PROYECTO FINAL

ESPECIALIDAD BIG DATA

BOOTCAMP MUJERES EN TECH

KEEPCODING & GLOVO (2022/23)

KEEPCODING
Tech School

Glovo

Autoras:

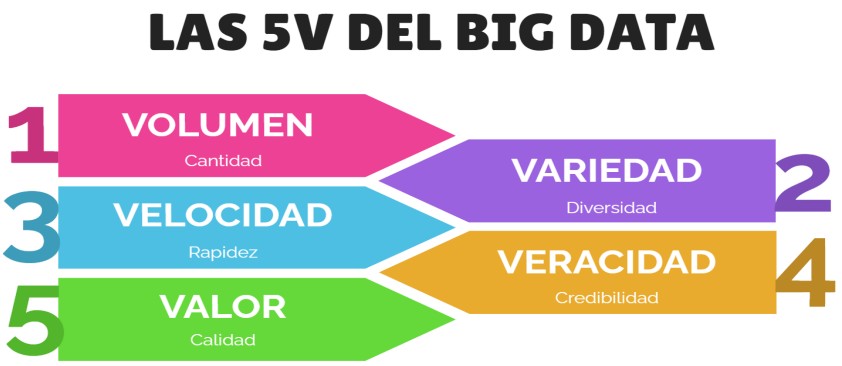
Elsa Cembrero Bonet
Diana Maria Toro López
María de Lluch Gual Pérez
Mireia Hernández Lozano
Sanja Aleksova
Sandra Julieth Castaño Reina

ÍNDICE

1. Introducción

El crecimiento en el volumen de datos generados por diferentes sistemas y actividades cotidianas en la sociedad han forjado la necesidad de modificar, optimizar, generar métodos y modelos de almacenamiento y tratamiento de datos que suplan las bases de datos y los sistemas de gestión de datos tradicionales.

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos – estructurados y no estructurados – cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización.



La utilidad de estos datos masivos es que pueden ofrecer soluciones a los problemas empresariales que antes no se hubieran podido resolver.

También podemos clasificar los datos según:

Es importante tener en cuenta algunos pasos para la implementación de Big Data, tales como:

- Entender el negocio y los datos.

- Determinar los problemas y cómo los datos pueden ayudar a resolverlos y evitarlos.

- Determinar la veracidad e integridad de los datos. Las fuentes de datos que usemos siempre deben ser fiables, veraces y estar actualizadas.

- Determinar si existe la necesidad de incorporar fuentes de datos externas que complementen las existentes en la empresa.

- La implementación del análisis.

- La ejecución del plan. Este procedimiento requiere que revisemos, analicemos y corrijamos aquellos parámetros que no sean correctos en cada caso. Este es un proceso activo que empieza cuando comienza el análisis.

- La distribución de la información obtenida. Una vez realizado el análisis debemos compartir la información con el departamento responsable para poner en marcha las medidas oportunas.

a) Presentación del equipo

Data Women's Company somos un grupo de mujeres que compartimos la misma pasión por el sector tecnológico y los datos. Cada una de nosotras aportamos nuestro granito de arena y como dice la expresión, cada pequeña acción, sumada a otras pequeñas acciones, da lugar a un conjunto mayor. Esto aplicado al grupo da lugar a un equipo extraordinario, donde constantemente nos retroalimentamos de conocimientos las unas de las

otras.

Diana es una mujer sumamente perseverante y audaz, cualificada en administración de base de datos, sistemas de información y desarrollo de software.
Lluch es una experta en sistemas de información geográfica, análisis espacial de datos y Data Science. Por si esto fuera poco, es una experta en estadística y probabilidades.
Sanja es especialista en lingüística computacional y una amante de los retos intelectuales. También es una erudita en chatbots y machine learning.
Sandra Julieth, o para nosotras July, es licenciada en sociología y domina Python. Su psique es increíblemente observadora y analítica.
Elsa está titulada en lingüismo y como apasionada de las letras es políglota. También ha trabajado como profesora y es altamente resolutiva.

Mireia está cualificada en administración y dirección de empresas, finanzas y trading. Es una mujer creativa y muy polivalente.

Funciones desempeñadas:

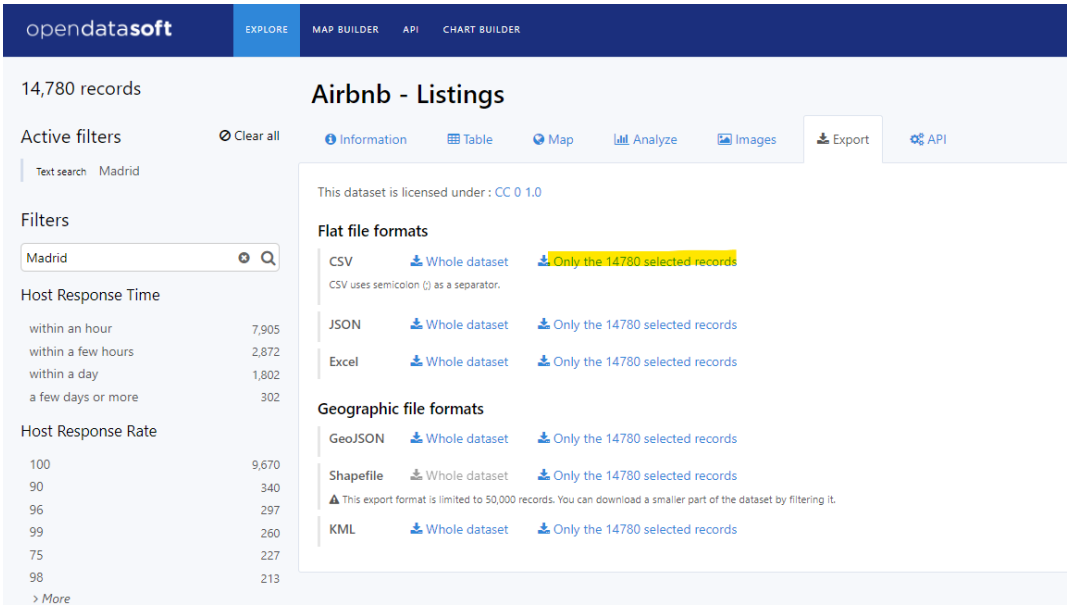
- Diana Maria Toro López : Análisis exploratorio inicial, limpieza de datos y Datawarehouse. Memoria y presentación.
- María de Lluch Gual Pérez : Análisis exploratorio en detalle, pre-procesamiento y modelado. Memoria, presentación e imagen corporativa.
- Sanja Aleksova : Visualización de métricas en Tableau y análisis de datos. Memoria y presentación.
- Sandra Julieth Castaño Reina : Análisis exploratorio inicial, limpieza de datos y Datawarehouse. Memoria y presentación.
- Elsa Cembrero Bonet : Visualización de métricas en Tableau y análisis de datos. Memoria y presentación.
- Mireia Hernández Lozano : Visualización de métricas en Tableau y análisis de datos. Memoria y presentación.

b) Objetivos

- Estudiar y poner en práctica los métodos y tecnologías aprendidas durante el Bootcamp Mujeres en Tech, organizado por Glovo y KeepCoding, para el tratamiento de datos no estructurados y que puedan ser aplicables a la información obtenida del scrapeado de Airbnb.
- Evaluar y seleccionar las técnicas de filtrado y de análisis de datos, aplicables a la información proveniente del dataset.
- Filtrar y analizar la información obtenida del dataset, haciendo uso de las metodologías adquiridas.
- Elaborar un dashboard para visualizar dicha información.
- Extraer conclusiones de valor.
- Plantear líneas futuras para ofrecerle una continuidad al proyecto.

