



EL POTENCIAL DEL DATO EN PLATAFORMAS DE ALOJAMIENTO TURÍSTICO

**Proyecto de final de máster
Ignacio Olcina y Héctor López Molas**

**Máster en Big Data y Marketing Analytics
EEME Bussines School**

ÍNDICE

1.- INTRODUCCIÓN.....	3
2.- CARGA Y TRANSFORMACIÓN DE DATOS.....	6
3.- ANÁLISIS Y ENTENDIMIENTO DE LOS DATOS.....	17
4.- VISUALIZACIÓN DE LOS DATOS.....	37
5.- PREDICCIÓN.....	41
5.1. PREDICCIÓN DEL PRECIO DE ALQUILER.....	41
5.2. PREDICCIÓN DE LA SATISFACCIÓN.....	45
6.- ALTERNATIVA DE ARQUITECTURA CLOUD.....	51
7.- CONCLUSIONES.....	56
8.- BIBLIOGRAFÍA.....	61
8.1. ANEXO DE EJECUCIÓN.....	61
8.2. ENLACES BIBLIOGRÁFICOS.....	61

1.- INTRODUCCIÓN

El modelo turístico de la sociedad mundial ha ido evolucionando con los distintos avances que se han ido dando en la sociedad con los años. Los cambios, como han sido los avances tecnológicos que han tenido como consecuencia una globalización tremenda de toda la población del mundo, en los cuales hace 20 años era impensable poder conectarse en microsegundos con otra parte del mundo remotamente. Otros cambios que han tenido bastante importancia han sido las mejoras en los medios de transporte tanto aéreo, terrestre como marítimo.

Estos cambios en la sociedad han tenido una repercusión muy positiva en el turismo haciendo posible a los usuarios viajar a cualquier parte del mundo para disfrutar de sus vacaciones o simplemente desplazarse por trabajo. Apareciendo así nuevas formas de alojamientos que todavía no eran conocidas, irrupciones en el mercado del turismo como el alquiler vacacional haciéndole competencia directa a las grandes cadenas hoteleras que dominaban tradicionalmente el sector turístico, algunos ejemplos de startups del sector son 9flats, Alterkeys, flatClub, Flat4Day, entre otras. La manera habitual de estas Startups para sacar beneficio de el alquiler de los apartamentos es mediante comisiones de los alquileres, que suelen oscilar entre el 16 y 20 por ciento del importe.

Una de estas plataformas de alquiler vacacional que sufrió un auge bestial entre 2010 y la actualidad fue AirBnB, que se situó como una de las opciones más utilizadas por los turistas para hospedarse, basándose en una comunidad colaborativa que ofrece alojamientos a través de Internet en más de 190 países adecuándose así a la nueva mercado global del turismo. Lo que hace distinto a estas nuevas plataformas como AirBnB es que no solamente son empresas como agencias de alquiler vacacional, son mucho más que eso, son empresas tecnológicas que basan sus decisiones en los datos que tienen para mejorar en la actualidad pero sobretodo pensando en el futuro.

Utilizan la cultura del dato para crear distintos modelos de negocio en su plataforma en función de lo que los clientes que utilizan la plataforma necesitan. Es tanta la importancia que le dan al dato en esta plataforma, que tienen una universidad llamada Data University Airbnb en la cual inculcan la cultura del dato a toda la empresa, teniendo

además Data Scientists en todas las unidades organizativas de la empresa. El uso de los modelos basados en el data los aplican para multitud de casos, para fijar los acuerdos de los precios para nuevos apartamentos, para recomendar los distintos apartamentos a los posibles clientes, para conocer la satisfacción que tendrá un nuevo apartamento, para estudiar la evolución del turismo y conocer donde hay más demanda de alojamientos, entre muchas otras cosas.

Dado el auge que ha tenido esta plataforma cada vez más apartamentos pasan a formar parte de la plataforma, que actualmente proporciona a sus usuarios una oferta de más de 2.000.000 de alojamientos distribuidos por el mundo. Por ello, se utiliza información de esta plataforma para cuantificar el potencial del dato en plataformas de alojamiento turístico, ya que es la plataforma del sector con más datos. Centrando nuestro análisis en datos que son de la ciudad de Madrid, España, tratándolos inicialmente desde la carga y transformación, análisis descriptivo y entendimiento de los mismos, modelos de predicción, visualización, hasta las conclusiones obtenidas durante esta ejemplificación del poder del dato en este sector. Se dispone las dos bases de datos nombradas anteriormente con sus respectivas tablas:

- Base de datos de AirBnB Madrid:
 - Calendar: Amplio histórico por días de cada apartamento, con su precio, disponibilidad, mínimo de noches y máximo.
 - Listings: Tabla con las características básicas de cada uno de los apartamentos
 - Listings detailed: Extendida de la anterior en la cual se añade toda la información recogida de cada uno de los apartamentos.
 - Neighbourhood: relación de las zonas que hay en cada barrio
 - Reviews: fecha de cuando se pone una review y a que apartamento
 - Reviews detailed: Es un extendido de la tabla anterior que añade quien hace la review, su nombre y el comentario.

- Base de datos del catastro de Madrid:
 - GSD_Madrid: Información muy amplia por cada sección censal de Madrid, desde habitantes, nacionalidades, cantidad de casas hasta la riqueza o el turismo.

2.- CARGA Y TRANSFORMACIÓN DE DATOS

Inicialmente para poder tener acceso libre a los datos todo el grupo que conforma el proyecto, se crea un servidor PostgreSQL donde cada uno de los integrantes utilizando PgAdmin 4 tengan acceso al servidor que tiene los datos. Para conseguirlo, se tiene que abrir los puertos en los que opera PostgreSQL del host y mediante permisos del host a las IP públicas del resto de miembros se hace posible la conexión al servidor.

Tras buscar la información, sacada de la página web insideairbnb.com portal donde se vuelca información mensual de AirBnB de distintas ciudades del mundo, se obtiene una serie de datos en formato csv y geojson. La información con la que se trabaja está actualizada del día 18 de febrero de 2020.

Trabajando con Pentaho, comenzamos haciendo la conexión con PostgreSQL para hacer uso así de la información almacenada en el datalake. En primer lugar, se modifica el archivo kettle.properties añadiendo en él las variables que se van a utilizar para la creación de las conexiones, que son *BBDD_HOST*, *BBDD_USER*, *BBDD_PASSWORD* y *BBDD_PORT*, teniendolas todos los integrantes igual. Después, se establece la conexión con la base de datos que se trabaja. Es importante tener en cuenta que la *BBDD_HOST* almacena la dirección ip a la que conectarse, y por tanto, será modificada cada vez que cambie su ip y así no habrá que modificar todos los lugares donde se hace uso de la ip, únicamente uno. La conexión se puede ver en la Figura 1:

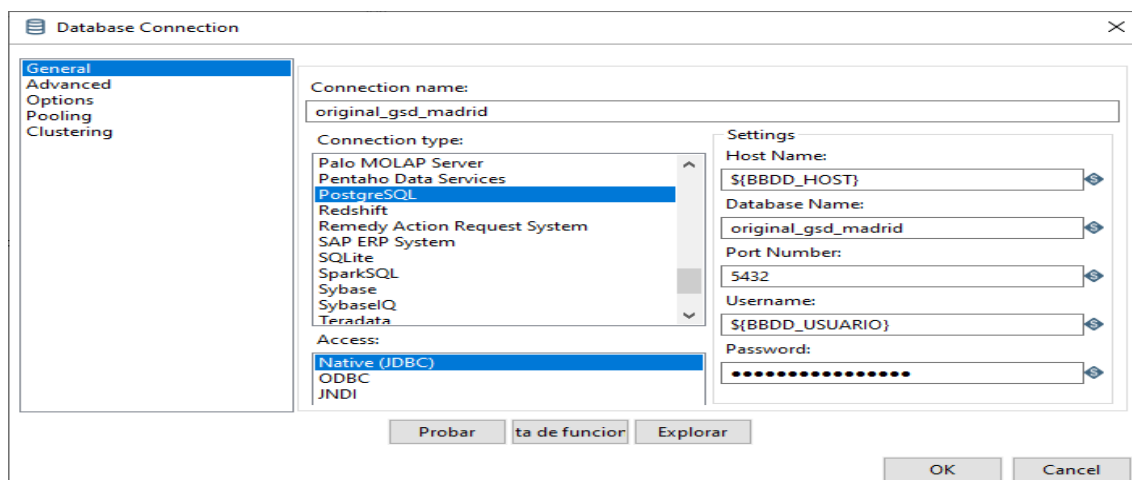


Figura 1. Conexión con la base de datos

Inicialmente, se plantea el siguiente modelo para el datawarehouse resultante de todas las transformaciones realizadas en pentaho, el primer datamart constaría de cuatro tablas, estando en el centro de este la tabla de mayor relevancia *calculada_listings_detailed*, en la que *number_host_verifications* es una métrica que indica cuantas verificaciones obtuvo el host en la página de airbnb. A continuación, se añade la dimensión *tratada_calendar* que dispone de los datos de cada día de los alojamientos y si estaba disponible ese día o no entre otros datos, a partir de esta tabla se plantea obtener el primer día que el alojamiento tuvo una estancia junto al día en el que acabó la estancia, su duración el precio total y la media de precio por día. También se plantea calcular los servicios de los que dispone cada AirBnB y almacenarlos en una nueva dimensión llamada *calculada_dias_ocupados*. La representación del modelo del datawarehouse se observa en la Figura 2:

