Facultad de CC. Económicas y Empresariales

Un estudio del mercado de alquiler en Madrid capital con R

Autor/a: Juan Pérez Llorente
Director/a: Patricia Soriano Machado

MADRID | Junio 2020

Resumen

En los últimos años la tecnología ha experimentado un enorme desarrollo, sobre todo en el ámbito digital, lo que ha provocado una generación masiva de datos de todo tipo a gran velocidad. Esta generación de datos recibe el nombre de *big* data, y "el proceso de cotejar, clasificar, procesar y estudiar datos de negocios, y usar modelos estadísticos y metodologías iterativas para transformar los datos en conocimientos de negocios" (MicroStrategy, 2020) es conocido como *business analytics.* Tal es el volumen y la variedad de los datos que se están generando, y la velocidad a la que esto sucede, que las herramientas utilizadas tradicionalmente han quedado obsoletas, ya que carecen de la capacidad para procesar cantidades tan grandes de información. Es por esta razón que se han desarrollado herramientas como Hadoop, Python o R.

En este trabajo se busca demostrar el potencial que este tipo de herramientas tienen para sacar información precisa y de calidad, y plantear algunos de los principales obstáculos que pueden aparecer al trabajar con datos, como puede ser la veracidad de la información. Para ello, haciendo uso del lenguaje de programación R, se va a analizar el mercado de la vivienda de alquiler de Madrid utilizando la información de la plataforma inmobiliaria Idealista. Tras limpiar y preparar la muestra seleccionada, se ha llevado a cabo un análisis descriptivo de cada una de las variables escogidas para el análisis y se han identificado los valores atípicos que pudiesen afectar a los resultados, para después desarrollar el modelo de regresión, con el objetivo de identificar las variables que más influyen en el precio mensual.

PALABRAS CLAVE: vivienda de alquiler, programación, dato, estadística descriptiva, modelo de regresión.

Abstract

In recent years, technology has undergone enormous development, especially in the digital sphere, which has caused a massive generation of data of all kinds at high speed. This generation of data is called big data, and "the process of collating, classifying, processing and studying business data, and using statistical models and iterative methodologies to transform data into business knowledge" (MicroStrategy, 2020) is known as business analytics. Such is the volume and variety of data being generated, and the speed at which this happens, that traditional tools have become obsolete, as they lack the ability to process such large amounts of information. It is for this reason that tools such as Hadoop, Pyton or R have been developed.

This paper seeks to demonstrate the potential that this type of tools have to obtain accurate and quality information, and to present some of the main obstacles that may appear when working with data, such as the veracity of the information. To do this, the Madrid rental housing market has been studied using the R programming language to analyze information from the Idealista platform. After cleaning and preparing the selected sample, a descriptive analysis of each of the variables chosen for the analysis has been carried out, and the outliers that could affect the results were identified. Then, the regression model has been developed, with the objective of identifying the variables that most influence the monthly price.

KEY WORDS: rental house, programming, data, descriptive statistics, regression model

Índice

Índice de tablas y figuras

1. Introducción

El mundo está siendo testigo de una de las mayores revoluciones tecnológicas de la historia, y en el centro se encuentra la digitalización de la sociedad y la generación de datos. Esta revolución está caracterizada por la dimensión de los cambios y la gran velocidad a la que estos ocurren, teniendo como resultado la generación constante y a gran velocidad de enormes volúmenes de datos con características muy variadas. Para poder hacer frente a este suceso se han ido desarrollando herramientas y programas capaces de seguir el ritmo de estos cambios para poder analizar la enorme cantidad de datos que se están generando de manera diaria para ofrecer soluciones a los distintos campos que componen la sociedad, desde la estrategia empresarial, hasta la campaña de comunicación de un partido político, o el despliegue de ayuda humanitaria. Tal es la importancia y potencial de este campo que se ha llegado a considerar un activo imprescindible, ya que tener la capacidad de manipular los datos y extraer información relevante es la única manera de poder utilizar este flujo constante de información. Aquellos organismos, empresas e individuos que no se adapten a este cambio de paradigma están destinados a desaparecer.

Un ejemplo de este tipo de herramientas es el lenguaje de programación R, un lenguaje informático perteneciente a la familia de las herramientas de *big data* y *business analytics* que ofrece una gran variedad de técnicas estadísticas y de visualización que permiten trabajar de manera efectiva y sencilla con grandes volúmenes de datos, facilitando la obtención de resultados y conclusiones aplicables a cualquier ámbito y área (R-project, s.f.).

El objetivo principal de este trabajo de investigación consiste en demostrar el enorme potencial que las herramientas de *big data* y *business analytics* tienen a la hora de obtener resultados coherentes y de calidad con el análisis de datos de los que puedan extraerse conclusiones relevantes. Para ello, se ha elaborado un caso práctico en el que se ha analizado el mercado madrileño de la vivienda en régimen de alquiler utilizando los datos de la plataforma inmobiliaria Idealista. Haciendo uso de los paquetes y funciones disponibles en R se va a estudiar la manera en la que el mercado de la vivienda de alquiler está distribuido, analizando los municipios más ofertados y el rango en el que se mueven las distintas variables que componen una vivienda, como el número de habitaciones o su superficie.

Este trabajo empieza con una breve explicación sobre la situación actual del mercado de alquiler en España para ofrecer unas nociones básicas de sus características principales, además de una comparativa de los precios entre comunidades autónomas. Si bien es verdad que el objetivo de este trabajo es demostrar la utilidad de R, se ha considerado importante realizar una pequeña mención al tema en el que se va a basar el análisis antes de entrar en materia. Seguidamente, se ofrece una sección sobre el *big data* y el *business analytics* en la que se exponen sus características, y las ventaja y obstáculos que plantea la revolución del dato. Una vez elaborada la parte teórica del trabajo, se procederá a la extracción y análisis de los datos. Esta sección está dividida en cuatro partes.

En primer lugar, se explica el proceso seguido para la extracción de los datos de la API de la plataforma inmobiliaria Idealista, seguido del procedimiento utilizado para realizar la primera limpieza de los datos, con el objetivo de crear el primer set en el que se va a basar el resto del análisis.

La segunda parte está compuesta por un análisis descriptivo de cada una da las variables seleccionadas, con la finalidad de ofrecer una primera instantánea de las características del mercado de alquiler según Idealista. Además, se aprovechará para eliminar aquellos valores que a primera vista parecen se demasiado atípicos al resto de la muestra, pero como se explicará, esto no es suficiente para identificar todos los valores atípicos que puedan perjudicar al análisis.

De manera muy relacionada con esta última idea y para conseguir una muestra que no presente demasiados valores atípicos, se ha llevado a cabo un proceso de identificación y eliminación de *outliers*. Este proceso se ha realizado para cada una de las variables para intentar conseguir una muestra que sea lo más adecuada posible para elaborar un buen modelo de regresión.

Finalmente, una vez la muestra está preparada, se procede a la elaboración del modelo de regresión, cuyo proceso está dividido en tres pasos:

1. Análisis de las relaciones entre variables mediante una matriz de correlaciones, con el objetivo de identificar indicios de heterocedasticidad o multicolinealidad.
2. Modelo de regresión simple de cada una de las variables respecto el precio mensual, para comprobar que los datos están distribuidos de manera aleatoria y siguen una distribución normal.

3. Construcción del modelo de regresión múltiple y mejora del modelo en base a los residuos.

2. El mercado de alquiler en España

2.1 Cómo se determina el precio de la vivienda de alquiler

Una de las principales críticas que está recibiendo el actual mercado de vivienda de alquiler es el elevado nivel de precios en el que estos activos se encuentran, consecuencia de la evolución que ha experimentado a lo largo de los años y cuyo mayor punto de inflexión fue la crisis de 2008. No corresponde a este trabajo de investigación explicar esta evolución, pero sí que sería interesante analizarlo en otro trabajo de investigación. Antes de plantear la situación actual del mercado de alquiler es importante conocer cuáles son los elementos que determinan el precio de una vivienda. Estos elementos son (Trecet, 2018):

- La localización, que es considerado el principal elemento que determina el valor de un piso.
- La superficie del piso. Esto no es ninguna novedad, cuantos más metros cuadrados tenga una casa, mayor será su valoración.
- Las características de la casa. En este segmento se incluyen el piso, si este es interior o exterior, si es un ático o un bajo, etc.
- El estado de la casa. Una casa renovada puede aumentar de manera considerable su valor.
- Otros servicios que la casa ofrece. Estos incluyen parking, piscina, jardín, ascensor, etc.…

Teniendo en cuenta estas variables podemos decir que el precio de alquiler de una vivienda se tiene que establecer de tal manera que se maximice la ganancia del alquiler, pero que también permita al arrendador ser selectivo con el inquilino que elije.

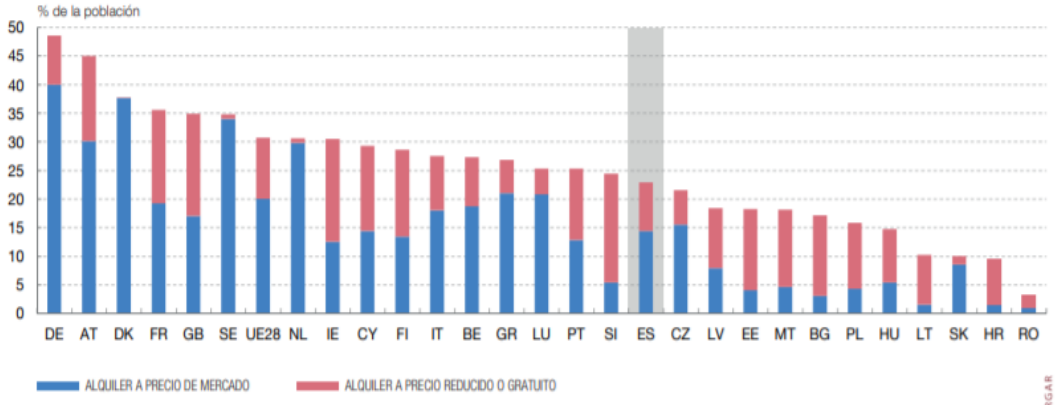
Existe un consenso entre plataformas y agentes inmobiliarios, e instituciones por el cual se aceptan solo estas variables a la hora de establecer el precio de una vivienda, pero es posible que existan otras que no están especificadas y que también pueden ser relevantes a la hora de determinar si una vivienda vale más o menos.

2.2 El mercado de alquiler en España

2.2.1 Situación del mercado de alquiler a partir de 2017

En España el número de hogares que residen en viviendas que no son de su propiedad es considerablemente reducido comparado con la media de la Unión Europea. Esto lo podemos comprobar si observamos la siguiente gráfica elaborada y publicada en 2018 por la Encuesta Europea de Ingresos y Condiciones de Vida sobre el porcentaje de residentes que habitaban una vivienda de alquiler en 2017 (López-Rodríguez & de los Llanos Mata , 2019):

Figura 1: Población residente en vivienda de alquiler en los países de la UE en 2017



Fuente: (López-Rodríguez & de los Llanos Mata , 2019)

Los resultados de la encuesta reflejaban que la proporción de población que vivía en vivienda alquilada en 2017 en España era del 22,9%, situándose muy por debajo de la media europea que se encontraba en el 30,7%. Los países que lideraron el ranking fueron Alemania (48,6%), Austria (45%) y Dinamarca (37,8%) (López-Rodríguez & de los Llanos Mata , 2019).

Aunque la vivienda de alquiler en España no tenga mucha relevancia, se ha podido observar una tendencia creciente en la demanda de este tipo de viviendas sobre todo tras la crisis económica de 2008. Según los datos obtenidos por el Instituto Nacional de Estadística, el peso del alquiler residencial ha aumentado del 19,4% en 2005 al 23,9% en 2018 (López-Rodríguez & de los Llanos Mata , 2019). Es importante mencionar que este incremento de ha dado sobre todo en el mercado de alquiler privado, ya que, por otra parte, el alquiler de vivienda social ha sufrido un descenso de un 0,8% desde 2005 (3,5%) a 2018 (2,7%). En 2018 había un total de 3 millones de hogares que habitaban en

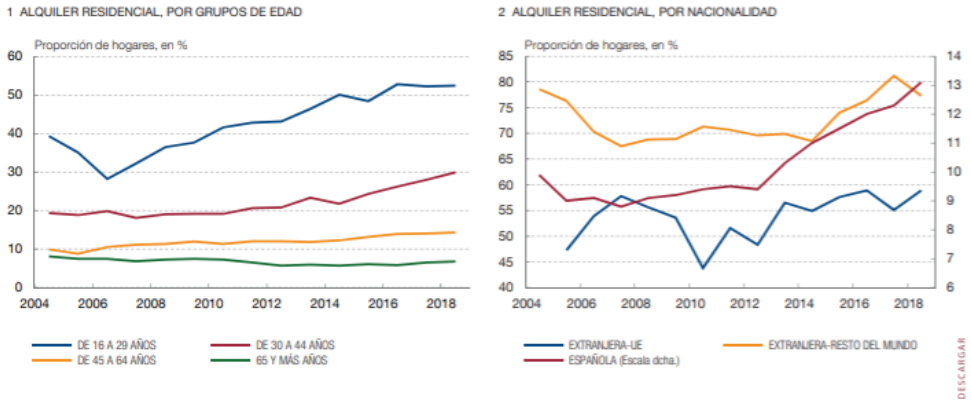
viviendas de alquiler, frente a los 2,4 millones de 2008 (López-Rodríguez & de los Llanos Mata , 2019).

2.2.2 Principales grupos y regiones

Los sucesos y razones que han provocado esta evolución del mercado de alquiler han condicionado también el perfil de la persona que busca alquilar una vivienda. Podemos distinguir dos grupos principalmente: jóvenes y extranjeros. Dentro del grupo de los jóvenes tienen especial relevancia aquellos que tienen entre 16 y 29 años, algo que se debe sobre todo a las precarias condiciones laborales con la que empiezan a trabajar. Se puede corroborar este dato con la información proveniente del Ministerio de Fomento, según el cual, siete de cada diez españoles menores de 30 años que se han emancipado residen en casa de alquiler.