2. Definiendo el dataset

El dataset seleccionado es un scrapeado de Airbnb y podemos acceder a él mediante este enlace.



Consta de 14780 registros correspondientes a los alojamientos registrados en la ciudad de Madrid en 2017. Descargamos el CSV y observamos las métricas que contiene para familiarizarnos con el dataset.

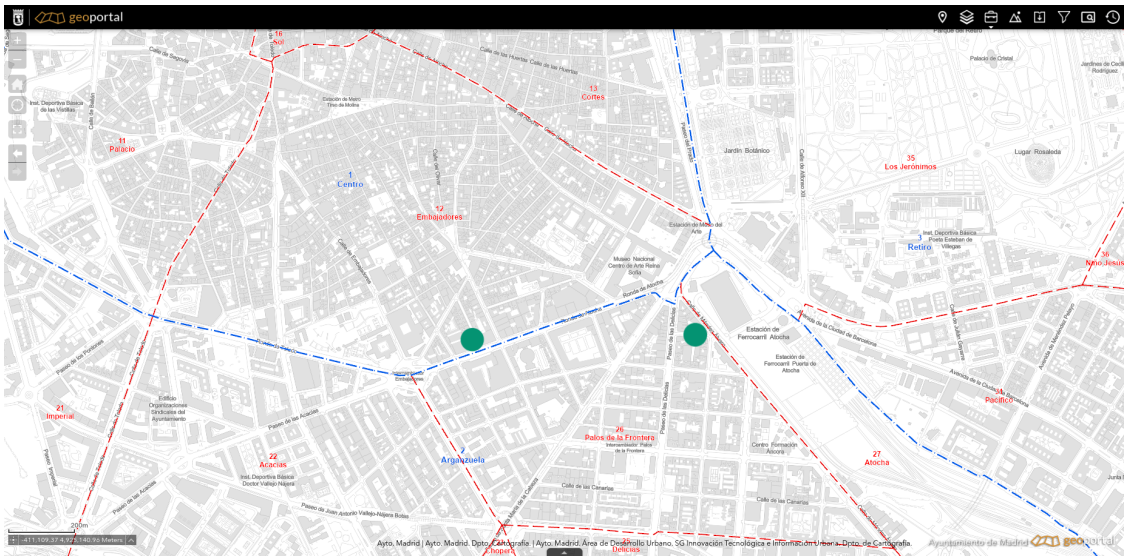
Durante esta etapa, tras observar los datos, organizamos las tareas del proyecto y las repartimos a cada integrante del equipo de Data Women's Company atendiendo a los puntos fuertes de cada una de ellas en el ciclo de vida del dato de Big Data y Data Science.

También planteamos la pregunta que debemos resolver en las siguientes etapas del proyecto, en las que: validaremos los datos, se definirá una estructura e implementará

Datawarehouse, se realizará una limpieza y un análisis exploratorio inicial de datos con Python y uno en profundidad con R donde seleccionaremos los datos de interés y se determinará outliers. Posteriormente se realizarán regresiones lineales mediante R y visualización de datos en Tableau para analizar los datos y obtener conclusiones en relación a la pregunta.

3. El servicio de consultoría

Data Women's Company recibe el encargo de un cliente. Acaba de heredar dos apartamentos de características similares ubicados en Madrid. Uno de ellos en la Calle de Amparo y el otro en la Calle de Méndez Álvaro.



Desea incorporar los apartamentos a la plataforma digital Airbnb, dedicada a la oferta de alojamientos particulares y turísticos mediante la cual los anfitriones pueden publicitar y contratar el arriendo de sus propiedades con sus huéspedes, para obtener rentabilidad a sus nuevas propiedades. Nos dice que las dos propiedades están muy bien comunicadas y cerca de la Estación Madrid-Puerta de Atocha.

Solicita que le indiquemos **a qué precio base debería publicar** cada uno de los alojamientos teniendo en cuenta las características de éstos. Ambos son apartamentos que serán rentados de manera íntegra. Tienen *dos habitaciones*, con *una cama matrimonial cada una*, y *un baño*. Además, los apartamentos cuentan con cocina totalmente equipada y sala de estar con sofá-cama donde podrían dormir hasta *dos huéspedes extra*. Nos comenta que se podrían alojar hasta *6 huéspedes*, pero sólo quiere *incluir en el precio base a un huésped* e incrementará el precio posteriormente en relación a la cantidad de huéspedes que reserven. Tampoco quiere que se incluya la tasa de limpieza ni el depósito por daños.

También desea conocer **cómo puede variar el precio según la cantidad de reviews que tenga** y en qué grado es recomendable que se implique en la gestión de los alojamientos. ¿**Es rentable ser superanfitrión**?

4. Arquitectura, validación de los datos y análisis exploratorio inicial

Lenguaje de programación aplicado: Python.

a) Primeros pasos

Para empezar el proyecto, lo primero que se debe hacer es analizar el dataset proporcionado, revisando el tipo de datos que hay y la calidad de estos mismos. Es importante ver también qué columnas del dataset son necesarias y relevantes, para así poder focalizar los datos y crear la base de datos necesaria para conseguir el objetivo de responder a la pregunta que hemos planteado anteriormente.

Una vez se tenga claro qué columnas son las necesarias, el siguiente paso es hacer el modelo de la base de datos. Inicialmente se hace una sola tabla con todas las columnas que se vayan a usar y el tipo de dato va a ser el dato inicial que viene ya con el dataset.

La base de datos ha sido creada en PostgreSQL con ayuda de Jupyter Notebook para hacer las conexiones con las bases de datos, para que no exista posibilidad de que ningún dato se quede perdido y poder cargarlos sin problema para su correcta visualización; apoyado por Talend para la una mejor integración de todos los datos y, por tanto, el uso de ETL para la extracción, transformación y carga de los datos en un solo repositorio; posteriormente esta herramienta ha facilitado a su vez, la creación del proceso Data Warehouse.

