualizar la descarga para ver si estos datasets tienen estructuras parecidas y son comparables, así como para entender las variables que los componen. Para realizar este proceso, vamos a continuar utilizando la herramienta de Python.

En primer lugar, vamos a mostrar un resumen sencillo de cada dataset, para observar el número de observaciones y variables que contiene cada uno de estos. Como se puede ver claramente en la tabla 2, el número de variables de los datasets de 2019 y 2020 es el mismo; sin embargo, el dataset de 2022 no tiene el mismo número de variables, aspecto que explicaremos en detalle más adelante.

Por otro lado, si nos fijamos en la tabla resumen de las características sociodemográficas de Nueva York, se ve claramente que el número total de las observaciones es mucho menor en comparación con las otras bases de datos. Además, hemos construido nuestro dataset solo utilizando los códigos postales en los que opera Citi Bike en Nueva York.

Tabla resumen del número de observaciones y variables por dataset

Tabla 2. Número de observaciones y variables por dataset

Nombre	Nº Observaciones	Nº Variables
Dataset 2019	4.110.339	19
Dataset 2020	3.901.371	19
Dataset 2022	6.123.801	17

Nombre	Nº Observaciones	Nº Variables
Datos Sociodemográficos de NY	89	21

Fuente: elaboración propia

Al observar el número de observaciones (viajes) que tienen los datasets de Citi Bike, podemos responder a una de las preguntas principales de este trabajo. Si comparamos la volumetría de viajes pre y post pandemia se aprecia claramente que ha habido un aumento del 50%, por lo que podemos inferir que el uso del sistema de bicicletas compartidas ha experimentado un crecimiento significativo durante estos años.

Habiendo explicado brevemente el tamaño de cada una de nuestras bases de datos es hora de explicar las variables que las componen. Para ello, se va a construir una tabla con una explicación detallada de cada una de las variables de los datos proporcionados por Citi Bike y por *Simple Maps*, que se muestra en la siguiente página.

Tabla resumen de las variables de los datasets de Citi Bike

Tabla 3. Resumen de las variables de los datasets de Citi Bike

Variable	**Tipo de Variable**	**Definición**
Tripduration	Numérica	Recoge la duración total del viaje hecha por el usuario
Startime	Fecha	Fecha en la que el usuario comenzó el trayecto. El formato de la fecha es año, mes, día y hora
Stoptime	Fecha	Fecha en la que el usuario finalizó el trayecto. El formato de la fecha es año, mes, día y hora
StartStation ID	Numérica	Clave identificativa de la estación *(dock)* en la que se inició el trayecto
StartStation name	Texto	Nombre de la estación en la que se inició el trayecto
StartStation latitude	Numérica	Latitud de la estación en la que se inició el trayecto
StartStation longitude	Numérica	Longitud de la estación en la que se inició el trayecto
EndStation ID	Numérica	Clave identificativa de la estación *(dock)* en la que se finalizó el trayecto
EndStation name	Texto	Nombre de la estación en la que se finalizó el trayecto
EndStation latitude	Numérica	Latitud de la estación en la que se finalizó el trayecto
EndStation longitude	Numérica	Longitud de la estación en la que se finalizó el trayecto
BikeID	Numérica	Clave identificativa de la bicicleta usada en cada trayecto
Usertype	Categórica	Tipo de usuario de cada trayecto. Pude tomar los valores "suscriber" o "customer" en función del plan contratado por el usuario
Birth Year	Numérica	Año de nacimiento del usuario del trayecto
Gender[2]	Categórica	Género del usuario del trayecto. Puede ser hombre (1), mujer (2) o desconocido (0)
Rideable type[3]	Categórica	Especifica si la bicicleta utilizada es eléctrica, clásica o de tipo *dock*

Fuente: elaboración propia

[2] Solo incluida en los ficheros anteriores a 2021
[3] Solo incluida en los ficheros de 2021 en adelante

Como se ha comentado anteriormente, los archivos que contienen los viajes realizados por los usuarios de Citi Bike en 2019, 2020 y 2022 tenían un número diferente de variables. Esta diferencia proviene principalmente por la falta de variables como Gender o Birth Year, que dejaron de incluirse en los datos para proteger cierta información personal de los usuarios. Por otro lado, la variable Tripduration tampoco viene incluida en los datos de 2022; sin embargo, en secciones posteriores explicaremos la metodología empleada para calcular esta variable.

Al igual que hemos explicado las variables de los ficheros de Citi Bike, vamos a explicar mediante la siguiente tabla las variables sociodemográficas que hemos descargado para la ciudad de Nueva York.

Tabla resumen de las variables sociodemográficas del dataset de Nueva York

Tabla 4. Resumen de las variables sociodemográficas de Nueva York

Variable	**Tipo de Variable**	**Definición**
Zipcode	Numérica	Código postal al que pertenecen los datos sociodemográficos
County	Texto	Estado del código postal
Density per meter squared	Numérica	Densidad por metro cuadrado
Median individual income	Numérica	Mediana del ingreso por individuo en dólares
Median household income	Numérica	Mediana del ingreso por vivienda en dólares
White	Numérica	Porcentaje de personas de raza blancas
Black	Numérica	Porcentaje de personas de raza negra
Asian	Numérica	Porcentaje de personas de raza asiática
Male	Numérica	Porcentaje de personas que son hombre
Rent	Numérica	Renta media en dólares
Median house value	Numérica	Mediana del valor de las viviendas
Median age	Numérica	Mediana de la edad
College degree or more	Numérica	Porcentaje de las personas que tienen un nivel de educación superior a la ESO
Married	Numérica	Porcentaje de las personas que están casadas

Labour force	Numérica	Porcentaje de las personas activamente ocupadas
Unemployed	Numérica	Porcentaje de desempleo
Commute time to work	Numérica	Tiempo medio que emplean las personas en ir a trabajar
No healthcare insurance	Numérica	Porcentaje de personas que no tienen seguro médico
Poverty	Numérica	Porcentaje de personas que son consideradas pobres

Fuente: elaboración propia

Exploración de las bases de datos

Una vez explicadas las principales variables que componen estas dos bases de datos del trabajo, es hora de conocer los diferentes valores que tienen cada una de ellas. Para ello, vamos a dividir este apartado en tres partes.

En primer lugar, mostraremos una imagen de las primeras líneas de cada base de datos para que se pueda ver de manera rápida como están estructuradas las variables y que tipo de información contienen cada una de ellas. En segundo lugar, se analizarán las principales estadísticas descriptivas de las variables más relevantes para nuestro análisis y así poder ver la tendencia central y la dispersión de estas. Finalmente, se mostrarán los diagramas de violín de estas variables con el objetivo de mostrar e identificar de manera gráfica los valores atípicos y comparar distribuciones (Buttarazzi et al., 2018).

Primeras líneas de cada base de datos

Como se ha comentado anteriormente, este apartado está dedicado a enseñar las primeras líneas de nuestras bases de datos. Esto va a servir de apoyo para mostrar los tipos de datos que se están manejando y cómo están estructurados. Además, también permite al lector visualizar de forma rápida y sencilla los datos con los que trabajaremos a lo largo del análisis. A continuación, se muestran las primeras líneas de los datos de Citi Bike, y las características sociodemográficas de Nueva York.

Ilustración 3. Primeras líneas del dataset de Citi Bike 2019

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	birth year	gender
456	2019-12-11 20:25:55.7620	2019-12-11 20:33:32.2410	469.0	Broadway & W 53 St	40.763441	-73.982681	442.0	W 27 St & 7 Ave	40.746647	-73.993915	28073	Subscriber	1974	1
1115	2019-06-10 08:38:33.8950	2019-06-10 08:57:09.6130	388.0	W 26 St & 10 Ave	40.749718	-74.002950	303.0	Mercer St & Spring St	40.723627	-73.999496	33226	Subscriber	1997	1
163	2019-04-10 22:55:33.5640	2019-04-10 22:58:17.2950	336.0	Sullivan St & Washington Sq	40.730477	-73.999061	229.0	Great Jones St	40.727434	-73.993790	31425	Subscriber	1984	1
1107	2019-05-27 12:32:14.5030	2019-05-27 12:50:41.9620	2008.0	Little West St & 1 Pl	40.705693	-74.016777	303.0	Mercer St & Spring St	40.723627	-73.999496	34253	Subscriber	1996	1

Fuente: elaboración propia

Ilustración 4. Primeras líneas del dataset de Citi Bike 2020

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	birth year	gender
6994	2020-06-10 16:32:59.2810	2020-06-10 18:29:33.3780	3437	Riverside Dr & W 91 St	40.793135	-73.977004	3135	E 75 St & 3 Ave	40.771129	-73.957723	43122	Customer	1997	2
1014	2020-11-06 13:45:08.4430	2020-11-06 14:02:02.7160	3553	Frederick Douglass Blvd & W 112 St	40.801694	-73.957145	4020	Bradhurst Ave & W 148 St	40.825125	-73.941616	38869	Subscriber	1996	2
504	2020-03-07 21:11:08.2940	2020-03-07 21:19:32.9090	456	E 53 St & Madison Ave	40.759711	-73.974023	447	8 Ave & W 52 St	40.763707	-73.985162	38566	Subscriber	1972	1
1613	2020-02-04 22:39:53.3130	2020-02-04 23:06:46.8440	519	Pershing Square North	40.751873	-73.977706	483	E 12 St & 3 Ave	40.732233	-73.988900	35386	Subscriber	1963	2

Fuente: elaboración propia

Ilustración 5. Primeras líneas del dataset de Citi Bike 2022

rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
classic_bike	2022-04-30 12:50:34	2022-04-30 12:58:29	Broadway & W 29 St	6289.06	W 42 St & 6 Ave	6517.08	40.746201	-73.988557	40.754920	-73.984550	member
electric_bike	2022-07-04 13:11:24	2022-07-04 13:12:51	6 Ave & W 34 St	6364.1	6 Ave & W 33 St	6364.07	40.749640	-73.988050	40.749013	-73.988484	member
classic_bike	2022-12-07 21:51:34	2022-12-07 21:56:34	35 Ave & 37 St	6563.12	Steinway St & 28 Ave	6915.02	40.755737	-73.923542	40.765625	-73.913669	member
classic_bike	2022-05-16 09:52:06	2022-05-16 09:53:09	W 56 St & 6 Ave	6809.07	W 56 St & 6 Ave	6809.07	40.763278	-73.977683	40.763406	-73.977225	member

Fuente: elaboración propia

Como se ha comentado anteriormente, el número de variables de los datos de 2022 era diferente al de 2019 y 2020. Si analizamos estas tres últimas ilustraciones, podemos ver claramente que los datos de los dos primeros años tienen estructuras idénticas; sin embargo, para 2022 estos son ligeramente diferentes incluyendo nuevas variables como

rideable_type que nos permite ver qué tipo de bicicleta ha usado cada cliente (clásica, eléctrica o *dockless*).

Una vez mostrado la estructura de los datasets de Citi Bike, vamos a repetir este análisis para los datos sociodemográficos de Nueva York. Como se ha mencionado anteriormente, nuestro dataset está compuesto por más de 20 variables, por lo que solo mostraremos las más relevantes en este apartado.

Ilustración 6. Primeras líneas del dataset sociodemográfico de Nueva York

zipcode	county	population	white	male	rent	median age	married	unemployed	family size	commute time to work	no healthcare insurance	poverty
10001	New York	26,966	63%	47%	22,690	35.8	28%	4%	2.9	26.3	3%	14%
10002	New York	76,807	32%	50%	1,047	43.7	36%	8%	3.1	32.5	6%	27%
10003	New York	54,447	71%	50%	2,478	32.0	26%	5%	2.7	24.4	3%	11%
10004	New York	4,795	58%	53%	3,501	33.9	58%	2%	3.1	26.3	1%	3%
10005	New York	8,637	75%	45%	3,501	30.3	32%	3%	2.8	24.0	1%	4%
10006	New York	3,894	65%	47%	3,398	31.9	39%	4%	2.9	30.4	11%	11%

Fuente: elaboración propia

Medidas estadísticas de las variables más relevantes

Una vez visualizada la estructura de cada base de datos es hora de calcular las principales estadísticas de algunas variables. En este caso no hemos optado por mostrar la de todas las variables ya que, dentro de cada dataset tenemos variables de texto o categóricas. Además, hay variables que muestran coordenadas de longitud o latitud en las que no es muy representativo mostrar la media o el valor máximo y mínimo.

Es por eso, que las variables que vamos a mostrar del dataset de Citi Bike van a ser *Tripduration, Birth Year* y *Gender* para los datasets de 2019 y 2020 puesto que para 2022 estas variables o no existen o habrá que calcularlas en secciones posteriores.

Tabla 5. Estadísticas principales de los datasets de Citi Bike para 2019 y 2020

	tripduration_2019	birth year_2019	gender_2019	tripduration_2020	birth year_2020	gender_2020
count	4110339.0	4110339.0	4110339.0	3901371.0	3901371.0	3901371.0
mean	978.6	1980.2	1.2	1313.2	1981.2	1.2
std	10495.8	12.1	0.5	14011.5	12.4	0.6
min	61.0	1874.0	0.0	61.0	1873.0	0.0
25%	362.0	1970.0	1.0	424.0	1969.0	1.0
50%	615.0	1983.0	1.0	780.0	1984.0	1.0
75%	1079.0	1990.0	1.0	1403.0	1991.0	2.0
max	3379585.0	2003.0	2.0	3716471.0	2004.0	2.0

Fuente: elaboración propia

Gracias a las estadísticas de estas variables podemos empezar a sacar las primeras conclusiones.

- Si nos fijamos en el número total de observaciones, vemos claramente que en 2019 había un mayor número de observaciones que en 2020, lo que nos indica que en ese año hubo un mayor volumen de viajes realizados por los usuarios.

- Cuando analizamos la variable *Tripduration* podemos ver que la media de segundos de cada viaje en 2020 era un 34% mayor que en 2019. Además, si nos fijamos en el valor máximo de esta variable, podemos ver que es de casi 4 millones de segundos, lo que equivaldría a más de 45 días seguidos de viaje. Lo más probable es que nos hayamos encontrado nuestro primer valor atípico en las bases de datos, que serán tratados en secciones posteriores.

- Si realizamos el mismo análisis para la variable *Birth Year*, podemos llegar a la conclusión de que, de media, la edad de los usuarios de Citi Bike de 2019 y 2020 está cerca de los 45 años. Por otro lado, si nos fijamos en los valores mínimos de esta variable, podemos ver que hay usuarios que nacieron alrededor de 1870, por lo que hoy tendrían más de 150 años. De nuevo, parece que nos hemos topado con otro valor atípico.

- Finalmente, si analizamos la desviación estándar de estas variables, podemos ver claramente que para la variable *Tripduration* los datos de 2019 están menos dispersos que los de 2020, y que para la variable *Birth Year*, la dispersión de los valores sobre la media es de aproximadamente 12 años.

Del mismo modo que hemos hecho para los datos de Citi Bike, vamos a replicarlo para los datos sociodemográficos de Nueva York por código postal. Como tenemos más de 20 variables, hemos optado por enseñar las medidas estadísticas de las variables que consideramos más relevantes como por ejemplo la población, renta o nivel de pobreza por código postal.