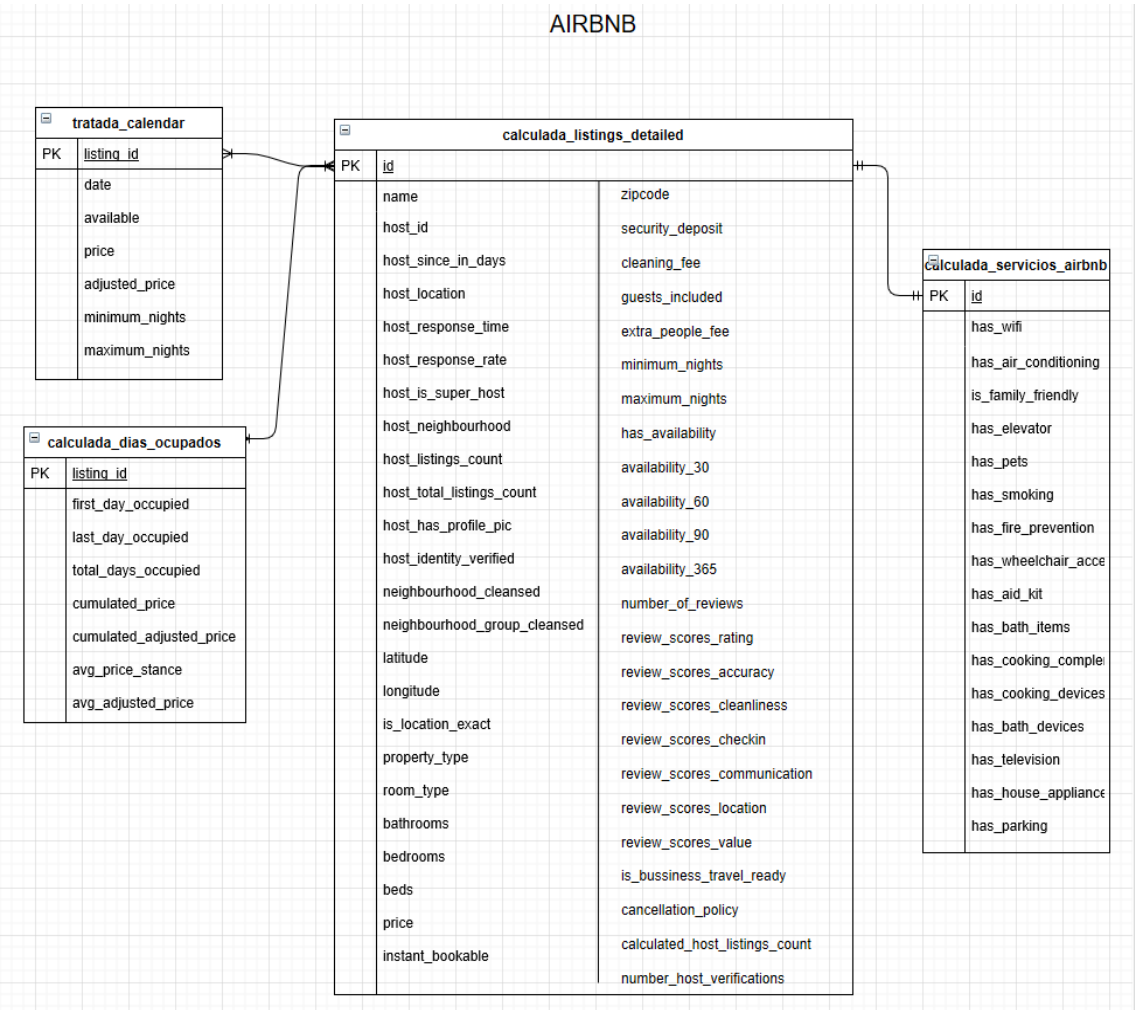


Figura 2. Modelo del datawarehouse

A continuación, se plantea el siguiente datamart que almacenará los datos espaciales obteniendo a partir de los datos iniciales algunas métricas como son la cantidad de habitantes de 0 a 14 años, la cantidad de 15 a 29 años, la cantidad de 30 a 44 años, la cantidad de 45 a 59 años, la cantidad 60 a 74 años, la cantidad de habitantes mayores de 75 años y un indicador numérico de la clase social.

calculada_gsd_madrid		
PK	gid	
	codigo_sc	separados
	cod_postal	divorciados
	municipio	viudos
	atractivo_comercial	habitantes_0_14
	riqueza	habitantes_15_29
	clase_social	habitantes_30_44
	transito	habitantes_45_59
	transito_laboral	habitantes_60_74
	transito_comercial	habitantes_75_90
	transito_ocio	poi_bus
	turismo	poi_bus_b1
	precio_viv	poi_bus_b3
	numero_de_viviendas	poi_metro
	numero_de_habitantes	poi_metr_1
	numero_de_hombres	poi_metro_
	numero_de_mujeres	poi_tren
	españoles	poi_tren_1
	extranjeros	poi_tren_b
	extranjeros_africanos	poi_tranvi
	extranjeros_americanos	poi_superm
	extranjeros_asiatcos	poi_superm
	extranjeros_europeos	poi_restau
	solteros	pasa_linea
	casados	geog

Figura 3. Datamart de los datos espaciales

En consecuencia, se ha creado un job global capaz de conseguir este modelo mediante procesos ETL, este job se llama *job global.kjb*, en el cual se ensamblan todos los procesos que se quieren hacer, todo ello con rutas relativas para que se pueda hacer desde cualquier ordenador por lo que es importante que estén bien ordenados los archivos. Además de las rutas relativas, se añade mensajes de error en cada una de las fases, se puede observar en la Figura 4.

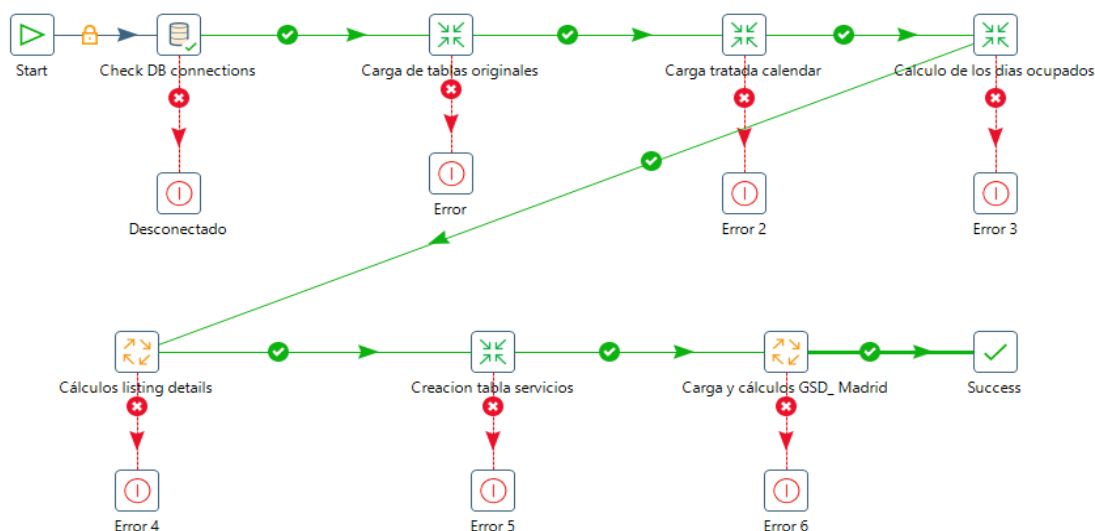


Figura 4. Job global de la ETL

En primer lugar, se comprueban las conexiones con los servidores PostgreSQL con los que se trabaja, almacenando los datalake y los datawarehouse. Tras ello, se cargan la información en el servidor común en un datalake con todas las tablas. En relación con la carga de datos, se ha creado un datalake formado únicamente por todos los datos originales que disponemos y en función de este, ir estructurando los datos. Para ello, se ha utilizado un script SQL con Pentaho, que será adjuntado en la entrega, el cual realiza un borrado de la tabla antes de crearla, para que no haya conflictos y falle la ejecución. Una vez borrada, se crea la tabla con las columnas de las distintas tablas en formato csv emparejando sus distintas columnas para subirlas a PostgreSQL, haciendo uso de Pentaho Data Integration para la carga mediante un archivo mostrado en la Figura 4, que lo que hace es cargarla de 50.000 en 50.000, pese a ello y debido a la volumetría de alrededor de 8M de registros, todo ello utilizando rutas relativas para la subida. En dicha transformación se crean las tablas las seis tablas de la siguiente manera:

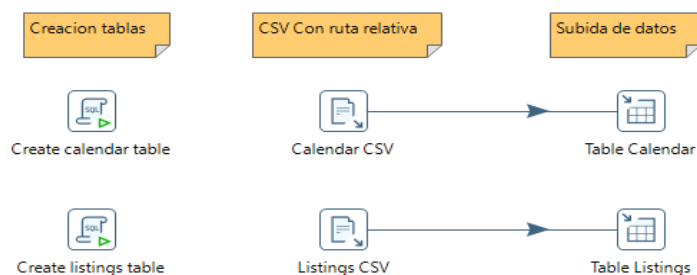


Figura 5. Ejemplo de la carga de dos tablas en PostgreSQL

Además de estos archivos, también se carga información del catastro de Madrid donde, se dispone información de la gente que vive en cada parte de la ciudad, rentas, turismo, entre otras cosas. En el caso del archivo `neighbourhood.geojson`, mediante la extensión PostGis de PostgreSQL se carga para hacer un análisis cruzando con los datos del catastro de Madrid.

El datalake inicial cargado en PostgreSQL tiene dos bases de datos diferenciadas, la base de datos formada por la información de AirBnB cargada con Pentaho formada por seis tablas y la segunda base de datos con la información de Madrid con dos tablas, una de ellas de información espacial.

Analizando la calidad de los datos, se van a realizar las transformaciones apropiadas para mejorar la calidad tanto en la información aportada por AirBnB como en los datos de Madrid. Acabada la creación del datalake, se empieza a trabajar la información de AirBnB, más concretamente por la tabla Calendar, que dispone de la información de cada alojamiento si ha estado ocupado o no, en que determinada fecha, el precio que tiene el piso por día y el precio que determina el propio AirBnB para poder adaptar el precio de los alojamientos automáticamente, esta última columna mencionada puede ser habilitada o deshabilitada por el anfitrión.

listing_id integer	date date	available character varying	price character varying	adjusted_price character varying	minimum_nights integer	maximum_nights integer
604632	2020-05-07	t	\$29.00	\$29.00	1	1000
604632	2020-05-08	t	\$30.00	\$30.00	1	1000
604632	2020-05-09	t	\$30.00	\$30.00	1	1000
604632	2020-05-10	t	\$29.00	\$29.00	1	1000
435010	2019-09-20	f	\$25.00	\$25.00	2	7

Figura 6. Tabla Calendar actualizada

En dicha columna, se han realizado las siguientes transformaciones, el cambio de tipo de dato de *price* y *adjusted_price* de tipo texto a numérico, la modificación del formato de la columna. Estas transformaciones se realizarán mediante un script SQL que retira el símbolo del dólar y las comas de los números para poder realizar dicha conversión posible. Continuando con Pentaho, se realiza una conversión de la columna *available* a tipo booleano utilizando pentaho sustituyendo las filas con valor 't' por un uno y las filas con valor 'f' por un cero.