En esta fase, como se comenta anteriormente, se introduce la librería que se va a usar para la conexión con la base de datos, se importa la librería de Pandas, se introduce la base de datos, se eliminan las columnas que no se vayan a utilizar en el desarrollo del trabajo y se establece la conexión con la base de datos.

Después de traer la información del dataset y agregarla a la base de datos, para verificar que esa información sí está en la base de datos, se hace

una desconexión para que no se generen errores. En PostgreSQL se puede verificar que el proceso de mover todos los datos se da sin ningún problema. Todo esto que se explica anteriormente se emplea para tomar dos caminos de desarrollo: por una parte se emplea para realizar el Data Warehouse y por otra parte para realizar el análisis de los datos.

Un paso para organizar, comprender mejor y acudir a la información seleccionada de manera sencilla, es crear un diccionario con la explicación de las columnas y comprobar cuál es el tipo de dato contiene cada una de estas.

b) Limpieza del dataset y base de datos

Para realizar el análisis y la exploración principal de los datos, se usa en este proyecto un cuaderno de Colab; para ello, se va a introducir la información del dataset inicial y las columnas que ya han sido elegidas en el paso anterior. A continuación se instalan todas las librerías que se requieran (mapas, geolocalización, Pandas, NumPy…).

En este paso se puede visualizar cómo son los datos, si son o no nulos, en el caso de que sean nulos, muchos de ellos se pueden eliminar puesto que podrían llevar a confusión para el posterior análisis. También existe la opción de imputar los datos, sobreescribiendo el dato por un valor que se considere probable para esa variable; para este paso se usa el dato que sea probabilísticamente mayor, en este como es el caso de Monthly Price, Weekly Price, Review Scores Value, Host, Zip Code, etc.

De las 88 columnas que eran, se eligen 36 columnas. Una vez que ya se ha seleccionado y acotado la información de las columnas, se revisa y se vuelve a comprobar para poder seguir editando pequeños matices, como por ejemplo, pasar todas las variables categóricas a letras minúsculas. Aún se pueden hacer más acotaciones, en este caso la ciudad objeto de estudio es Madrid, por lo tanto se filtran los datos que sean sólo de Madrid, ya que el dataset tiene información de otras ciudades nacionales e internacionales.

- Almacenamiento de la base de datos

A la hora de hacer el almacenamiento de los datos surgió un problema, la idea era usar el Data Warehouse para esta parte como se menciona anteriormente en los primeros pasos; sin embargo, no ha sido posible. Se intentó hacer el almacén de los datos con el dataset inicial, pero en cuanto se hacía la prueba de quitar las columnas que no habían sido seleccionadas, el dataset generaba error. Por tanto, se tomó la decisión de recurrir a SQL, en donde se hizo un job sencillo en el que cargamos la tabla inicial donde están los datos, esa tabla inicial la alimentamos desde el Jupyter Notebook. Los datos vienen de Lectura de datos, pasan por el tMap y llegan a la tabla ODS, se pasan los 14 780 datos y es así como se puede ver y comprobar que los datos - tanto en la base de datos original como en el respaldo - quedan grabados finalmente en SQL.

c) Análisis inicial de variables

- Análisis de las variables categóricas

En el análisis de los datos categóricos se le da especial importancia a las columnas *host* y *superhost*, en donde *host* se convierte en *false* y *superhost* en *true*, a continuación se revierte el proceso: se reemplaza *true* por *superhost* y *false* por *host*. Este paso se realiza para evitar confusiones y para que se pueda hacer el cambio automático cuando busquemos los datos para el análisis posterior. Para R se convertirá en *0* para *Host* y *1* para *Superhost*, para facilitar su uso en la regresión lineal múltiple.

Se analiza la columna *Neighbourhood Group Cleansed* con el fin de crear después un diccionario, para que los valores sean únicos y tengan su código postal correspondiente. Se agrupan *Neighbourhood Cleansed* con *Zipcode* y se trabaja con esos valores. Luego, se van filtrando los errores y valores nulos hasta que ya queden limpias las columnas y los datos de estas sean claros y correctos, para poder trabajar de una manera más cómoda y ordenada con ellos.

- **Análisis de las variables numéricas**

Se lleva a cabo un procedimiento parecido al de las variables categóricas, en dónde se va acotando y filtrando los valores erróneos y nulos hasta conseguir la limpieza total de los datos. Se da especial importancia al precio de las rentas de los alojamientos.

La detección de los datos atípicos juega un papel muy importante en la toma de decisiones, puesto que distorsionan los resultados al cambiar el comportamiento del resto de los casos, afectando las medidas estadísticas utilizadas para representar a la muestra elegida (p.ej. el promedio) y porque afectan considerablemente a otras técnicas de análisis de datos, cuya base es la presencia de normalidad.

En ambos casos, en el análisis de las variables, aunque es una práctica habitual el tomar datos similares calculados con los métodos estadísticos robustos, para este proyecto en particular ha resultado más lógico prescindir de datos que no fueran a resultar importantes, así como limpiar con exhaustividad el dataset proporcionado, ya que podría dar margen a errores de veracidad y precisión de los resultados finales.

También se ha limpiado rigurosamente el dataset y la información de sus columnas para poder visualizar los diagramas de barras generados con mayor claridad, ya que por el aspecto de estos en una fase inicial, habría sido complicado analizar la información; puesto que es prácticamente imposible ver los datos que reflejan, como es en el caso del diagrama de geolocalización y/o funciones.

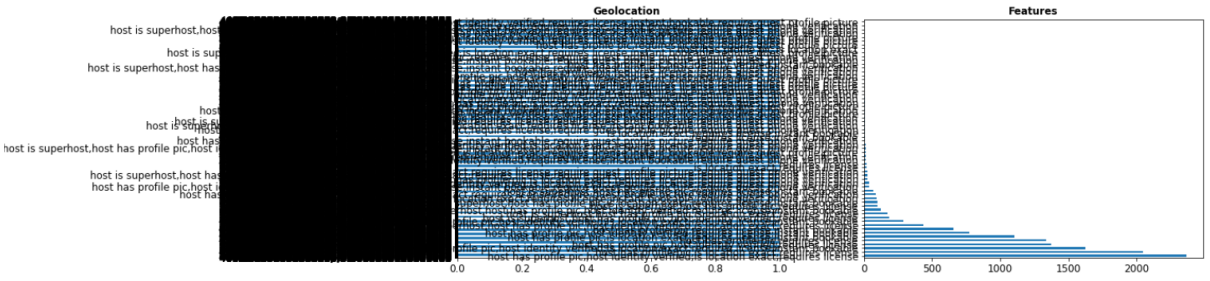


Diagrama de barras generado con Colab a partir del dataset de Airbnb. Diagrama_1_Geolocalización. Diagrama_2_Funciones.

d) Conclusiones

En todo este proceso han ido surgiendo más preguntas, además de las planteadas a raíz del análisis de datos, habiendo marcado un objetivo principal, preguntas que se han ido contestando a lo largo de la realización del proyecto, para mayor calidad y satisfacción del cliente que ha requerido nuestros servicios.