Tabla 6. Estadísticas principales de las características sociodemográficas de Nueva York

	population	median_age	rent	commute_time_to_work
count	89.0	89.0	89.0	89.0
mean	49490.0	36.2	2202.7	36.3
std	27085.2	3.9	2304.5	7.1
min	3894.0	29.5	908.0	24.0
25%	29780.0	33.9	1392.0	30.4
50%	45771.0	35.7	1835.0	36.9
75%	68777.0	38.7	2455.0	42.3
max	109111.0	49.3	22690.0	47.8

Fuente: elaboración propia

De nuevo, gracias a estas medidas somos capaces de llegar a ciertas conclusiones a cerca de las características sociales, demográficas y económicas de Nueva York:

- Si nos fijamos en la variable que marca la población por código postal, podemos observar una desviación estándar que representa casi un 50% del valor de la media, lo que significa la alta variabilidad poblacional de los códigos postales de Nueva York.

- Otra variable incluida en nuestro conjunto de datos es la renta media que pagan los individuos por código postal. Similar a la variable de población, esta variable

muestra una desviación significativa. Esto se puede explicar porque hay zonas donde la renta mensual supera los $20.000.

- Finalmente, al analizar el tiempo promedio que cada persona tarda en desplazarse a su lugar de trabajo, observamos que la media es de aproximadamente 36 minutos. Esta variable está estrechamente relacionada con el uso del *bike-sharing* ya que, a medida que aumenta el tiempo de desplazamiento, los usuarios tienden a optar por otros medios de transporte, como autobuses o coches (Kou & Cai, 2019).

Conociendo las principales medidas estadísticas de las variables más relevantes en nuestro análisis, es momento de representar su distribución gráficamente. En lugar de usar los diagramas de caja tradicionales, utilizaremos diagramas de violín, que combinan los diagramas de bigote con un gráfico de densidad de probabilidad (Blumenschein et al., 2020). Este enfoque nos permitirá identificar visualmente cualquier valor atípico en las variables analizadas.

Antes de explicar los resultados, vamos a explicar mediante la siguiente tabla la gama de colores que se va a utilizar a lo largo del trabajo para comparar los datos de diferentes años.

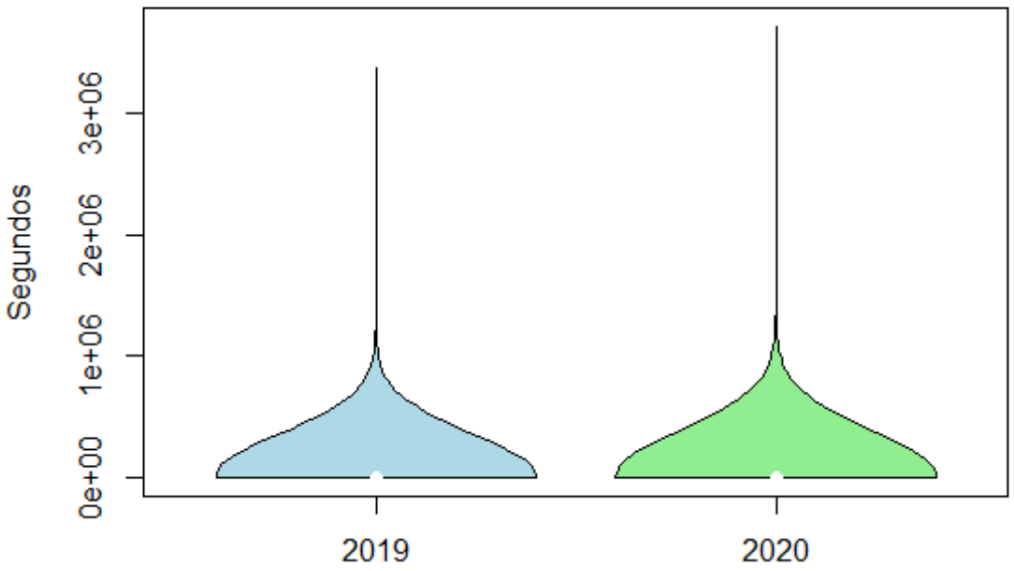
Tabla 7. Descripción de colores para comparar los datasets de nuestro análisis

Dataset	Color
Datos Citi Bike 2019	
Datos Citi Bike 2020	
Datos Citi Bike 2022	
Datos sociodemográficos NY	

Fuente: elaboración propia

Diagramas de violín para la variable *Tripduration* para los años 2019 y 2020

Ilustración 4. Diagramas de violín para la variable Tripduration para los años 2019 y 2020



Fuente: elaboración propia

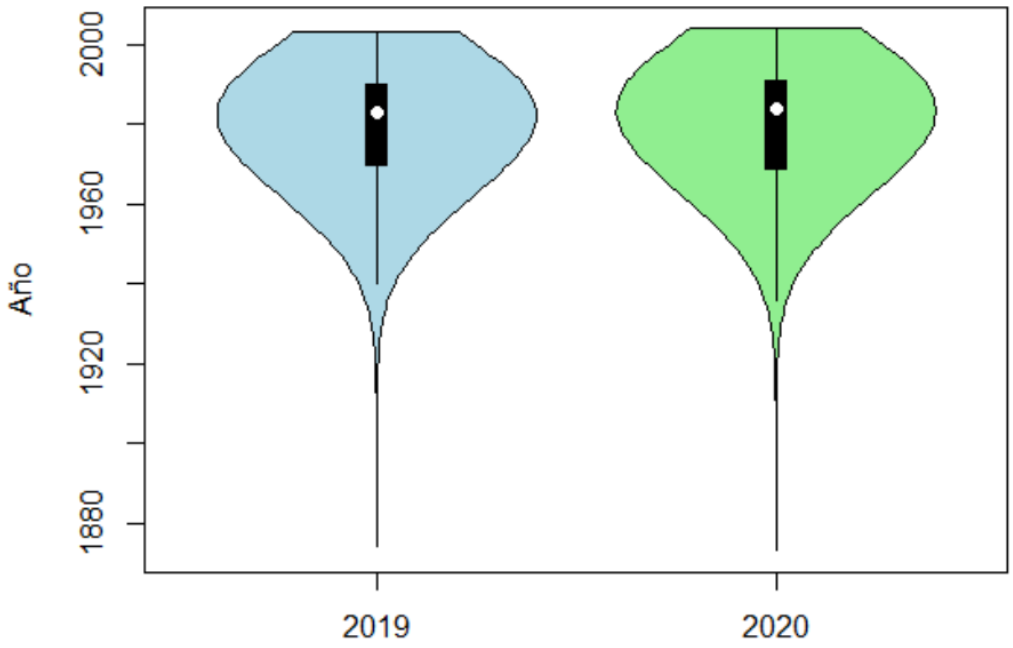
Como se mencionó anteriormente en el análisis estadístico, detectamos algunos valores atípicos en nuestros datos ya que, había viajes que superaban los tres millones de segundos. Al interpretar esta variable mediante el diagrama de violín, observamos que, debido a la gran escala de algunos valores, la caja de los cuartiles no es visible en el gráfico, lo que confirma nuevamente la presencia de valores atípicos.

Por otro lado, podemos notar que la distribución de la variable para ambos años es similar, aunque en el segundo año se registran valores máximos aún mayores. Además, es evidente que la distribución se corta a partir de los 0 segundos ya que, no puede haber viajes con una duración negativa.

Estas distribuciones se concentran en torno a la mediana (punto blanco), que para los años 2019 y 2020 fue de 615 y 780 segundos, respectivamente. Por lo tanto, en secciones posteriores, necesitaremos tratar estos valores atípicos para evitar interferencias en nuestro análisis.

Diagramas de violín para la variable *Birth Year* para los años 2019 y 2020

Ilustración 5. Diagramas de violín para la variable Birth Year para los años 2019 y 2020



Fuente: elaboración propia

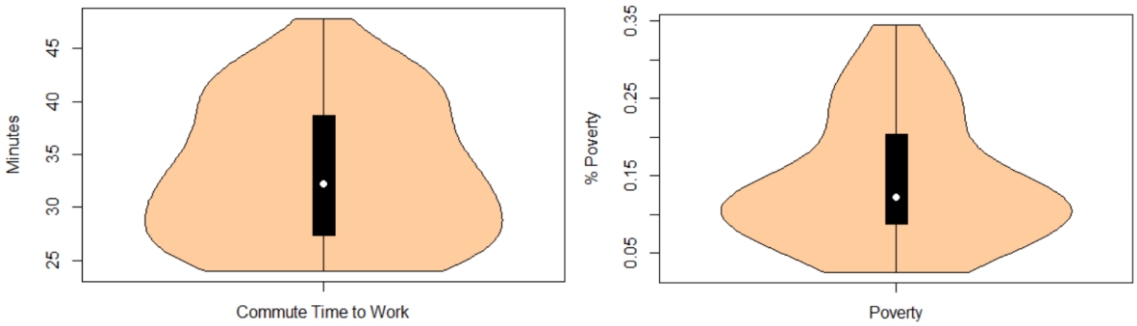
Si repetimos el proceso para la variable *Birth Year* vemos que en este gráfico podemos apreciar la caja intercuartílica propia de un diagrama de bigotes. Vemos como la mayor parte de la distribución ronda los valores de la mediana, que eran de 1983 (41 años) para 2019 y de 1.984 (40 años) para 2020.

Por otro lado, volvemos a confirmar la presencia de outliers ya que, para ambos años, hay observaciones que toman los valores cercanos a 1870, que significaría que los usuarios de Citi Bike tendrían 150 años.

Finalmente, podemos apreciar que a medida que la variable se acerca al año 2000, la distribución se corta. Esto se puede explicar ya que, Citi Bike solo deja utilizar sus bicicletas a aquellas personas mayores de 20 años.

Diagramas de violín para los datos sociodemográficos de Nueva York

Ilustración 6. Diagramas de violín para los datos sociodemográficos de Nueva York



Fuente: elaboración propia

Al igual que hemos hecho los diagramas de violín para alguna variable de los datos de Citi Bike, vamos a repetir este proceso para alguna variable del dataset que contiene las características sociodemográficas de Nueva York. En este caso, hemos elegido las variables de *Commute time to work*, que reflejaba la media de tiempo en minutos que tardan las personas en ir a trabajar, y *Poverty*, que indicaba el porcentaje de pobreza presente en cada uno de los códigos postales.

Gracias a estos dos gráficos podemos extraer las siguientes conclusiones:

- Si nos fijamos en el gráfico de la variable *Commute time to work* podemos observar que el grueso de su distribución estaba en torno a los 27 minutos, un 37% menor que la mediana de la variable (37 minutos). Además, si nos fijamos en el límite superior e inferior de la variable, vemos que no hay ningún valor que supere estos, por lo que no hay presencia de valores atípicos.

- Por otro lado, en el gráfico de la variable *Poverty* se observa claramente que la mayor parte de la distribución de la variable se encuentra entorno al 12%, muy cercano a la mediana de la variable (13%). Del mismo modo que pasaba con la variable anterior, tampoco hay presencia de ningún valor atípico que haya que tratar.

Tratamiento de las variables

Una vez explicadas y mostradas las variables que componen nuestras bases de datos, así como su estructura, es hora de limpiar los datos para eliminar cualquier tipo de ruido en nuestro análisis. En esta sección, nos enfocaremos exclusivamente en los conjuntos de datos que registran los viajes realizados por los usuarios de Citi Bike ya que, los datos sociodemográficos que hemos descargado ya vienen completamente limpios de la fuente.

Tabla 8. Resumen del tratamiento de variables

Acciones	**Variables Afectadas**
Tratamiento de valores nulos	Todas
Modificación de las variables	Nombres de las variables, *Tripduration* y *Usertype*
Tratamiento de valores atípicos	Variables Numéricas

Fuente: elaboración propia

Hemos dividido este apartado en tres partes: tratamiento de valor nulos, corrección manual de determinadas variables y tratamiento de valores atípicos. Para esta sección utilizaremos la herramienta de R Studio, que también hemos utilizado a lo largo de carrera.

Tratamiento de valores nulos

Para garantizar la mayor precisión posible en nuestros datos, debemos asegurarnos de que todas las observaciones contengan valores para todas las variables de nuestras bases de datos. Utilizando unas líneas de código, identificamos las observaciones que presentan valores faltantes y las eliminamos de nuestros datos. A continuación, se muestra el número de observaciones eliminadas de cada base de datos.

Tabla 9. Resumen del número de valores nulos por dataset

Dataset	Número de observaciones eliminadas	% Total observaciones
Citi Bike 2019	312	0,074%
Citi Bike 2020	0	0%
Citi Bike 2022	2.571	0,041%

Fuente: elaboración propia

Hemos optado por eliminar los valores nulos ya que, como se muestra en la tabla, su cantidad es muy pequeña en comparación con el número total de observaciones. Por lo tanto, la eliminación de estos valores no representa una pérdida significativa de datos para nuestro análisis.

Modificación de las variables

Después de eliminar las observaciones que no contenían todos los valores para las diferentes variables, vamos a modificar ciertas variables para asegurar la comparabilidad de los datos en las tres bases de datos de Citi Bike. A continuación, se explican los cambios que se han hecho.

- **Nombres de las variables:** al utilizar R Studio, hemos encontrado que los nombres de variables con espacios pueden ocasionar errores en ciertas líneas de código. Para resolver este problema, hemos decidido reemplazar todos los espacios por barras bajas (_). Para 2022, las variables ya vienen todas con esta estructura, por tanto, a partir de ahora se van a modificar los nombres de las variables de 2019 y 2022 de acuerdo con esta estructura.

- **Variable *Tripduration* del dataset de 2022:** anteriormente, no se podían mostrar las principales medidas estadísticas ni el diagrama de violín de la variable *Tripduration* para los datos de Citi Bike de 2022. ara obtener esta variable, observamos que este conjunto de datos incorpora las variables de inicio y fin de viaje en formato fecha y hora (aaaa/mm/dd hh:mm). Por lo tanto, para calcular

esta variable, restamos la fecha en que el usuario dejó la bicicleta a la fecha en que inició su viaje.

Hemos decidido no volver a mostrar sus principales estadísticas ya que, siguen estructuras muy parecidas a las de los años anteriores, que ya han sido mostradas.

- **Variable Usertype para el dataset de 2022:** al analizar los valores de la variable Usertype en las diferentes bases de datos, se observa que tanto en 2019 como en 2020 los valores son "subscriber" o "customer". Sin embargo, en el dataset de 2022, esta variable presenta una estructura diferente al utilizar los valores "member" o "subscriber". Para asegurar la homogeneidad entre los tres archivos, hemos reclasificado los valores de la variable de 2022, sustituyendo "member" por "customer".