La proporción de extranjeros que residen en España y que alquilan una vivienda es bastante elevado. Para ciudadanos de la UE se sitúa en un 58,9%, mientras que para otras nacionalidades alcanza el 77,3%. A continuación, se presenta un gráfico elaborado por el INE con datos extraídos de la Encuesta de Condiciones de Vida en el que se refleja toda esta información:

Figura 2: Evoluciónd del alquiler residencial en España (2004-2018)

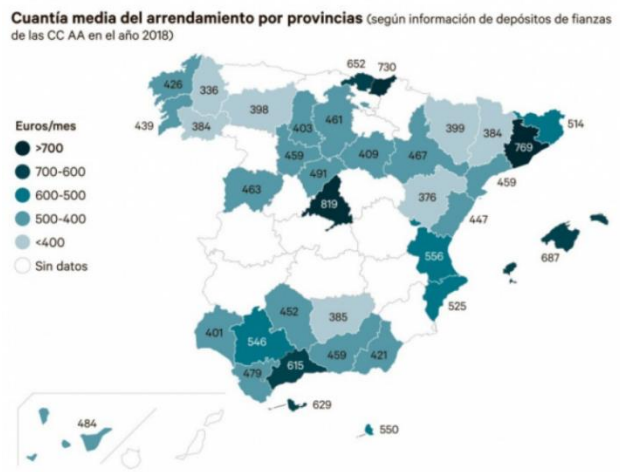


Fuente: (López-Rodríguez & de los Llanos Mata , 2019)

En lo que respecta a las zonas de España en las que hay una mayor actividad arrendataria se pueden destacar cuatro comunidades autónomas en las que el alquiler representa un elevado porcentaje sobre el total de inmuebles del mercado inmobiliario: islas Canarias (34,5%), Baleares (34,1%), Cataluña (30,4%) y Madrid (26,6%). Todas estas, salvo las islas Canarias, entran dentro del ranking de las comunidades autónomas

en las que el precio del alquiler es más elevado en proporción de precio por metro cuadrado, al que se une el País Vasco.

Figura 3: Cuantía media del arrendamiento por provincias



Fuente: (Tragacete, 2019)

En este mapa elaborado por el INE se pueden observar **los precios medios de 34 provincias y dos ciudades autónomas**. La Comunidad de Madrid se encuentra a la cabeza de la lista con un precio medio de 819 euros, seguida por Barcelona y Mallorca con un precio medio de 769 y 687 euros, respectivamente (Tragacete, 2019). Es importante mencionar que dada la naturaleza y objetivos de este trabajo de investigación, la información obtenida es suficiente para su desarrollo; no obstante, no hay que olvidar que esta se ha obtenido utilizando unas variables concretas, y no está teniendo en cuenta todas las opciones disponibles, como pueden ser los alquileres que se dan fuera de la legalidad, o cuyas partes han acordado un precio por debajo o por encima del precio de mercado. Por lo tanto, si se quisiera realizar un estudio más profundo y entrar más en detalle sería necesario acudir a otras fuentes y comparar resultados. Por otras fuentes se hace referencia a otros portales inmobiliarios como Fotocasa, o informes elaborados por otras entidades como pueden ser empresas (PricewaterhouseCoopers, Deloitte, CBRE, etc.) o instituciones.

La siguiente sección es la introducción al tema central de este trabajo de investigación. En primer lugar, se ofrece con una explicación sobre el *big data* y *business analytics*, para después analizar el mercado de alquiler madrileño utilizando el lenguaje de programación R.

3. *Big data* y *business analytics*

3.1 *Big Data*

El análisis del mercado de alquileres no es un tema especialmente innovador ni original, existen decenas de estudios, informes y otras investigaciones que analizan diferentes aspectos de este sector. Es por esta razón que lo innovador de este trabajo reside en el uso del programa R para llevar a cabo el análisis de los datos sobre el mercado de alquileres.

Este programa pertenece a la familia de las herramientas digitales que se han ido desarrollando para tratar el *big data* haciendo uso del *business analytics.* Este término hace referencia a la enorme cantidad de datos estructurados y no estructurados que inundan los negocios cada día (Powerdata, s.f.). El avance tecnológico, sobre todo en el ámbito digital con la aplicación de herramientas de código como Hadoop (conjunto de softwares que permite el almacenamiento y manipulación de volúmenes masivos de datos para resolver problemas complejos), ha resultado en una generación masiva de datos a una gran velocidad que son imposibles de tratar con las herramientas tradicionales como Excel. Cuando se acuñó este término se hizo utilizando las tres v's de Doug Laney (SAS, 2020):

- Volumen
- Velocidad
- Variedad

Además de estas tres primeras características, en los últimos años se han añadido otras dos: variabilidad y veracidad (SAS, 2020). Al existir infinidad de fuentes de datos, ya no solo se están generando enormes volúmenes de datos a una velocidad vertiginosa, si no que estos cambian constantemente de acuerdo a tendencias, por lo que es muy importante para las empresas ser capaces de igualar el ritmo de estos cambios para poder extraer el máximo potencial de los datos. Muy ligado a esto, las empresas se enfrentan al reto de extraer datos verídicos y de calidad. Al haber tantas fuentes de información, las empresas tienen que corroborar los datos con diferentes sistemas y asegurarse de que están extrayendo información relevante para su negocio.

3.2 *Business analytics*

La extracción de datos de calidad es cada vez más importante en el mundo empresarial, pero lo realmente esencial es el uso que se da a esos datos. Para ello, se ha desarrollado el *business analytics*, que es "el proceso de cotejar, clasificar, procesar y estudiar datos de negocios, y usar modelos estadísticos y metodologías iterativas para transformar los datos en conocimientos de negocios" (MicroStrategy, 2020). Dicho de manera resumida, el *business analytics* permite procesar los enormes volúmenes de datos que se están generando para obtener información sintetizada y resumida de ellos, facilitando así la toma de decisiones para el negocio.

Dentro del *business analytics* se pueden identificar tres áreas (Itelligent, 2018):

- Analítica descriptiva: consiste en el análisis de datos históricos para identificar patrones y tendencias
- Analítica predictiva: utilizando tanto los datos pasados como los generados en el momento presente permite determinar la probabilidad de que un evento tenga lugar.
- Analítica prescriptiva: determina nuevas formas de operar para que la empresa obtenga sus objetivos

Estas tres áreas permiten a la empresa desarrollar su actividad de una manera óptima y eficiente, identificar tendencias y adelantarse a los cambios que puedan darse en el mercado. En el contexto actual, el buen uso de los datos se ha convertido en un activo esencial para que las empresas puedan obtener una ventaja competitiva creando así más valor. Se pueden utilizar estos procesos para la reducción de costes, desarrollo de productos, investigación de mercados y *decisión making*, entre otros.

3.3 Desafíos del *business analytics* y el *big data*

El potencial de estas herramientas para las empresas es impresionante, sobre todo si tenemos en cuenta que estas están poco desarrolladas y aún hay un gran margen de mejora. No obstante, los avances que están teniendo lugar también implican una serie de desafíos.

Al haber tantas fuentes y tipos de datos se vuelve muy difícil el trabajar con estos. Además, el tamaño de los datos que se generan es enorme, de hecho, se esperaba que para 2020 cada persona generase 1,7 MB de datos cada segundo (Grupo BIT, 2020). Esto complica mucho ejecutar un proceso de calidad en un periodo de tiempo razonable. Por

último, los datos cambian constantemente y a gran velocidad, lo que implica que la validez de estos es corta. Se puede observar como las mismas características que dan tanto valor a los datos presentan al mismo tiempo grandes problemas, ya que al no tratar los datos de manera adecuada se pueden obtener resultados erróneos que puedan resultar en una mala toma de decisiones.

A continuación, se empieza a elaborar el proceso necesario para extraer, limpiar y manipular los datos de la plataforma Idealista para poder desarrollar el modelo de regresión y así analizar el mercado de alquiler. Al haber utilizado el lenguaje de programación R es posible que hay términos y pasos que no sean fáciles de entender, por lo que se ha intentado explicarlo de la mejor manera posible. Además, al final del trabajo se puede encontrar todo el código utilizado.

4. Metodología

Lo primero que hay que hacer es escoger la fuente desde la que queremos extraer los datos con los que vamos a trabajar. Como ya se ha explicado en el apartado de *big data* y *business analytics*, hay un gran número de fuentes a nuestra disposición, y para realizar un estudio adecuado que ofrezca resultados que se puedan utilizar es esencial escoger una que nos vaya a aportar datos de calidad. Es importante mencionar que no es recomendable utilizar como base de este estudio los datos provenientes del INE, ya que son el resultado de analizar una muestra de viviendas que se mantiene estable en el tiempo, y no tiene en cuenta las nuevas incorporaciones al mercado de alquiler, ni las características que estas tienen en comparación con las que ya estaban incluidas. Por esta razón, vamos a extraer los datos de la plataforma Idealista, por ser "el portal con mayor cobertura del mercado de alquiler para el conjunto de España […]" (López-Rodríguez & de los Llanos Mata , 2019), para desarrollar esta investigación.

Recientemente, la Comisión Nacional de los Mercados y la Competencia (CNMC) ha iniciado una investigación a siete portales inmobiliarios, entre los que se encuentra Idealista, por "posible coordinación de precios y otras condiciones comerciales en sus actividades". El organismo acusa a estas empresas de haber usado algoritmos y programas informáticos que habrían provocado un aumento del precio de las viviendas, conducta considerada ilícita. Idealista ha negado tajantemente haber participado en esta actividad

y ha asegurado que su portal solo funciona como punto de conexión entre arrendador e interesados, y que son los primeros quienes fijan los precios y condiciones del alquiler, algo que la empresa no manipula en ningún momento (Expansión, 2020). Este suceso es un ejemplo de lo complicado que es establecer un índice de precios que refleje fielmente la situación del mercado de alquileres, algo que puede dificultar su estudio y, que como se ha dicho anteriormente, necesita la comparación entre plataformas para obtener resultados lo más precisos posibles.

4.1 Extracción de datos

Se podría pensar que utilizando la propia página web de Idealista sería suficiente para obtener la información necesaria; sin embargo, eso no funcionaría por temas de seguridad y privacidad. La información disponible en la página web es pública, todo el mundo puede visitarla y comparar precios. No obstante, poder trabajar con esa información es una realidad completamente distinta, ya que se trata de información que solo la propia empresa tiene permitido manipular, algo que sucede en todos los sectores. Es por esta razón que se crea lo que se conoce como APIs, acrónimo para *Application Programming Interface*. Estos sistemas permiten a los usuarios acceder a los datos de la propia plataforma para extraer y trabajar con esa información de una manera mucho más rápida y eficaz utilizando el código de la propia empresa. Es importante mencionar que la gran mayoría de estas API no son de libre acceso, sino que hay que pedir permiso a la entidad en cuestión para que permita bucear por su plataforma, e incluso cuando se tiene acceso, este es limitado por cuestiones de seguridad.

Una vez se tiene acceso a la API de Idealista, tenemos que seguir las instrucciones de código que ofrecen para poder extraer aquellos datos que nos interesan. El mercado de viviendas de alquiler español es muy amplio, además de que cada comunidad autónoma presenta características muy diferentes, por lo que para poder realizar un análisis preciso y de calidad nos vamos a centrar en la Comunidad de Madrid. Además, siendo conscientes de que la vivienda más demandada para alquilar son los pisos, vamos a utilizar la información que hace referencia exclusivamente a este tipo de viviendas. Por lo tanto, hay que crear un código que extraiga de la API de Idealista todos aquellos pisos que estén en alquiler dentro de Madrid capital.

4.2 Traducción a lenguaje R y preparación de los datos

El siguiente paso consiste en trasladar los datos extraídos al programa R para poder manipularlos. R, dicho de manera simplificada, es un tipo de lenguaje informático

que ofrece una gran variedad de técnicas estadísticas y de visualización que permiten trabajar de manera efectiva y sencilla con grandes volúmenes de datos, facilitando la obtención de resultados y conclusiones aplicables a cualquier ámbito y área (R-project, s.f.). Una de las cualidades de este lenguaje, que comparte con otros similares, es su capacidad de evolución. Estos programas están abiertos para que todo el mundo los pueda modificar y adaptar, siempre dentro de unos límites y con ciertas restricciones, lo que permite su desarrollo de manera constante. Otra de las principales características de R es que funciona con paquetes, cada uno de los cuales se utiliza para una tarea específica. Por ejemplo, el paquete *ggplot2* proporciona herramientas para generar gráficos muy variados, mientras que el paquete *cluster* permite realizar, como su propio nombre indica, análisis de clústeres para identificar observaciones que comparten ciertas características.

Se han utilizado varios de estos paquetes para llevar a cabo el análisis de datos; sin embargo, para no sobrecargar esta parte de la investigación con exceso de información sobre el programa utilizado, se irá profundizando en cada uno de ellos a medida que se vayan mencionando. Además, de esta manera se facilitará la explicación del procedimiento seguido para la obtención de resultados.

Para traducir los datos a lenguaje de R utilizamos el paquete *json*, con el que obtenemos el fichero **data**. Este set de datos presenta 10.412 observaciones, que identifican cada una de las viviendas que vamos a estudiar; y 34 variables, que son las características que Idealista utiliza para describir cada una de las viviendas.

Muchas de las variables que se pueden encontrar en el set original no van a aportar nada a la investigación, por lo que procedemos a eliminar todas aquellas que no vamos a utilizar. Algunos ejemplos de estas variables son: *propertycode, thumbnail o numPhotos*, que presentan información sobre el número de fotos disponibles en el portal, el código de referencia, o si el dueño de la vivienda es un particular o una inmobiliaria. La tabla resultante sigue manteniendo el mismo número de observaciones, solo que ahora, presenta solo 14 variables que hemos ordenado para facilitar el trabajo:

- Categóricas
    - *adress*: especifica la dirección de la vivienda
    - *district*: especifica el municipio de Madrid en el que se encuentra la vivienda. Este término hace referencia a las demarcaciones en la que se

subdivide Madrid, entendiendo por estas los distritos de la Moncloa, Tetuán, Hortaleza, Retiro y similares.