Las conclusiones que podemos obtener de esta limpieza y corrección de los datos son:

- Podemos encontrar una diferencia notable entre la cantidad de *superhost* y *host* que existen, registrándose así 1554 tipos de usuario *superhost* con el 11.72% de los alojamientos y 11710 tipos de usuario tipo *host* con el 88.28% los alojamientos. Los Host tienen ratios de respuesta más bajos que los Superhosts.

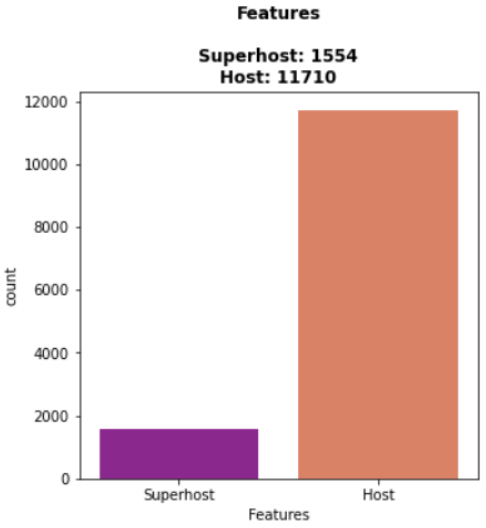


Diagrama de barras generado a partir de la limpieza del dataset de Airbnb acerca de la cantidad de *host* y *superhost* que hay en Madrid.

- Según este análisis de los datos, el precio del alquiler depende de la zona donde se encuentre el *host*; la media del costo por alquiler tiende a ser más alta en Fuentelarreina y más baja en Pueblo Nuevo.

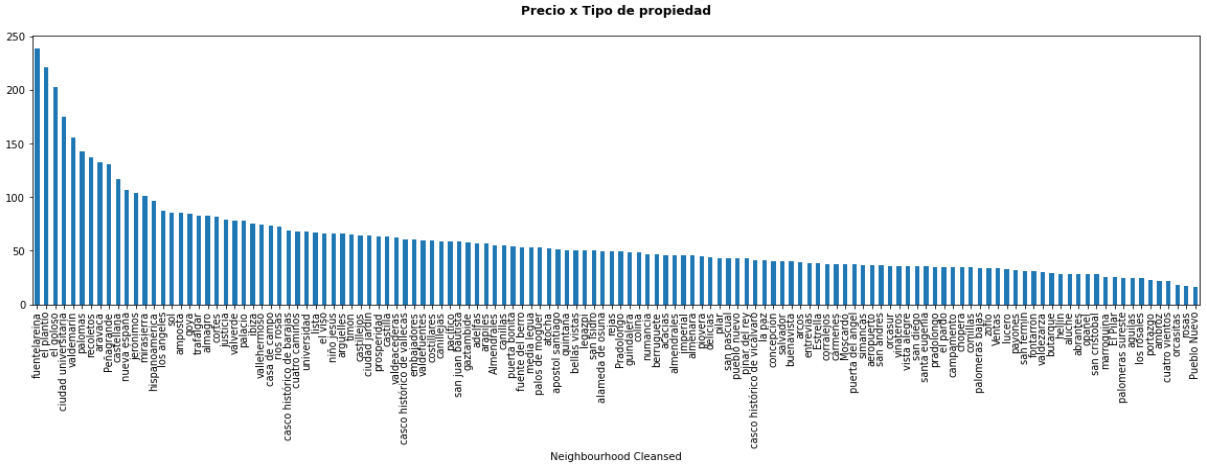


Diagrama de barras generado a partir de la limpieza del dataset de Airbnb acerca del tipo de propiedad que hay en Madrid.

- La cantidad de alojamientos es más alta en el distrito Centro de Madrid con 6735 *Neighbourhood Group Cleansed* y la más baja es en el Distrito Vicálvaro 34 *Neighbourhood Group Cleansed.*

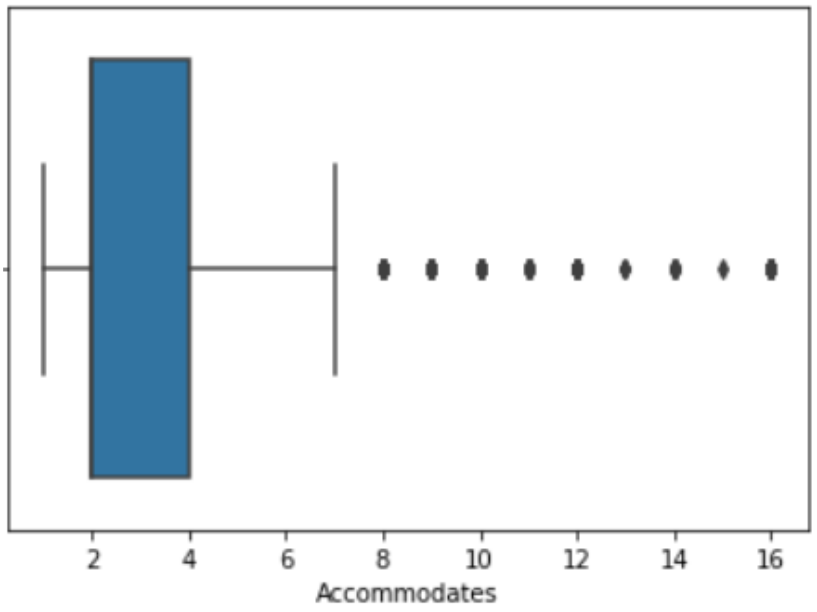
```
========================================
Distrito       ||    Cantidad Alojamientos
========================================

centro                  6735
chamberí                 957
salamanca                876
arganzuela               791
tetuán                   461
moncloa - aravaca        447
retiro                   436
latina                   378
chamartín                359
carabanchel              358
ciudad lineal            309
puente de vallecas       220
hortaleza                186
fuencarral - el pardo    170
usera                    145
san blas - canillejas    118
villaverde                83
barajas                   81
moratalaz                 74
villa de vallecas         46
vicálvaro                 34
Name: Neighbourhood Group Cleansed, dtype: int64
```

Cuadro extraído de Colab del dataset de Airbnb acerca la cantidad de alojamientos que hay en los distintos distritos de Madrid

- El tipo de alojamiento más común es el apartamento con 10910. Se evidencia una gran diferencia con el resto de tipos de

alojamiento siendo el segundo mayor requerido con un número de 999, las casas.

- En la variable *Accommodates*, se evidencian valores atípicos para las acomodaciones de más de 6 personas en adelante, las habitaciones más comunes son *entire home/apt* con una cantidad de 7875, la cantidad de habitaciones más común es una por vivienda y la menos común es 8, presentándose valores atípicos.



Cuadro extraído de Colab del dataset de Airbnb acerca de los valores atípicos de la variable Accommodates.