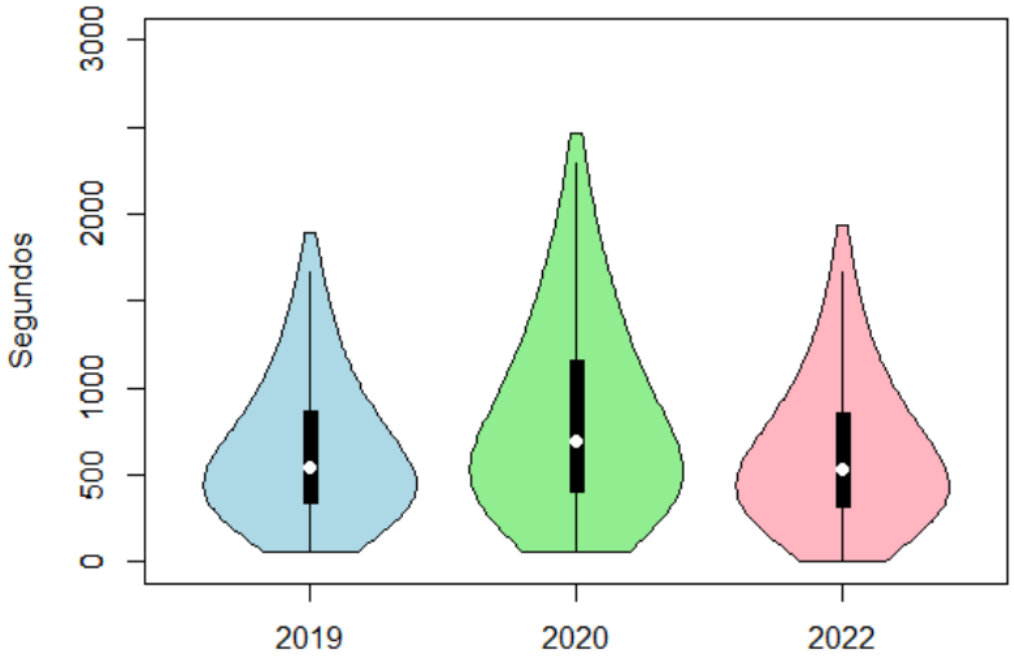
Tratamiento de valores atípicos

Una vez hemos modificado estas variables, es hora de tratar todos aquellos valores atípicos que hemos identificado previamente en los diagramas de violín. Existen diferentes métodos para poder tratar estos datos; sin embargo, para nuestro análisis hemos decidido fijarnos en el número de desviaciones que tiene cada valor sobre la media. Es decir, vamos a escalar las variables numéricas de nuestro *dataset* y si el número de desviaciones (en valor absoluto) es superior a 2, trataremos esos valores como *outliers.*

Gracias a este método, somos capaces de identificar los valores atípicos de nuestro dataset de manera robusta y con menor propensión a sesgos causados por valores extremos en variables con escalas muy diferentes. En el anexo A del trabajo, se muestra la ilustración de las líneas de código utilizadas para poder tratar estos *outliers* en los datos de 2019, que se replicarán para los siguientes años.

Finalmente, para comprobar que nuestro código ha funcionado vamos a volver a mostrar los diagramas de violín de la variable *Tripduration* para los tres años. Como se puede observar en la siguiente ilustración, nuestro dataset está limpio de valores atípicos al no tener ninguna observación cuya desviación esté dos veces por encima o por debajo de la media.

Ilustración 7. Comprobación de que los outliers han sido tratados



Fuente: elaboración propia

3.3.Creación de nuevas variables para nuestro análisis

En los apartados anteriores, hemos ido limpiando nuestras bases de datos para que la información fuese lo más precisa posible. Además, para poder llevar nuestro análisis un paso más allá, hemos optado por crear nuevas variables en función de las que ya teníamos. Durante este apartado explicaremos la manera en la que se han construido estas nuevas variables y la utilidad que les vamos a dar durante el análisis.

Tabla 10. Resumen de las variables creadas

Variables Creadas	**Datasets Afectados**
Distancia Media (km)	2019, 2020 y 2022
Velocidad Media (km/h)	2019, 2020 y 2022
Códigos Postales	2019, 2020 y 2022

Fuente: elaboración propia

Las variables que se van a crear son la distancia media de los trayectos, velocidad media de cada viaje, y a qué código postal corresponden las estaciones de inicio y fin de viaje. Para esta sección volveremos a utilizar Python.

Distancia media

Como se ha mostrado en secciones anteriores, las bases de datos de Citi Bike, estaban compuestas por variables geográficas como la calle en la que se inició y se finalizó el viaje, así como las coordenadas que corresponden a esas direcciones. Sin embargo, a pesar de tener toda esta información, no tenemos ninguna variable que nos indique la distancia recorrida en cada viaje. Es por eso, que hemos decidido calcular la distancia media recorrida por cada usuario para saber el rango de distancias que se suelen recorrer en los viajes, y analizar si esta ha disminuido o aumentado después de la pandemia.

Para poder calcular esta distancia, hemos utilizado la fórmula de Haversine. Esta comenzó a utilizarse en 1801 aunque no fue hasta el año 1805 cuando fue publicada por el astrónomo James Andrew. Esta fórmula revolucionó el mundo de la navegación marítima al permitir calcular la distancia circular máxima entre dos puntos de un globo, sabiendo su latitud y longitud (Dauni et al., 2019). Como hemos mencionado previamente, tenemos las coordenadas de latitud y longitud de las estaciones del inicio y fin de viaje. Por tanto, podremos calcular la distancia media que hay entre una estación y la otra con la que podremos tener una aproximación de la distancia recorrida en cada viaje.

Sin embargo, no podemos calcular la distancia real que ha recorrido cada usuario ya que, puede haber casos en los que el usuario haya cogido una bicicleta en la parada más cercana a su casa, se haya ido a dar una vuelta, y luego haya vuelto a dejar la bicicleta en la misma parada. En este caso, la distancia que nos saldría sería de 0, aunque en verdad tendría que ser mucho más.

A continuación, se explicarán las principales fórmulas que hemos llevado a cabo en Python, aunque en el anexo B se muestra de forma visual las líneas de código utilizadas.

Los pasos que hemos ido haciendo para poder calcular estas distancias han sido:

1. Crear una función "Haversine" en la que introducimos los parámetros de latitud y longitud de las estaciones de inicio y final de viaje.
2. Definimos la distancia del radio de la tierra para luego utiliza en la fórmula, que es de 6.371 kilómetros.
3. Debido a que la mayoría de las funciones trigonométricas como los senos y cosenos esperan que los ángulos estén en radianes, hemos convertido los ángulos formados por las coordenadas de grados a radianes
4. Calculamos la diferencia entre las diferentes coordenadas y las almacenamos en las variables *dlat* y *dlon*.
5. Almacenamos en la variable "a", un término medio de la fórmula de Haversine utilizando la siguiente fórmula:

$$\mathbf{sin^2}\left(\frac{\mathbf{dlon}}{\mathbf{2}}\right) + \boldsymbol{cos}(\boldsymbol{lat1}) * \boldsymbol{cos}(\boldsymbol{lat2}) * \boldsymbol{sin^2}(\frac{\boldsymbol{dlon}}{\mathbf{2}})$$

6. Almacenamos en la variable "c" el ángulo central de la esfera entre los dos puntos utilizando la siguiente fórmula.

$$\mathbf{2} * \mathbf{atan\,2}\left(\sqrt{\mathbf{a}}\right), \sqrt{\mathbf{1-a}}$$

7. Finalmente, calculamos cuanto es la distancia multiplicando el ángulo formado por la esfera por el radio de la tierra y le pedimos a la función que nos devuelva ese valor.

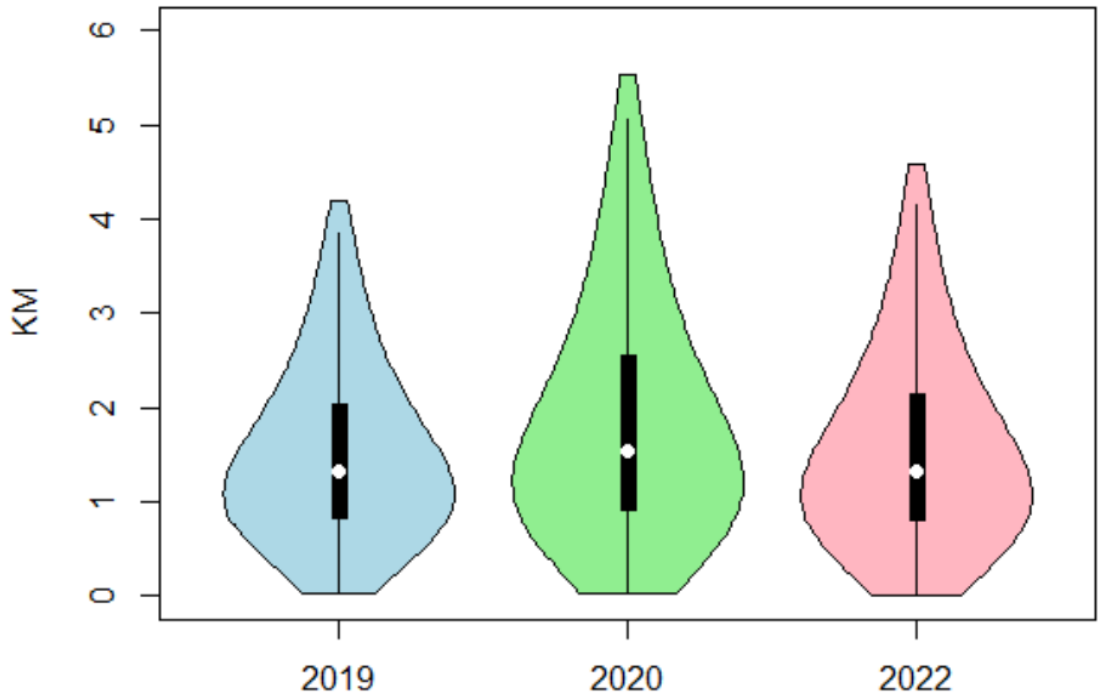
Repetimos este proceso para las otras dos bases de datos para calcular la distancia en los tres conjuntos de datos del análisis. Para limpiar la variable y abordar los valores atípicos que pueda contener, realizaremos dos iteraciones:

- Para los valores que estén por encima del límite superior, vamos a llevar a cabo un procedimiento similar al de la sección del tratamiento de *outilers*. Es decir, eliminar las observaciones cuya desviación se encuentre dos veces por encima de la media.
- Para aquellas observaciones que tengan distancia de 0 kilómetros pero que tengan una duración del viaje superior a 60 segundos, calcularemos su distancia teniendo en cuenta el siguiente cálculo: asumimos que una persona poco experimentada en

el mundo de las bicicletas consigue recorrer en 3 minutos un kilómetro, pero al tratarse de una ciudad con tanto tráfico y volumen de personas como Nueva York, añadiremos un minuto más a este tiempo. Por tanto, asumimos que cada 4 minutos se recorre 1 kilómetro.

Finalmente, para comprobar que nuestra variable está libre de errores, mostraremos en la siguiente ilustración sus diagramas de violín para los tres años del análisis.

Ilustración 8. Diagrama de violín de las distancias recorridas por los usuarios en cada dataset



Fuente: elaboración propia

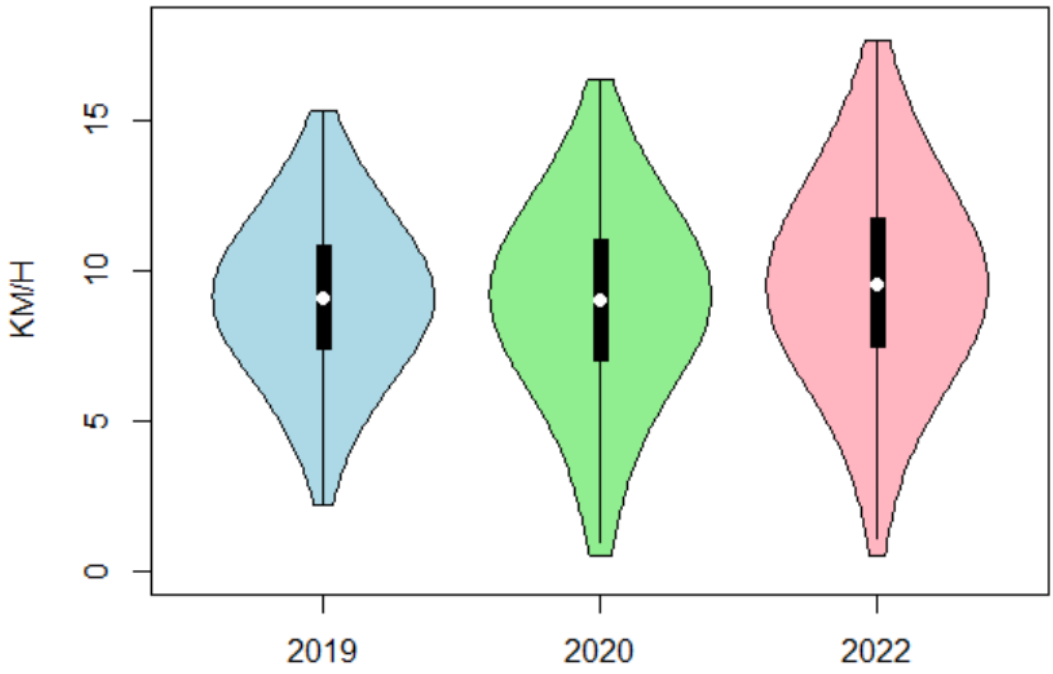
Gracias a estos gráficos, podemos concluir que el año en el que de media los usuarios realizaron un mayor número de kilómetros fue en 2020. Además, también podemos concluir que, de media, el percentil 50 de los tres datasets se encuentra entre 1,4 y 1,6 kilómetros de distancia.

Velocidad

Después de calcular cuál había sido la distancia media de cada viaje y teniendo también la duración de este, podemos calcular cuál ha sido la velocidad media a la que ha ido cada usuario en cada uno de los trayectos.

El cálculo de esta variable es muy sencillo. En primer lugar, hemos pasado a horas la duración de cada trayecto (como estaba en segundos hemos dividido por 3.600). Teniendo esta variable en este formato, hemos dividido los kilómetros recorridos por las horas empleadas para obtener la variable de kilómetros por hora. A continuación, se vuelve a mostrar los diagramas de caja de esta nueva variable.

Ilustración 9. Diagrama de violín de la velocidad media de los usuarios por año



Fuente: elaboración propia

Como ya hemos limpiado la variable *Tripduration* y Distancia, no debería haber ningún valor atípico en esta nueva variable. Si nos fijamos en los diagramas de violín de esta variable para cada uno de los años, confirmamos que la variable está limpia.

Código postal de cada estación

Como se ha comentado anteriormente, uno de los objetivos de este trabajo consistía en identificar si las características sociodemográficas de Nueva York tienen algo que ver con un mayor o menor uso de los sistemas de *bike-sharing*.

Gracias a las variables "Start Station Name" y "End Station Name", podemos conocer la estación donde el usuario inició y finalizó el servicio. Sin embargo, no disponemos de una variable que nos indique el código postal asociado a estas estaciones. Por lo tanto, en esta sección explicaremos cómo hemos procedido para obtener esta información.

Mediante el proceso de *Geocoding*, convertimos direcciones como "Calle Nuria 63, 28034, Madrid, España" en coordenadas geográficas (40.49397961, -3.70891693). Estas coordenadas pueden ser utilizadas para colocar marcadores en un mapa o para ubicar de manera precisa una ubicación en la superficie terrestre. De esta manera, transformamos direcciones en puntos geográficos que facilitan el análisis y la visualización de datos.

Al igual que podemos obtener coordenadas a partir de una dirección, existe un proceso inverso conocido como Reverse Geocoding. Como mencionamos anteriormente, nuestros conjuntos de datos incluían tanto la dirección de las estaciones como sus coordenadas, por lo que podríamos haber utilizado cualquiera de estos dos enfoques.