- *neighborhood*: especifica los barrios, es decir, las partes en las que se dividen los municipios
- *propertyType:* se pueden identificar tres tipos de pisos: *flat*, que consiste en una única planta compuesta por varias habitaciones; *penthouse*, "último piso de un edificio, más bajo de techo que los inferiores, que se construye para cubrir el arranque de las techumbres y a veces como ornamento"; y duplex, "conjunto de dos pisos superpuestos y unidos por una escalera interior, destinado a vivienda independiente dentro de un edificio de varias plantas. Estas dos definiciones se han extraído del Diccionario de la Real Academia Española (Real Acaedmia Española, 2019). Como se explicará más adelante, se trabajará solo con aquellas viviendas pertenecientes a la tipología *flat.*

- Numéricas
  - *price*: el precio publicado en la plataforma de Idealista
  - *size*: superficie del suelo de la vivienda en metros cuadrados
  - *priceByArea*: precio por metro cuadrado
  - *floor*: piso en el que se encuentra (1º, 2º, 3º, …)
  - *rooms*: número de habitaciones de las que dispone la vivienda
  - *bathrooms*: número de baños de los que dispone la vivienda
- Dicotómicas
  - *exterior*: si se trata de un piso exterior (1) o interior (0). En un principio esta variable era lógica, por lo que en vez de 1 y 0 utilizaba los valores TRUE y FALSE. Se ha convertido esta variable en numérica para facilitar su estudio y poder incluirla en el modelo de regresión que se planteará más adelante.
  - *hasLift*: si tiene ascensor (1) o no (0)

Por cuestiones de formato, a partir de ahora se hará referencia a estas variables utilizando su traducción al español, de tal manera que los términos serán los siguientes:

- Calle
- Tipo
- Precio

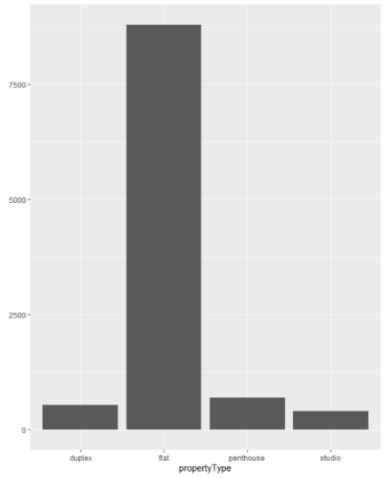
- Tamaño
- Precio por metro cuadrado
- Planta
- Habitaciones
- Baños
- Exterior
- Ascensor

Se ha explicado en la sección de *big data* y *business analytics* que uno de los principales obstáculos que se puede encontrar a la hora de trabajar con datos es la tipología de estos, algo que es muy difícil de controlar al provenir de una fuente externa. El problema más común derivado de esta cuestión es el de los NA, constantes lógicas que aparecen cuando falta un valor (RDocumentation, s.f.) y que impiden trabajar con el conjunto de datos. Para solucionar este problema utilizamos la función **datos[is.na(datos)] <- 0** para asignar a estos el valor 0. De esta manera podemos continuar con el análisis, no obstante, surge otro problema que no podemos resolver. Los NA aparecen en cualquier campo, por lo que al asignar el valor 0 estamos perdiendo información ya que no sabemos que es lo que tendría que haber en su lugar, restando precisión al análisis. Por ejemplo, puede darse el caso de que una vivienda se encuentre en el municipio 0 o que tenga una superficie de 0 metros cuadrados. La manera de abordar este problema se explicará para cada caso.

4.3 Limpiar subcategorías

Los datos descargados ofrecen información de todas aquellas viviendas que Idealista considera pisos, incluyendo dúplex, estudios (studios) y áticos (*penthouse*). En el siguiente gráfico se puede observar que el tipo de vivienda que predomina en el mercado son los pisos normales (*flats*), con una gran diferencia respecto a los otros tres:

Figura 4: Distribución de las viviendas de alquiler en Madrid según tipología



Fuente: Elaboración propia con datos extraídos de Idealista

Para obtener este gráfico se ha utilizado el paquete *ggplot2*, que pone a disposición del usuario una serie de funciones y herramientas para elaborar distintas gráficas dependiendo de las características de los datos que se quieren visualizar. En este caso, al tratarse de una única variable categórica (*propertyType*), se ha generado un gráfico de barras utilizando el comando **ggplot(datos)+ geom_bar(aes(x=propertyType)).** Este comando se utilizará a lo largo de la investigación para representar variables categóricas, ya que en ella se pueden observar patrones y sacar conclusiones de manera rápida y sencilla. Se ha mencionado antes que R permite visualizar datos de una manera sencilla, no obstante, el estilo no es especialmente atractivo.

Haciendo uso de cifras, en la siguiente tabla se puede observar que los pisos normales representan el 84,42% del total de todas las viviendas ofertadas, muy por encima de los dúplex, áticos y estudios que representan el 5%, el 6,68% y el 3,79% del total de la vivienda ofertada, respectivamente.

Tabla 1: Distribución de frecuencias de las viviendas de Madrid según tipología

Property Type	Freq
Duplex	526
Piso	8.795
Ático	696
Estudio	395

Fuente: Elaboración propia con datos extraídos de Idealista

Para centrar el estudio en los pisos es necesario crear otro dataset en el que no aparezcan ni *studio, ni penthouse* ni *duplex*. Para ello, hay que utilizar la función *filter*, que identifica las filas que cumplen una determinada condición, en este caso que aparezca el campo *flat*, y genera un *subset* en el que solo se incluyen dichas observaciones. Para identificar este *subset* se le ha dado el nombre de **trabajo**. El comando quedaría de la siguiente manera: **trabajo<-datos%>% filter(propertyType=="flat")**. Esta función pertenece al paquete *Dplyr*, que ofrece un listado de verbos (acciones/funciones) que permiten al usuario resolver los problemas más comunes de la manipulación de datos (RDocumentation, s.f.).

El último problema que presentan los datos extraídos se encuentra en la columna de la variable planta. Se puede observar que hay varias observaciones en las que en vez de indicar la planta en la que se encuentra la vivienda se utilizan las nomenclaturas bj, sb, et y st. Al no conocerse el significado de esta terminología se presentan dos posibles alternativas. La primera consiste en la eliminación de todas las observaciones que presenten estos términos, lo que supondría la pérdida de un número importante de observaciones; la segunda implica interpretar su significado. La página web de Idealista ofrece una serie de filtros entre los que se encuentra el de planta, que presenta tres opciones: última planta, plantas intermedias o bajos. Para el buen desarrollo de esta investigación se ha trabajado asumiendo que los cuatro valores mencionados hacen referencia a la última categoría. Por consiguiente, es requerido asignar a los campos en los que aparecen alguno de estos valores el valor 0. Es importante recordar que se trata de una variable categórica, por lo que antes de realizar cualquier cambio es necesario convertirla a numérica, con la consecuencia de que los valores que queremos sustituir aparecerán como NA al ser letras y no números. Haciendo uso del comando explicado anteriormente se asignan a estos el valor 0. La razón por la que se ha hecho la conversión de la variable a numérica es porque se respeta el orden de las plantas, es decir, que la más baja será la 0, seguida de la 1 y así hasta llegar a la más alta, algo que facilitará la ejecución de los comandos y la interpretación de los resultados. El resultado de esta limpieza del set de datos es que las observaciones con las que se van a trabajar a partir de ahora se han reducido a 8.795, manteniendo las 14 variables seleccionadas antes.

Resulta muy importante entender que es inevitable encontrar valoras extremos o atípicos dentro de una base de datos. Estos pueden restar precisión al análisis o distorsionar los resultados que se obtengan, por lo que es de vital importancia identificar

estos valores y eliminarlos para poder proceder con la investigación. Con esto se quiere explicar que a medida que se proceda con el análisis del set de datos es muy probable que se vayan perdiendo observaciones. No obstante, al ser valores atípicos su eliminación no debería suponer un problema.

4.4 Análisis descriptivo individual de variables

Hasta ahora solo se han manipulado los datos extraídos de la API de Idealista para, puesto de manera sencilla, volverlos aptos para trabajar con ellos. A continuación, se procede con lo que es realmente el análisis de los datos para obtener resultados y conclusiones.

En la introducción se presentaba un listado de los elementos que más influyen a la hora de establecer el precio de la vivienda de alquiler, por lo que se va a continuar la investigación con el análisis de estas variables con el objetivo de identificar patrones, elementos destacados u otros componentes que merezca la pena resaltar para construir el modelo de regresión múltiple y sacar las conclusiones correspondientes. Cabe mencionar que el análisis de las 14 variables no va a ser necesario, ya que hay algunas que no van a aportar información que sea relevante. Por esta razón, las siguientes variables se van a omitir para esta parte de la investigación: calle, precio por metro cuadrado, tipo de propiedad.

1. Variable distrito

Antes de proceder con el análisis propiamente dicho de los datos, se había tenido que llevar un proceso de limpieza para para conseguir que el set tuviese una características y calidad adecuadas para el análisis, y como consecuencia de dicho proceso se sustituyeron varios campos en los que aparecía NA por el valor 0. Es ahora cuando salta a la vista uno de los principales problemas del *big data* y el *business analytics*, que es la tipología de los datos extraídos. Como consecuencia de la limpieza hecha antes se puede encontrar un municipio 0, lo que supone una anomalía y un problema, ya que no hay ninguna manera de saber a qué municipio pertenecen las viviendas correspondientes. Una manera de obtener esta información sería revisando el campo barrio de cada una de estas observaciones; no obstante, debido a falta de conocimiento en el momento de desarrollar este trabajo solo se podría llevar a cabo de manera manual. Por lo tanto, se procederá a eliminar estas observaciones, aunque suponga una pérdida considerable de elementos a

estudiar (se eliminan 99 observaciones). En el análisis de la variable barrio aparece el mismo problema, pero la manera de abordarlo será distinta.

Sin tener en cuenta aquellas observaciones que tiene un municipio 0, de los datos extraídos de Idealista se obtienen 182 municipios. Esto resulta un poco sorprendente ya que, de acuerdo con el INE, Madrid está dividido en 179 municipios, lo que significa un desfase de tres municipios. Esta diferencia no es especialmente significativa, pero pone de manifiesto otro de los problemas de la trata de datos, que es la fiabilidad de los datos que se extraen de terceros. Convenía hacer este apunte para evitar posibles confusiones, pero para que el análisis de los datos fuese óptimo habría que revisar esto para descubrir posibles duplicidades o errores.

Volviendo al análisis de la variable, al haber tantos municipios resulta poco práctico analizar todos y cada uno de los distritos, por lo que para el análisis descriptivo se van a escoger los 10 municipios en los que hay una mayor oferta de viviendas de alquiler. En alguna otra situación tendría sentido identificar aquellos municipios que tienen le menor número de viviendas ofertadas, pero al haber tantos municipios que tan solo ofrecen una o dos viviendas para alquilar, no aportaría nada al estudio, por lo que no se van a tratar.

Lo primero que hay que hacer es obtener la tabla de frecuencias utilizando el comando ya mencionado al principio de esta sección del trabajo. A continuación, se presentan los resultados obtenidos en R a través de una tabla.

Tabla 2: Distribución de frecuencias de los municipios con mayor oferta de viviendas de alquiler en Madrid

Distrito	Freq
Salamanca	1.574
Centro	1.543
Chamberí	830
Chamartín	721
Tetuán	369
Moncloa	355
Retiro	355
Arganzuela	256
Fuencarral	253
Hortaleza	253

Fuente: Elaboración propia con datos extraídos de Idealista

Para comprobar si los datos de Idealista coinciden con otras plataformas se van a exponer los valores medios del precio, el número de habitaciones y baños, al igual que la planta más ofertada en estos distritos, y si predominan las viviendas exteriores y con ascensor. Este estudio se realizará sobre los cinco municipios más ofertados en la plataforma inmobiliaria de Idealista:

Tabla 3: valores medios de las viviendas situadas en los barrios más ofertados

Municipio	Precio (€)	Superficie(m2)	Habitaciones (u)	Baños (u)	Planta	Ascensor	Exterior
Salamanca	2.771	137,20	2,57	2,12	3	0,76	0,94
Centro	1.897	91,38	1,86	1,86	1	0,71	0,69
Chamberí	2.117	118,90	2,32	2,32	3	0,92	0,73
Chamartín	2.173	127,20	2,44	2,44	1	0,93	0,89
Tetuán	1.478	95,58	2,05	2,05	1	0,83	0,83

Fuente: Elaboración propia con datos extraídos de Idealista

Uno de los principales problemas que surgen en el análisis de datos son los *outliers* o valores extremos. Estas observaciones presentan valores atípicos al resto de la muestra, lo que puede resultar en una distorsión de los resultados una vez realizado el modelo de regresión. Por lo tanto, a lo largo del análisis descriptivo se van a ir eliminado estos valores de acuerdo con los resultados observables. Además, hay valores extremos que no se pueden identificar a primera vista, por lo que antes de empezar el desarrollo del modelo de regresión se utilizarán una serie de comandos en R para identificar y eliminar aquellos *outliers* que surgen de la combinación del conjunto de datos.

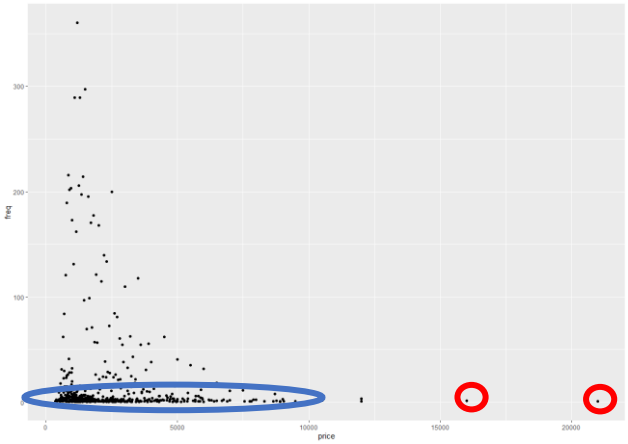
Para conseguir la mayor precisión posible en los resultados del modelo, el resto del análisis se va a realizar con aquellas viviendas localizadas en municipios que tienen una frecuencia superior a 50, ya que por debajo de ese número la representación es mínima, e incluirlos en el estudio podría distorsionar los resultados.

2. Variable precio

Como ya se ha mencionado el principal objetivo de este trabajo es estudiar cuales son las variables que más influyen en el precio de la vivienda del mercado de alquiler, por lo que algo esencial es analizar como varían los precios de dicho mercado. En el set de datos analizado el precio de las viviendas se encuentra entre los 350 y los 21.000 euros mensuales. Al ser un rango tan amplio se puede afirmar que los precios demasiado elevados van a ser casos muy particulares, por lo que habría que eliminarlos. En la

siguiente gráfica se puede observar que, en efecto, a partir de 5.000 euros al mes el número de viviendas se reduce de manera drástica:

Figura 5: Distribución de la variable precio mensual



Fuente: Elaboración propia con datos extraídos de Idealista

En el gráfico se pueden identificar claramente dos *outliers* que se han marcado en rojo, por lo que es necesario eliminarlos para que no perjudiquen los resultados del modelo. Es importante explicar que, a pesar de que haya varios precios cuya frecuencia sea 1, solo convienen eliminar aquellos que tengan un precio demasiado elevado respecto al resto de las observaciones ya que estos son los valores que pueden distorsionar los resultados. Esta es la razón por la que no se eliminan todas las observaciones que se encuentran en la zona azul.

3. Variable planta

La variable planta ofrece valores entre el -2 y el 60 (ver anexo 1). A parte de que este rango es demasiado amplio, es muy probable que las viviendas que se encuentran en plantas negativas o superiores a la planta 20 sean muy pocas, por lo que se va a realizar un primer reconocimiento con el objetivo de corroborar esta idea mediante una tabla de frecuencias en la que se presentan aquellas viviendas que presenten une frecuencia superior a 50.

Tabla 4: Distribución de frecuencias de la variable planta

Planta	Freq
1	1430
2	1418
3	1287
4	1093
5	668
0	659
6	418
7	250
8	103
10	63

Fuente: Elaboración propia con datos extraídos de Idealista

Se puede observar claramente que la mayor parte de los pisos se encuentran entre la planta 0, o bajo, y la planta 10. Por lo tanto, se procederá a eliminar aquellas observaciones que se encuentren en plantas superiores a la 11 e inferiores a la 0. Resulta curioso encontrar pisos por debajo del nivel 0, ya que podría tratarse de un fallo de la plataforma. De todas maneras, al ser solamente dos observaciones las que se encuentran en estas plantas, no afectaría al estudio su eliminación.