- Las habitaciones menos comunes son *shared room* con una cantidad de 192; la cantidad de camas más común es 1 por vivienda y la menos común es de 15 por vivienda. La cantidad de baños más común es 1 y la menos común es 7 baños, la cantidad de huéspedes más común que se presenta en esta limpieza de datos es 1 y la menos común es 14, presentándose valores atípicos en estas últimas tres variables.
- La mediana que se presenta en la cantidad de personas extra es de 7; el mínimo de noches tiene una mediana de 2 y el máximo tiene una mediana de 1125, presentando en los dos casos, valores atípicos.

- La política de cancelación más común es *Strict* con 5004 viviendas con esta política, la menos común es *super_strict_30*, con 2 viviendas que usan esta política
- El *Host Id* nos indica que hay varios anfitriones con varios hospedajes a su nombre y con una mediana de 2.75; es decir, los anfitriones tienden a tener dos alojamientos registrados.

A este punto, se deja todo preparado para la siguiente parte del proyecto y así poder implementar las distintas herramientas que se han usado en la formación, en este caso R para el análisis exploratorio y el modelado de la regresión lineal y para Tableau en la que se hace un mapa ubicando los superanfitriones, número, y nombre del distrito.

5. Análisis exploratorio con R: selección de datos

Lenguaje de programación aplicado: R.

Recapitulando…

De acuerdo al documento “Práctica Final Glovo Big Data”, en este apartado se nos solicita hacer un estudio estadístico con R o Python, según preferencia personal, y averiguar cuáles son las métricas adecuadas para el dataset. No olvidemos:

- Revisión de la calidad de los datos
- Detección outliers (rango de variables), imputación valores nulos.
- Boxplots, histogramas, etc.
- Normalización de los valores de las tablas (quitar tildes, "dobles espacios", etc.)

Mediante Python se ha realizado la limpieza, normalización e imputación de datos y análisis inicial de datos. También se ha preparado el dataset limpio que se emplea en la definición e implementación del datawarehouse.

Se han realizado gráficas para familiarizarnos con los datos, explorarlos y tomar decisiones iniciales .

Una vez conocidos los datos y familiarizados con ellos, se planteaba un problema. Lo resolveremos seleccionando métricas de interés, muestreando el dataset limpio. Este problema está relacionado con la tarea de modelado: hacer un algoritmo de regresión lineal que prediga el precio de un inmueble en función de las características que elijamos.

Por lo tanto, el objetivo del ANÁLISIS EXPLORATORIO CON R es el de seleccionar las métricas de interés para resolver las preguntas del problema planteado, detectar los outliers de dichas métricas y generar dataframes adecuados para generar un modelado de más calidad. Nos apoyamos del análisis visual realizado paralelamente en Tableau**.**

a) ¿Qué datos debemos seleccionar?

A continuación, se indican las métricas que se seleccionan tras el análisis inicial en Python y de acuerdo a lo que nos solicita el cliente.

- Ubicación: investigamos en qué barrios y distritos se ubican los alojamientos.
    - Calle de Amparo:
        - Distrito: *Centro - selección 1*
        - Barrio: *Embajadores - subselección 2*

    - Calle de Méndez Álvaro
        - Distrito: *Arganzuela - selección 2*
        - Barrio: *Palos de Moguer - subselección 2*
- Características de los alojamientos
    - Selección de datos:
        - Tipo de propiedad: *apartamento - selección 1 y 2*
        - Tipo de arriendo: *completo - selección 1 y 2*
        - El número de baños no es representativo, tal como observamos en el análisis inicial.

- Métricas variables:
    - Número de habitaciones: 2 - *Variable independiente.* Vamos a considerar variable para que sea más generalista el modelo y se pueda emplear para apartamentos de n número de habitaciones.
    - *Número de camas*: *2*. - *Variable independiente*. La vamos a considerar variable independiente para que sea lo más generalista posible. Está relacionada con el número de habitaciones, por lo que puede haber colinealidad y podría alterar el modelo
    - *Número máximo de huéspedes*: 6 - Variable independiente, está relacionado con el número de camas.
    - *Número de huéspedes incluidos*: 1 - Variable independiente
    - Otras métricas de interés:
        - *Número de reviews* - Variable independiente
        - *Programa de superanfitrión* - Variable independiente
    - *Precio - Variable dependiente*

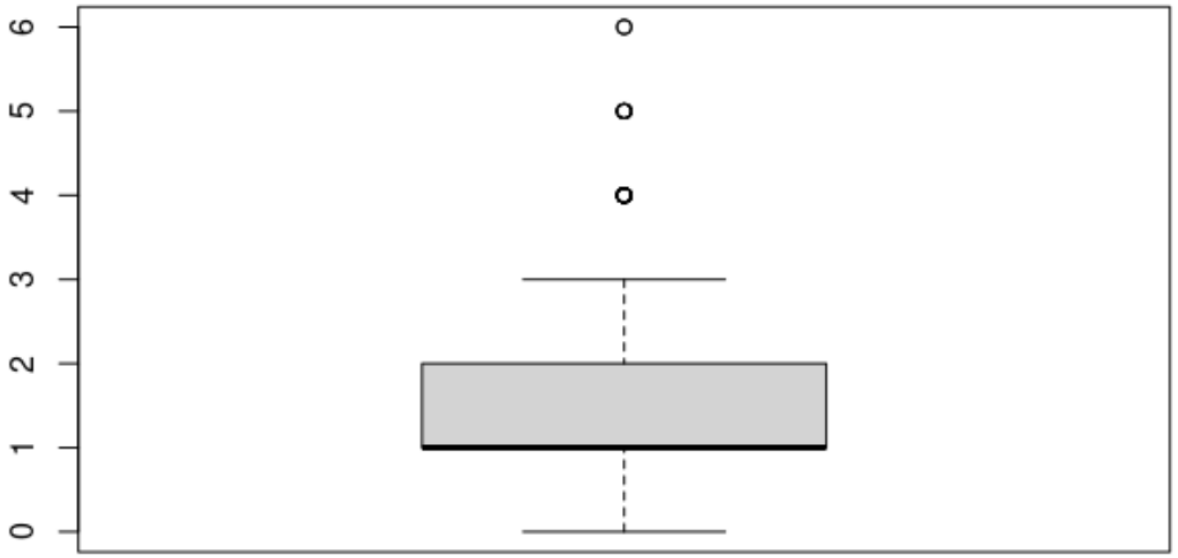
b) Selección de datos

El dataframe está limpio y los tipos de variables son correctas para poder proceder a seleccionar los datos que nos interesan, analizarlos (detectar outliers, decidir su utilidad) y realizar el modelo.

En primer lugar vamos a seleccionar todos los apartamentos de Madrid que se rentan íntegramente y analizaremos los outliers mediante boxplots. El número de datos con outliers es muy bajo para la cantidad de listings que se tienen inicialmente.