Sin embargo, optamos por realizar Reverse Geocoding porque las coordenadas son únicas, lo que minimiza significativamente el margen de error en comparación con la geolocalización basada en direcciones. Las líneas de código utilizadas para obtener los códigos postales se encuentran detalladas en el anexo C.

A continuación, vamos a mostrar cuál es el valor que toman estas variables y a comparar si nuestro código ha funcionado correctamente.

Ilustración 10. Comprobación de la extracción de códigos postales a partir de coordenadas

start_station_name	start_zip	start_city	end_station_name	end_zip	end_city
Broadway & W 53 St	10019	NY	W 27 St & 7 Ave	10001	NY
W 26 St & 10 Ave	10001	NY	Mercer St & Spring St	10012	NY
Sullivan St & Washington Sq	10012	NY	Great Jones St	10012	NY
Little West St & 1 Pl	10004	NY	Mercer St & Spring St	10012	NY
Central Park West & W 102 St	10025	NY	Broadway & W 56 St	10019	NY
Suffolk St & Stanton St	10002	NY	Cleveland Pl & Spring St	10012	NY

Fuente: elaboración propia

Si verificamos manualmente la primera ubicación, observamos que la calle "Broadway & W 53 St" corresponde al código postal 10019, ubicado en Nueva York. Al repetir el proceso con la calle "W 27 St & 7 Ave", confirmamos nuevamente que corresponde al código postal 10001. En ambas ubicaciones, encontramos dos estaciones de Citi Bike.

Capítulo 4: Análisis de la evolución de los hábitos de uso del servicio de Citi Bike

Después de haber explicado el proceso de creación de variables a partir de otras existentes en nuestras bases de datos, vamos a analizar la evolución de algunos aspectos entre 2019, 2020 y 2022. Pare ello, vamos a dividir el análisis en dos partes. En primer lugar, se compararán las variables que solo están presentes para los años 2019 y 2020, así como las que solo existen para 2022. En segundo lugar, se compararán aquellas variables existentes para los tres años del análisis con las que podremos observar si ha habido una evolución de cifras pre y post pandemia.

4.1. Análisis de las variables presentes únicamente en los datasets 2019-2020 y 2022

Como hemos comentado a lo largo del trabajo, hay ciertas variables que solo están presentes para alguno de los años. En este apartado analizaremos la evolución de estas variables entre 2019 y 2020, para ver si la pandemia ha podido inferir en la evolución de sus cifras. Además, también analizaremos variables presentes en los datos de 2022 que puedan ofrecernos más información acerca del uso de estas bicicletas. Las variables que se van a comentar son el género y la edad de los usuarios, bicicletas más usadas y tipo de bicicletas usadas por los usuarios.

4.1.1. Uso de las bicicletas por género (solo para 2019 y 2020)

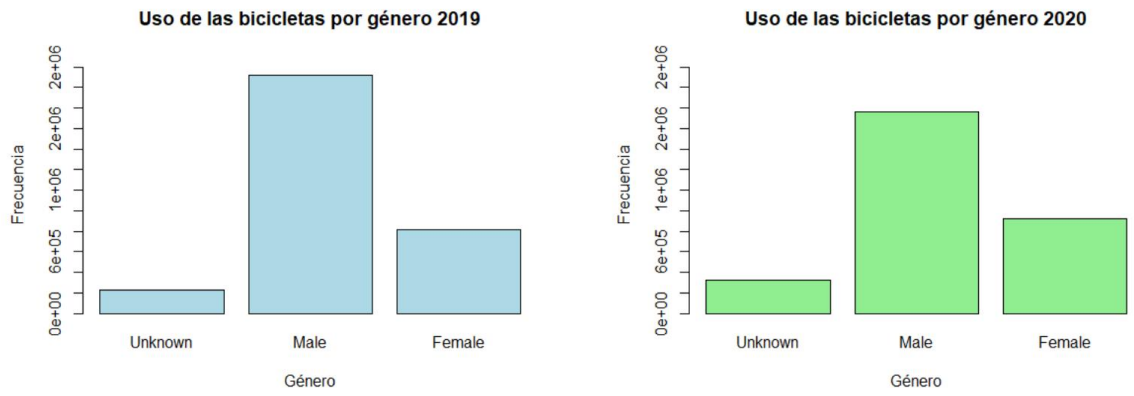
Cuando explicábamos las variables que componían cada dataset, resaltábamos que para 2022 no estaba disponible la variable de género debido a que, a partir de 2021, se dejó de publicar para cumplir con la política de protección de datos de NYCBS. Es por eso, que en esta sección solo se va a comparar la evolución del uso de bicicletas por género para los años 2019 y 2020. Como se ha mencionado anteriormente, los valores que podía tomar esta variable eran los siguientes:

Tabla 11. Resumen de la variable Gender

Valor	Descripción
0	Género desconocido
1	Hombre
2	Mujer

Fuente: elaboración propia

Ilustración 11. Uso de las bicicletas por género para 2019 y 2020



Fuente: elaboración propia

Para estudiar si ha habido algún cambio en el uso de las bicicletas por género nos vamos a fijar en la proporción de hombres y de mujeres en cada uno de los años. Para 2019, los viajes realizados por los hombres representaban el 69% del total, los realizados por mujeres ascendían al 24% y finalmente, un 7% de los viajes habían sido realizados por usuarios de los que se desconocía su género. Si ahora analizamos las cifras de 2020

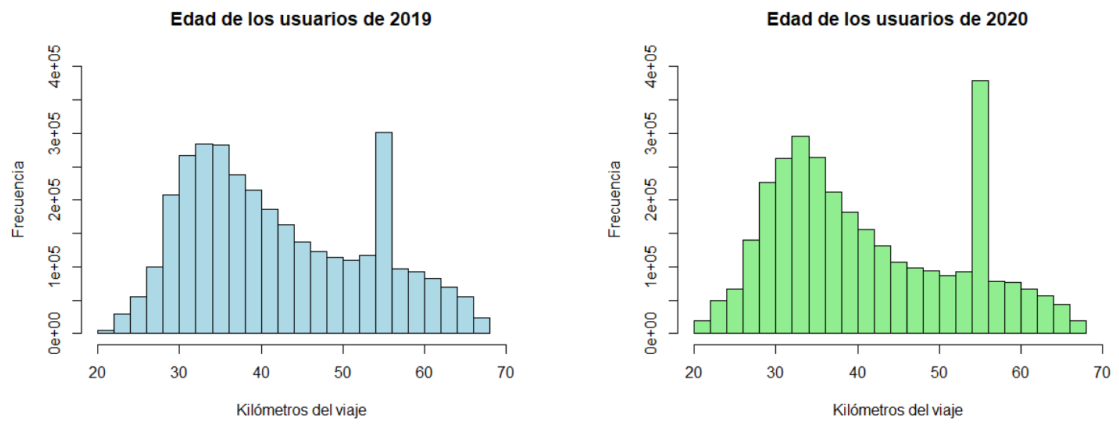
resaltamos que un 61% de los viajes fueron realizados por hombres, un 28% por mujeres, y un 11% por los del género desconocido.

Por tanto, destacamos que la proporción de mujeres en 2020 incrementó en un 4% en comparación a las cifras de 2019. Estas cifras están muy alineadas al estudio realizado por Hosford y Witners (2020) en el que usaron los datos de Citi Bike de 2013 hasta 2018, en donde se mostraba que la proporción de mujeres que utilizaban los servicios de *bike-sharing* de Citi Bike, no superaba en ninguno de los años más del 30%. Sin embargo, en los últimos años, se está reduciendo cada vez más el hueco que hay entre el uso de los sistemas de *bike-sharing* por género (Hosford & Winters, 2020).

4.1.2. Uso de las bicicletas por edad de usuario (solo para 2019 y para 2020)

Al igual que la variable de género no estaba presente en todos los datasets debido a las políticas de datos, la variable que mostraba la edad de cada usuario tampoco lo estaba. Como sabemos, esta variable recogía el año de nacimiento de cada usuario que había realizado un viaje. Por tanto, para saber qué edad tenía cada uno, hemos restado la fecha actual (2024) a la de la variable. Los resultados han sido los siguientes:

Ilustración 12. Edad de los usuarios para 2019 y 2020



Fuente: elaboración propia

Si nos fijamos en la distribución de esta variable para los años 2019 y 2020, podemos destacar que ambas presentan una asimetría positiva en la que la media es más pequeña que la mediana de la variable.

Por otro lado, se puede ver a simple vista que la frecuencia de viajes para aquellas personas con 55 años es bastante superior que el resto de los valores. La literatura indica que las personas pertenecientes a la generación X (nacidos a partir de 1965) tienden a utilizar este tipo de servicios con mayor frecuencia debido a diferentes variables entre las que destacamos los beneficios a la actividad física que presentan este tipo de servicios o como forma de luchar contra la soledad y el aislamiento (Smith et al., 2017), lo que podría explicar la alta frecuencia de viajes de este intervalo de edad.

Del mismo modo que para la variable de género, vamos a estudiar cómo ha variado la proporción de edad entre 2019 y 2020. Para ello, vamos a dividir las edades en estas seis categorías:

Tabla 12. Resumen de los intervalos de edad que vamos a analizar

Intervalo de edad	Proporción 2019	Proporción 2020
< 25 años	1,1%	2,2%
25-34 años	27,2%	30,9%
35-44 años	32,3%	29,4%
45-54 años	17,9%	14,9%
55-64 años	19,2%	20,5%
> 65 años	2,4%	2,1%

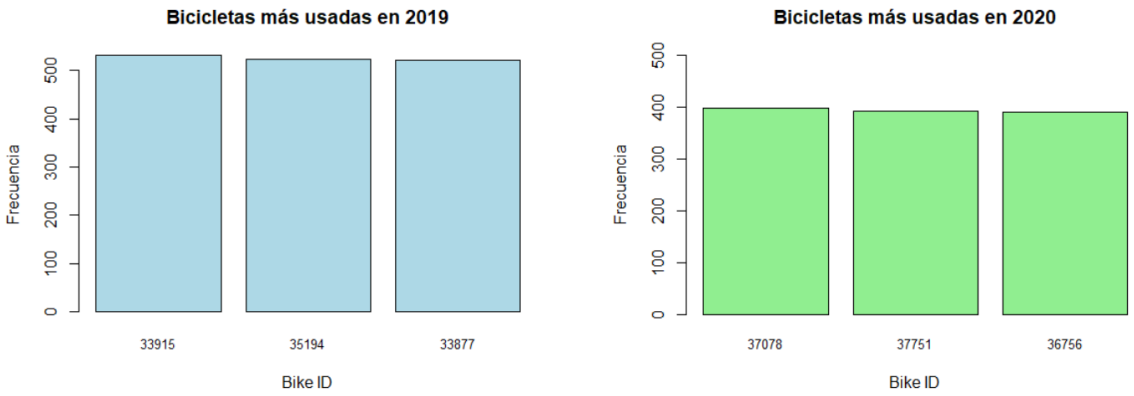
Fuente: elaboración propia

Si observamos las proporciones para los diferentes intervalos de edad para los años 2019 y 2020, podemos destacar que la proporción de personas con edades entre 20 y 34 años era menor para 2019 que para 2020. Además, si observamos la proporción de personas entre 45 y 64 años, para 2019 era mayor que para 2020. Por tanto, podemos concluir que los usuarios de 2020 eran más “jóvenes” que los de 2019.

4.1.3. Uso de bicicleta por ID

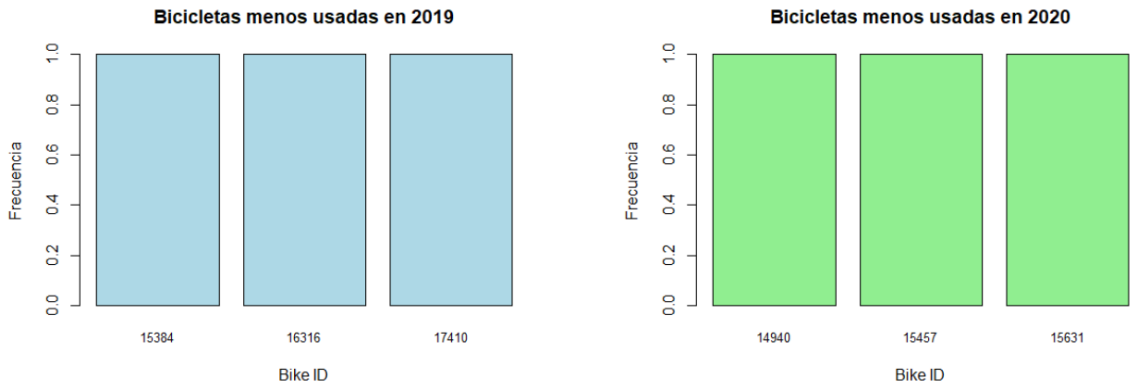
Habiendo explicado la diferencia de género y de edad de los usuarios entre 2019 y 2020, es hora de explicar alguna característica de las bicicletas. A continuación, vamos a estudiar qué bicicletas han tenido un mayor y menor uso apoyándonos en la variable *"bikeid"*, que contenía un código identificativo única de cada bicicleta. Los resultados obtenidos son los siguientes:

Ilustración 13. Bicicletas que más se han usado durante 2019 y 2020



Fuente: elaboración propia

Ilustración 14. Bicicletas que menos se han usado durante 2019 y 2020



Fuente: elaboración propia

De entre todos los códigos identificativos de las bicicletas, hemos decidido mostrar las tres bicicletas que más y que menos se han usado durante estos años. La bicicleta más usada de 2019 había recorrido un total de 531 viajes, mientas que la de 2020 había

recorrido 399, que se traducía en una bajada del 25%. Del mismo modo, si nos fijamos en las tres bicicletas que menos se ha usado durante estos años podemos observar que, solo acumulaban un viaje. Esto puede ser un indicativo de que o se están utilizando poco los servicios de Citi Bike en las zonas de esas bicicletas, o que hay una escasa rotación de bicicletas en todo el sistema.

Gracias a este análisis Citi Bike puede conocer qué bicicletas son las que tienen más y menos uso, lo que puede ayudar a la empresa en diferentes aspectos como en: la planificación de mantenimiento, mejora de la experiencia de los usuarios al saber qué bicicletas están seleccionado los clientes o incluso en la eficiencia de gestión de recursos al asignar mayor recursos o personal a las zonas en las que se están utilizando más las bicicletas.

4.2. Análisis de las variables presentes en los tres datasets (2019, 2020 y 2022)

En las secciones anteriores hemos analizado las variables que solo estaban presentes en alguno de los tres datasets. A continuación, analizaremos aquellas variables que están presentes para los tres años del análisis y que por tanto podemos comparar. Para poder comparar la evolución de estas variables, vamos a mostrar los diferentes histogramas para los tres años del análisis.