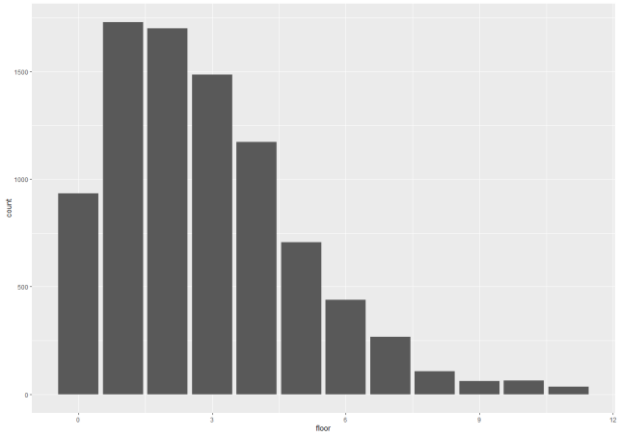
Esta tabla, al igual que todas aquellas en las que se quiere mostrar la frecuencia de cada valor, se ha generado utilizando las funciones **group_by()** y **summarise()**, ambas pertenecientes al paquete **Dplyr** mencionado anteriormente. El comando resultante es el siguiente: **floor_freq<-datos_pisos%>% group_by(floor) %>% summarise(freq=n())**. Este comando se utilizará siempre que se quiera obtener la frecuencia de los valores de una variable, es decir, las veces que cada valor está repetido dentro de un *dataset*.

Para crear el *subset* que incluya solo las viviendas que se encuentran en las plantas mencionadas utilizamos los siguientes comandos:

- **datos_pisos<-trabajo%>% filter(floor<12)**
- **datos_pisos<-datos_pisos%>%filter(floor>-1)**

Tras ejecutar estos dos comandos se obtiene otro *dataset* al que se asigna el identificador **datos_pisos** y que tiene 8.700 observaciones. En la siguiente tabla se pueden observar las frecuencias correspondientes, es decir, las veces que cada planta se repiten dentro del *dataset*:

Figura 6: Distribución de frecuencias de la variable planta



Fuente: Elaboración propia con datos extraídos de Idealista

Se puede observar que la mayor parte de las viviendas del mercado de alquiler en Madrid, según Idealista, se encuentran entre la primera y la quinta planta, acumulando el 69,2% del total de viviendas.

4. Variable habitaciones

Tras realizar un primer análisis de la variable habitaciones se puede observar que el rango de habitaciones se encuentra entre 1 y 12 por vivienda. Además, se pueden identificar claramente los *outliers*, ya que el número de viviendas con un número de habitaciones superior a 6 es prácticamente inexistente en comparación con el conjunto de datos, pudiendo identificar seis viviendas con siete habitaciones, una con ocho habitaciones, una con nueve y otra con doce.

Tabla 5: Distribución de frecuencias de la variable habitaciones

Habitación	Freq
2	2559
1	1949
3	1812
4	832
5	258
6	59
7	6
8	1
9	1
12	1

Fuente: Elaboración propia con datos extraídos de Idealista

Por lo tanto, para evitar que estas observaciones perjudiquen al desarrollo de la investigación, de procederá a su eliminación. El *dataset* resultante, **datos_pisos_rooms,** presenta 7469 observaciones. Como se puede observar, y se ha mencionado al principio de este apartado, el número de observaciones que componen la muestra que se está analizando va reduciendo su número como consecuencia del propio análisis.

5. Variable baños

Las viviendas que se están analizando presentan entre 1 y 7 baños, como se puede observar a la tabla inferior. No obstante, en la siguiente tabla se puede ver claramente que la gran mayoría de las viviendas tienen 1 o 2 baños, representando el 83,91% sobre el total:

Tabla 6: Distribución de frecuencias de la variable baños

Baño	Freq
2	3295
1	2842
3	827
4	352
5	128
6	23
7	2

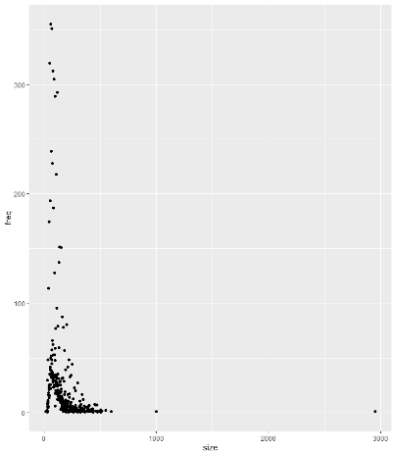
Fuente: Elaboración propia con datos extraídos de Idealista

En este caso no se van a eliminar los valores extremos porque es posible que haya muchos cuartos de baño en una vivienda que tenga una gran superficie. De todas maneras, en la siguiente sección se va a utilizar R para identificar los *outliers* que no se han considerado como tales en este primer análisis descriptivo.

6. Variable superficie

La superficie es probablemente una de las variables que más impacto tienen a la hora de establecer el precio de la vivienda de alquiler, aunque dicha hipótesis se comprobará en el siguiente apartado. Dentro del set de datos que se está analizando, el rango de las superficies de las viviendas se encuentra entre los 15 y los 2.950 metros cuadrados. Sin embargo, la media de todas las observaciones es de 111,8 metros cuadrados y la mediana se encuentra en la vivienda de 133 metros cuadrados. Esto implica que hay que acotar las observaciones para eliminar todas aquellas que sean atípicas.

Figura 7: Distribución de la variable superficie



Fuente: Elaboración propia con datos extraídos de Idealista

En este gráfico de puntos, generado utilizando la función **geom_jitter**, se puede observar claramente que la mayor parte de los valores se encuentran entre los 15 y los 500 metros cuadrados de superficie. En el eje Y se encuentra la frecuencia de cada una de las superficies, lo que pueda dar una primera instantánea de cuál es el tamaño de los pisos más ofertados en el mercado de vivienda de alquiler madrileño. De entre todas las posibilidades, las viviendas más ofertadas son aquellas que tienen entre 40 y 120 metros cuadrados. La tabla de frecuencias se encuentra en el anexo 2.

7. Variable barrio

Al realizar el análisis preliminar de esta variable surge el mismo problema que surge al analizar los municipios, que hay campos en los que aparece un cero en vez de un barrio. En este caso el problema es un poco más serio, ya que al igual que se podía identificar el municipio a través del barrio, no se puede hacer lo mismo a la inversa, por lo que no hay manera alguna de conocer a que barrio pertenece la vivienda. Se podría utilizar la calle de cada observación para situar la vivienda en un barrio haciendo uso de un mapa, pero esto requeriría demasiado tiempo.

Esta explicación se ha dado para mostrar uno de los problemas que surgen al trabajar con datos de fuentes externas, y es que no se puede controlar que toda la información esté completa o que sea correcta. A lo largo del análisis se han presentado varios de estos ejemplos que, o bien se han modificado o se han eliminado. En este caso no va a ser necesario tomar ninguna de estas medidas, ya que tan solo se busca identificar los barrios más ofertados en el portal de Idealista, por lo que basta con ignorar aquellas

observaciones que tengan un 0 en ese campo. Además, eliminar estas observaciones supondría una pérdida muy grande de información que afectaría al desarrollo del modelo de regresión.

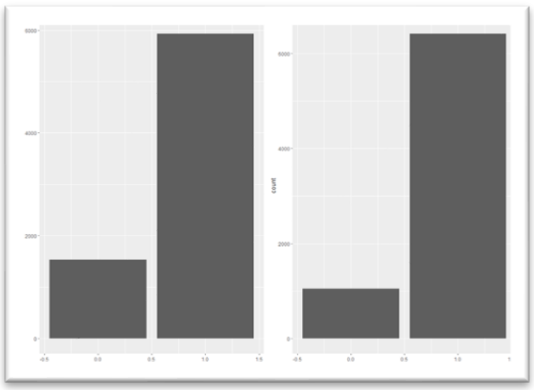
Siguiendo los mismos pasos utilizados para analizar la variable municipios y se obtienen los diez barrios en los que hay una mayor oferta de vivienda en alquiler (ver anexo):

1. Recoletos
2. Chueca-Castilla
3. Castellana
4. Goya
5. Malasaña-Universidad
6. Lavapiés-Embajadores
7. Palacio
8. Almagro
9. Bernabéu-Hispanoamérica
10. Lista

8. Variable exterior y variable ascensor

Las variables que quedan por analizar son aquellas que determinan si una vivienda está situada en el exterior o interior del edificio, y si tiene o no ascensor. Al tratarse de variables dicotómica (1,0), su análisis es mucho más sencillo y simplemente se va a mostrar una gráfica conjunta para identificar cuál de las características predomina en las viviendas estudiadas.

Figura 8: Distribución de las variables exterior (derecha) y ascensor (izquierda)



Fuente: Elaboración propia con datos extraídos de Idealista

Se puede observar claramente como la mayoría de las estudiadas tienen ascensor y se encuentran en el exterior.

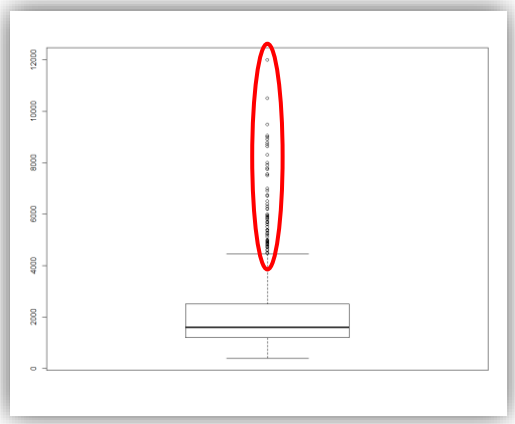
4.5 Identificación de valores atípicos

Antes de proceder con el desarrollo del modelo de regresión es importante puntualizar que la muestra que se va a utilizar es considerablemente más pequeña que la muestra original, presentando 7.469 observaciones frente a las 10.412 iniciales. Además, en esta sección se van a identificar todos los valores extremos que no se han podido identificar con el estudio descriptivo, por lo que seguramente esta muestra se reduzca aún más.

Para el desarrollo de esta sección se han utilizado las funciones pertenecientes al paquete **DMwr**, utilizado para examinar grandes volúmenes de datos con el objetivo de generar nueva información que se pueda analizar. Esto también se conoce como *data mining*.

Una de las maneras más sencillas para identificar valores extremos es mediante la representación gráfica de los datos en un diagrama de caja. Este consiste en una representación gráfica de una variable cuantitativa elaborada con el primer y tercer cuartil, los valores máximo y mínimo, y los datos atípicos. A continuación, se presenta el diagrama de caja de la variable precio para ofrecer un ejemplo. Los valores atípicos se encuentran marcados en rojo.

Figura 9: Gráfico de caja de la variable precio para identificar outliers



Fuente: Elaboración propia con datos extraídos de Idealista

Al igual que en el análisis descriptivo se han eliminado aquellas observaciones que se diferenciaban mucho del resto de los datos, a partir de ahora no va a ser tan sencillo ni automático. Para decidir eliminar una observación, hay que estudiar si afecta al modelo, para lo que se va a aplicar el método *Interquartile Range* (IQR) (R-bloggers, 2020). Este método calcula la diferencia entre el cuartil 1 y el cuartil 3, por lo que identifica como valores atípicos todos aquellos que se encuentren por encima del 75% de los datos o por debajo del 25%. La lista de comando se presenta a continuación:

Q<-quantile(datos$price,probs=c(.25,.75))

iqr<-IQR(datos$price)

up<-Q[2]+1.5*iqr

low<-Q[1]-1.5*iqr

Para eliminar los *outliers,* se genera un *subset* en base al segundo y tercer cuartil generando el comando **datos_sin<- subset(datos, datos$price > (Q[1] - 1.5*iqr) & datos$price < (Q[2]+1.5*iqr))** (R-bloggers, 2020) que tienen como resultado un *dataset* de 5.966 observaciones.

Una vez completado, solo falta revisar que no quede ningún valor atípico dentro de los datos, para lo que se va a ejecutar el comando **which(datos_sin$price %in% boxplot.stats(datos_sin$price)$out)** con cada una de las variables numéricas, mostrando aquellas observaciones que siguen siendo *outlier*s. Al ejecutar el comando se presentan el número de cada observación dentro de la base del propio *dataset,* es decir, que no sale toda la información, por lo que para identificar los valores aplicamos la función **print()** y se limpian los datos en base a los valores de las observaciones.

En este caso, basta con eliminar aquellas viviendas que tienen un precio superior a los 3.100 euros. Al repetir el mismo proceso con el resto de las variables numéricas se puede observar que no quedan más valores atípicos, por lo que se puede continuar con el desarrollo del modelo.

4.6 Modelo de regresión

Como se ha explicado en la introducción, el objetivo de este trabajo es estudiar cual es el impacto que tienen las variables de una vivienda a la hora de establecer su precio de alquiler. Para ello, se va a desarrollar un modelo de regresión lineal, que consiste en hacer uso de un conjunto de herramientas estadísticas para explorar y cuantificar la

relación de dependencia entre la variable dependiente o respuesta, y una o más variables independientes.

Existen muchos métodos de estimar los parámetros de la ecuación del modelo, pero para este trabajo se va a utilizar el método de mínimos cuadrados por ser el más aceptado y utilizado. El desarrollo de esta sección de ha basado en los trabajos de Joaquín Amat Rodrigo, *Ejemplo práctico de regresión lineal simple, múltiple, polinomial e interacción entre predictores*, publicado en agosto de 2016 en RPubs (Rodrigo, RPubs, 2016) e *Introducción a la Regresión Lineal Múltiple* publicado en julio de 2016 (Rodrigo, 2016); y el trabajo de María Elvira Ferre Jaén, *FEIR 40: Modelos de Regresión*, actualizado por última vez el 4 de abril de 2019 (Ferre Jaén, 2019). Además, para el desarrollo del código se han utilizado otras publicaciones de la web RPubs, en la que se pueden encontrar ejemplos de código aplicables a una gran variedad de temas.

La manera más sencilla de plantear la modelización es a través de una ecuación lineal de la forma $Y=\beta0+\beta1X1+\cdots+\beta nXkY=\beta0+\beta1X1+\cdots+\beta nXk$. Lo que se busca es crear un modelo que permita predecir el valor de la variable dependiente a partir de las variables independientes. La finalidad del método de mínimos cuadrados es que los residuos sean pequeños, lo que significa que su media sea cero. Dicho de una manera más sencilla, lo que se busca es que la diferencia entre los datos reales y las predicciones sea lo más pequeña posible.