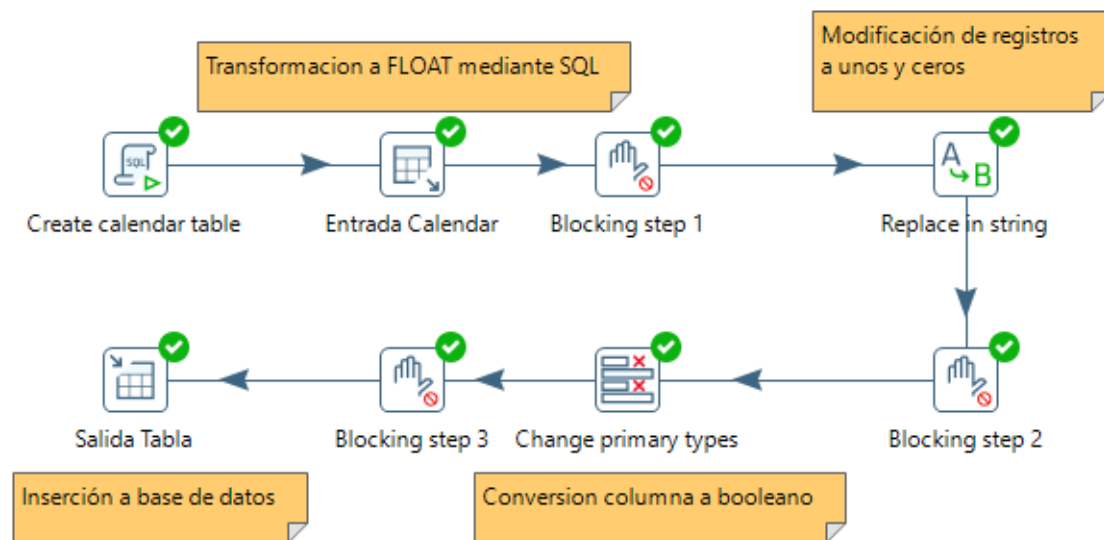


Figura 7. Proceso ETL sobre la tabla Calendar

A partir de la tabla *calendar* y mediante la transformación *Cálculo días ocupados* del job global, se va generar una nueva tabla a creando nuevas métricas. Al disponer simplemente de los datos que tiene por día se han agrupado estos datos en función a las estancias de los clientes, la nueva tabla va a tener la siguiente estructura: El *id_listing* para conocer el alojamiento que tendrá formato *integer*, el *first_day_occupied* para conocer el primer día que comenzó una estancia del alojamiento, el *last_day_occupied* para saber cuando terminó el alojamiento del cliente, estos dos últimos mencionados con formato *date*, *total_days_occupied* en formato *integer* para conocer la duración de la estancia medida en días, *cumulated_price* y *cumulated_adjusted_price* en formato *float* que van a servir para conocer el precio total, tanto el normal como el precio que recomienda el propio AirBnB respectivamente y por último, sus medias por día *avg_price_stance* y *avg_adjusted_price*. Toda esta carga de datos se ha realizado mediante un procedimiento SQL llamado *calculada_dias_ocupados* el cual su funcionamiento es el siguiente.

Primero realiza la declaración de variables, las cuales conforman los campos que se van a insertar en nuestra nueva tabla compuesta por métricas. La variable registro servirá para acceder a las columnas que va recorriendo el cursor llamado “BUSCADOR”, el cual obtiene todos los campos de la tabla *calendar* excepto *minimum_nights* y *maximum_nights*. La variable *first_day_chosen* servirá para prevenir que se sobrescriba el primer día que la estancia estuvo ocupada. Tras ello, se comprueba la

disponibilidad de la estancia para comenzar a crear las métricas si la disponibilidad del alojamiento está ocupada.

En cambio, si el alojamiento pasa a estar disponible se guarda en la variable *last_day_occupied* la fecha de la fila que está recorriendo el cursor menos un día, para obtener realmente el día en el que el cliente abandonó el alojamiento. A continuación, se calcula para la variable *time_occupied*, que es el tiempo de estancia en días, y las medias de precio pagado por día tanto el precio adaptable de AirBnB como el normal. En el caso de que la estancia haya sido de un día, se le asigna el precio que pagó por ese día, evitando así errores de división por cero. Finalmente, se procede a realizar la carga en la tabla y a reiniciar las variables. El script se ejecuta en pentaho mediante una simple transformación que proporciona la siguiente tabla con un total de ochenta y cinco mil registros y ocho columnas como resultado.

listing_id integer	first_day_occupied, date	last_day_occupied, date	total_days_occupied, integer	cumulated_price double precision	cumulated_adjusted_price double precision	avg_price_stance double precision	avg_adjusted_price double precision
6369	2019-09-19	2019-09-21	2	220	220	110	110
6369	2019-09-25	2019-09-26	1	140	140	140	140
6369	2019-09-28	2019-10-02	4	355	355	88.75	88.75
6369	2019-10-13	2019-11-08	26	1925	1925	74.0384615384615	74.0384615384615
21853	2019-12-18	2020-07-31	226	25330	25330	112.079646017699	112.079646017699
24805	2019-09-19	2019-09-21	2	380	380	190	190

Figura 8. Tabla resultante de los días ocupados

Una vez tratada la tabla calendar, se estudia la tabla *listings_detailed* que visualmente es la que más contenido aporta en nuestra investigación. En primer lugar, se ha eliminado las columnas que no aportan ningún valor al posterior análisis del negocio con el fin de reducir su dimensionalidad antes de realizar las transformaciones. Para ello, se han reducido las columnas haciendo una consulta SQL y seleccionando las columnas más interesantes en nuestro estudio, se han obtenido un total de cuarenta y nueve columnas contando las métricas de ciento seis iniciales. Las columnas que se han borrado ha sido como consecuencia de que contenían un excesivo número de valores nulos, columnas relacionadas con el scrapping (urls origen, fotos, etc.), de las cuales la mayoría redireccionan a la página principal de AirBnB o columnas que contenían descripciones desestructuradas por los usuarios que en algunas ocasiones era nulo.

Posteriormente, se ha realizado la categorización de las columnas como el tipo de propiedad o tipo de habitación, para reducir la varianza de estas columnas también se han transformado a entero estas columnas junto con las políticas de cancelación, el tiempo de respuesta del anfitrión. Además de estas, se han obtenido las modas y las medias de la mayoría de las columnas como el número de camas, habitaciones, las puntuaciones obtenidas de los alojamientos o los precios. Estas medias y modas se han obtenido mediante una consulta SQL y han sido almacenadas en variables de pentaho para su posterior uso en la transformación que sustituye nulos por otros valores especificados.

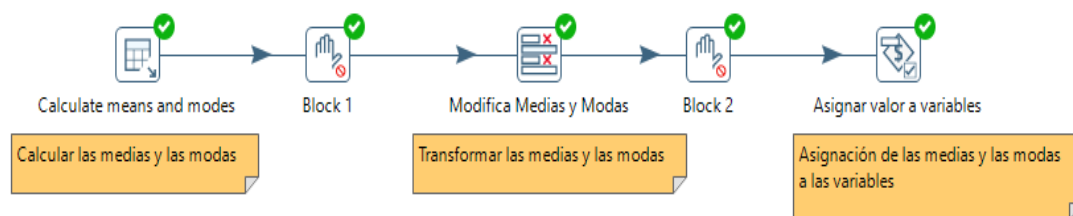


Figura 9. Proceso de cálculo de medias y modas

A continuación, se han obtenido la latitud y la longitud del csv original puesto que en la base de datos viene redondeado. Por tanto, mediante pentaho se van a obtener estas columnas con una precisión de decimales de seis para realizar su correcta inserción. También se calcula la métrica *number_host_verifications* para obtener todas las verificaciones que el anfitrión ha obtenido en la web. Finalmente, utilizando pentaho se van a combinar ambas fuentes de datos para realizar las transformaciones apropiadas y su correcta carga de datos en la nueva tabla. Además de ello, se ha agrupado el tipo de propiedad clasificandolo en cinco grupos numerados de 1 a 5 que son, apartamentos, casas, bungalows y casas de campo, hostelería y en el último grupo los demás. Otra agrupación ha sido la política de cancelación que se categoriza de menos a más siendo 1 la menos estricta hasta 5 siendo la más estricta. La clasificación del tiempo de respuesta de los caseros también se ha clasificado entre 1 y 5, clasificando los valores menos de una hora, varias horas, un día, varios días y otros valores respectivamente. Por último en relación con esta tabla, se han modificado las columnas de variables booleanas pasándolas a sus respectivos valores numéricos. Estas variables describen si

la localización es exacta, si es considerado al casero como superhost, si el casero tiene imagen de perfil y si está verificado.

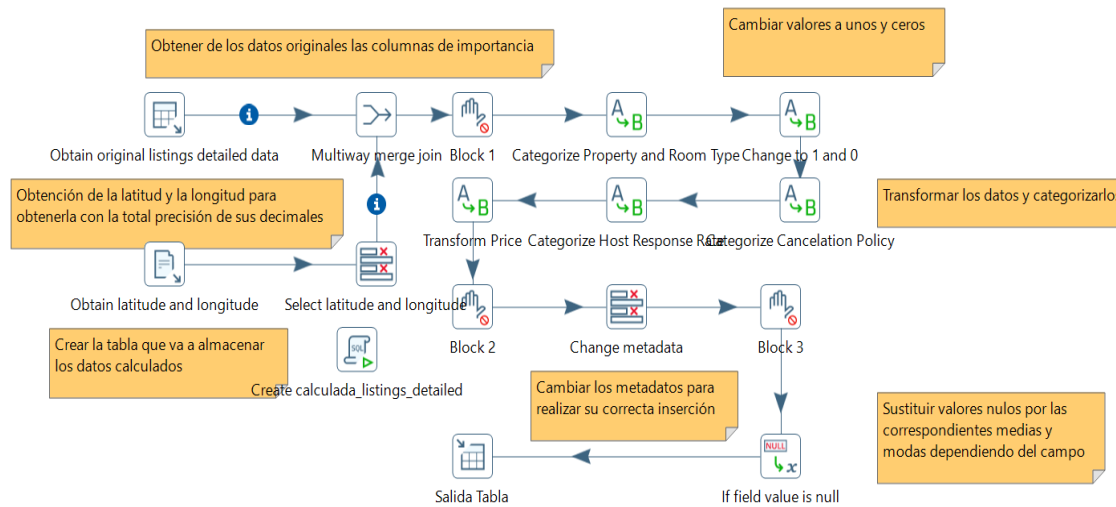


Figura 10. Proceso ETL sobre listing_detailed

En esta tabla, se ha considerado muy importante la columna *amenities*, que aunque estando en formato texto se cree que puede tener un valor importante. Por ello, se estructuran los datos de la columna *amenities* para obtener los servicios que aporta un Airbnb como por ejemplo si tiene televisión o si se puede fumar, entre otros. Tras esto, se traslada a una tabla que contiene el id y el resto de columnas en formato booleano indicando si dispone o no de determinado servicio. Algunos de estos datos se agregan en un mismo campo, por ejemplo Internet y Pocket Wifi se agrupan en la nueva columna Wifi. Para realizar estas transformaciones se ataca a la columna *amenities*, utilizando de nuevo consultas SQL, obteniendo por ejemplo los valores que contienen “Wifi” y los que no lo tienen y asignándoles unos o ceros respectivamente. Con los servicios de mayor relevancia se crea la nueva tabla llamada *calculada_servicios_airbnb*, en la cual hay un total de diecisiete columnas, que son numéricas teniendo en cuenta que si el hospedaje ofrece el servicio su valor es 1 y si no se ofrece ese servicio su valor es 0.