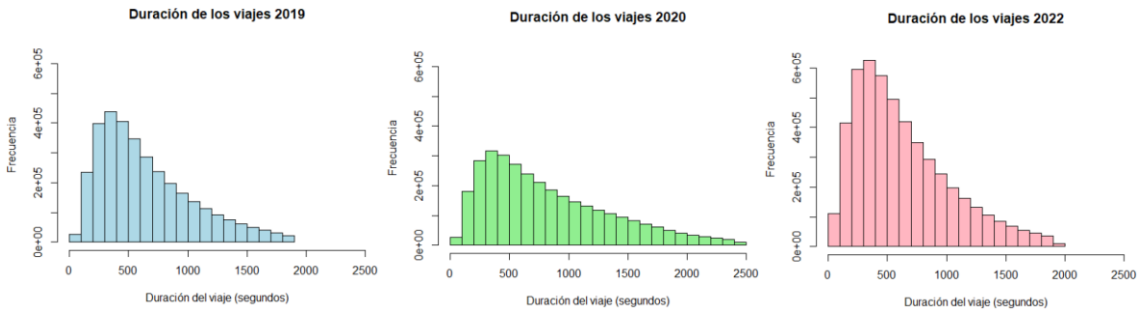
Como se ha mencionado previamente, el número de viajes realizados en 2022 es significativamente mayor al del resto de años del análisis por lo que podemos confirmar que después de la pandemia, el número total de viajes ha aumentado. Por tanto, el objetivo de este apartado no es medir si ha habido un menor o mayor número de viajes, si no analizar si los hábitos de los consumidores han variado después de la pandemia.

Las variables que vamos a comparar en este apartado son las siguientes: duración de los viajes, distancia media recorrida en cada año, velocidad media por cada año, número de viajes por mes, número de viajes por día de la semana, número de viajes por franja horaria y número de viajes por tipo de usuario.

4.2.1. Duración de cada viaje

Como hemos comentado en secciones anteriores, una de las variables más relevantes de nuestros datos es la duración de cada viaje. Para los años de 2019 y 2020 ya venía calculada; sin embargo, para 2022 la teníamos que calcular restando la fecha en la que se había finalizado e iniciado el viaje. Los resultados después de crear un histograma por año son los siguientes:

Ilustración 15. Evolución de la duración de cada viaje por año



Fuente: elaboración propia

Si empezamos analizando la distribución de esta variable en los diferentes años, podemos concluir que en los tres casos presenta una asimetría positiva. Se ve claramente que en estos tres escenarios la mayor parte de la distribución tiende a estar muy cercana a los 500 segundos por viaje, lo que se traduciría en casi 7 minutos pedaleando.

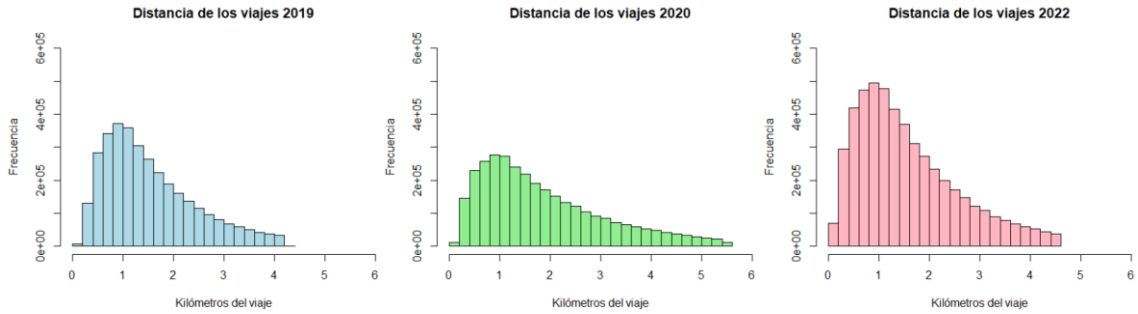
En cuanto a la evolución de esta variable durante estos tres años, podemos observar cómo en 2020 las frecuencias de las duraciones eran ligeramente menores a las de 2019, principalmente por el confinamiento derivado de la COVID-19. Sin embargo, a medida que han pasado los años, se han recuperado las cifras previas a la pandemia como se muestra en el gráfico de 2022, donde la estructura de la duración de los viajes es muy similar a la que había en 2019 incorporando cada vez viajes más largos.

4.2.2. Distancia recorrida de cada viaje

Una vez analizada la evolución de la duración de cada viaje, vamos a estudiar cuál ha sido la evolución de la distancia recorrida por viaje para los años 2019, 2020 y 2022. Como explicábamos en secciones anteriores, esta variable ha sido calculado midiendo la distancia que había entre la estación de origen y de destino de cada viaje. Para calcular esta variable, utilizábamos la función Haversine, que nos permitía saber la distancia entre dos coordenadas.

Para este análisis intuimos que la mayoría de los usuarios no se habían desviado en exceso de la ruta entre estación de inicio y de fin de viaje. Los resultados que hemos obtenido han sido los siguientes:

Ilustración 16. Evolución de la distancia de cada viaje por año



Fuente: elaboración propia

Al igual que en la variable de duración, esta variable también presenta una asimetría a la derecha en los tres años del análisis. La mayor parte de la distribución de la variable está comprendida entre los 800-900 metros de viaje y llega a unos valores máximos entre 4 y 5 kilómetros. Por otro lado, también se puede apreciar que, durante 2020, los usuarios tendieron a realizar viajes cada vez más largos sobrepasando en algunos casos los 5 kilómetros.

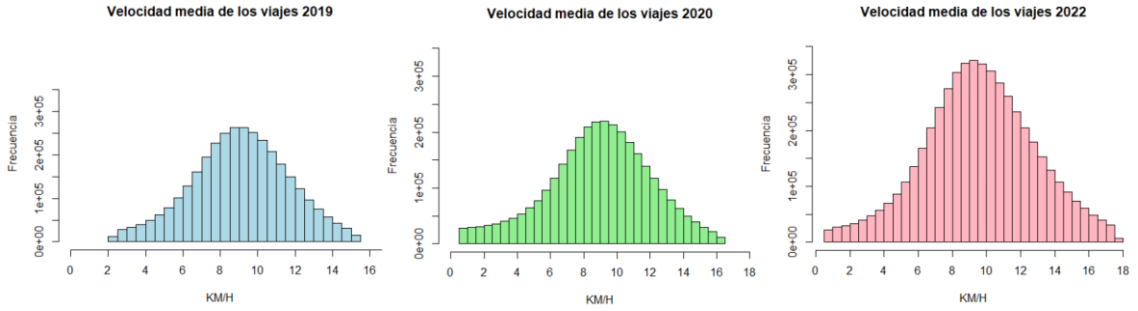
Si comparamos la estructura entre los años 2019 y 2022, podemos observar que la distribución de la frecuencia sigue patrones similares entre los dos años, y que al igual que pasaba con la variable de duración, los usuarios también están recorriendo distancias cada vez más larga, algo que es lógico debido a que las dos variables tienden a estar correlacionadas.

4.2.3. Velocidad media de cada viaje

En secciones anteriores, hemos analizado la evolución de la duración y distancia de los viajes realizados por los usuarios. Con estas dos variables éramos capaces de calcular una tercera que nos ayudase a estimar la velocidad media a la que iba cada usuario por viaje.

Esta variable podría sernos útil para saber aspectos como la eficiencia del servicio o calidad de la ruta o incluso para mejorar la seguridad vial en caso de que haya zonas por la que se vaya muy despacio o excesivamente rápido. Al interpretar gráficamente los resultados hemos obtenido lo siguiente:

Ilustración 17. Evolución de la velocidad media de cada viaja por año



Fuente: elaboración propia

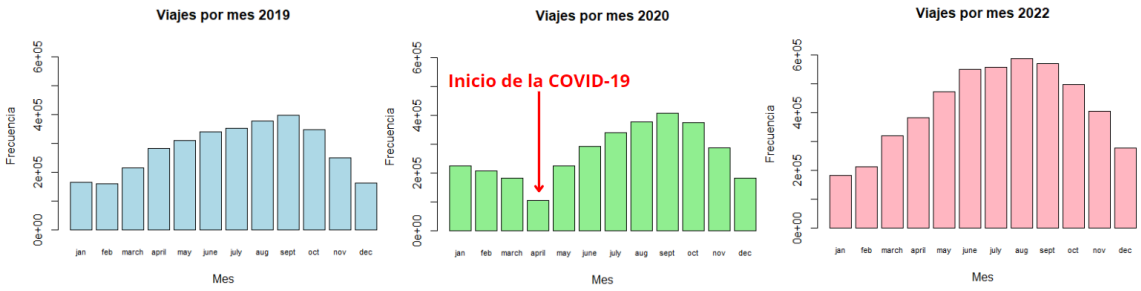
A diferencia de los análisis anteriores, nos encontramos ante una variable con una distribución simétrica en la que la media de la variable tiene su valor muy cercano a la mediana. Podemos observar que la velocidad media de los tres datasets ronda los 9 kilómetros por hora, cifra ligeramente inferior a la media mundial de bicicleta por ciudad, que asciende a los 12 kilómetros por hora (Langford, 2015).

Algunas variables externas que puedan explicar estos podrían ser: la cantidad de tráfico que se forma en una de las ciudades más grandes del mundo, la escasez de carriles especiales para bicicletas y otros elementos de transporte como patinetes o el mal estado de las vías que puede imposibilitar ir un poco más rápido en este tipo de transportes.

4.2.4. Número de viajes por mes del año

Como se ha mencionado previamente, nuestros datasets incorporaban una variable en formato fecha que nos indicaba cuando comenzaba cada viaje. Gracias a esta, podemos analizar el mes en el que se produjo cada viaje a lo largo del año. Los resultados se muestran a continuación.

Ilustración 18. Número de viajes por mes del año por año



Fuente: elaboración propia

Si nos fijamos en la frecuencia de viajes por mes en 2019, podemos ver claramente un crecimiento gradual desde enero hasta septiembre, en donde se producen la mayor cantidad de viajes (el 12% del total de viajes del año). Si comparamos esta tendencia con los años posteriores podemos concluir lo siguiente:

- En 2020 la tendencia sigue siendo parecida excepto para el mes de abril, momento en el que empezó la pandemia y el confinamiento, lo que supuso una reducción del número total de viajes (Xin et al, 2022). Desde ese mes hasta mayo, hubo un crecimiento de viajes del 116% y el mes con más frecuencia volvió a ser en septiembre. Este incremento pudo haberse debido a que los usuarios preferían utilizar elementos de transporte como las bicicletas para moverse por la ciudad en vez del transporte público ya que, había menos riesgo de contagio.

- En 2022, la tendencia fue muy similar a la 2019 aunque con algunas diferencias. En los primeros meses del año el uso de este servicio fue de un 18% menor en comparación con las cifras de 2019 y el mes en el que los usuarios utilizaron más estas bicicletas fue en agosto.

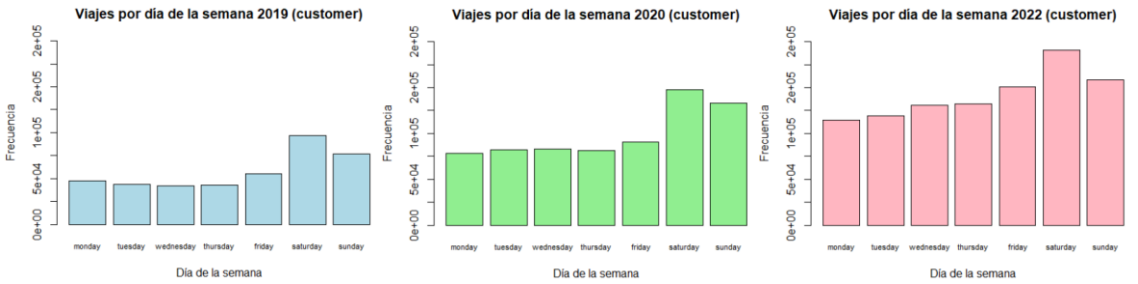
4.2.5. Número de viajes por día de la semana

En apartados anteriores, concluíamos que ha habido una evolución en cuanto al uso de las bicicletas por franja horaria entre 2019 y 2022. Es por eso por lo que en esta sección vamos a analizar si este patrón también se repite por día de la semana. Este análisis lo hemos dividido en dos partes en función del tipo de cliente que realiza cada viaje.

En primer lugar, vamos a estudiar el patrón de los usuarios "customer", que corresponde a aquellos que no pagan una suscripción, y por tanto utilizan este servicio de manera esporádica. En segundo lugar, vamos a repetir este proceso para los usuarios "suscriber", que tienden a usar este servicio con mayor frecuencia que el otro tipo de usuarios.

Para ello, hemos calculado el día de la semana a la que correspondía la fecha de inicio de viaje y hemos extraído los siguientes resultados:

Ilustración 19. Número de viajes realizados por día de la semana por año (customer)

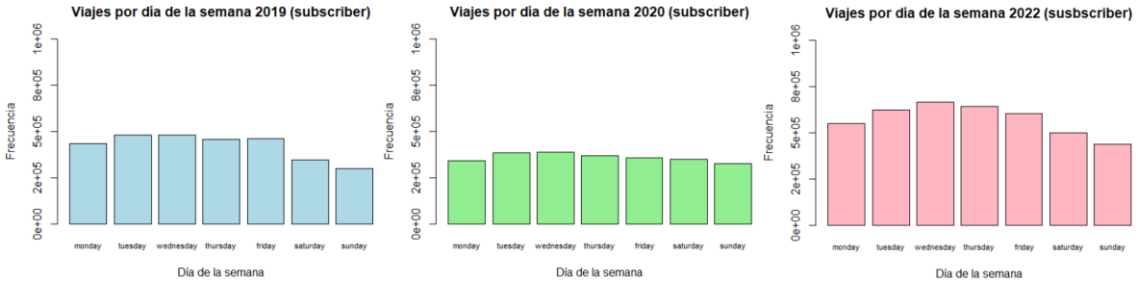


Fuente: elaboración propia

Al observar los gráficos, podemos concluir que el usuario *"customer"* tiende a utilizar este servicio principalmente los días del fin de semana (sábado y domingo). De hecho, con estos días se explican aproximadamente el 40% de viajes de toda el año, lo que puede indicar que este tipo de usuarios utiliza este servicio como medida de desplazamiento dedicado al ocio.

Si repetimos este proceso para los usuarios suscritos a este servicio obtenemos los siguientes resultados:

Ilustración 20. Número de viajes realizados por día de la semana por año (suscriber)



Fuente: elaboración propia

Al observar el número de viajes por día de la semana de 2019 podemos concluir que los usuarios suscritos al servicio de Citi Bike utilizaban este con más frecuencia durante los días lectivos de la semana (especialmente a partir del lunes), mientras que, en los fines de semana, el uso de estos disminuía.

Además, también se nota una clara diferencia en la frecuencia del uso de este servicio entre los lunes de 2019 y 2022, pudiendo ser un motivo de esto la modalidad del teletrabajo que están brindando las empresas a raíz de la COVID-19 y que según una encuesta llevada a cabo por Barrero, Bloom y Davis (2021), muchas personas lo prefieren hacer ese mismo día.