4.6.1 Correlación lineal

El primer paso que hay que dar consiste en analizar la correlación lineal, que es una medida que busca cuantificar el grado de variación entre dos variables y cuyos valores oscilan entre -1 y 1 (Ucha, 2020). Lo que se busca es definir el grado de intensidad y el sentido que la relación de dos variables tiene. Existen tres posibles resultados:

- Ambas variables se mueven en la misma dirección, aunque no en la misma medida. Esta sería una relación positiva o directa ($r>0$)
- Las variables varían de manera inversa, es decir, cuando una crece la otra cae. Esta sería una relación inversa o negativa ($r<0$)
- Que no haya ninguna correlación y cada variable siga su propia evolución ($r=0$)

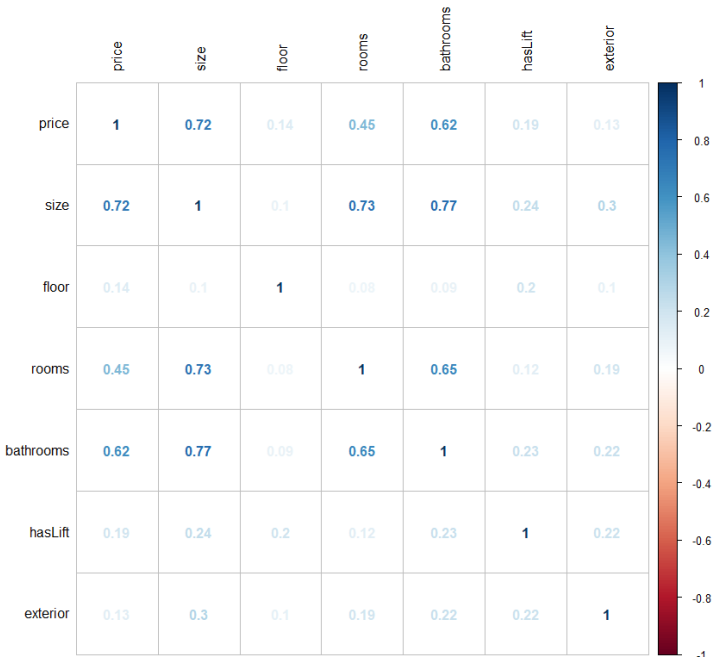
Es necesario hacer una serie de apuntes antes de continuar. El primero es que las relaciones no suelen ser perfectas, razón por la que se han utilizado los símbolos de menor y mayor para explicar las posibles relaciones. No obstante, en caso de que una relación sea perfecta, r sería igual 1 o -1. El segundo, y más importante, es que los valores obtenidos hay que interpretarlos utilizando su valor absoluto; es decir, que la relación entre dos variables será más fuerte cuanto más cerca este su r de los extremos (-1 o 1).

La manera más sencilla de identificar el tipo de relación existente entre dos variables es a través de un diagrama de dispersión, en el que se plasma una nube de puntos que representa la relación. No obstante, para ofrecer un estudio más concreto, esta manera de analizarlo se va a omitir, utilizando otra representación gráfica en su lugar.

Uno de los elementos básicos para poder proceder con esta parte del análisis es el coeficiente de correlación. Existen varios coeficientes como el Rho de Spearman y el Tau de Kendall; no obstante, se va a utilizar el de Pearson, que es el más adecuado para trabajar con variables cuantitativas y que sigan la distribución normal.

Para calcular la correlación en R se va a utilizar la función **cor()**, que se encuentra dentro del paquete **stats**. El comando utilizado queda de la siguiente manera: **cor(cor$price, cor$size, method = "pearson")**. Los resultados se presentan en la siguiente tabla de correlaciones:

Figura 10: Matriz de correlaciones



Fuente: Elaboración propia con datos extraídos de Idealista

Se puede observar que las variables que presentan una mayor correlación son el tamaño (0,69), el número de habitaciones (0,56) y el número de baños (0,73). Las otras tres variables tienen una correlación bastante pequeña por debajo de 0,15.

Una vez obtenidos los índices de correlación es necesario realizar un test para evaluar su significación estadística. Para ello se va a utilizar la función **cor.tes()**. Para no mostrar los resultados de manera individual se va a utilizar otra función en la que los p-valores muy pequeños se redondean a cero. En los resultados podemos observar que todas las variables tienen un p-valor menor a 0,05, por lo que son significativos y se acepta que los cambios en estas variables tienen un efecto, por pequeño que sea, sobre los cambios en la variable precio.

4.6.2. Regresión lineal simple

Siguiendo la misma línea de analizar la relación entre la variable dependiente (precio) y cada una de las demás variables de manera individual, se va a proceder con el desarrollo de la regresión lineal simple. Es importante mencionar, que cumplir este paso va a servir para asentar las bases del modelo múltiple, ya que va a permitir determinar si los datos cumplen los requisitos necesarios para desarrollar el modelo múltiple. Este modelo sigue la expresión $Y=\beta0+\beta1X+ \varepsilon$, en la que $\beta0$ y $\beta1$ son los coeficientes de regresión, además:

- $\beta0$ es la ordenada en el origen
- $\beta1$ es la pendiente de la recta de regresión
- ε es la variable que representa todos aquellos factores que influyen en la variable dependiente en pequeña magnitud, y a la que se conoce como el error aleatorio. Esta variable es muy importante ya que representa la existencia de incertidumbre dentro de la relación entre dos variables. En el caso del precio de la vivienda de alquiler podrían incluirse el perfil del arrendador, la presencia de un fondo que busque maximizar su rentabilidad o que el suelo de la vivienda esté inclinado.

Este modelo se basa en las siguientes hipótesis:

- Los errores son totalmente independientes, es decir, que su correlación es 0
- El conjunto de los residuos tiene media 0
- La varianza de los residuos es constante
- Los residuos se comportan como variables normales

La situación ideal sería aquella en la que todos los puntos de la gráfica se encontrasen en la recta. La realidad es que existen muchas rectas dentro de un mismo modelo, por lo que el objetivo, como se ha mencionado antes, consiste en encontrar la recta en la que la diferencia entre los valores reales y las predicciones se la mínima. Para valorar la validez de los modelos se utilizará el p-valor de la t de *student* y el coeficiente de determinación (R2).

Para generar el modelo de regresión lineal por mínimos cuadrados se va a utilizar la función **lm()**, perteneciente al paquete mismo paquete **stats** con el que se han calculado los índices correlación para cada una de las variables respecto al precio. Tras ejecutar el comando **modelo_simple<-lm(data=cor,formula=price~size),** se obtienen la función del modelo para las variables precio y tamaño. Habría que ejecutar el comando **modelo_simple<-lm(data=cor,formula=price~size)** para cada una de las variables que se quieren estudiar. Para no incluir demasiadas tablas y facilitar la interpretación de los resultados se presenta, a continuación, una tabla en la que se incluyen todos los valores obtenidos en R, seguida de la explicación pertinente. (Consultar el anexo para ver los resultados obtenido en R)

Tabla 7: Resultados individuales de los modelos simples

Variable	p-value	Multiple R2
tamaño	2,00E-16	0,512
habitaciones	2,00E-16	0,198
planta	2,00E-16	0,0196
baños	2,00E-16	0,382
ascensor	2,00E-16	0,0164
exterior	2,00E-16	0,0361

Fuente: Elaboración propia con datos extraídos de Idealista

El primer parámetro que se va a analizar es el p-valor que muestra el nivel de significación para cada variable. Esto quiere decir que comprueba la hipótesis nula de que el coeficiente es igual a 0, es decir, que una variable no tiene efecto sobre la respuesta. Cuando un p-valor es menor a 0,05 se puede rechazar la hipótesis nula, lo que significa que los cambios que se producen en la variable dependiente están relacionados con los cambios que se producen en la variable independiente. En la tabla de arriba se puede observar que las cuatro variables independientes presentan un valor menor al 0,05, por lo que podemos afirmar que las variables están significativamente relacionadas con el

precio, que los cambios en el precio de la vivienda de alquiler están relacionados con los cambios que se dan en el tamaño, el número de habitaciones y de baños, la planta en la que se encuentra, y si tiene ascensor o es exterior.

Para comprobar la validez de modelo hay que utilizar el coeficiente de determinación (R2), que mide la bondad de ajuste de un modelo, es decir, "la proporción de la variabilidad de la variable dependiente explicada por la variabilidad de las variables independientes" (Martínez de Ibarreta, Álvarez Fernández, Budría Rodríguez, Curto González, & Escobar Torres, 2017). En la tabla de arriba se puede observar que solo el tamaño de la vivienda explica un porcentaje elevado de la variabilidad del precio, un 51%. El resto de las variables tiene un R2 muy pequeño, aunque no eso no quiere decir que no sirvan para el desarrollo del modelo conjunto, simplemente significa que de manera individual no explican la variabilidad del precio.

Por último, al analizar los residuos de los seis modelos de regresión simples, se puede observar que los datos se distribuyen de forma lineal en los seis casos. Además, se puede observar que los residuos están prácticamente alineados, por lo que se puede confirmar que su distribución es prácticamente normal. Para corregir esto, habría que identificar de nuevo los valores atípicos y eliminarlos. No obstante, se va a profundizar más en esta parte al estudiar el modelo múltiple. La regresión lineal también se puede representar gráficamente, pero para evitar ocupar demasiado espacio se han situado en el Anexo3.

4.6.3 Modelo de regresión múltiple

De la misma manera que el modelo de regresión simple se ha utilizado para explicar el precio en función de cada una de las otras variables de manera individual, ahora se va a proceder a analizar el valor del precio en función del resto de las variables en su conjunto con el objetivo de determinar cuales tienen una mayor influencia en la determinación del precio de la vivienda de alquiler. El modelo sigue la siguiente ecuación (Rodrigo, 2016): $Yi=(\beta0+\beta1X1i+\beta2X2i+\cdots+\beta nXni) +ei$.

- Bo representa el valor de la variable dependiente, manteniéndose constantes el resto de las variables
- Bi representa el efecto promedio que tiene el incremento en una unidad de la variable predictora sobre la variable dependiente, manteniéndose constantes el resto de las variables.

- ei es el residuo o error.

A diferencia del modelo simple, el modelo múltiple necesita cumplir una serie de condiciones para ser válido:

- No colinealidad o multicolinealidad, es decir, que las variables predictoras no estén linealmente relacionadas entre sí. En caso de existir este fenómeno, no se podría identificar de manera precisa el efecto individual de cada variable. Esto ocurre cuando el índice de correlación entre alguna de las variables independientes es igual a 1, y como se ha podido observar en la tabla de correlaciones, esto no sucede en los datos que se están estudiando.
- Las variables independientes tienen que estar linealmente relacionadas con la variable independiente mientras los demás predictores se mantienen constantes. De nuevo, como se ha explicado en el apartado anterior, esta condición se cumple para todas las variables.
- Los residuos deben estar distribuidos de forma normal, también explicado y demostrado en la sección anterior.
- La varianza de los residuos debe ser constante en todo el rango de observaciones, lo que demuestra la aleatoriedad de los datos, también conocido como homocedasticidad. En caso de que las varianzas sean iguales los datos presentarían heterocedasticidad, que supone un gran problema para el desarrollo del modelo.

Una vez establecidas las condiciones se puede plantear el modelo y analizar su validez. El comando que se va a utilizar es el mismo que el del modelo simple, solo que ahora se incluyen todas las variables dentro de la función **lm(formula = price ~ size + rooms + bathrooms + floor + exterior + hasLift + district, data = datos_sin)** Conviene recordar que la variable municipio es categórica, y por lo tanto, cada uno de los municipios seleccionados van a considerarse como una variable más. Véase a continuación la tabla resultante:

Tabla 8: Resultados del modelo múltiple

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	311.9384	26.0550	11.972	< 2e-16	***
size	9.4670	0.2143	44.186	< 2e-16	***
rooms	-58.9116	7.2486	-8.127	5.32e-16	***
bathrooms	190.2764	11.4786	16.577	< 2e-16	***
floor	18.8728	2.3648	7.981	1.74e-15	***
exterior	-34.9480	11.1906	-3.123	0.001799	**
hasLift	-33.4430	13.1081	-2.551	0.010757	*
districtBarajas	-234.3155	45.1240	-5.193	2.14e-07	***
districtCarabanchel	-203.2026	34.1150	-5.956	2.73e-09	***
districtCentro	292.8998	22.8843	12.799	< 2e-16	***
districtChamartín	230.7670	25.2100	9.154	< 2e-16	***
districtChamberí	312.6750	24.4826	12.771	< 2e-16	***
districtCiudad Lineal	-73.0133	31.1183	-2.346	0.018994	*
districtEl Soto de la Moraleja	324.6688	50.7951	6.392	1.77e-10	***
districtEncinar de los Reyes	569.3733	43.1176	13.205	< 2e-16	***
districtFuencarral	-143.0065	30.8160	-4.641	3.55e-06	***
districtHortaleza	-142.6462	29.7768	-4.791	1.71e-06	***
districtLatina	-42.0006	41.9276	-1.002	0.316510	
districtMoncloa	103.2802	28.5818	3.613	0.000305	***
districtPuente de Vallecas	-196.1199	46.2819	-4.238	2.30e-05	***
districtRetiro	178.1806	28.8300	6.180	6.83e-10	***
districtSalamanca	432.9570	23.2443	18.626	< 2e-16	***
districtSan Blas	-231.9872	35.9677	-6.450	1.21e-10	***
districtTetuán	77.4385	27.7708	2.788	0.005313	**
districtZona Prado de Somosaguas - La Finca	-118.6292	58.1818	-2.039	0.041501	*

Fuente: Elaboración propia con datos extraídos de Idealista

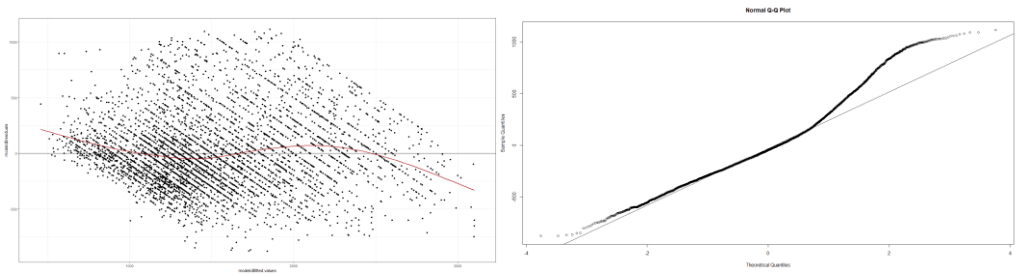
Se puede observar claramente que prácticamente todas las variables son significativas, con excepción de los distritos de Prado Somosaguas, Barajas y Ciudad Lineal, y si la vivienda tiene ascensor o si es exterior. Esto último puede explicarse porque al ser considerados como extras, es posible que no condicionen demasiado el precio final de la vivienda, aunque si tienen un mínimo de influencia por ser una característica que añade valor al activo. Además, se pueden extraer los siguientes resultados:

- Por cada metro cuadrado extra que tenga una vivienda, el precio sube en 9,47 EUR
- Por cada baño extra, una vivienda cuesta 190 EUR más y por cada habitación extra, el precio se reduce en 58,9 EUR
- Si la vivienda se encuentra en una planta más alta, el precio aumenta en 18,87 EUR

Los resultados obtenidos para las variables superficie y planta pueden ser muy útiles para futuros estudios o para sacar conclusiones respecto el precio de las viviendas de alquiler. No obstante, los resultados referentes al número de habitaciones y de baños habría que estudiarlos más en profundidad porque no se sabe a partir de qué número el precio cambia según lo determinado.