id	has_wifi	has_air_conditioning	is_family_friendly	has_elevator	has_pets	has_smoking	has_fire_prevention	has_wheelchair_access	has_aid_kit
integer	integer	integer	integer	integer	integer	integer	integer	integer	integer
6369	1	1	1	1	0	0	0	0	0
21853	1	1	0	1	0	0	0	0	1
24805	1	1	1	1	0	0	0	0	0
24836	1	1	1	1	0	0	0	0	0

Figura 11. Resultado de la tabla que almacena los servicios

Las tablas dos tablas de la base de datos de AirBnB relacionadas con las reviews no se van a tener en cuenta ya que la información sobre las reviews que hay en la tabla *listings_detailed* aporta mucho más valor e información sobre el apartamento. En lo que refiere a la tabla *neighbourhood*, no ha hecho falta modificarla ya que únicamente son dos columnas que describen qué barrios hay en cada zona.

Finalizado el trabajo con la información de AirBnB, se continúa con la información del catastro de Madrid, donde se tenía inicialmente 668 columnas con información muy diversa, desde la cantidad de colegios, institutos, viviendas sociales, hasta la cantidad de habitantes y sus nacionalidades o los niveles de renta y atractivo comercial. Una vez conocido el negocio, se ha hecho una primera selección de las columnas que podrían tener algún tipo de relación con el negocio de AirBnB por mínimo que fuese, eliminando así columnas repetidas como la cantidad de hombres y mujeres que viven en cada sección o columnas que no aportan ninguna información como podría ser la comunidad autónoma y provincia que para todos los registros es la misma. Acabada esta primera selección, se ha eliminado las columnas que no se prevén importantes mediante una segunda selección de los datos importantes en relación con el negocio, con el fin de tratarlos posteriormente para cruzarlos con los datos de AirBnB.

Después de estos cambios, se ha modificado la tabla dejándola lista para su uso con pentaho, aplicando diversas transformaciones. Se han eliminado las columnas que describen características de las casas, lo que no aportaba nada de valor porque conocemos las características de las casas, algunas de las columnas relacionadas eran, el número de habitaciones o antigüedad de los edificios. Otro de los cambios, ha sido quedarse con la información relativa al barrio suficiente para hacerse una idea del barrio y manteniendo información de las comunicaciones, restaurantes, supermercados, riqueza, atractivo comercial y tipo de gente que vive en la zona. Además de esto, se ha transformado la clase social a entero de menor a mayor teniendo en cuenta la economía de la sección censal, se ha agrupado las edades de quince en quince años, y se ha reemplazado todos los valores null de los indicadores por ceros.

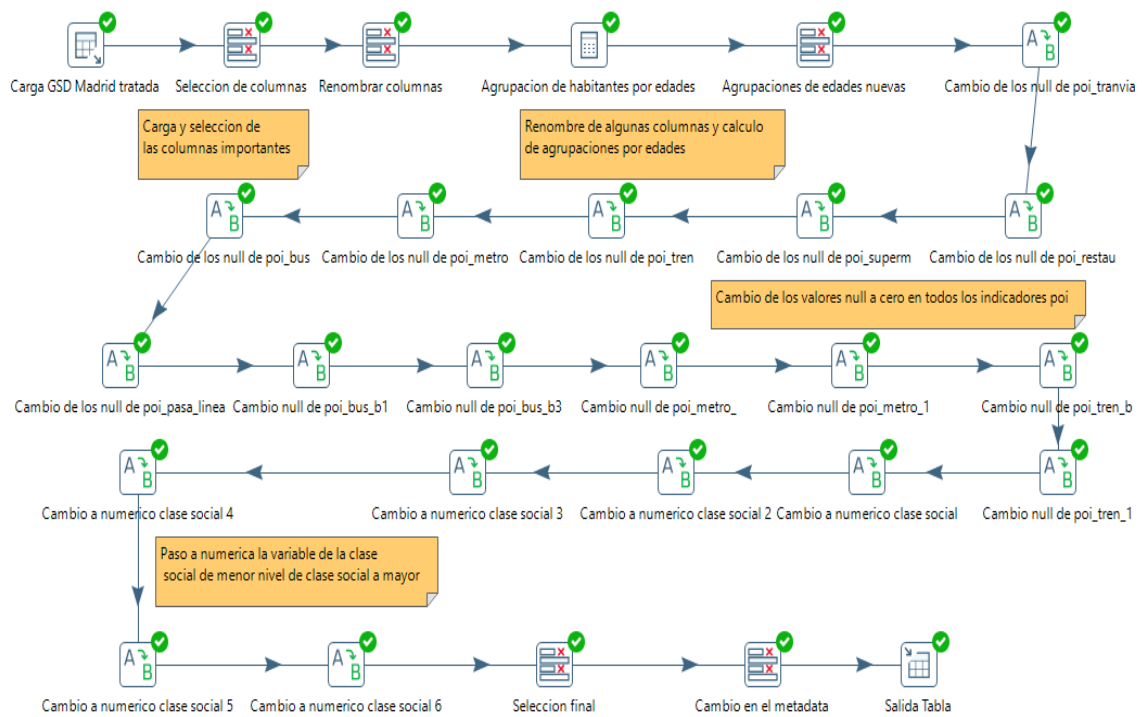


Figura 12. Proceso ETL sobre los datos de Madrid

3.- ANÁLISIS Y ENTENDIMIENTO DEL DATO

Una vez tratados los datos, se hace un análisis de los datos y entendimiento de los mismos, tanto de la información de Madrid, como de la información de AirBnB y la relación entre las dos fuentes de información. Para ello, se ha hecho uso del programa QGIS, PostgreSQL que es donde tenemos almacenada la información y PostGIS 2.0, que es la extensión que permite conectar ambas.

En relación con el estudio de la información de Madrid, la cual dispone de una columna de geometría llamada geom, que describe los distintos polígonos formados por cada sección censal correspondiente a cada fila de la base de datos. Haciendo uso de las conexiones de QGIS, se cargan los datos para aplicarles distintos filtros extrayendo así más valor, que además se van a situar en capas encima de mapas de la ciudad de Madrid, concretamente el mapa OpenStreetMap que ofrece QGIS.

Una vez creadas las capas, se pretende sacar más valor a la información conociendo la distribución de las distintas zonas según su economía. Para ello, se categoriza la capa de las distintas secciones censales en función de la columna de la variable *clase social*, dando así color azul a las zonas que son de clase social baja, llegando hasta el color rojo en las secciones de clase social alta, dando lugar así a claras conclusiones acerca de los distintos niveles de vida en cada zona. Destacan como zonas de alta clase social, la zona centro de Madrid y la zona está caracterizada por ser zonas residenciales de chalets como en Pozuelo de Alarcón, Las Rozas, entre otros municipios. Por el contrario, hacia la zona sur del centro de Madrid, donde se encuentran municipios como Entrevias, Villaverde, Getafe, Leganés, son zonas en las que la gente que habita son de clase social baja. También, las zonas del externas de la comunidad de Madrid exceptuando en la parte de la Sierra por el este, son generalmente zonas de clase media baja. Se puede observar en la siguiente imagen.

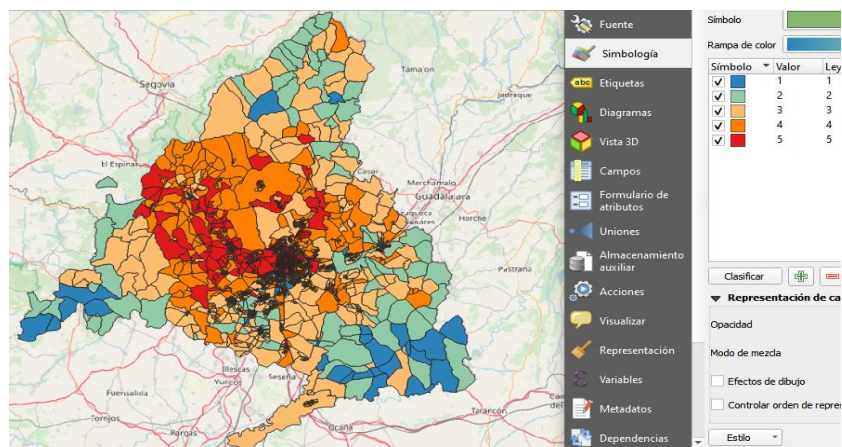


Figura 13. Clasificación de las distintas zonas de Madrid según clase social

Seguidamente, se ha estudiado la distribución del tránsito de las personas por la comunidad, con un resultado esperado destacando las centro de cada una de las localidades como Madrid centro, Leganés o Getafe. Pero si se especifica el tipo de tránsito, como en el caso del tránsito comercial y de ocio donde tiene valores más altos es en ciertas zonas del centro de la ciudad de Madrid, y son muy similares.