Luego seleccionamos también los datos de hosts y superhosts para hacer dos grupos diferentes de análisis para evaluar la diferencia de precio.

Se seleccionan los datos teniendo en cuenta el rango de variables que han detectado los boxplots como outliers.



Boxplot de número de habitaciones en apartamentos de Madrid que se rentan completamente

Por otro lado, se realiza el mismo proceso pero seleccionando datos más restrictivos: teniendo en cuenta el barrio y el distrito de cada alojamiento. Las ubicaciones por barrio y distrito se deben investigar previamente mediante herramientas como Google Maps.

Para los 7 dataframe generados, generamos otros 7 en los que seleccionan rango de valores descartando los outliers.

En el siguiente cuadro podremos ver el resultado de la selección tras el análisis. La cantidad de listings desciende mucho conforme avanzamos en la selección y exclusión de rango de datos. Esto podría ser un problema a la hora de obtener un buen modelo al no contar con suficientes datos, sobre todo en los barrios.

	NOMBRE / CANTIDAD DE LISTINGS			
	Base		Limpio / Sin outliers	
Selección	**Nombre**	**Total**	**Nombre**	**Total**
Dataset	*airbnb-listings.csv*	14780	*airbnb-listings_clean_R .csv*	13264
Madrid	*df_madrid*	6970	*madrid_s*	5229
Host	*df_host*	6068	*host_s*	4497
Superhost	*df_superhost*	902	*super_s*	644
Dº Centro (c/ de Amparo)	*df_amparo*	4311	*amparo_s*	3495
Bº Embajadores (c/ de Amparo)	*df_amparito*	1099	*amparito_s*	922
Dº Arganzuela (c/ de Méndez Álvaro)	*df_alvaro*	303	*alvaro_s*	219
Bº Palos de Moguer (c/ de Méndez Álvaro)	*df_alvarito*	114	*alvarito_s*	86

6. Visualización de métricas

Herramienta aplicada: Tableau.

Para crear las visualizaciones a utilizar en la presentación de conclusiones a nuestro cliente, usando Tableau, decidimos seleccionar cuatro constantes:

- Tipo de propiedad: apartamentos (*Property Type*),
- el precio por día (*Price*),
- 2 habitaciones (*Bedrooms*)
- Distritos (*Neighbourhood Group Cleansed*).

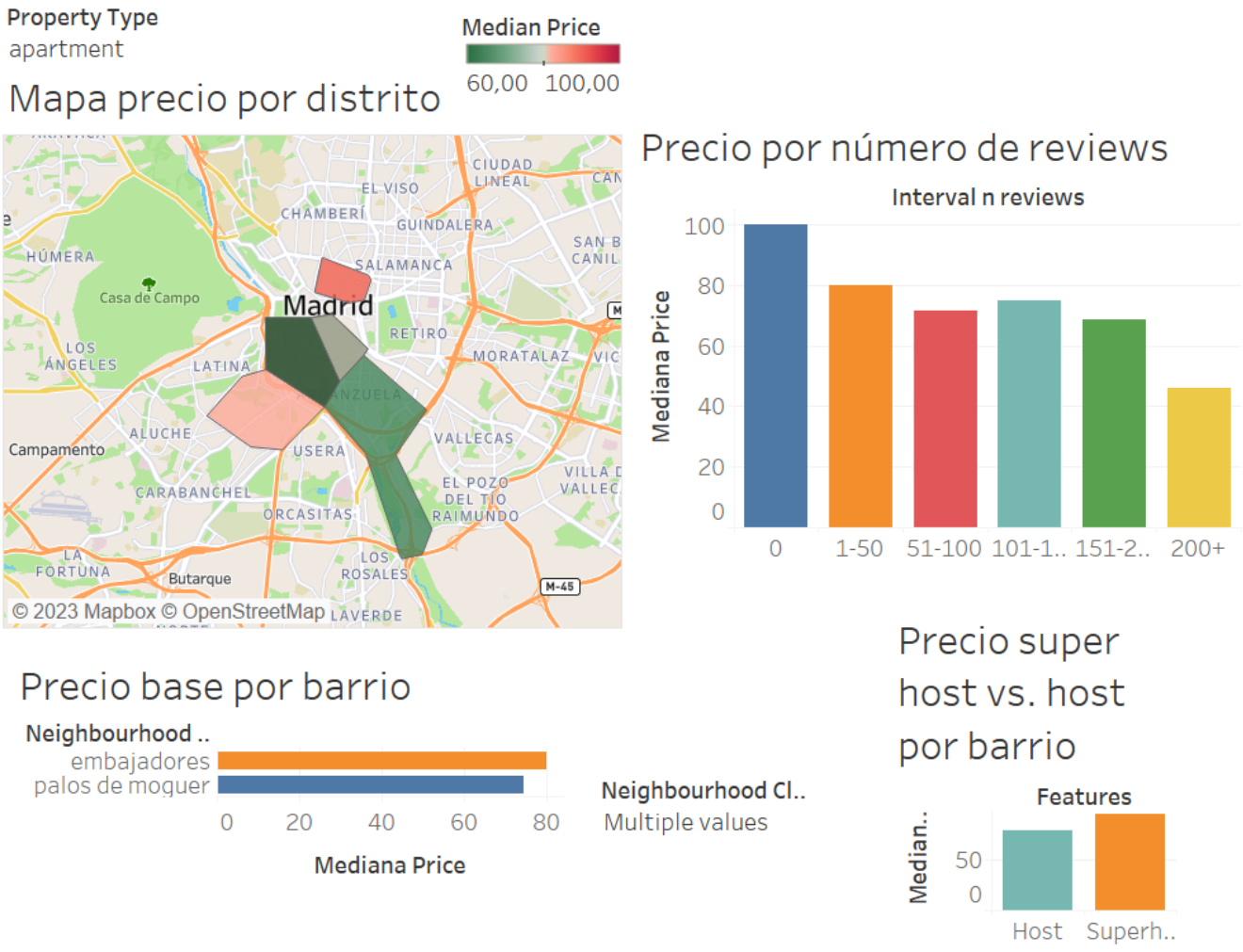
En el mapa se pueden visualizar los distritos de interés (Centro y Arganzuela), en rojo se muestran las zonas con la mediana del precio más alta, y en verde las zonas con la mediana del precio más baja.

En el dashboard decidimos utilizar dos gráficos de barras verticales y uno horizontal, ya que estos nos ayudan a detectar patrones, tendencias y relaciones en los datos. También hemos tenido en cuenta la paleta de colores para hacer que el diseño sea más efectivo y perceptible a primera vista.