Al analizar el coeficiente de determinación se puede observar que el modelo es capaz de explicar el 67,56% de la variabilidad del precio mensual de la vivienda de alquiler. Este es un porcentaje muy elevado, aunque como se verá más adelante, es posible aumentarlo para conseguir un modelo capaz de explicar un mayor cambio.

Figura 11: Gráficos de residuos (izquierda) y su distribución (derecha)



Fuente: Elaboración propia con datos extraídos de Idealista

Lo primero que hay que analizar es la normalidad de la distribución de los residuos. En la gráfica de la derecha se puede observar que, en efecto, los datos presentan una distribución normal, aunque en la parte final se altera esta distribución, por lo que habrá que identificar esos valores atípicos y eliminarlos. Para demostrar que la distribución de los residuos es normal, se puede utilizar también, y probablemente sea más preciso, el test de Shaphiro, aunque por razones desconocidas, en el momento en el que se realizó el análisis, al ejecutar el comando en R sale un error que indica que la muestra tienen que tener entre 3 y 5.000 observaciones, y se está utilizando una muestra superior, por lo que no da ningún resultado.

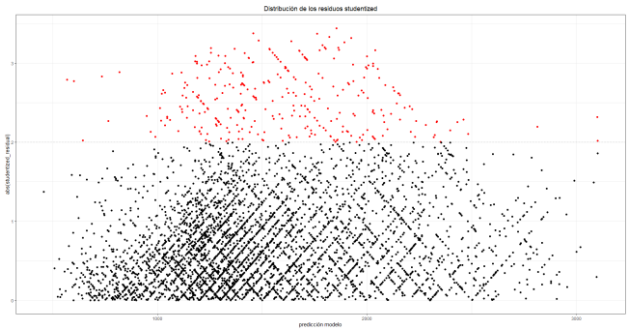
Para poder demostrar que los valores están distribuidos de manera aleatoria, es decir, que presenten homocedasticidad y no están distribuidos siguiendo un patrón claro. En el gráfico de la izquierda se puede observar que se cumple esta condición.

Por último, queda demostrar la existencia o no de multicolinealidad, es decir, que las variables predictoras estén linealmente relacionadas entre sí. Ya se ha demostrado en al plantear la matriz de correlaciones que ninguna de las variables tiene una relación demasiado fuerte con las demás, por lo que también se cumple este requisito

4.6.3.1 Mejora del modelo

Ya se ha demostrado que el modelo es válido, por lo que solo queda buscar la manera de mejorarlo. En este caso se van a identificar los valores atípicos dentro del análisis de residuos. El resultado gráfico es el siguiente:

Figura 12: Representación gráfica de los outliers en los residuos



Fuente: Elaboración propia con datos extraídos de Idealista

Los puntos rojos son los valores atípicos identificados de acuerdo al límite puesto, que en este caso ha sido k igual a 2,5. Se procede a su eliminación siguiendo el mismo proceso utilizado en el apartado de modelo de regresión simple, quedando como resultado un *dataset* de 5.472 observaciones. De nuevo, se puede observar como la muestra original de 10.412 observaciones se ha acabado reduciendo de manera muy considerable. Una vez eliminados los valores atípicos, si se vuelve a realizar todo el proceso del modelo múltiple se puede observar que se ha obtenido un mejor modelo capaz de explicar un 76,93% de la variabilidad del precio de la vivienda de alquiler. Es muy probable que al volver a buscar valores atípicos aparezcan otros, pero este valor del R2 es suficiente para el objetivo de este trabajo.

Como se han eliminado varias observaciones para mejorar el modelo, es necesario revisar que la información de los datos finales coincide con la información y los resultados obtenidos en el análisis descriptivo. Al comprobarlo, se pueden observar los siguientes cambios:

- En la variable municipios, el barrio más ofertado pasa a ser Centro y no Salamanca.
- Los precios de las viviendas están más distribuidos, encontrando la mayoría de las viviendas entre los 800 y los 2.500 EUR.
- La distribución de las plantas sigue siendo prácticamente la misma
- El número de baño sigue manteniendo el mismo orden, solo que no hay viviendas que tengan más de tres baños.
- Las variables superficie, número de habitaciones, exterior y ascensor apenas muestran cambios. Solo es necesario mencionar que se han eliminado observaciones como consecuencia del análisis de los *outliers.*

El último apunte que conviene realizar es, exponer los valores medios de una vivienda en régimen de alquiler en Madrid, pero para no repetir la información se va a presentar directamente en las conclusiones.

5. Conclusión

En los últimos años, la tecnología ha experimentado un increíble avance en todos los sentidos, provocando importantes cambios en todas las industrias y sectores, resultando en una mejora considerable de los procesos productivos, desde el diseño de nuevos productos hasta su comercialización y venta. En este sentido, la digitalización está ocupando un lugar central en esta revolución, ya que está permitiendo que empresas, organizaciones e instituciones puedan extraer información relevante de los enormes volúmenes de datos que se están generando a gran velocidad y de manera constante.

Como ya se ha mencionado en este trabajo, las características de estos datos (volumen, velocidad, variedad, veracidad y variabilidad) han tenido como consecuencia que las herramientas tradicionales como Microsoft Excel hayan quedado de alguna manera obsoletas, ya que carecen de la potencia necesaria para soportar volúmenes de datos de gran envergadura, y de las funcionalidades necesarias para manipular la gran variedad de datos que se están generando. Por esta razón, se ha ido desarrollando otras herramientas y programas que permiten trabajar con estos datos, para poder extraer resultados y conclusiones aplicables a cualquier ámbito de la sociedad, desde la estrategia de marketing de una empresa, hasta la detección de redes de corrupción.

Para este trabajo se ha utilizado el lenguaje de programación R para analizar el mercado de la vivienda en régimen de alquiler en la Comunidad de Madrid, centrando el estudio en Madrid capital. Mediante el uso de distintos paquetes, funciones y comandos, se ha conseguido desarrollar un modelo de regresión lineal capaz de explicar el 76,93% de la variabilidad del precio mensual de la vivienda en base a las variables tamaño, planta, nº de habitaciones y baños, municipio, ascensor y exterior, con datos extraídos de la API de la plataforma inmobiliaria Idealista.

Con los paquetes básicos de R, como **Dplyr, ggplot2** o **factoextra,** se ha llevado a cabo un primer proceso de limpieza y estructuración de los datos extraídos con el objetivo de preparar la muestra para su manipulación, de tal manera que solo quedan aquellas viviendas que pertenecen a la categoría de piso, es decir, una única planta compuesta por varias habitaciones. A este *dataset* se la ha denominado **datos**, compuesto por 8.975 observaciones y 10 variables.

El siguiente paso ha consistido en el desarrollo de un análisis descriptivo de cada una de las variables para obtener una primera imagen de la distribución del mercado de vivienda de alquiler en Madrid según Idealista, creando el set de datos denominado **datos_pisos_rooms,** en el cual, el número de observaciones se ha reducido a 7.469. En esta parte se ha eliminado aquellos valores que distanciaban demasiado del resto de las observaciones, y se han realizado modificaciones para adecuar los datos para el desarrollo del modelo de regresión. Un ejemplo de este tipo de modificaciones se puede encontrar en la variable planta, en la que se podían observar varias observaciones que presentaban siglas en sus valores y que, haciendo uso de la información de Idealista y la intuición, se han cambiado por el valor 0 para facilitar el desarrollo del análisis. Los resultados obtenidos se muestran a continuación:

- Los cinco barrios más ofertados son Salamanca, Centro, Chamberí, Chamartín y Tetuán
- Están situadas entre la sexta planta y el bajo
- Tienen entre una habitación y cinco, y entre cuatro y un baño
- Tienen una superficie que se sitúa entre los 25 y los 500 metros cuadrados
- Tienen ascensor y son exteriores

Una vez finalizado este primer análisis, se ha empezado llevado a cabo un proceso de identificación de valores atípicos o *outliers*, para eliminar aquellas observaciones que pudiesen perjudicar la precisión de los resultados. El paquete utilizado para esta parte es el **DmwR**, que permite identificar todos los valores atípicos de cada variable en función a toda la muestra. En esta parte del proceso se ha llevado a cabo una eliminación considerable de las observaciones, resultando en un *dataset* con 6.126 observaciones.

Con los datos completamente limpios y preparados, se ha analizado mediante una matriz de correlación las relaciones y su intensidad entre las distintas variables, y se ha podido observar que los datos no muestran signos de heterocedasticidad ni multicolinealidad.

El modelo de regresión se ha dividido en dos partes, y para su elaboración se han utilizado los paquetes **lmtest** y **psych,** . Primero, se ha llevado han construido modelos de regresión lineal simple entre cada una de las variables y el precio mensual, con el objetivo de identificar su comportamiento, si los datos estaban distribuidos de manera aleatoria y si los residuos presentaban una distribución normal. Nuevamente, los

resultados han confirmado estas primicias, lo que ha permitido proceder con la construcción del modelo múltiple. El proceso es el mismo que para los modelos simples, solo que se han incluido todas las variables dentro del modelo. Los resultados muestran que este primer modelo es capaz de explicar el 67,56% de la variabilidad del precio. Aunque se es un coeficiente de determinación bastante elevado, se ha intentado mejorar el modelo mediante un análisis de los residuos con el objetivo de identificar y eliminar aquellas observaciones que se considerasen atípicas, haciendo uso del paquete **car**. El resultado es el ya mencionado, un modelo de regresión capaz de explicar el 76,93% de la variabilidad del precio mensual. Es importante mencionar que el set resultante tiene 5.472 observaciones, casi un 50% menos que las observaciones de la muestra original, lo que muestra la dificultad que tiene trabajar con datos extraídos de fuentes externas.

Respecto a la vivienda en régimen de alquiler en Madrid, se puede concluir que la vivienda de alquiler media cuesta 1.485 EUR, ocupa 90,89 metros cuadrados (17,33EUR/m2), se encuentra entre la segunda y la tercera planta, tiene dos habitaciones y dos baños, es exterior y tiene ascensor (ver anexo). Dicho esto, según el propio Idealista, en marzo de 2020 el metro cuadrado costaba 16,3 EUR de media (Idealista, 2020); y según la plataforma EnAlquiler, el precio medio del alquiler en marzo de 2020 ha sido de 1.745 EUR (enAlquiler, 2020).

Por último, se ha considerado importante plantear tres conclusiones que han surgido con la elaboración de este trabajo, y que hacen referencia al potencial que las herramientas de *big data* y *analytics* tienen para plantear soluciones a problemas de toda índole, y los obstáculos a los que se pueden enfrentar.

Lo primero que merece la pena resaltar es la rapidez y eficiencia que estás herramientas ofrecen para el análisis de enormes volúmenes de datos que pueden tener características muy variadas. El proceso que más tiempo ha llevado es el de limpiar y preparar los datos, ya que es necesario analizar la estructura de los datos para poder decidir qué variable son importantes para el análisis que se quiere llevar a cabo, y que planteamiento se va a hacer para llevarlo a cabo. Las demás partes que componen el trabajo en R se han podido extraer sin ninguna dificultad de internet, eso sí, realizando los cambios necesarios para adaptar el código al trabajo que se quiere desarrollar.

Siguiendo este planteamiento, también es importante resaltar la facilidad con la que se pueden utilizar estas herramientas. No hay que cometer el error de pensar que todos

los programas son iguales ni que son igual de exigentes o desafiantes, pero si es verdad que, al ser programas de código abierto, es posible encontrar códigos ya desarrollados para cualquier estudio que se quiera realizar, por lo que solo hace falta encontrar el que mejor se adapte a las necesidades particulares y adaptarlo a los datos que se quieren analizar. El modelo de regresión de este trabajo se ha basado en tres trabajos de terceros: *Ejemplo práctico de regresión lineal simple, múltiple, polinomial e interacción entre predictores* y *Introducción a la Regresión Lineal Múltiple* de Joaquín Amat Rodrigo (Rodrigo, RPubs, 2016); *y FEIR 40: Modelos de Regresión*, de María Elvira Ferre Jaén (Ferre Jaén, 2019). Además, gran parte del código empleado de ha extraído de publicaciones que han hecho otros usuarios en sitios web como RPubs, y que se ha adaptado para poder desarrollar este trabajo de investigación. Lo que se quiere explicar, es que cualquier persona, dentro de unos límites de formación y capacidades, puede aprender a utilizar estos programas, siempre y cuando se tenga interés real.

Directamente ligado a esta idea, se encuentra el carácter democrático que estas herramientas pueden tener. Como ya se ha mencionado, cualquier persona con un mínimo de formación puede aprender a utilizar este tipo de programas, y todo lo necesario para ello está disponible en internet. Esto puede brindar oportunidades a muchas personas, reduciendo la brecha existente entre estratos de la sociedad. Puede resultar muy interesante estudiar el posible impacto que estas herramientas puedan tener en reducir la brecha de desigualdad que existe en la actualidad.

Finalmente, conviene mencionar que apenas se ha empezado a explotar este tipo de herramientas, y que hay un gran margen de mejora, sobre todo en lo referente a la categorización de los datos y la manera en la que se presenta la información. En este trabajo se han visto varios ejemplos de los problemas que se pueden encontrar al utilizar datos extraídos de fuentes externas, como es el encontrar valores vacíos o con información que no se puede interpretar directamente. Todo este proceso de identificar, modificar y eliminar valores que puedan perjudicar el desarrollo del modelo, y la precisión de los resultados obtenidos puede implicar una pérdida considerable de observaciones, y por lo tanto de información que podría ser utilizada.

6. Bibliografía

enAlquiler. (marzo de 2020). *enAlquiler*. Obtenido de enAlquiler: https://www.enalquiler.com/precios/precio-alquiler-vivienda-madrid_31-30-27745-0.html

Expansión. (19 de febrero de 2020). La CNMC investiga si Look & Find, Idealista y Remax inflaron con algoritmos el precio de la vivienda. *Expansión*. Obtenido de https://www.expansion.com/empresas/inmobiliario/2020/02/19/5e4d3cf7468aeb66618b4629.html

Ferre Jaén, M. E. (4 de abril de 2019). *FEIR 40: Modelos de Regresión.* Obtenido de http://gauss.inf.um.es/feir/40/

Grupo BIT. (2020). *Grupo BIT Business Analytics*. Obtenido de Grupo BIT Business Analytics: https://business-intelligence.grupobit.net/blog/cuantos-datos-se-producen-en-un-minuto

Idealista. (marzo de 2020). *Idealista*. Obtenido de Idealista: https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/alquiler/madrid-comunidad/madrid-provincia/madrid/

Itelligent. (8 de agosto de 2018). *Itelligent* . Obtenido de Itelligent : https://itelligent.es/es/que-es-business-analytics/

López, J. R. (2016). *Las viviendas que pudieron hundir la economía española. La caída del mercado de vivienda y sus consecuecias.* Madrid : Ediciones Complutense.