Tras el estudio de tránsito, otro aspecto importante es el precio del metro cuadrado de las viviendas por sección censal. Por lo tanto, se gradua por el precio, sacando en claro que la zona centro de Madrid junto con la parte norte del centro de la comunidad son las zonas en las cuales el precio por metro cuadrado es más caro. El precio generalmente varía junto con la distancia al centro de la ciudad y al centro económico de la ciudad. La ilustración que se ha obtenido con QGis es siguiente resultado:

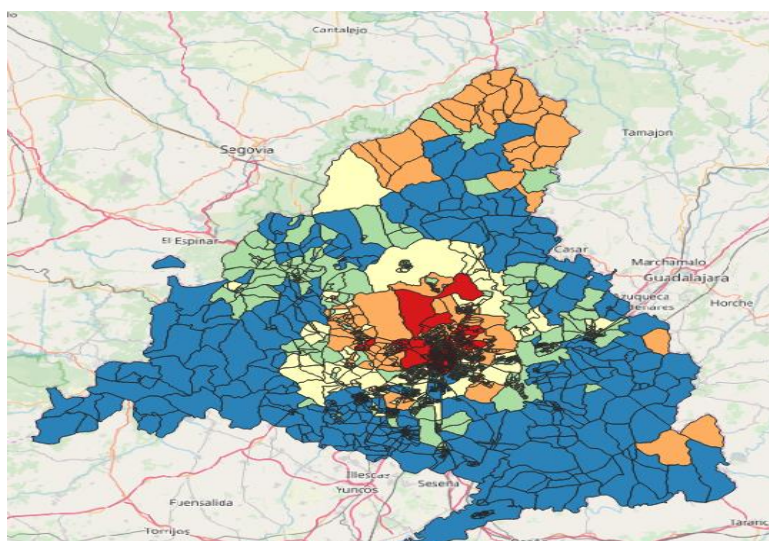


Figura 14. Clasificación de las zonas en función del precio del suelo

Para finalizar con los datos de Madrid por separado, la distribución de los extranjeros y españoles por el territorio es puede aportar una descripción de los distintos barrios. Para ello, se gradúa el número de extranjeros de cada zona. Como se puede observar en la siguiente imagen, la cantidad de extranjeros en las zonas más caras de Madrid es mucho menor que en el extrarradio, donde hay algunas secciones censales de comunidades de gente extranjera como en Villanueva de la Cañada, de un perfil más adinerado, o en Coslada y alrededores de un perfil menos adinerado.

Después de tratar superficialmente los datos de Madrid se ha realizado el estudio de los pisos de Airbnb, creando su geometría a partir de la latitud y longitud que se encuentran en la tabla de listings_detailed. Así situando cada Airbnb se comienza su estudio.

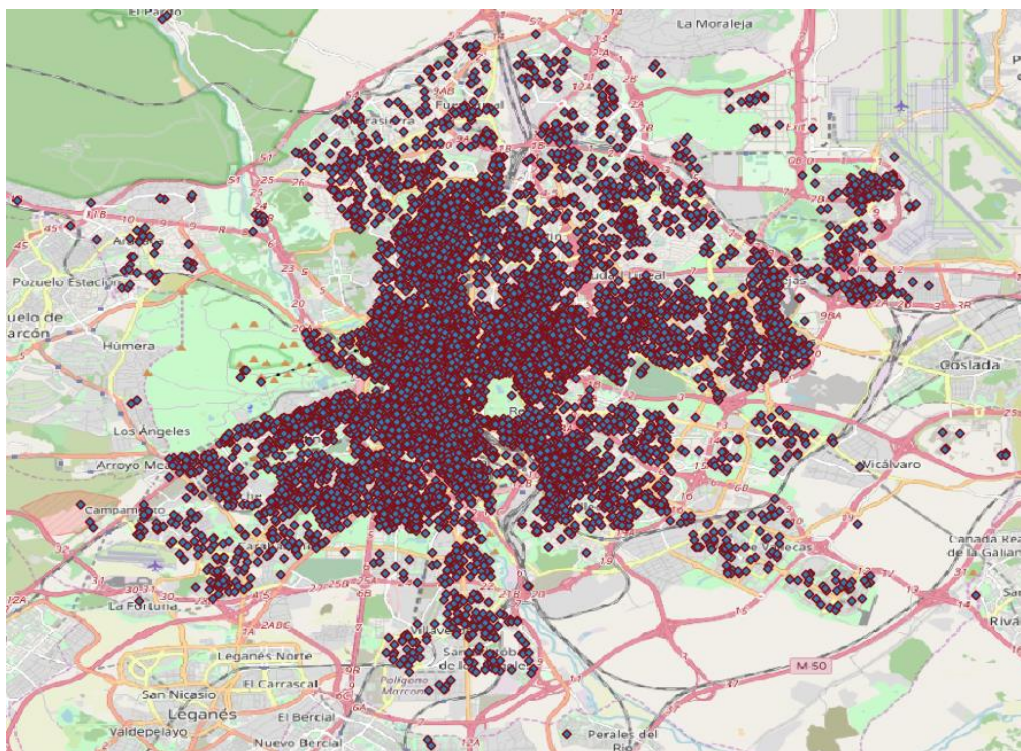


Figura 15. Localización geográfica de cada uno de los alojamientos

Una vez creadas estas capas, se analiza los datos que se consideran más interesantes como el precio del alquiler, los pisos con mayor puntuación, los valores más comunes dentro de nuestros datos entre otros rasgos, primero se realiza un estudio superficial de los precio de los Airbnb's.

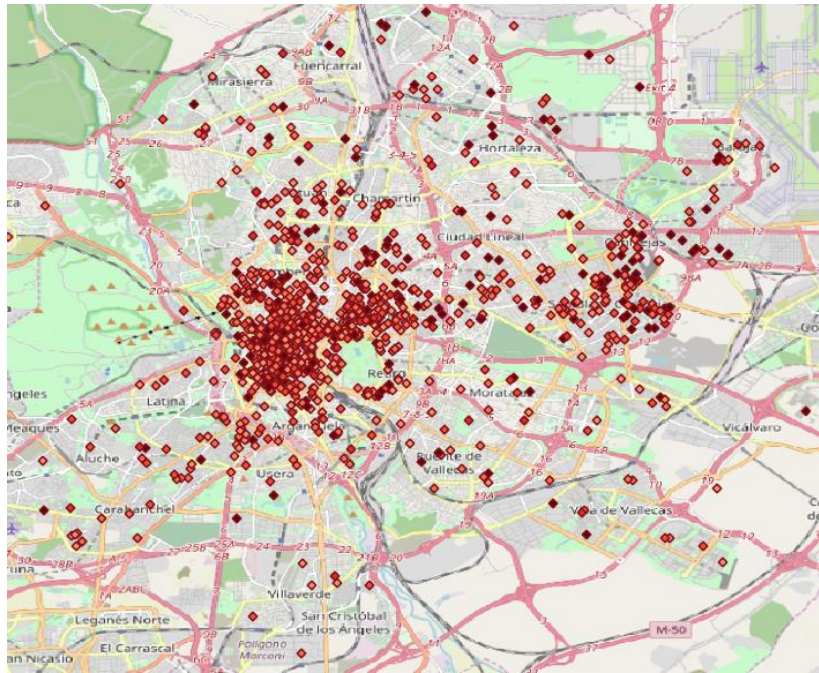


Figura 16. Localización de los alojamientos con precios altos

Como se podía prever, los pisos con mayor precio de alquiler (por encima de 210 €) se encuentran en el centro de Madrid y en otros barrios con mayor riqueza, aunque también se ha visto en la anterior imagen que la acumulación de los pisos es mucho mayor en el centro, ahora se realiza un pequeño estudio de los pisos más baratos por debajo de (210 €).

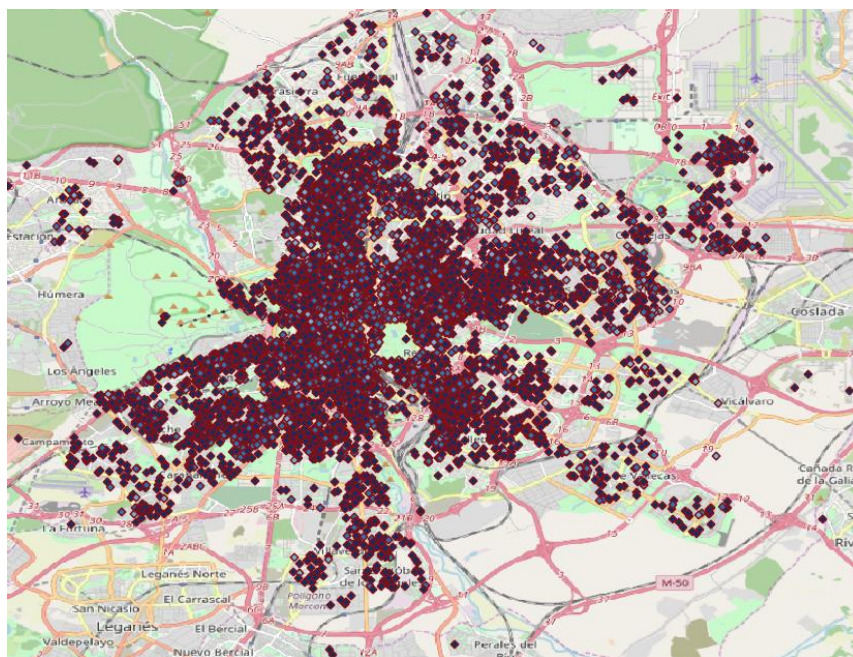


Figura 17. Localización de alojamientos más económicos

Aparecen bastantes más AirBnB en barrios más humildes con menor alquiler, pero se aprecia una gran acumulación en el centro de Madrid. Además, se puede observar como la mayoría de los pisos se encuentran en el rango de menor riqueza.

Tras ello, se analiza qué rango de precios ocupa la mayoría de los pisos. Unificando los códigos postales de cada alojamiento se ha hecho un promedio del precio por código postal. La gran mayoría de los pisos se encuentra en torno a los rangos de 15 a 130 euros.

Posteriormente, analizando los posibles factores que afectan se ha comenzado por los servicios que aportan los pisos en cuestión empezando por los pisos de mayor valor. Esta información de cada piso se nutre de la tabla *calculada_servicios_airbnb*. En la cual se estudia los servicios que ofrece cada alojamiento siendo 1 si lo ofrece y 0 en caso de no ofrecerlo. En la siguiente gráfica, se puede observar la dispersión de precios de los alojamientos en función de los servicios, sacando conclusiones como tener una localización exacta del alojamiento (marrón) tiende a subir más su valor, un factor influyente también es si tiene wifi (morado claro) o no, si es o no apto para un ambiente familiar (resaltado en verde claro).