En las gráficas podemos observar que las variables que afectan al precio positivamente son: el barrio, distrito y ser superhost. Por otro lado, la variable que afecta negativamente es el número de reviews. Observamos que la mediana del precio está más baja para los alojamientos con más reseñas. La mediana de precio para un apartamento de 2 habitaciones en el barrio de Embajadores (Centro) es 80 dólares y para un piso con las mismas características en el barrio de Palos de Moguer (Arganzuela) la mediana del precio es de 74.5 dólares.

Conclusiones:

El precio base por día recomendado para el piso ubicado en el barrio Embajadores, en distrito Centro, es de 80 dólares por alojamiento de las características citadas. Para el piso ubicado en el barrio Palos de Moguer, en distrito Arganzuela, es de 74.5 dólares. En ningún caso se incluye la fianza y el servicio de limpieza. Recomendamos el programa de superhost, porque se observa que la mediana del precio de los pisos en estos dos distritos para pisos de 2 habitaciones con anfitrión superhost es de 95 dólares y para anfitrión normales de 79 dólares.



Cuadro extraído del dashboard con un mapa y tres gráficos de barras: de relación de precio de precio base por barrio, la mediana en relación a las reviews y el precio según se es Host/Superhosts.

7. Preprocesamiento y modelado

Lenguaje de programación aplicado: R.

Una vez analizados los datos con R y con Tableau, hemos determinado que las variables seleccionadas pueden tener una buena correlación con para poder determinar el precio de sus alojamientos, tal como nos solicita el cliente.

Los modelos de correlación lineal múltiple requieren de las mismas condiciones que los modelos lineales simples más otras adicionales. Según la fuente Regresión Lineal Múltiple en R - RPubs, debemos tener en cuenta ciertas condiciones.

En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinialidad entre ellos. La colinialidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores. Como consecuencia de la colinialidad no se puede identificar de forma precisa el efecto individual que tiene cada una de las variables colineales sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto que resulta prácticamente imposible establecer su significancia estadística. Además, pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes.

Por lo tanto, debemos observar la relación entre las variables.

- Si el coeficiente de determinación R2 cuadrado es alto pero ninguno de los predictores resulta significativo, hay indicios de colinialidad.
- Calcular una matriz de correlación en la que se estudia la relación lineal entre cada par de predictores. Es importante tener en cuenta que, a pesar de no obtenerse ningún coeficiente de correlación alto, no está

asegurado que no exista multicolinialidad. Se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y que las correlaciones simples entre pares de estas mismas variables no sean mayores que 0.5.

- Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el coeficiente de determinación R2 es alto, estaría señalando a una posible colinialidad.

El mejor modelo es aquel capaz de explicar con mayor precisión la variabilidad observada en la variable respuesta empleando el menor número de predictores, por lo tanto, con menos asunciones. A esto se le llama parsimonia.

Cada predictor numérico tiene que estar linealmente relacionado con la variable respuesta mientras los demás predictores se mantienen constantes, de lo contrario no se puede introducir en el modelo. También, es importante detectar los valores atípicos para descartarlos del modelo.

Cuando se introduce una variable categórica como predictor, un nivel se considera el de referencia (normalmente codificado como 0, en este caso Host) y el resto de niveles se comparan con él.

a) Analizar la relación entre variables y modelado

El primer paso a la hora de establecer un modelo lineal múltiple es estudiar la relación que existe entre variables. Esta información es crítica a la hora de identificar cuáles pueden ser los mejores predictores para el modelo, qué variables presentan relaciones de tipo no lineal (por lo que no pueden ser incluidas) y para identificar colinialidad entre predictores. Se emplean las funciones:

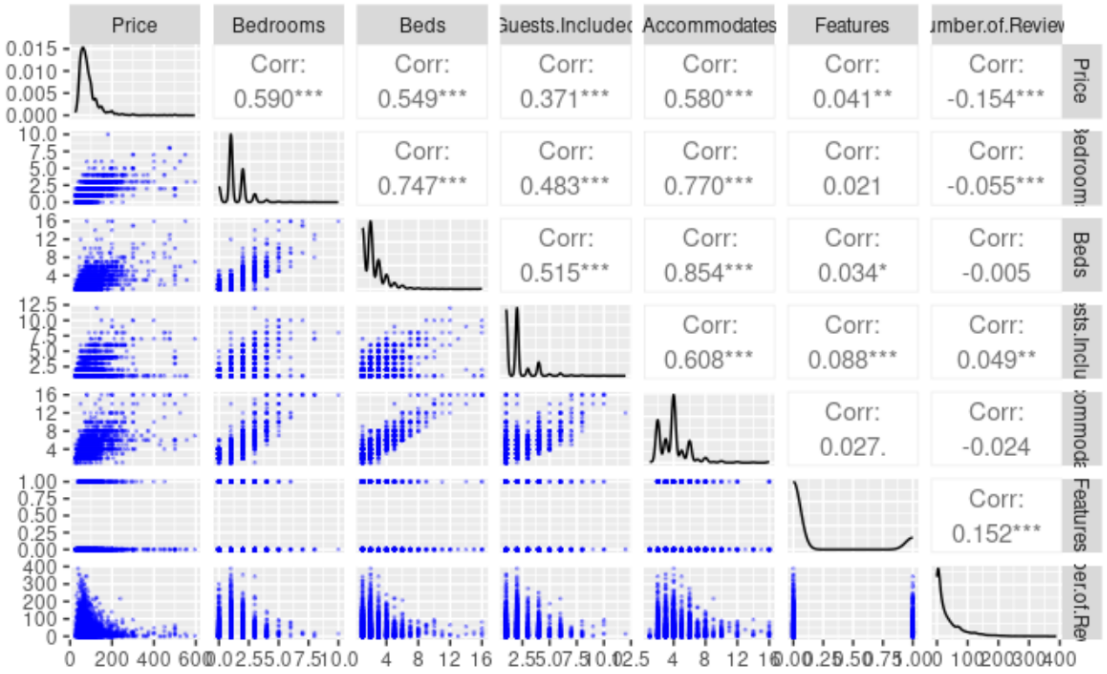
cor(X = data, method = “pearson”) *cor(X = data, method = “spearman”)*

ggpairs(data, lower = list(continuous = ….)

model<- lm(data= x , formula = Price~ …. + ……+.....)

Se observa que para todos 14 dataframes los predictores Beds y Accommodates parecen guardar colinealidad al tener alta correlación. Se decide mantenerlos hasta evaluar en el modelo de regresión lineal.