4.2.6. Número de viajes por franja horaria

Además de analizar la duración, distancia o velocidad media de cada viaje, también vamos a analizar el número total de viajes que se realizan en función de la franja horaria. Para este análisis, hemos creado cinco categorías que representan las horas más importantes del día:

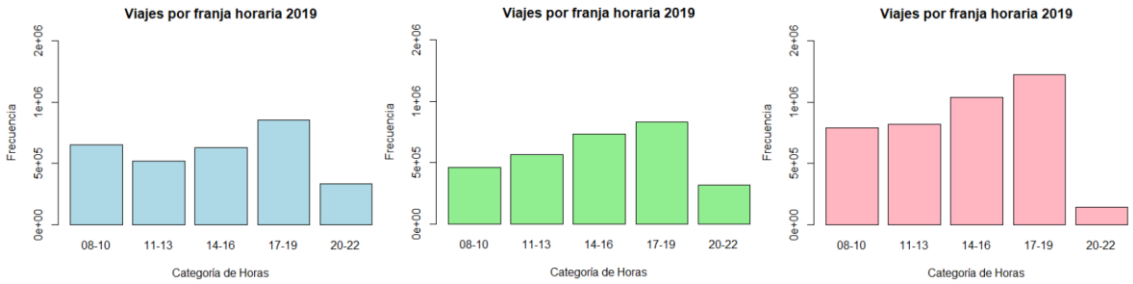
Tabla 13. Resumen de las franjas horarias para analizar

Franja horaria	Descripción
08-10	Comienzo de la jornada laboral
11-13	Desplazamientos de ocio
14-16	Hora de la comida
17-19	Fin de la jornada laboral
20-22	Desplazamientos de ocio

Fuente: elaboración propia

Una vez hemos analizado la frecuencia de viajes por mes y por día de la semana, vamos a estudiar en que franja horaria se producen más viajes de este servicio. Somos capaces de conocer en estas franjas ya que, como se ha mencionado anteriormente, la variable que indica el inicio de cada viaje tiene formato de fecha y hora. Después de representarla gráficamente, obtenemos los siguientes resultados:

Ilustración 21. Evolución de los viajes por franja horaria por año



Fuente: elaboración propia

Después de observar el gráfico, podemos concluir que la franja horaria en la que más viajes se producen es cuando finaliza la jornada laboral. De hecho, en 2022 esta franja representaba un 35% del total de viajes realizados a lo largo del año.

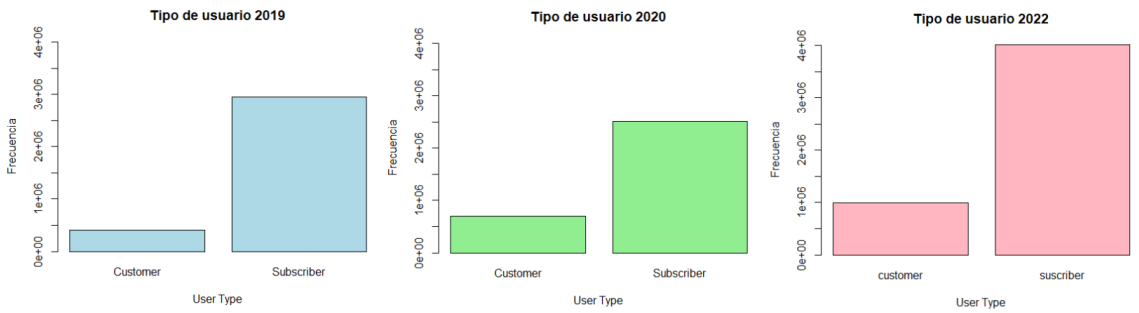
Por otro lado, si nos centramos en los viajes realizados al inicio de la jornada laboral, podemos ver claramente, que la frecuencia para 2019 es mayor que la de 2020 y 2022. Como consecuencia del confinamiento, las personas tuvieron que adecuarse a un estilo de vida con un 100% de teletrabajo. Después de haber pasado más de 5 meses trabajando desde casa, es muy probable que haya personas que prefieran la modalidad del teletrabajo en vez de ir a la oficina, lo que podría explicar porque la frecuencia de viajes al comienzo de la jornada laboral es menor para los años 2020 y 2022 en proporción a las cifras de 2019.

4.2.7. Tipo de usuario que usa las bicicletas

En secciones anteriores, analizábamos la frecuencia de viajes por día de la semana y por tipo de usuario. Es por eso que para este apartado vamos a analizar la volumetría que presenta cada tipo de usuario al servicio de Citi Bike. Para ello, vamos a utilizar la variable "User Type" que nos decía si un usuario era "customer" o "suscriber".

Como se comentaba en secciones anteriores, el primer término hace referencia a aquellos usuarios que no tienen ningún plan contratado y son cobrados $4,79 por 30 minutos de viaje. Por el otro lado, Citi Bike tiene una serie de planes diarios y anuales que permiten al usuario que la contrata gozar de ciertas condiciones más flexibles como reducción de tarifas o desbloqueo de bicicletas gratis. Los resultados fueron los siguientes:

Ilustración 22. Evolución del tipo de usuario por año



Fuente: elaboración propia

Si nos fijamos en estos tres gráficos, podemos ver claramente que la mayor parte de los usuarios prefiere contratar un plan de suscripción que les permita obtener mejores

condiciones en el servicio. Si mostramos la proporción que representa cada categoría a lo largo de los tres años, obtenemos la siguiente tabla:

Tabla 14. Evolución de la variable User Type por año

Valor	2019	2020	2022
Customer	12%	21,6%	19,8%
Suscriber	88%	78,4%	80,2%

Fuente: elaboración propia

Para 2019 la diferencia en proporción de “customer” a “suscriber” era mucho mayor que para los os años siguientes. Una de las razones que apoyan esta idea podría ser que las personas dejaron de pagar sus planes de suscripción de esta plataforma al ver que se estaban imponiendo nuevas medidas para mitigar la crisis sanitaria generada por la COVID-19, y que a medida que pasaron los años, las personas optaron por volver a contratar estos planes de suscripción.

Capítulo 5: Análisis geográfico de la distribución de viajes de Citi Bike en NY

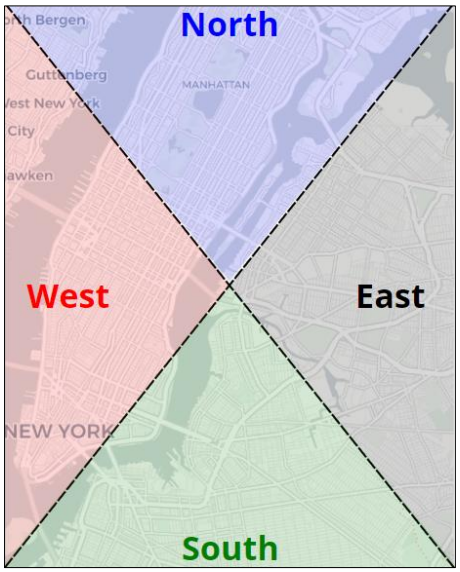
Una vez analizada la evolución en los hábitos de uso de los usuarios de Citi Bike pre y post pandemia, hemos podido concluir que los patrones que había en 2019 sufrieron unos cambios durante 2020, principalmente debido a las consecuencias que generó la pandemia. Sin embargo, en 2022 la situación volvía a recuperar la mayor parte de los patrones que se daban durante 2019.

En este apartado, vamos a llevar a cabo un análisis geográfico para conocer de primera mano cuáles son las zonas dentro de Nueva York en las que se realizan un mayor número de viajes. Para ello, hemos dividido el proceso en las siguientes partes:

División del mapa de Nueva York

Para llevar a cabo un análisis geográfico detallado, hemos dividido el mapa de Nueva York en cuatro secciones, aplicando una forma de X. El triángulo superior de esta X correspondería a la zona norte de la ciudad, el inferior al sur, el de la derecha al este y el de la izquierda al oeste. A continuación, se muestra el mapa de Nueva York con la división realizado y una tabla resumen de los barrios que integran estas cuatro secciones.

Ilustración 23. División del mapa de Nueva York



Fuente: elaboración propia

Tabla 15. Resumen de los barrios comprendidos en las zonas geográficas de NY

Región geográfica	Barrios comprendidos
North	Bronx, Harlem, Morris Heights, High Central Park
South	Brooklyn
East	Long Island City, Greenpoint, Maspeth
West	Chinatown, Mid/Lower Manhattan, Union Square, Low Central Park

Fuente: elaboración propia

Clasificación de las coordenadas

Una vez definidos los barrios comprendidos dentro de estas cuatro secciones, tenemos que asignar a cada viaje una zona. Para ello, hemos realizado los siguientes pasos:

1. Creamos dos variables en cada dataset de Citi Bike con el nombre de *"start_geography"* y *"end_geography"*, que pueden tomar valores de North, South, East o West.
2. Como sabemos, tenemos en nuestros datos unas coordenadas en donde se inició y se finalizó el viaje. Utilizaremos estos datos para asignar cada viaje a una zona. Es decir, creamos un código para que, si las coordenadas de inicio y fin de viaje caen en una de las cuatro secciones de Nueva York, que tomen el valor correspondiente
3. Finalmente, cuando ya tenemos clasificados nuestros viajes, vamos a comprobar que el código ha funcionado y que la asignación geográfica está bien hecha mostrando la siguiente tabla.

Ilustración 24. Valores comprendidos de las variables Start y End geography

start station name	end station name	start_geography	end_geography
Broadway & W 53 St	W 27 St & 7 Ave	West	West
W 26 St & 10 Ave	Mercer St & Spring St	West	West
Sullivan St & Washington Sq	Great Jones St	West	West
Little West St & 1 Pl	Mercer St & Spring St	West	West
Central Park West & W 102 St	Broadway & W 56 St	North	West

Fuente: elaboración propia

Si comprobamos la primera observación, podemos confirmar que la asignación geográfica se ha hecho correctamente. La calle de "Broadway & w 53 St" pertenece a la parte media de Manhattan, localizada al Oeste la ciudad, así como la calle "W 27 St & 7 Ave".

4. Una vez hemos comprobado que la asignación de zonas es correcta, vamos a hacer recuento de todas las combinaciones de viajes posibles. Es decir, vamos a ver la frecuencia de viajes de una sección a otra (Norte a Norte, Norte a Sur, Norte a Este…)

Representación gráfica

En base a los datos obtenidos de las variables *"start_geography"*, *"end_geography* y al recuento de los viajes de todas las combinaciones posibles, vamos a realizar un gráfico que nos permita ver visualmente la frecuencia de viajes que se realizan de una zona a otra, o incluso dentro de la misma zona. Para este proceso, volveremos a utilizar la herramienta de Python y nos centraremos en los viajes realizados en 2022.

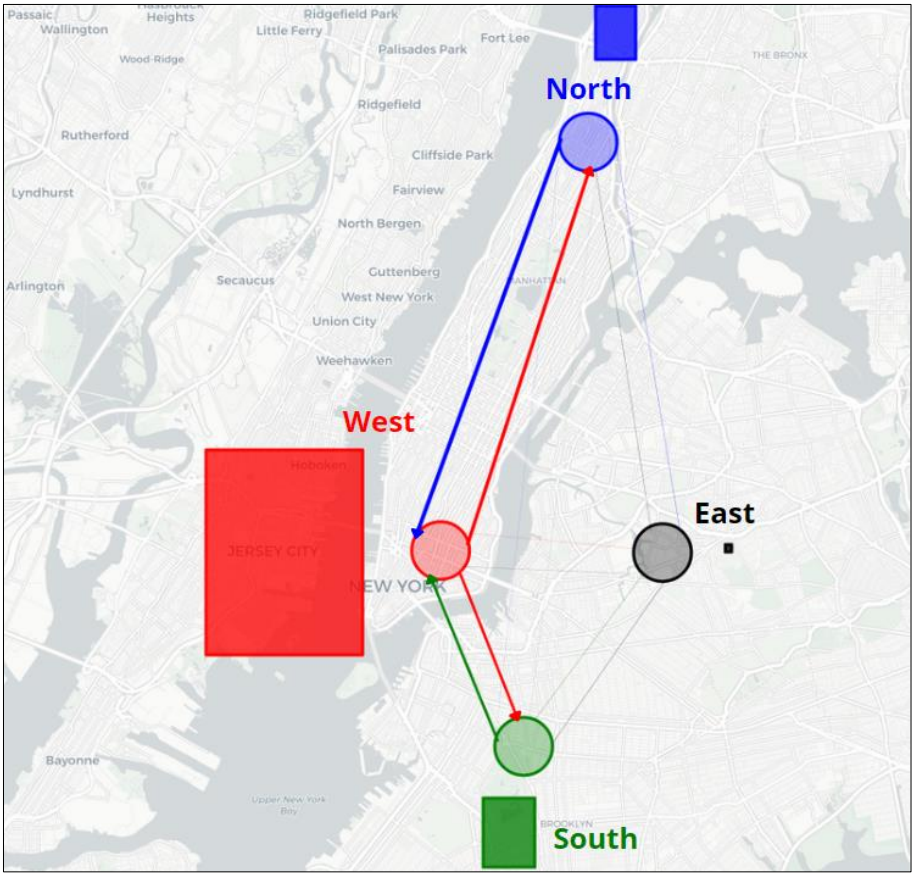
Como se ha mencionado anteriormente, hemos dividido el mapa de Nueva York en cuatro secciones. Para poder diferenciar estas cuatro secciones en el mapa, hemos dibujado un círculo en la parte central de estas de acuerdo con los colores de cada sección (azul para el Norte, verde para el Sur, negro para el Este y rojo para el Oeste).

Por otro lado, mediante el código realizado en Python, hemos pintado una serie de flechas de sección a sección cuyo grosor depende directamente de la frecuencia de los viajes realizada de una geografía a otra. De hecho, si nos fijamos en el gráfico a

continuación, podemos ver que los principales viajes realizados se dan entre *North-West* y *South-West*.

Finalmente, para poder representar gráficamente aquellos viajes que empiezan y terminan en la misma sección, hemos dibujado cuatro rectángulos de colores distintos que representan a cada zona. El tamaño de estos cuadrados depende directamente de la frecuencia de viajes de esos trayectos.

Ilustración 25. Viajes realizados por geografía en Nueva York (2022)



Fuente: elaboración propia

Conclusiones

Después de analizar este gráfico, llegamos a las siguientes conclusiones:

- En un primer lugar, se puede apreciar que las secciones en las que se realizan un mayor número de viajes son principalmente de Norte a Oeste (y viceversa), y de Sur a Oeste (y viceversa). Esta distribución puede explicarse debido a que en el Oeste de Nueva York es donde se encuentran el mayor núcleo de oficinas de la ciudad, por lo que podemos inferir que la mayoría de estos viajes se realizan para ir a trabajar y para volver a sus casas.
- En segundo lugar, si nos fijamos en el grosor de las líneas de estos trayectos, podemos concluir que la frecuencia de los viajes de Oeste a Norte y de Oeste a Sur es muy similar (173.000 vs 142.000).
- Por otro lado, si nos fijamos los rectángulos del mapa, se puede ver claramente que la mayor parte de viajes entre mismas geografías se vuelve a dar en el oeste de la ciudad. De nuevo, esto vuelve a confirmar el uso de estos sistemas de *bike-sharing* para ir a trabajar.

Capítulo 6: Análisis para ver si hay algún patrón con los códigos postales de NY

En la sección anterior, analizamos la distribución geográfica de los viajes de Citi Bike en Nueva York, concluyendo que la mayoría de los viajes se iniciaban y terminaban en el Oeste, donde se encuentran la mayoría de las oficinas de la ciudad.

Además de investigar el impacto de la pandemia en el uso de Citi Bike, también queríamos examinar si alguna característica sociodemográfica de Nueva York estaba relacionada con este uso. Por ello, hemos dedicado este apartado a elabora dicho análisis, utilizando el dataset de Citi Bike de 2022 por ser el más actualizado.