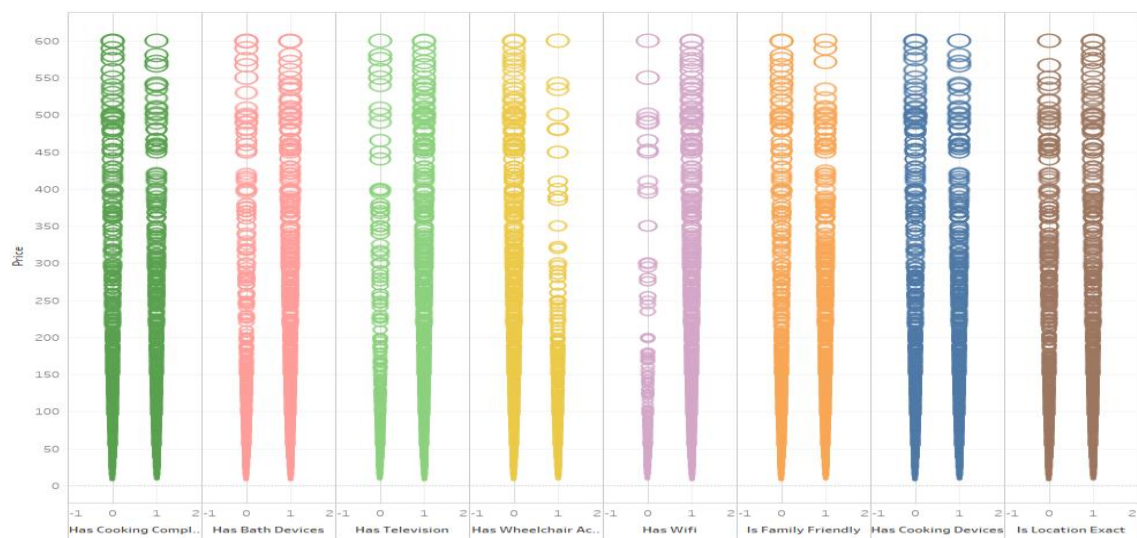


Figura 18. Precio de los alojamientos según sus servicios

Aunque para verificar que esta dispersión no son valores atípicos se realiza un estudio un poco más minucioso respecto a la disponibilidad de los servicios frente a los precios. Analizando los servicios que se consideren más relevantes de los generados por la ETL, empezando primero por los pisos económicos con un rango de hasta 70€.

Respecto a los alojamientos de precio más económico no se aprecia una gran diferencia de precios respecto a los servicios. Los únicos dos aspectos que aportan relevancia en el precio son si la casa incorpora dispositivos de baño como secadores, o si el Airbnb dispone de televisión.

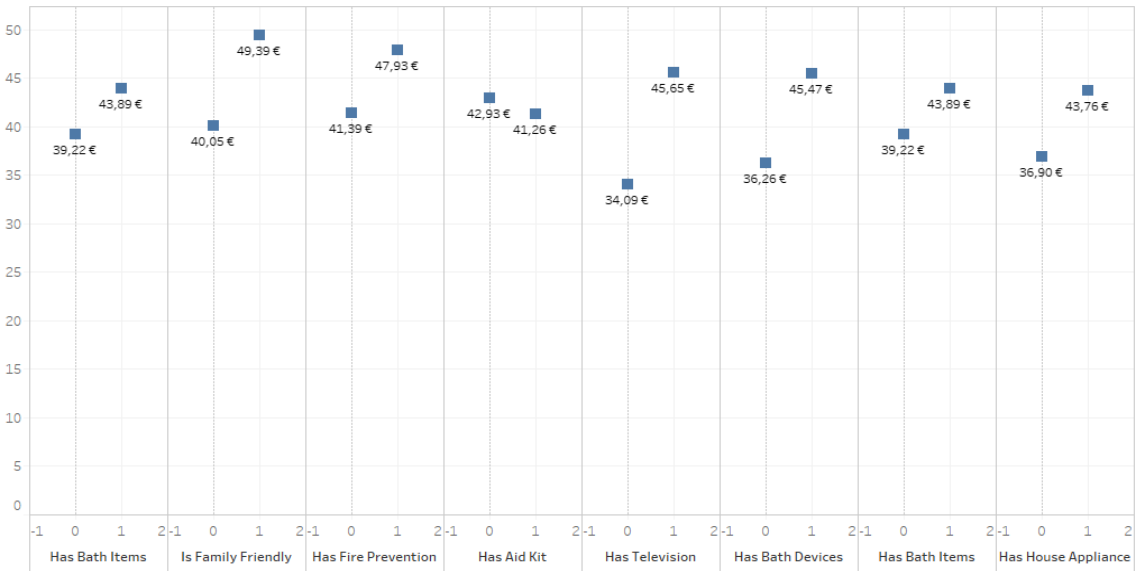


Figura 19. Cambio de precio por servicios en los alojamientos con precio menor a 70€

Aumentando a un rango de precios un poco superior, de 70€ hasta 170€, los servicios que disponga un alquiler prácticamente no afecta a los precios, esto se tendrá cuenta cuando se realicen las predicciones de precios puesto que alrededor de un tercio de los pisos se encuentra en este rango de precios.

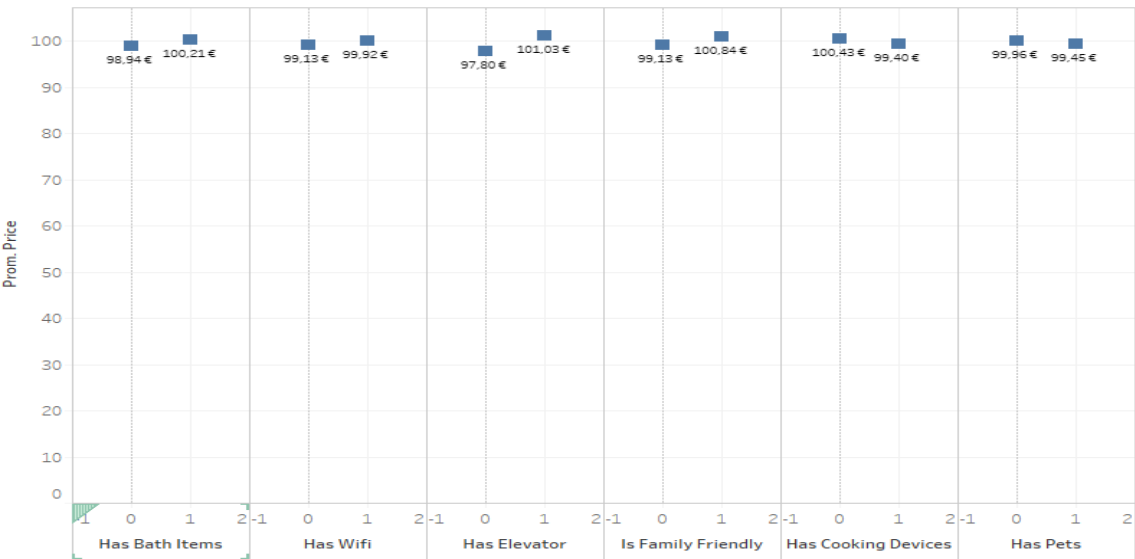


Figura 20. Cambio de precio por servicio en alojamientos entre 70€ y 150€

En cambio, en el conjunto de pisos algo más caro si que se empiezan a notar pequeñas diferencias, de entre diez a veinte euros en el precio, siendo de mayor importancia que el piso incorpore televisión, que disponga de silla de ruedas y que traiga aparatos de baño o vajilla.

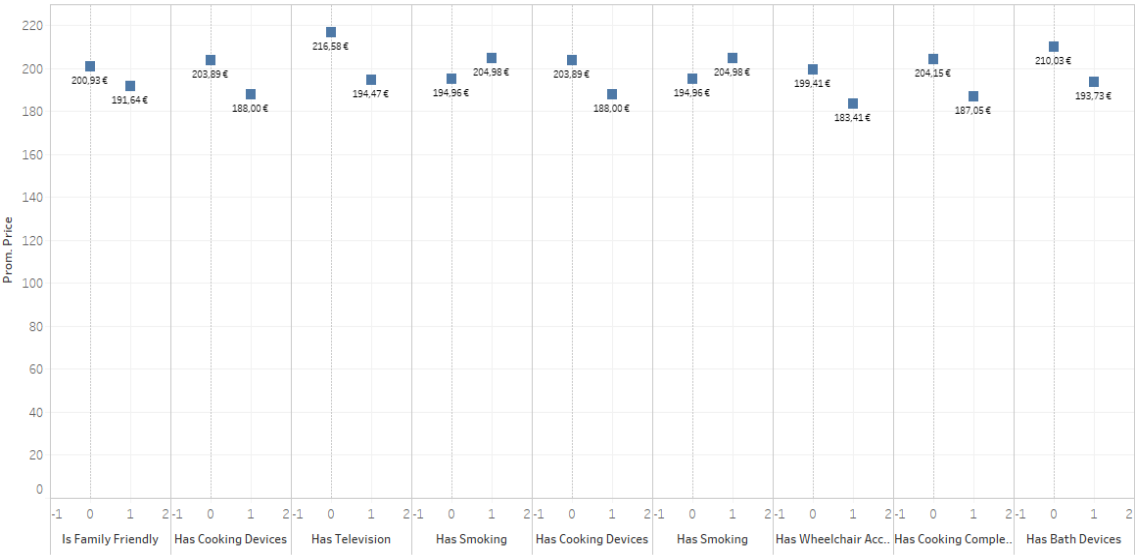


Figura 21. Cambio de precio frente a los servicios en alojamientos de 150€ o superiores

Finalmente, estudiando los servicios en general frente a cualquier tipo de hospedaje, su influencia se concluye que los servicios más importantes son si dispone de televisión, si tiene wifi, si es un entorno amigable para ir familias y si está equipado con complementos de baño y cocina.