La correlación de los predictores con price es buena en todos los dataframes. El que menor correlación tiene con el precio es Guests Included.



b) Resultados

Se muestran los resultados de la regresión lineal teniendo en cuenta los 14 dataframes:

Selección	REGRESIÓN LINEAL				
	Predictores aplicados	Fórmula $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_kX_{ki}+\epsilon_i$	Multiple R^{2*} - p-value	Precio apartamento	
				Host	Superhost
Dataset	-	-	-	-	
Madrid	Bedrooms + Beds + Accommodates + Features + Number.of.Reviews	*Y=30.59119+17.32385*Br+3.96598*B+6.30426*A+6.22138*F-0.15766*NR*	**0.3389** / < 2.2e-16	$111,00	$117,22
Host	Bedrooms + Beds + Guests.Included + Accommodates + Number.of.Reviews	*Y=31.55054+17.24586*Br+4.24305*B-1.18408*GI +6.58241*A-0.17421*NR*	**0.3306** / < 2.2e-16	$111,65	-
Superhost	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=30.67782+17.98233*Br+5.59964*GI+6.02448*A-0.07431*NR*	**0.4383** / < 2.2e-16	-	$108,39
Dº Centro (c/ de Amparo)	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=32.3670+17.8795*Br +3.3104*B+1.3557*GI+6.1332*A+6.2986*F-0.1518*NR*	**0.4078** / < 2.2e-16	**$112,90**	**$119,20**
Bº Embajadores (c/ de Amparo)	Bedrooms + Beds + Accommodates + Features + Number.of.Reviews	*Y=29.1843+14.1807*Br +2.5654*B+6.0246*A +10.7265*F-0.1191*NR*	**0.4636** / < 2.2e-16	**$98,82**	**$109,55**
Dº Arganzuela (c/ de Méndez Álvaro)	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=35.2937+9.3598*Br-2.9502*GI+6.9069*A-0.1186*NR*	0.2925 / < 2.2e-16	**$92,50**	**$92,50**
Bº Palos de Moguer (c/ de Méndez Álvaro)	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=35.8010+10.6005*Br-8.5652*GI+7.8809*A-0.2023*NR*	**0.3888** / 4.947e-11	**$95,75**	**$95,75**

REGRESIÓN LINEAL - SIN OUTLIERS

Selección	Predictores aplicados	Fórmula $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_kX_{ki}+\epsilon_i$	Multiple R^{2*} - p-value	Precio apartamento	
				Host	Superhost
Dataset	-	-	-	-	
Madrid	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=49.34626+7.18268*Br+1.77273*B-1.73606Gl+4.91038*A+8.93240*F-0.23564*NR*	**0.2012** < 2.2e-16	$94,98	$103,92
Host	Bedrooms + Beds + Guests.Included + Accommodates + Number.of.Reviews	*Y=48.66326+7.25955*Br+2.06097*B-1.11485*Gl+4.85718*A-0.29265*NR*	**0.2058** < 2.2e-16	$95,33	
Superhost	Bedrooms + Guests.Included + Accommodates	*Y=57.976+8.573*Br+-4.891*Gl+4.984*A*	**0.1562** < 2.2e-16		$100,14
D° Centro (c/ de Amparo)	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=45.2834+9.4199*Br+1.4329*B+2.2162*Gl+4.1568*A+7.0679*F-0.1714*NR*	**0.305** < 2.2e-16	$96,15	$101,21
B° Embajadores (c/ de Amparo)	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=38.4211+9.0340*Br+2.5862*B+1.2239*Gl+3.6121*A+9.7052*F-0.1477*NR*	**0.3544** < 2.2e-16	**$84,56**	**$94,26**
D° Arganzuela (c/ de Méndez Álvaro)	Bedrooms + Beds + Features + Number.of.Reviews	*Y=46.1461+4.1607*Br+4.6348*B+6.3421*F-0.1417*NR*	**0.1387** 1.666e-05	$63,74	$70,08
B° Palos de Moguer (c/ de Méndez Álvaro)	Bedrooms + Beds+ Guests.Included + Accommodates	*Y=37.900+7.713*Br+5.142*B-7.518*Gl+4.947*A*	**0.3454** 5.275e-07	**$85,77**	**$85,77**

Respondemos a la pregunta de nuestro cliente, basándonos en datos:

1. ¿**A qué precio base debería publicar** cada uno de los alojamientos teniendo en cuenta las características de éstos?

a. C/ de Amparo: Min: $84,56
b. C/ de Méndez Álvaro: Min: $85,77

Si no tenemos en cuenta los alojamientos fuera de rango, el precio asciende. Podría no ser competitivo en el mercado.
En ese caso, sería a

a. C/ de Amparo: Min: $98,82
b. C/ de Méndez Álvaro: Min: $95,75

2. ¿Cómo puede **variar el precio según la cantidad de reviews que tenga**?
    a. C/ de Amparo: afectaría negativamente.

        Y=38.4211+9.0340*Br+2.5862*B+
        +1.2239*Gl+3.6121*A+9.7052*F-0.1477*NR
        A razón de $ - 0,1477 /review

    b. C/ de Méndez Álvaro: afecta negativamente
        Y=35.8010+10.6005*Br-8.5652*Gl+7.8809*A-0.2023*NR
        A razón de $ - 0,2023 /review

3. En qué grado es recomendable que se implique en la gestión de los alojamientos. ¿**Es rentable ser superanfitrión**?

    a. C/ de Amparo: sí, puede poner un precio base más elevado.

        $94,26

    b. C/ de Méndez Álvaro: no afectan al precio la condición de superhost en la regresión lineal. En Tableau incrementa a

        $95,00

8. Conclusiones

Tras analizar los resultados, se observa relación con lo analizado en la visualización de datos con Tableau. Los precios son ligeramente más altos, pero se debe al análisis más general de las métricas. Se han incorporado más variables a tener en cuenta, como es el número de habitaciones.

También se observa que si se toma en cuenta los barrios el R^2 es más elevado.

Los resultados son similares a los analizados en Tableau. La relación de reviews-precio es negativa. A más reviews menor tiende a ser el precio de los alojamientos. Los huéspedes suelen seleccionar alojamientos más baratos.

- Barrio Embajadores:

Y=29.1843+14.1807*Br+2.5654*B+6.0246*A+10.7265*F-0.1191*NR

- Barrio Palos de Moguer:

Y=35.8010+10.6005*Br-8.5652*GI+7.8809*A-0.2023*NR

¿Qué te ha aportado el desarrollar este proyecto?

- Satisfacción de superar retos y ser resolutivas
- Consolidación de conocimientos mediante práctica
- Trabajo en equipo implementando SCRUM
- Hacernos amigas de la frustración,

¿Qué has aprendido?

- Saber cómo y en qué momento aplicar las herramientas
- A trabajar en un proyecto pseudorreal
- A conocer los límites personales
- A implementar nuevas funciones de herramientas

¿Qué es lo que no volverías a hacer de la misma manera?

- Big Data es aprendizaje diario, tenemos mucho que descubrir y aprender

HARÍAMOS IGUAL PERO CON MÁS TIEMPO

¿Qué cosas seguirías haciendo en el futuro para mejorar el proyecto?

- Continuar con el modelado realizando regresiones logísticas.
- Conseguir más métricas continuas o métricas temporales para mejorar los resultados

¡SEGUIR TRABAJANDO Y MEJORANDO NUESTRAS SKILLS!

Enlace al repositorio de GitHub.