López-Rodríguez, D., & de los Llanos Mata , M. (2019). *Evolución reciente del mercado de alquiler de vivienda en España.* Boletín económico, Banco de España.

Martínez de Ibarreta, C., Álvarez Fernández, C., Budría Rodríguez, S., Curto González, T., & Escobar Torres, L. S. (2017). Modelos cuantitativos para la economía y la Empresa en 101 ejemplos. Madrid: EV Services.

Máxima formación. (2018). *Máxima formación*. Obtenido de Máxima formación: https://www.maximaformacion.es/blog-dat/analisis-de-correlacion-en-r/

MicroStrategy. (2020). *MicroStrategy*. Obtenido de MicroStrategy: https://www.microstrategy.com/us/resources/introductory-guides/business-analytics-everything-you-need-to-know

Powerdata. (s.f.). *Powerdata*. Obtenido de Powerdata: https://www.powerdata.es/big-data

R-bloggers. (19 de enero de 2020). *R-bloggers*. Obtenido de R-blogger: https://www.r-bloggers.com/how-to-remove-outliers-in-r/

RDocumentation. (s.f.). Obtenido de https://www.rdocumentation.org/packages/dplyr/versions/0.7.8

RDocumentation. (s.f.). Obtenido de https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/NA

Real Academia Española. (2019). *Real Academia Esapñola* . Obtenido de Real Academia Española : https://dle.rae.es/d%C3%BAplex?m=form

Real Acaedmia Española. (2019). *Real Academia Española*. Obtenido de Real Academia Española: https://dle.rae.es/%C3%A1tico%20?m=form

Rodrigo, J. A. (julio de 2016). *Introducción a la Regresión Lineal Múltiple*. Obtenido de https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple

Rodrigo, J. A. (agosto de 2016). *RPubs. Ejemplo práctico de regresión lineal simple, múltiple, polinomial e interacción entre predictores*. Obtenido de RPubs: https://rpubs.com/Joaquin_AR/254575

R-project. (s.f.). *R-project*. Obtenido de R-project: https://www.r-project.org/about.html

SAS. (2020). *SAS*. Obtenido de SAS: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Tragacete, M. (26 de septiembre de 2019). Los alquileres más caros de España están en Madrid, donde se disparan los precios un 15%. *20 minutos*. Obtenido de https://www.20minutos.es/noticia/3778329/0/alquiler-subio-15-madrid-6-cataluna-2018-fomento/

Trecet, J. (15 de marzo de 2018). *Finect* . Obtenido de FInect : https://www.finect.com/blogs/vivienda_e_inmobiliario/articulos/que-precio-vivienda-alquiler

Ucha, A. P. (2020). *Economipedia* . Obtenido de Economipedia : https://economipedia.com/definiciones/coeficiente-de-correlacion-lineal.html

7. Anexos

Anexo 1. Tabla de frecuencias de la variable planta

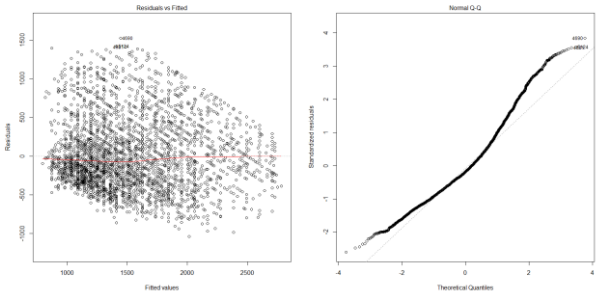
floor	freq
1	1430
2	1418
3	1287
4	1093
5	668
0	659
6	418
7	250
8	103
10	63
9	54
11	35
13	18
16	17
12	11
14	9
15	5
17	5
24	5
18	4
20	3
23	3
30	2
-2	1
-1	1
22	1
26	1
28	1
29	1
31	1
60	1

Anexo 2. Tabla de frecuenias de la variable tamaño (al haber 345 observaciones solo se ofrece una imagen de las superficies que se repiten más de 50 veces, ya que de lo contrario ocuparían demasiado espacio)

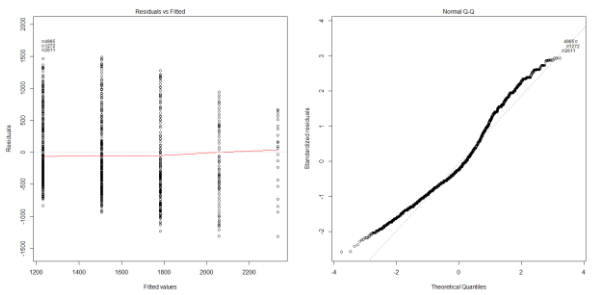
size	freq
60	305
50	292
70	286
80	265
120	243
100	234
90	222
65	203
75	194
55	180
110	173
45	164
85	157
140	137
130	119
150	119
40	103
95	95
200	79
160	78
115	75
170	69
105	68
125	64
72	55

Anexo 3. Representación gráfica de los residuos de los modelos de regresión simple

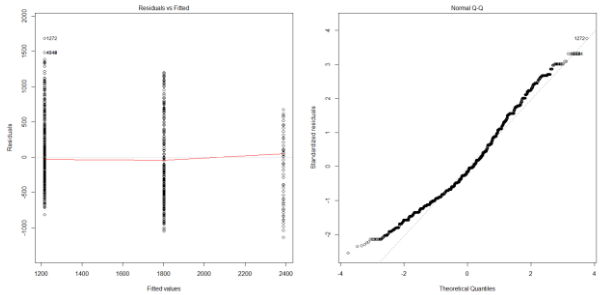
Precio – superficie



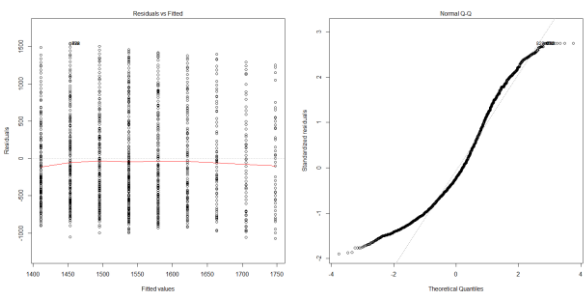
Precio – nº habitaciones



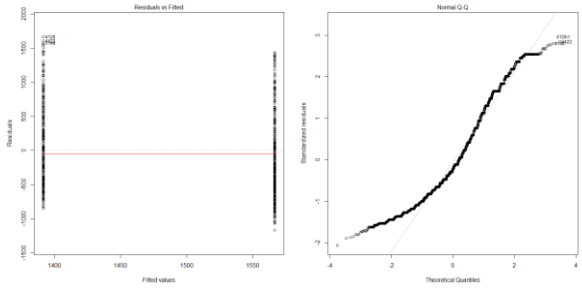
Precio – Nº de baños



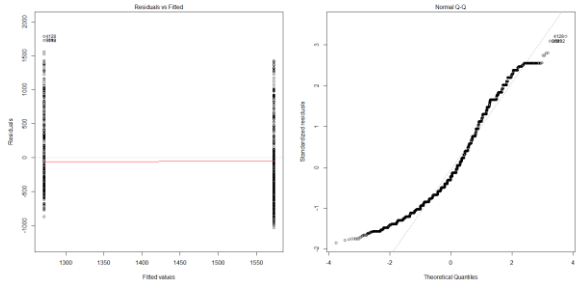
Precio - Planta



Precio - Ascensor



Precio - Exterior

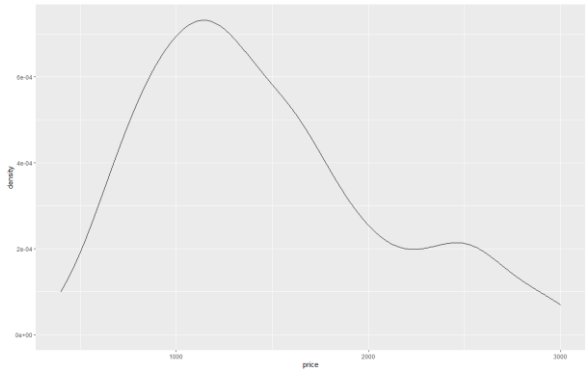


Anexo 4. Tablas de frecuencias y gráficas del modelo de regresión final

Anexo 4.1. Tabla de frecuencias de los municipios

district	freq
Centro	1123
Salamanca	938
Chamberí	582
Chamartín	494
Tetuán	301
Moncloa	270
Retiro	250
Arganzuela	248
Hortaleza	233
Fuencarral	205

Anexo 4.2. Gráfica de densidad de los precios de la vivienda de alquiler



Anexo 4.3 Tabla de frecuencias de la superficie de la vivienda de alquiler

size	freq
60	268
50	261
70	244
80	215
120	202
100	195
90	194
65	176
75	176

Anexo 4.4. Tablas de frecuencias del número de habitaciones y número de baños

rooms	freq
1	1657
2	2112
3	1322
4	343
5	38

bathrooms	freq
1	2886
2	2280
3	306

Anexo 5. Código de R

5.1. Limpieza de datos y análisis descriptivo

```
idealista<-data[,c(5,6,7,9,10,11,12,13,14,15,16,18,19,20,26,27,28,29)]

datos<-idealista[,c(8,11,12,3,2,4,16,1,5,6,7,13,14,15)]

datos[is.na(datos)] <- 0

install.packages("factoextra")

install.packages("sf")

install.packages("textshape")

install.packages("MASS")

install.packages("mapproj")

install.packages("PerformanceAnalytics")

install.packages("ISLR")

library(mapproj)

library(factoextra)

library(sf)

library(textshape)

library(MASS)

library(PerformanceAnalytics)

library(ISLR)

str(datos)
```

```
summary(datos)

###limpiar subcategorías###

ggplot(datos)+

 geom_bar(aes(x=propertyType))  ###podemos que ver que sigue habiendo mucha diferencia entre los sutipos de piso, por lo que podemos eliminar duplex, penthouse y studio; o hacer un estudio separado para estos subgrupos

type_freq<-datos%>% group_by(propertyType) %>% summarise(freq=n())

trabajo<-datos%>% filter(propertyType=="flat")

trabajo$floor<-as.numeric(trabajo$floor) ###los hacemos numéricos para que queden en orden de mayor a menor y así facilitar la limpieza de los más altos###

trabajo$exterior<-as.numeric(trabajo$exterior)

trabajo[is.na(trabajo)] <- 0

summary(trabajo$floor)

datos_pisos_rooms<-trabajo

###vamos a dividir los distritos en dos grupos, aquellos con más oferta y aquellos con menos, para centrar el estudio en los más ofertados###

###distritos con más oferta###

datos_pisos_rooms<-datos_pisos_rooms%>%filter(datos_pisos_rooms$district!=0)

district_freq<-datos_pisos_rooms%>% group_by(district) %>% summarise(freq=n())

summary(district_freq)

district_freq_most<-district_freq%>% filter(freq>50)

ggplot(district_freq_most, aes(x=district, y=freq))+  ###hay muchos distritos, por lo que podemos quitar aquellos en los que haya una frecuencia inferior a x###

 geom_col()+

 coord_flip()

###precio medio según idealista en estos distritos###

salamanca<-datos_pisos_room%>%filter(st$district=="Salamanca")

summary(salamanca)
```

```
ggplot(salamanca, aes(x=floor))+
  geom_bar()
ggplot(salamanca, aes(x=hasLift))+
  geom_bar()
ggplot(salamanca, aes(x=exterior))+
  geom_bar()
centro<-datos_pisos_rooms%>%filter(st$district=="Centro")
summary(centro)
ggplot(centro, aes(x=floor))+
  geom_bar()
ggplot(centro, aes(x=hasLift))+
  geom_bar()
ggplot(centro, aes(x=exterior))+
  geom_bar()
chamberi<-datos_pisos_rooms%>%filter(st$district=="Chamberí")
summary(chamberi)
ggplot(chamberi, aes(x=floor))+
  geom_bar()
ggplot(chamberi, aes(x=hasLift))+
  geom_bar()
ggplot(chamberi ,aes(x=exterior))+
  geom_bar()
chamartin<-datos_pisos_rooms%>%filter(st$district=="Chamartín")
summary(chamartin)
ggplot(chamartin, aes(x=floor))+
  geom_bar()
ggplot(chamartin, aes(x=hasLift))+
```

```
 geom_bar()

ggplot(chamartin, aes(x=exterior))+

 geom_bar()

tetuan<-datos_pisos_rooms%>%filter(st$district=="Tetuán")

summary(tetuan)

ggplot(tetuan, aes(x=floor))+

 geom_bar()

ggplot(tetuan, aes(x=hasLift))+

 geom_bar()

ggplot(tetuan, aes(x=exterior))+

 geom_bar()

###distritos con menos oferta###

district_freq_least<-district_freq%>% filter(freq<6)

ggplot(district_freq_least, aes(x=district, y=freq))+  ###hay muchos distritos, por lo que podemos quitar aquellos en los que haya una frecuencia inferior a x###

 geom_col()+

 coord_flip()

datos_pisos_rooms<-merge(datos_pisos_rooms,

        district_freq_most,

        by.x="district",

        by.y="district",

        all.x=TRUE)

datos_pisos_rooms[is.na(datos_pisos_rooms)]<-0

datos_pisos_rooms<-datos_pisos_rooms%>%filter(datos_pisos_rooms$freq>0

datos_pisos_rooms<-datos_pisos_rooms[,-c(2,4,12,13,15)]

###selección de lo más ofretado###

ggplot(datos_pisos_rooms,aes(x=district))+
```

```
  geom_bar()+
  coord_flip()
###miramos en que rango se concentran los precios mensuales del alquiler###
###de todas formas habría que revisarlo y aserguar para centrar el estudio###
ggplot(datos_pisos_rooms)+
  geom_freqpoly(aes(x=price))
summary(datos_pisos_rooms$price)
price_freq<-datos_pisos_rooms%>% group_by(price) %>% summarise(freq=n())
ggplot(price)+
  geom_jitter(aes(x=freq, y=price))+
  coord_flip()
datos_pisos_rooms<-datos_pisos_rooms%>%filter(datos_pisos_rooms$price <13000)
###limpiar piso###
###Replace###con esto sustituimos las iniciales de bj(bajo) por 0 ####
ggplot(datos_pisos_rooms)+
  geom_bar(aes(x=floor))
summary(datos_pisos_rooms$floor)
floor_freq<-datos_pisos_rooms%>% group_by(floor) %>% summarise(freq=n())
datos_pisos_rooms<-datos_pisos_rooms%>% filter(floor<12)
datos_pisos_rooms<-datos_pisos_rooms%>%filter(floor>-1)
ggplot(datos_pisos_rooms)+
  geom_bar(aes(x=floor))
###ahora vamos a analizar el número de habitaciones###
ggplot(datos_pisos_rooms)+
  geom_bar(aes(x=rooms))  ###en el gráfico podemos ver que hay habitaciones con 0 habitaciones, y hay muy pocas a partir de 5 habitaciones, por lo que podemos eliminar los outliers###
                ###hay que adpatar la gráfica para que aparezcan núneros enteros, no decimales###
```

```
summary(datos_pisos_rooms$rooms)

room_freq<-datos_pisos_rooms%>% group_by(rooms) %>% summarise(freq=n())

datos_pisos_rooms<-datos_pisos_rooms %>% filter(rooms<7)

ggplot(datos_pisos_rooms)+

 geom_bar(aes(x=rooms))  ###gráfico con el número de habitaciones###

###analizar el número de baños###

ggplot(datos_pisos_rooms)+

 geom_bar(aes(x=bathrooms))

summary(datos_pisos_rooms$bathrooms)

bathroom_freq<-datos_pisos_rooms%>% group_by(bathrooms) %>% summarise(freq=n())

###anaálisis de la superficie de las viviendas###

summary(datos_pisos_rooms$size)

size_freq<-datos_pisos_rooms%>% group_by(size) %>% summarise(freq=n())

size<-arrange(size_freq,freq)

summary(size)

ggplot(size)+

 geom_jitter(aes(x=freq,y=size))+

 coord_flip()

###barrios###

neighborhood_freq<-datos_pisos_rooms%>% group_by(neighborhood) %>% summarise(freq=n())

###análisis de la variable exterio y ascensor###

s<-ggplot(datos_pisos_rooms)+

 geom_bar(aes(x=exterior))

exterior_freq<-datos_pisos_rooms%>% group_by(exterior) %>% summarise(freq=n())

t<-ggplot(datos_pisos_rooms)+

 geom_bar(aes(x=hasLift))

ggarrange(s,t)
```