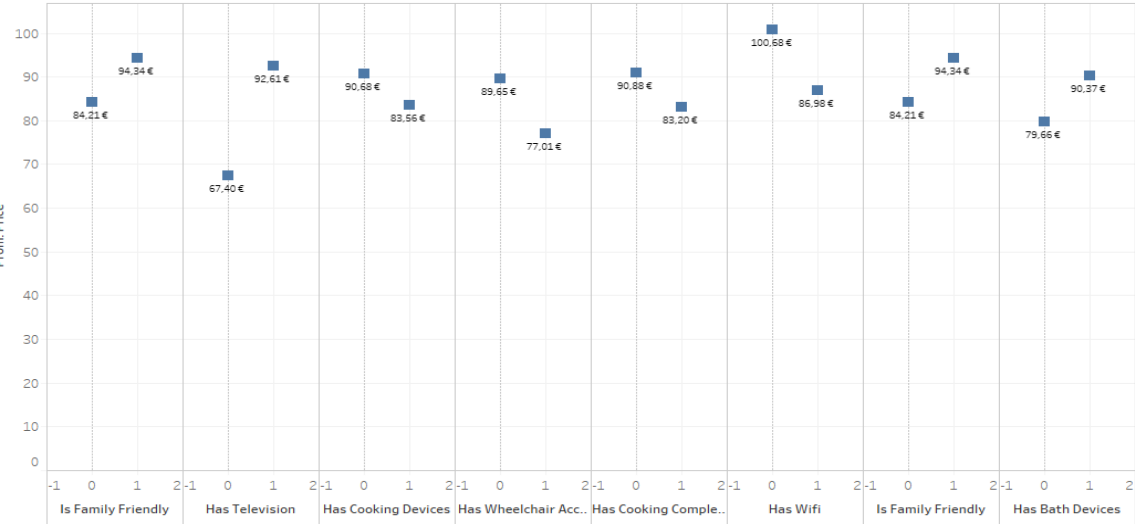


Figura 22. Cambios de precio más significativos debido a los servicios

A continuación se ha procedido a analizar las tasas de pago por dañar objetos del piso, tasas de limpieza y tasas de invitados agrupado por los distintos distritos de Madrid, y se observa como los promedios de las tasas pueden estar correlacionados con los precio de cada piso.

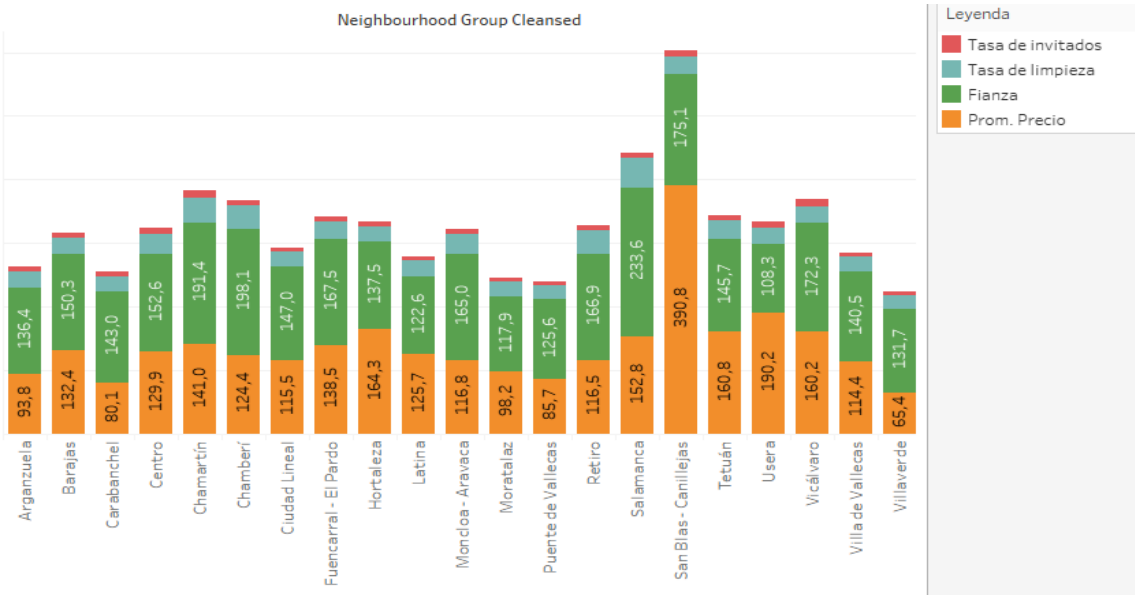


Figura 23. Relación entre los distintos pagos por barrio

Además se ha analizado la tasa de invitados frente al promedio del precio agrupado por el número de invitados (eje x) incluidos con el precio promedio tasa de invitados y se puede observar como siguen tendencias parecidas.

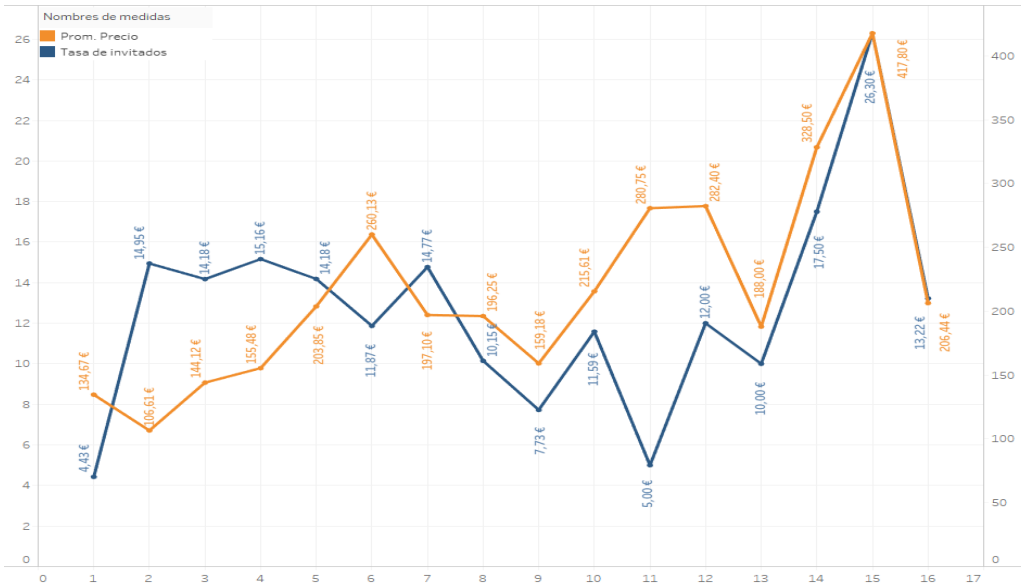


Figura 24. Comparativa de tendencias de la tasa de invitados y el precio

Acabado el análisis superficial del precio, se pasa a estudiar el comportamiento de la satisfacción de los clientes mediante la valoración que le dan a los alojamientos de 0 a 100, siendo 100 la mejor experiencia. Inicialmente se estudia las valoraciones de los clientes organizada por código postal. La mayoría de códigos postales tienen muy buena puntuación excepto el 28524.

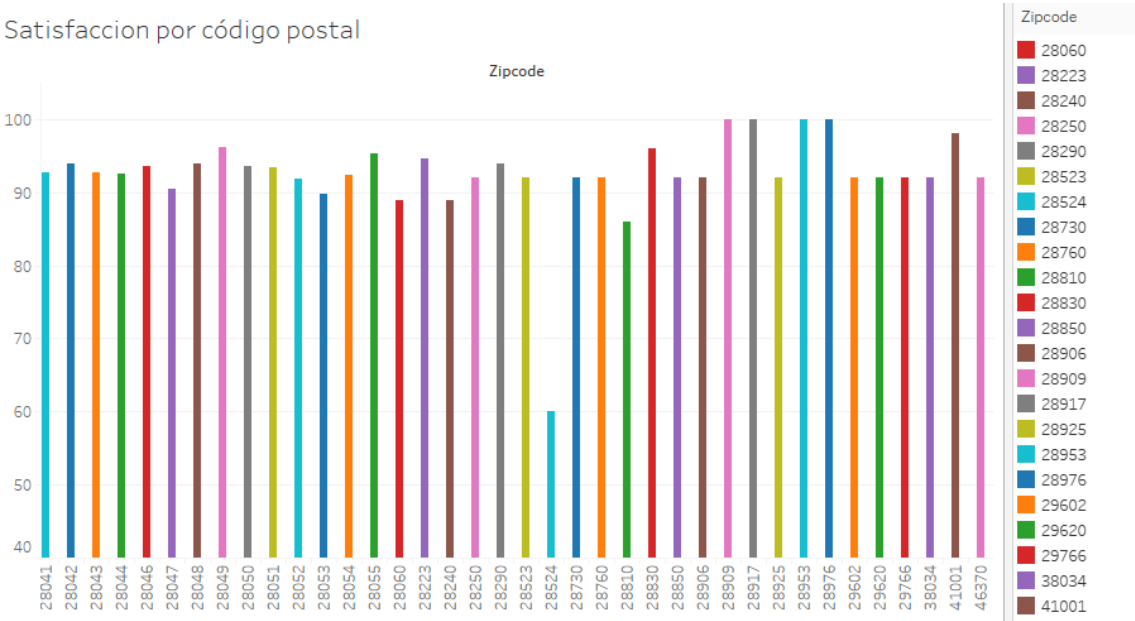


Figura 25. Valoración promedio sobre 100 por código postal

Otra gráfica que puede tener relevancia es la valoración de los clientes por tipo de propiedad. En dicha gráfica se puede ver como las casas y apartamentos tienen una media un poco mayor que el resto, pero los promedios son similares.

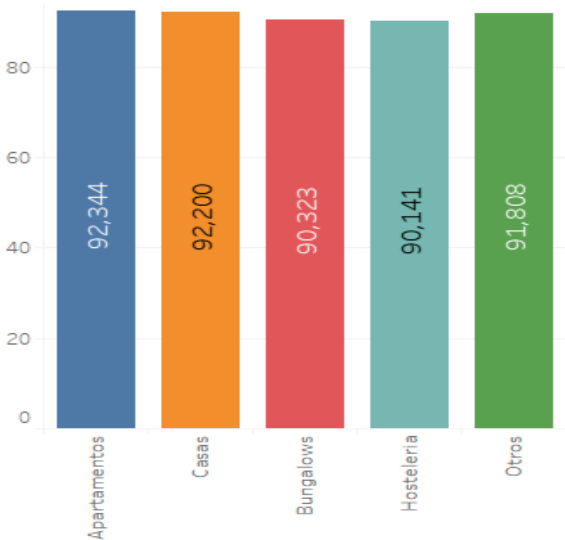


Figura 26. Valoración promedio por tipo de propiedad

Después de varias gráficas, sin sacar grandes conclusiones de cómo cambia la satisfacción que tienen los clientes es necesario conocer las características de la zona del alojamiento, lo que puede ser algo determinante. Para ello, con el fin de hacer un análisis descriptivo más profundo, se han juntado las dos fuentes de datos, la que tiene información de las secciones censales de Madrid y la información de los apartamentos de AirBnB en Madrid, haciendo uso de las herramientas de QGIS para unir atributos mediante la localización. Una vez cargadas las dos capas, se duplica la capa de los apartamentos que es sobre la cual se hace la unión. En este paso, se hace el join teniendo en cuenta si el punto descrito de con la localización del apartamento está contenido en la sección censal de Madrid. Esta agrupación proporciona los datos unión de los datos en una única capa lista para su filtrado y análisis.