Dado que nuestro dataset sociodemográfico por código postal contiene más de 20 variables, acotamos el análisis a tres variables relevantes, mostradas en la siguiente tabla:

Tabla 16. Variables sociodemográficas para analizar con los datos de Citi Bike

Variable Sociodemográficas	Descripción
Population	Población media por código postal
Commute Time to Work	Tiempo medio (minutos) que tardan las personas en ir al trabajo
No Healtchare Insurance	Porcentaje de la población que no tiene seguro médico

Fuente: elaboración propia

Para poder estudiar la relación entre ambos datasets, hemos tenido que añadir una nueva columna a nuestro dataset con las características sociodemográficas de Nueva York con el número de viajes realizados por código postal en 2022. Este análisis se divide en dos partes, en primer lugar, examinaremos la correlación entre las variables sociodemográficas y el número de viajes de 2022, en segundo lugar, desarrollaremos un modelo sencillo para determinar si estas variables son significativas y pueden explicar la variabilidad en el número de viajes.

Correlación entre las variables sociodemográficas y la volumetría de viajes de 2022

Como se ha comentado anteriormente, vamos a estudiar la correlación que existe entre el número de viajes realizados por código postal y las tres variables sociodemográficas explicadas anteriormente. Estos han sido los resultados:

Ilustración 26. Correlaciones entre los viajes y las variables sociodemográficas de NY

	trips_2022
population	0.064
commute_time_to_work	-0.646
no_healthcare_insurance	-0.475

Fuente: elaboración propia

En un primer vistazo, se puede apreciar que la variable *Population* no guarda ningún tipo de correlación con el número de viajes. Sin embargo, las otras dos variables tienen

una correlación negativa moderada, que a simple vista tienen sentido. Por ejemplo, en apartados anteriores se mencionaba que cuanto mayor es la distancia al trabajo, las personas tienden a utilizar otro tipo de transportes complementarios a las bicicletas como el transporte público o los coches, por lo que la correlación es negativa

Modelo de regresión lineal múltiple de las características sociodemográficas

Después se extraer las correlaciones de las variables del análisis con el número de viajes de 2022, hemos podido concluir que alguna de estas variables tenía una correlación negativa con el número de viajes efectuados. Sin embargo, para un análisis más detallado, vamos a realizar un modelo de regresión múltiple con el que saber cómo de significativas son estas variables para el modelo.

Un modelo de regresión es una herramienta estadística utilizada para entender y predecir la relación que hay entre una o más variables independientes y una variable dependiente. En este caso, nuestras variables independientes son *"Population", "Commute Time to Work" y "No Healthcare Insurance"*, mientras que la variable dependiente será el número de viajes de 2022. Para poder correr este modelo utilizaremos la herramienta de R Studio.

Después de elaborar este modelo los resultados han sido los siguientes:

Ilustración 27. Modelo de regresión múltiple para predecir el número de viajes realizados

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.120e+05  2.092e+04  10.137 6.15e-15 ***
population               8.759e-01  1.558e-01   5.621 4.46e-07 ***
commute_time_to_work    -6.065e+03  8.208e+02  -7.389 3.84e-10 ***
no_healthcare_insurance  5.162e+04  1.721e+05   0.300    0.765
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28590 on 64 degrees of freedom
Multiple R-squared:  0.6097,    Adjusted R-squared:  0.5914
F-statistic: 33.33 on 3 and 64 DF,  p-value: 4.282e-13
```

Fuente: elaboración propia

Para evaluar la importancia de estas variables en la predicción del número de viajes, observamos el p-valor. Este indica la probabilidad de que los resultados sean aleatorios si la hipótesis nula es cierta, normalmente siendo significativos cuando son menores a 0,05.

Al analizar los p-valores de nuestro estudio, notamos que tanto para la variable *"Population"* como para *"Commute Time to Work"*, estos valores están muy próximos a 0. Esto sugiere que estas variables son altamente significativas para el modelo y son buenos predictores del número total de viajes. En contraste, al observar el p-valor para la variable *"No Healthcare Insurance"*, vemos que es significativamente alto, indicando que esta variable no es estadísticamente significativa para el modelo.

Finalmente, al considerar el coeficiente R cuadrado (R-squared) del modelo de regresión múltiple, el cual indica qué proporción de la variabilidad de la variable dependiente es explicada por nuestro modelo, encontramos que el 60% de la variabilidad en los viajes realizados puede ser explicada por las tres variables sociodemográficas analizadas. Esto demuestra que nuestro modelo tiene una capacidad razonable para predecir nuestra variable dependiente.

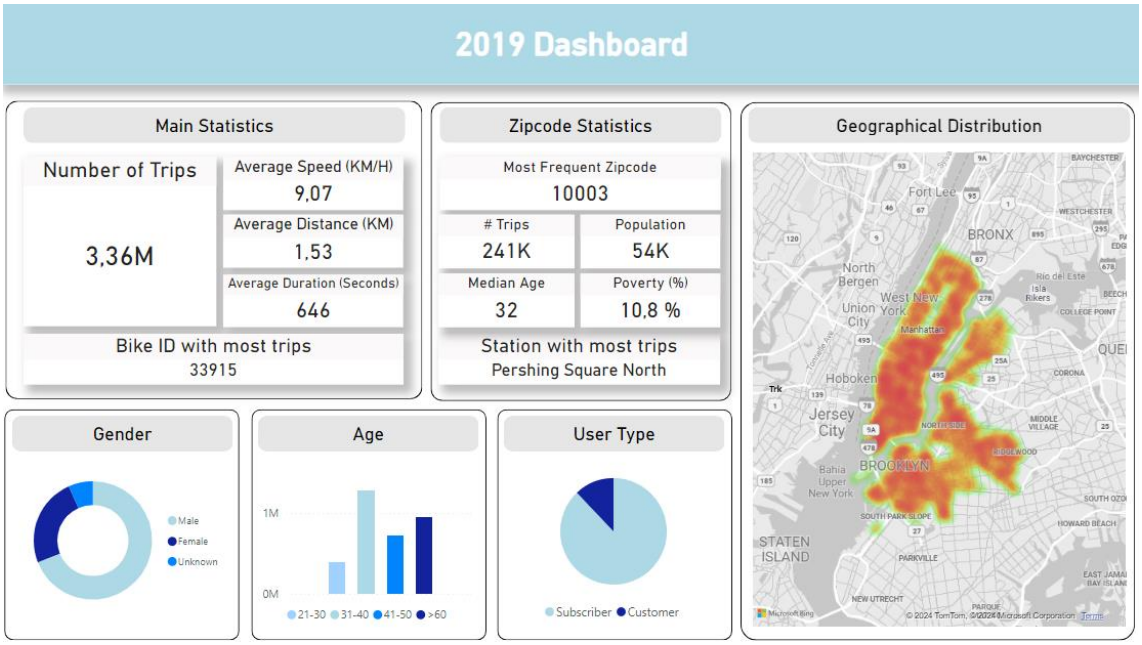
Capítulo 7: Visualización gráfica en PowerBI

Después de llevar a cabo un análisis para conocer si había alguna variable sociodemográfica de Nueva York que demostrase un mayor o menor uso de estos sistemas, hemos llegado a la conclusión de la significancia que tienen las variables *"Population"* y *"Commute Time to Work"* para predecir la cantidad de viajes efectuados en un futuro.

Como último apartado de este análisis, vamos a elaborar diferentes *dashboards* en Power Bi que permitan a la compañía de Citi Bike observar de manera fácil y rápida cuáles han sido las principales cifras durante los años de 2019, 2020 y 2022, así como una comparación entre ellos que permita analizar si la evolución del negocio se ha visto interrumpida por la pandemia o si ha visto reforzada.

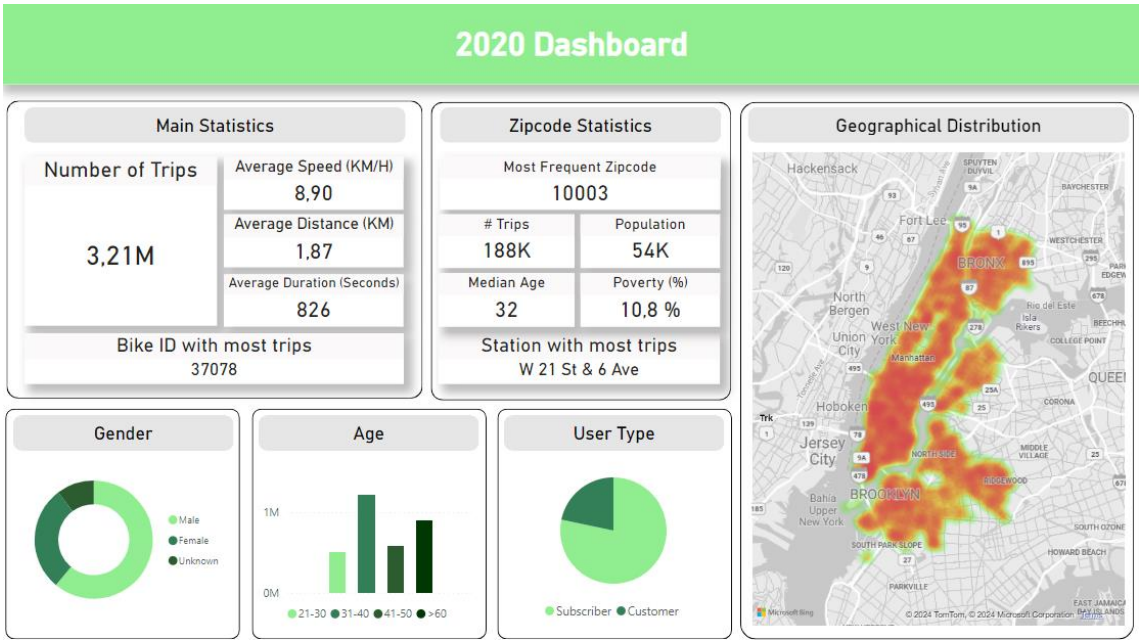
A continuación, se muestran los diferentes *dashboards* creados a partir de los datasets de Citi Bike de 2019, 2020 y 2022.

Ilustración 28. Dashboard de Power BI para el año 2019 de Citi Bike



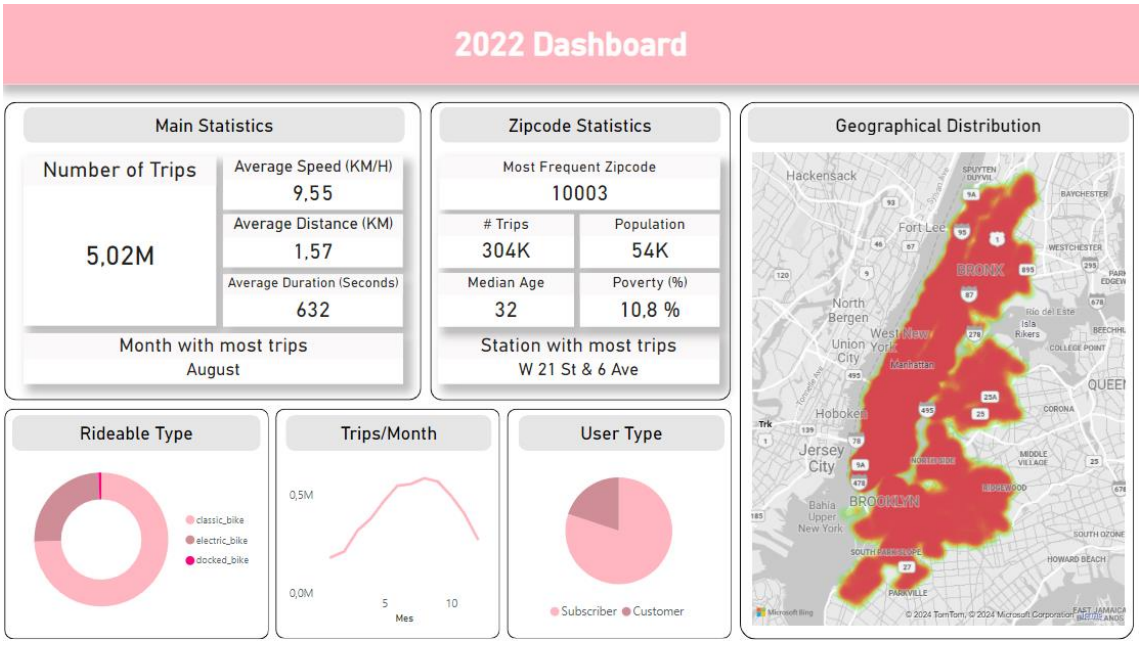
Fuente: elaboración propia

Ilustración 29. Dashboard de Power BI para el año 2020 de Citi Bike



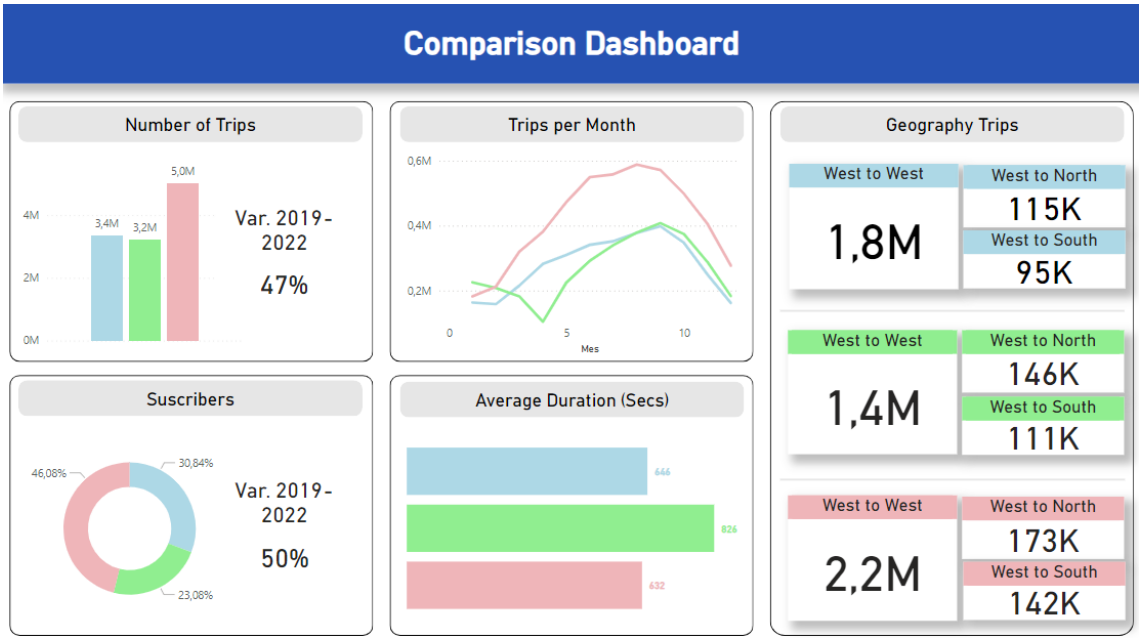
Fuente: elaboración propia

Ilustración 30. Dashboard de Power BI para el año 2022 de Citi Bike



Fuente: elaboración propia

Ilustración 31. Dashboard de Power BI comparando las cifras de los distintos años



Fuente: elaboración propia

Antes de abordar las conclusiones de estos *dashboards* hay que explicar de nuevo que para 2022, el *dashboard* es ligeramente diferente al de 2019 y 2020 debido a que no incorpora las variables de género y edad. Sin embargo, hemos incluido dos gráficos que permiten conocer el tipo de bicicleta usada por viaje y cuál fue la distribución de viajes por mes en ese año.

Tras esta breve explicación, vamos a proceder con alguna conclusión:

- Uno de los gráficos más llamativos de este análisis, es el mapa de calor. Si observamos la evolución en estos tres años, podemos observar que las zonas de calor se van expandiendo, lo que confirma que Citi Bike se está expandiendo cada vez más su servicio a otras partes de Nueva York.
- Finalmente, si observamos la comparación del número de suscriptores de la plataforma, podemos apreciar que la variación entre 2019 y 2022 fue del 50%, lo que demuestra el gran crecimiento de usuarios que ha tenido Citi Bike en tan solo tres años.

Conclusiones

Los sistemas de *bike-sharing*, que forman parte de la micro movilidad, ofrecen bicicletas convencionales y eléctricas ideales para trayectos cortos. Estos sistemas permiten el alquiler de bicicletas a corto plazo a través de plataformas tecnológicas, ofreciendo comodidad y flexibilidad sin los gastos asociados a la propiedad. Su crecimiento se atribuye a su capacidad para reducir emisiones, mejorar el tráfico urbano e integrar la actividad física en la rutina diaria. Además, proporcionan beneficios significativos para la salud pública, como la reducción del estrés y la promoción de la pérdida de peso.

Durante este estudio, hemos analizado el impacto de la pandemia en el uso de estos servicios, y hemos llegado a la conclusión de que:

En 2022, el uso de Citi Bike experimentó un notable incremento del 50% en comparación con 2019, recuperándose rápidamente después de la caída inicial en 2020 causada por las restricciones de movilidad debido a la pandemia. Este aumento se pudo atribuir a la búsqueda de alternativas de transporte que minimizasen el riesgo de contagio ya que, las bicicletas ofrecen una opción segura y al aire libre.

En términos de hábitos de uso, la duración media de los viajes aumentó un 34% en 2020 en comparación con 2019. Esto sugiere que los usuarios estaban dispuestos a realizar trayectos más largos para evitar el transporte público y reducir el riesgo de exposición al virus. En 2022, la mayoría de los viajes se concentraron al final de la jornada laboral, representando un 35% del total de los viajes realizados en ese año, lo que sugiere un aumento del uso de bicicletas en horas de ocio y de regreso a casa. Además, se observó un incremento en la adopción de bicicletas eléctricas, reflejando una preferencia por opciones de transporte más rápidas y menos exigentes físicamente.

También se observó una proporción diferente en el uso de las bicicletas por parte de hombres y mujeres. En 2019, los viajes realizados por hombres representaban el 69% del total, mientras que los realizados por mujeres ascendían al 24%. En 2020, se incrementó la proporción de viajes realizados por mujeres, alcanzando el 28%, mientras que la proporción de viajes realizados por hombres disminuyó al 61%. Este cambio junto con las cifras de los últimos años, sugieren una reducción gradual en la brecha de género en el uso de Citi Bike.

Además, se identificaron diferencias significativas en el uso de bicicletas según la edad de los usuarios. En 2019 y 2020, la mayoría de los usuarios de Citi Bike tenían entre 35 y 44 años, seguidos por aquellos en el rango de 25 a 34 años. Sin embargo, en 2020 se observó un aumento en la proporción de usuarios menores de 25 años, alcanzando el 22%, en comparación con el 11% en 2019. Esto sugiere que los usuarios más jóvenes comenzaron a utilizar más el servicio durante la pandemia, posiblemente debido a la búsqueda de alternativas de transporte seguras y económicas.

La distribución geográfica del uso de Citi Bike también mostró cambios significativos. El servicio se expandió a más áreas de Nueva York, especialmente en zonas con alta densidad de oficinas. Los patrones de uso revelaron que las zonas Oeste y Norte de la ciudad concentraban la mayoría de los viajes, especialmente en áreas con alta densidad de oficinas. Esto sugiere que muchos de estos viajes estaban relacionados con desplazamientos laborales, subrayando la importancia del sistema de *bike-sharing* en la movilidad urbana diaria.

El análisis sociodemográfico identificó factores significativos que influyen en el uso de Citi Bike. La regresión múltiple mostró que la densidad de población y el tiempo de desplazamiento al trabajo son determinantes en el número de viajes realizados. Una

mayor densidad de población en áreas urbanas aumenta la probabilidad de uso del servicio de *bike-sharing*, y un menor tiempo de desplazamiento al trabajo está correlacionado con un mayor uso de bicicletas compartidas, ofreciendo una alternativa eficiente. En contraste, la posesión o falta de seguro médico no mostró una significancia estadística relevante en el uso de Citi Bike.

Gracias a este análisis hemos podido confirmar que la pandemia ha tenido un impacto notable en el uso de los sistemas de *bike-sharing* de Citi Bike, incrementando el volumen de viajes y modificando los hábitos de los usuarios.

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, Alfredo González-Izquierdo Antón, estudiante de Administración de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis de los viajes de bike sharing de Citi Bike. ¿Cómo ha afectado la pandemia a esta plataforma?", 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:

1. **Brainstorming de ideas de investigación**: Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
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.

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.

Anexos

Anexo A: Líneas de código para tratar los outliers de nuestros datos