5.2. Identificación de valores atípicos

```
datos<-datos_pisos_rooms

install.packages("DMwR")

library(DMwR)

outlier.scores <- lofactor(datos, k=20)

outliers <- order(outlier.scores, decreasing = T)[1:10]

print(datos[outliers,])

plot(density(outlier.scores))

outlier.scores

summary(datos)

outliers

n<-nrow(datos)

labels<-1:n

labels[-outliers]<-"."

biplot(prcomp(datos))

pch<-rep(".",n)

pch[outliers]<-"+"

col<-rep("black",n)

col[outliers]<-"red"

pairs(datos,pch=pch,col=col)

boxplot(datos$price)$out

Q<-quantile(datos$price,probs=c(.25,.75))

iqr<-IQR(datos$price)

up<-Q[2]+1.5*iqr

low<-Q[1]-1.5*iqr
```

```
datos_sin<- subset(datos, datos$price > (Q[1] - 1.5*iqr) & datos$price < (Q[2]+1.5*iqr))


###outliers del precio###

a <- which(datos_sin$price %in% boxplot.stats(datos_sin$price)$out) ###valores de price que están fuera del boxplot###

a

print(datos_sin[a,]

boxplot(datos_sin$price)

datos_sin<-datos_sin%>%filter(datos_sin$price<3100)

###outliers de la superficie###

b <- which(datos_sin$size %in% boxplot.stats(datos_sin$size)$out)

b

print(datos_sin[b,])

summary(datos_sin)

datos_sin<-datos_sin[-
c(389,562,661,743,784,813,831,833,851,884,920,961,979,987,1062,1065,1121,1148,1233,126
0,
1261,1326,1391,1405,1613,1620,1754,1770,1772,1834,1902,1917,1918,1932,1961,1992,2015
,2060,2066,2080,2083,2086,2088,2095,2102,2107,2118,2122,2126,2128,2136,2176,2190,220
4,2209,2219,2229,2259,2270,2278,
2303,2329,2332,2354,2357,2368,2375,2406,2437,2440,2453,2466,2475,2490,2546,2554,2576
,2604,2605,2631,
2673,2674,2680,2688,2695,2709,2732,2738,2759,2820,2830,2831,2834,2857,2859,2871,2873
,2890,2908,2928,
2952,3001,3008,3020,3042,3074,3123,3157,3159,3174,3182,3192,3288,3291,3319,3478,3483
,3517,3557,3578
3603,3611,3640,3676,3725,3762,3785,3799,3815,4007,4038,4150,4158,4169,4204,4237,4243
,4244,4277,4365,
4377,4399,4412,4415,4533,4589,4605,4644,4650,4697,4709,4714,4731,4768,4792,4846,4853
```

```
,4856,4903,4918
4933,4952,4999,5003,5008,5015,5026,5034,5042,5095,5107,5111,5123,5159,5171,5232,5276
,5296,5325,5326,
5403,5412,5426,5448,5460,5521,5524,5559,5571,5584,5618,5633,5635,5646,5660,5668,5709
,5722,5728,5744,
5764,5780,5790,5802,5803,5816,5853,5957,6029,6070,6074,6206,6209,6210,6218,6264,6267
,6300,6322,6358,6438,6439,6446,6450,6523,6534
,6543,6550,6563,6564,6565,6566,6570,6571,6572,6573,6579,6581,6583,6586),

###outliers de las habitaciones###

c<- which(datos_sin$rooms %in% boxplot.stats(datos_sin$rooms)$out)

c

###outliers de los baños###

d<-which(datos_sin$bathrooms %in% boxplot.stats(datos_sin$bathrooms)$out)

d

print(datos_sin[d,])

datos_sin<-datos_sin%>%filter(datos_sin$bathrooms<4)

###outliers de planta###

e<-which(datos_sin$floor %in% boxplot.stats(datos_sin$floor)$out)

e

print(datos_sin[e,])

datos_sin<-datos_sin%>%filter(datos_sin$floor<9)

###outliers de la intersección tamaño y baños###

outlier_intersect <- intersect(a,d)

outlier_intersect

print(datos[outlier_intersect,])

outlier_intersect <- intersect(b,d)

outlier_intersect
```

```
print(datos[outlier_intersect,])

###quitar outliers###

a <- which(datos_sin$price %in% boxplot.stats(datos_sin$price)$out) ###valores de price que están fuera del boxplot###

a

print(datos_sin[a,])

b <- which(datos_sin$size %in% boxplot.stats(datos_sin$size)$out)

b

print(datos_sin[b,])

c<- which(datos_sin$rooms %in% boxplot.stats(datos_sin$rooms)$out)

c

print(datos_sin[c,])

d<-which(datos_sin$bathrooms %in% boxplot.stats(datos_sin$bathrooms)$out)

d

print(datos[d,])
```

5.3. Modelo simple

```
install.packages("psych")

install.packages("ggstatsplot")

require(psych)

cor<-datos_sin

###correlación de Pearson###

cor(datos$price, datos$size, method = "pearson")

cor(datos$price, datos$rooms, method = "pearson")

cor(datos$price, datos$floor, method = "pearson")

cor(datos$price, datos$exterior, method = "pearson")
```

```
cor(datos$price, datos$bathrooms, method = "pearson")

cor(datos$price, datos$hasLift, method = "pearson")

### significación estadística###

cor.test(datos$price, datos$size, method = "pearson")

cor.test(datos$price, datos$rooms, method = "pearson")

cor.test(datos$price, datos$bathrooms, method = "pearson")

cor.test(datos$price, datos$floor, method = "pearson")

cor.test(datos$price, datos$exterior, method = "pearson")

cor.test(datos$price, datos$hasLift, method = "pearson")

seis<- data.frame(datos_sin$price, datos_sin$size, datos_sin$rooms, datos_sin$bathrooms,
datos_sin$exterior, datos_sin$hasLift, datos_sin$floor)

corr.test(seis, use = "complete", method = "pearson")

multi.hist(x = cor[,3:4], dcol = c("blue","red"), dlty = c("dotted", "solid"),
        main = "")

multi.hist(x = cor[,5:7], dcol = c("blue","red"), dlty = c("dotted", "solid"),
        main = "")

### regresión lineal simple ###

###size###

modelo_simple<-lm(data=datos_sin,formula=price~size)

names(modelo_simple)

summary(modelo_simple)

###ecuación del tamaño### Y = 579.347 + 10.104X

###gráfico de correlación###

plot(datos_sin$size, datos_sin$price, xlab='size', ylab='price')

abline(modelo_simple, lwd = 3, col = "red")
```

```
  ###diagnóstico del modelo###
residuos<-rstandard(modelo_simple)
valores.ajustados<-fitted(modelo_simple)
plot(valores.ajustados,residuos)
qqnorm(residuos)
qqline(residuos)
par(mfrow = c(1,2))
plot(modelo_simple)
par(mfrow = c(2,2))
###rooms###
modelo_simple<-lm(data=datos_sin,formula=price~rooms)
names(modelo_simple)
summary(modelo_simple)
###la ecuación de la recta de mínimos cuadrados quedaría: Y = 1089.70 + 189.29X###
plot(datos_sin$rooms, datos_sin$price, xlab='rooms', ylab='price')
abline(modelo_simple,lwd = 3, col = "red")
###diagnóstico del modelo###
residuos<-rstandard(modelo_simple)
valores.ajustados<-fitted(modelo_simple)
plot(valores.ajustados,residuos)
qqnorm(residuos)
qqline(residuos)
par(mfrow = c(1,2))
plot(modelo_simple)
```

```
par(mfrow = c(2,2))

###bathrooms###

modelo_simple<-lm(data=datos_sin,formula=price~bathrooms)

names(modelo_simple)

summary(modelo_simple)

###la ecuación de la recta de mínimos cuadrados quedaría: Y=674.99 + 530.64X###

plot(datos_sin$bathrooms, datos_sin$price, xlab='bathrooms', ylab='price')

abline(modelo_simple,lwd = 3, col = "red")

###diagnóstico del modelo###

residuos<-rstandard(modelo_simple)

valores.ajustados<-fitted(modelo_simple)

plot(valores.ajustados,residuos)

qqnorm(residuos)

qqline(residuos)

par(mfrow = c(1,2))

plot(modelo_simple)

par(mfrow = c(2,2))

###planta###

modelo_simple<-lm(data=datos_sin,formula=price~floor)

names(modelo_simple)

summary(modelo_simple)

###Ecuacion: Y = 1348.783 + 54.908X

plot(datos_sin$floor,datos_sin$price, xlab='floor', ylab='price')

abline(modelo_simple,lwd = 3, col = "red")
```

```
###diagnóstico del modelo###

residuos<-rstandard(modelo_simple)

valores.ajustados<-fitted(modelo_simple)

plot(valores.ajustados,residuos)

qqnorm(residuos)

qqline(residuos)

par(mfrow = c(1,2))

plot(modelo_simple)

par(mfrow = c(2,2))

###exterior###

###planta###

modelo_simple<-lm(data=datos_sin,formula=price~exterior)

names(modelo_simple)

summary(modelo_simple)

###Ecuacion: Y = 1348.783 + 54.908X

plot(datos_sin$exterior,datos_sin$price, xlab='exterior', ylab='price')

abline(modelo_simple,lwd = 3, col = "red")

###diagnóstico del modelo###

residuos<-rstandard(modelo_simple)

valores.ajustados<-fitted(modelo_simple)

plot(valores.ajustados,residuos)

qqnorm(residuos)

qqline(residuos)

par(mfrow = c(1,2))
```

```
plot(modelo_simple)

par(mfrow = c(2,2))

###ascensor###

###planta###

modelo_simple<-lm(data=datos_sin,formula=price~hasLift)

names(modelo_simple)

summary(modelo_simple)

###Ecuacion: Y = 1348.783 + 54.908X

plot(datos_sin$hasLift,datos_sin$price, xlab='floor', ylab='price')

abline(modelo_simple,lwd = 3, col = "red")

###diagnóstico del modelo###

residuos<-rstandard(modelo_simple)

valores.ajustados<-fitted(modelo_simple)

plot(valores.ajustados,residuos)

qqnorm(residuos)

qqline(residuos)

par(mfrow = c(1,2))

plot(modelo_simple)

par(mfrow = c(2,2))
```

5.4. Modelo múltiple

```
###analizar relación entre variables###

multi.hist(x = datos_sin, dcol = c("blue", "red"), dlty = c("dotted", "solid"),

         main = "")

ggpairs(datos_sin, lower = list(continuous = "smooth"),

       diag = list(continuous = "bar"), axisLabels = "none")
```

```
###modelo###
modelo <- lm(price ~ size + rooms + bathrooms + floor + exterior+hasLift + district
          , data = datos_sin)
summary(modelo)
###intervalo de confianza###
confint(lm(formula=price ~ size + rooms + bathrooms+floor + district + hasLift+exterior
                , data = datos_sin))
###validación###
plot1 <- ggplot(data = datos_sin, aes(size, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot2 <- ggplot(data = datos_sin, aes(rooms, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot3 <- ggplot(data = datos_sin, aes(bathrooms, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot4 <- ggplot(data = datos_sin, aes(floor, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot5 <- ggplot(data = datos_sin, aes(district, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot6 <- ggplot(data = datos_sin, aes(hasLift, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
```

```
  theme_bw()
plot7 <- ggplot(data = datos_sin, aes(exterior, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
grid.arrange(plot1, plot2, plot3, plot4)
grid.arrange(plot5,plot6,plot7)
qqnorm(modelo$residuals)
qqline(modelo$residuals)
shapiro.test(modelo$residuals)
ggplot(data = datos_sin, aes(modelo$fitted.values, modelo$residuals)) +
  geom_point() +
  geom_smooth(color = "firebrick", se = FALSE) +
  geom_hline(yintercept = 0) +
  theme_bw()
bptest(modelo)
ggarrange(s,t)
corrplot(cor(dplyr::select(datos_sin, price, size, floor, rooms, bathrooms, hasLift, exterior)),
       method = "number", tl.col = "black")
library(car)
vif(modelo)
dwt(modelo, alternative = "two.sided")
datos_sin$studentized_residual <- rstudent(modelo)
ggplot(data = datos_sin, aes(x = predict(modelo), y = abs(studentized_residual))) +
  geom_hline(yintercept = 2, color = "grey", linetype = "dashed") +
  geom_point(aes(color = ifelse(abs(studentized_residual) > 2, 'red', 'black'))) +
```

```
  scale_color_identity() +
  labs(title = "Distribución de los residuos studentized",
       x = "predicción modelo") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
which(abs(datos_sin$studentized_residual) > 2)
print(datos_sin[r,])
summary(datos_sin)
```