```
for (variable_name in names(datos_2019_outliers)) {

  variable_estandarizada_19 <- scale(datos_2019_outliers[[variable_name]])
  condicion_19 <- abs(variable_estandarizada_19) <= num_desviaciones_std
  new <- datos_2019_outliers[[variable_name]][!condicion_19] <- NA
}


datos_2019_graph <- datos_2019 %>% select(-one_of(outlier_variable_2019))
datos_2019_graph <- na.omit(cbind(datos_2019_outliers, datos_2019_graph))
```

Fuente: elaboración propia

Anexo B: Función Haversine para calcular la distancia entre dos coordenadas

```
def haversine(lat1, lon1, lat2, lon2):
    radio_tierra = 6371.0
    lat1, lon1, lat2, lon2 = map(radians, [lat1, lon1, lat2, lon2])
    dlat = lat2 - lat1
    dlon = lon2 - lon1
    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))
    distancia = radio_tierra * c
    return distancia
```

Fuente: elaboración propia

Anexo C: Código empleado para extraer el código postal de nuestras coordenadas

```
from geopy.geocoders import Bing

api_key = "[redacted]"

geolocator = Bing(api_key=api_key)

def get_zipcodes(row):
    location = geolocator.reverse((row['latitude'], row['longitude']), exactly_one=True)
    zipcode = location.raw['address'].get('postalCode', None)
    city = location.raw['address'].get('adminDistrict', None)
    return zipcode, city

data_19[['zipcode', 'city']] = data_19.apply(get_zipcodes, axis=1, result_type='expand')
data_19
```

Fuente: elaboración propia

Para almacenar todos los archivos y el código utilizado a lo largo de este trabajo, vamos a utilizar OneDrive, que nos permite cargar y almacenar nuestros archivos en la nuebe de manera segura. Hemos optado por este repositorio y no por otros más conocidos como GitHub o GitLab debido al gran tamaño de nuestros archivos.

Puedes acceder al repositorio donde está almacenado todo el código a través de la siguiente URL: Repositorio TFG Citi Bike Alfredo González-Izquierdo Antón.

Referencias

Arciniegas, Y. (2021). Estados Unidos se compromete a reducir las emisiones de carbono en un 50% para 2030. FRANCE 24. Recuperado de: https://www.france24.com/es/ee-uu-y-canad%C3%A1/20210422-cambio-climatico-cumbre-joe-biden-neutralidad-carbono-2030

Barrero, Jose Maria, Nicholas Bloom, and Steven J. Davis. (2021). "Why working from home will stick," National Bureau of Economic Research Working Paper 28731.

Bloomberg. (2014). Citi Bike supplier rides into bankruptcy. Recuperado de: https://www.crainsnewyork.com/article/20140122/TRANSPORTATION/140129956/citi-bike-supplier-rides-into-bankruptcy

Blumenschein, M., Debbeler, L. J., Lages, N. C., Renner, B., Keim, D. A., & El-Assady, M. (2020). V-plots: Designing hybrid charts for the comparative analysis of data distributions. Vol. 39 No. 3, pp. 565-577, doi: 10.1111/cgf.14002

Bozzi, A. D., & Aguilera, A. (2021). Shared E-scooters: A review of uses, health and environmental impacts, and policy implications of a new micro-mobility service. Vol. 13 No. 16, pp. 8676, doi: 10.3390/su13168676

Buttarazzi, D., Pandolfo, G., & Porzio, G. C. (2018). A boxplot for circular data. *Biometrics*. Vol. 74 No. 4, pp. 1492-1501, doi: 10.1111/biom.12889

Campbell, A. A., Cherry, C. R., Ryerson, M. S., & Yang, X. (2016). Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Vol. 67, pp. 399-414, doi: 10.1016/j.trc.2016.03.004

Chen, Y., Zhang, Y., Coffman, D., & Mi, Z. (2022). An environmental benefit analysis of bike-sharing in New York City. Vol. 121, pp. 0264-2751, doi: 10.1016/j.cities.2021.103475

CycleHop. (s.f.). Recuperado de: https://cyclehop.com/

Dauni, P., Firdaus, M. D., Asfariani, R., Saputra, M. I. N., Hidayat, A. A., & Zulfikar, W. B. (2019). Implementation of Haversine formula for school location tracking. Vol. 1402 No. 7, pp. 1742-6596, doi: 10.1088/1742-6596/1402/7/077028

Doğanlar, M., Mike, F., Kızılkaya, O., & Kardaşlar, A. (2024). Temperature, precipitation and economic growth: The case of the most polluting countries. Vol. 18 No. 4, pp. 2008-2304, doi: 10.1007/s41742-023-00555-5

EPA. (2022). Sources of greenhouse gas emissions. Recuperado de: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. Vol. 54, pp. 2210-6707, doi: 10.1016/j.scs.2019.101882

Naturgy. (2023). Evolución de las emisiones sectoriales de gases efecto invernadero en el contexto europeo 2017-2021. Recuperado de: https://www.fundacionnaturgy.org/publicacion/evolucion-de-las-emisiones-sectoriales-de-gases-efecto-invernadero-en-el-contexto-europeo-2017-2021/

Faghih-Imani, A., & Eluru, N. (2015). Analysing bicycle-sharing system user destination choice preferences: Chicago's Divvy system. Vol. 44, pp. 53-64, doi: 10.1016/j.jtrangeo.2015.03.005

Galatoulas, N.-F., Genikomsakis, K. N., & Ioakimidis, C. S. (2020). Spatio-temporal trends of E-bike-sharing system deployment: A review in Europe, North America and Asia. Vol. 12 No. 11, doi: 10.3390/su12114611

Galpin, T., Whitttington, J. L., & Bell, G. (2015). Is your sustainability strategy sustainable? Creating a culture of sustainability. Vol. 15 No. 1, pp. 1-17, doi: 10.1108/cg-01-2013-0004

GeoPy. (s/f). Welcome to GeoPy's documentation! — GeoPy 2.4.1 documentation. Recuperado de: https://geopy.readthedocs.io/en/stable/

Hannah Ritchie, Pablo Rosado and Max Roser (2020) - "Emissions by sector: where do greenhouse gases come from?" Published online at OurWorldInData.org. Recuperado de: https://ourworldindata.org/emissions-by-sector

Hosford, K., & Winters, M. (2019). Quantifying the bicycle share gender gap. Transport Findings. Doi: 10.32866/10802

IPSOS. (2023). What worries the world? Recuperado de: https://www.ipsos.com/sites/default/files/ct/news/documents/2023-12/Global-Report-What-Worries-the-World-December-23.pdf

JOCO. (s/f). Recuperado de: https://ridejoco.com/

Kou, Z., & Cai, H. (2019). Understanding bike-sharing travel patterns: An analysis of trip data from eight cities. Vol. 515, pp. 785-797, doi: 10.1016/j.physa.2018.09.123

Langford, B. C., Chen, J., & Cherry, C. R. (2015). Risky riding: Naturalistic methods comparing safety behavior from conventional bicycle riders and electric bike riders. Vol. 82, pp. 220-226, doi: 10.1016/j.aap.2015.05.016

NABSA. (2022). 2022 shared micromobility state of the industry report - north American bikeshare & scootershare association. North American Bikeshare & Scootershare Association. Recuperado de: https://nabsa.net/2023/08/10/2022industryreport/

Niu, Z., & Chai, L. (2022). Carbon emission reduction by bicycle-sharing in China. Vol. 15 No. 14, pp. 5136, doi: 10.3390/en15145136

Oeschger, G., Carroll, P., & Caulfield, B. (2020). Micromobility and public transport integration: The current state of knowledge. Transportation Research. Part Vol. 89, pp. 1361-9209, doi: 10.1016/j.trd.2020.102628

Ricci, M. (2015). Bike-sharing: A review of evidence on impacts and processes of implementation and operation. Vol. 15, pp. 28-38, doi: 10.1016/j.rtbm.2015.03.003

Ritchie, H., Rosado, P., & Roser, M. (2023). Breakdown of carbon dioxide, methane and nitrous oxide emissions by sector How much does electricity, transport and land use contribute to different greenhouse gas emissions? Recuperado de: https://ourworldindata.org/emissions-by-sector

Sanders, B. F., Feldman, D. L., Sweet, W., Matthew, R. A., Luke, A., Moftakhari, H. R., & AghaKouchak, A. (2015). Increased nuisance flooding along the coasts of the United States due to sea level rise: Past and future. Vol. 42, pp. 9846-9852, doi: 10.1002/2015gl066072

Smith, G., Banting, L., Eime, R., O'Sullivan, G., & van Uffelen, J. G. Z. (2017). The association between social support and physical activity in older adults: a systematic review. Vol. 14 No. 56, doi: 10.1186/s12966-017-0509-8

Spinlister. (s/f). Recuperado de: https://es.spinlister.com/

Wessler, H., Chan, C.-H., Wozniak, A., Wessler, H., & Chan, C.-H. (2021). The event-centered nature of global public spheres: The UN climate change conferences, Fridays for Future, and the (limited) transnationalization of media debates. Recuperado de: https://www.un.org/es/climatechange/un-climate-conferences

Xin, R., Ai, T., Ding, L., Zhu, R., & Meng, L. (2022). Impact of the COVID-19 pandemic on urban human mobility - A multiscale geospatial network analysis using New York bike-sharing data. Vol. 126, doi: 10.1016/j.cities.2022.103677

Zhang, Y., & Mi, Z. (2018). Environmental benefits of bike-sharing: A big data-based analysis. Vol. 220, pp. 296-301, doi: 10.1016/j.apenergy.2018.03.101

Zheng, L., Meng, F., Ding, T., Yang, Q., Xie, Z., & Jiang, Z. (2022). The effect of traffic status on dockless bicycle-sharing: Evidence from Shanghai, China. Vol. 381 No. 1, doi: 10.1016/j.jclepro.2022.135207

Zhu, D., Lan, J., Thornton, T., Mangalagiu, D., & Ma, Y. (2018). Challenges of collaborative governance in the sharing economy: The case of free-floating bike-sharing in Shanghai. Vol. 197 No. 1, pp. 356-365, doi: 10.1016/j.jclepro.2018.06.213