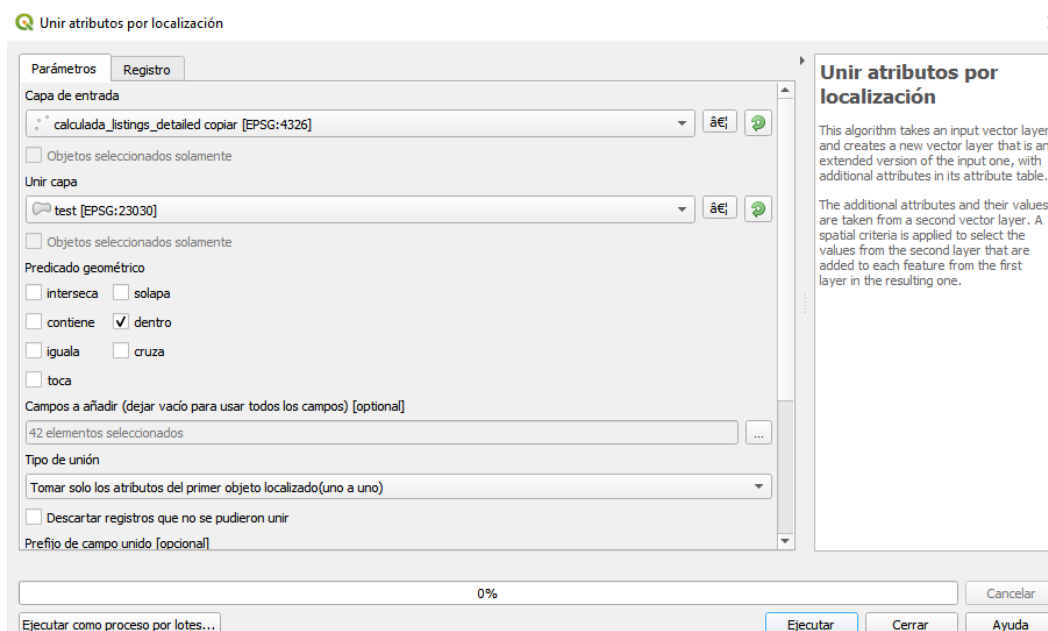


Figura 27. Configuración del joint espacial

Una vez juntas, se comienza aplicando una rampa de color basada en el precio para la categorización de las distintas casas según lo que cuestan. Tras ello, se han aplicado una serie de filtros para ver cómo influyen dichos filtros en los cambios. Entre dichos filtros visualmente se aprecia claramente la influencia de la cantidad de habitaciones y de camas en el precio de alquiler. Otra característica que proporciona valor a las casas es la proximidad a las zonas de ocio, y atractivo comercial como son las zonas del centro de la ciudad. Sin embargo, la riqueza del barrio influye menos de lo esperado, por ejemplo,

con una riqueza de 0.75, se observa una cantidad bastante grande de casas “baratas”. En las siguientes imágenes se puede observar dichos cambios que inician con el mapeado del precio por noche se obtiene:

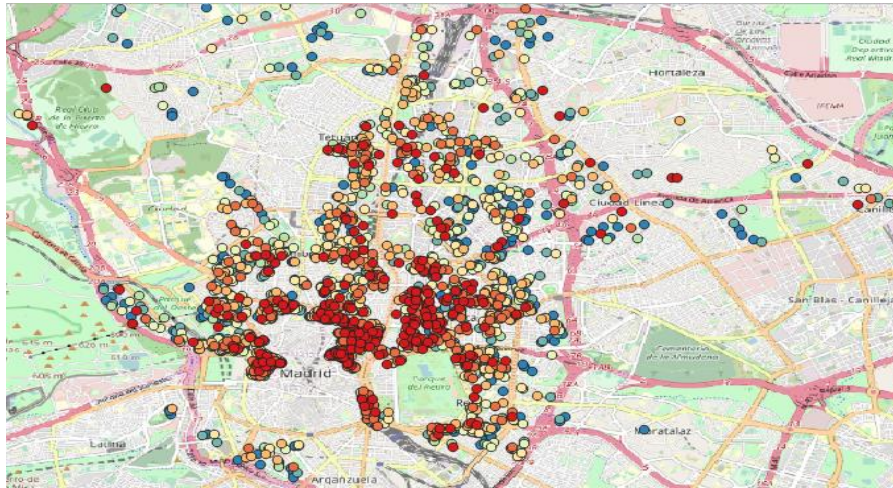


Figura 28. Alojamientos en zonas de riqueza mayores a 0.75

No obstante, para conseguir mayor entendimiento de las dos variables más importantes del sector que son el precio del alquiler y la satisfacción del propio cliente, se ha enfrentado dichas variables frente a las variables que tenemos en la base de datos. En primer lugar se presenta la evolución gráfica de las variables relacionadas con el precio de alquiler por noche mediante el programa Tableau Desktop. De todas las gráficas iniciales las que aportan más relevancia son las siguientes:

1.- Número de camas frente al precio de alquiler: En la gráfica, se aprecia el crecimiento del precio con el aumento de las camas hasta cierto punto.

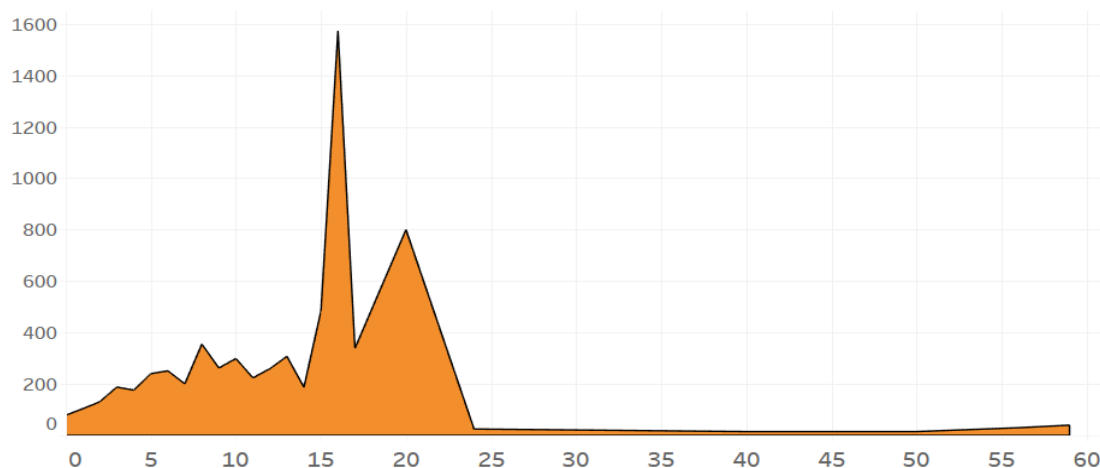


Figura 29. Precio frente al número de camas

2.- Número de habitaciones frente al precio de alquiler: De igual manera que en el caso anterior, la tendencia es que cuantas más habitaciones tenga la casa cueste más. Es importante destacar que en el dataset que se tiene no hay ninguna casa de entre 11 y 49 habitaciones respectivamente.

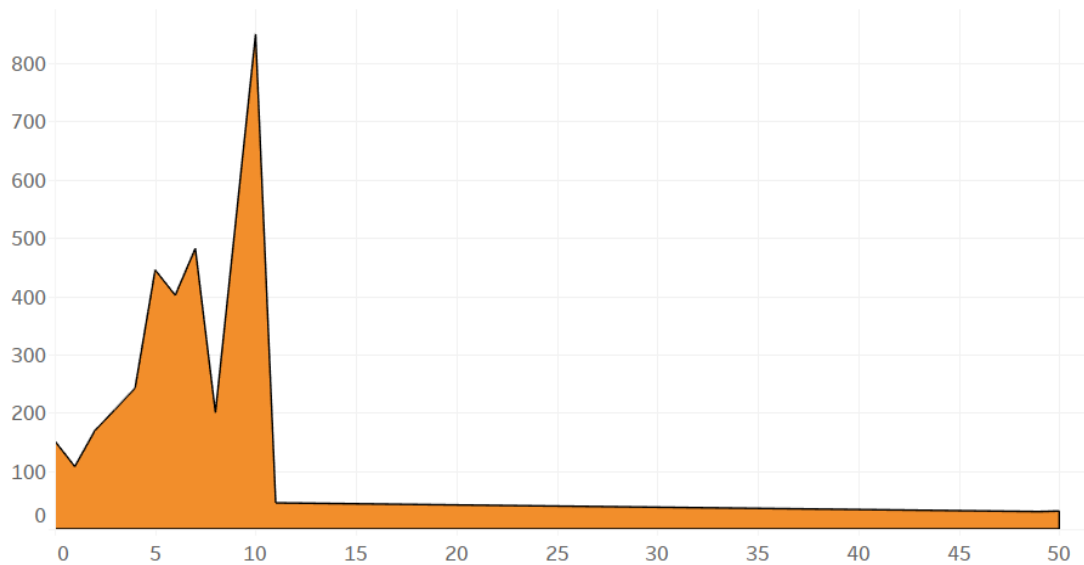


Figura 30. Precio frente al número de habitaciones

3.- Precio de alquiler frente al número de reviews: El precio de las casas va descendiendo en función de la cantidad de reviews que tiene cada casa. Lo cual describe o que las casas con más reviews dejan a los huéspedes muy satisfechos en función de lo que han pagado, o por el contrario, que no se han quedado contentos con el hospedaje esperando mucho más por el alquiler contratado.

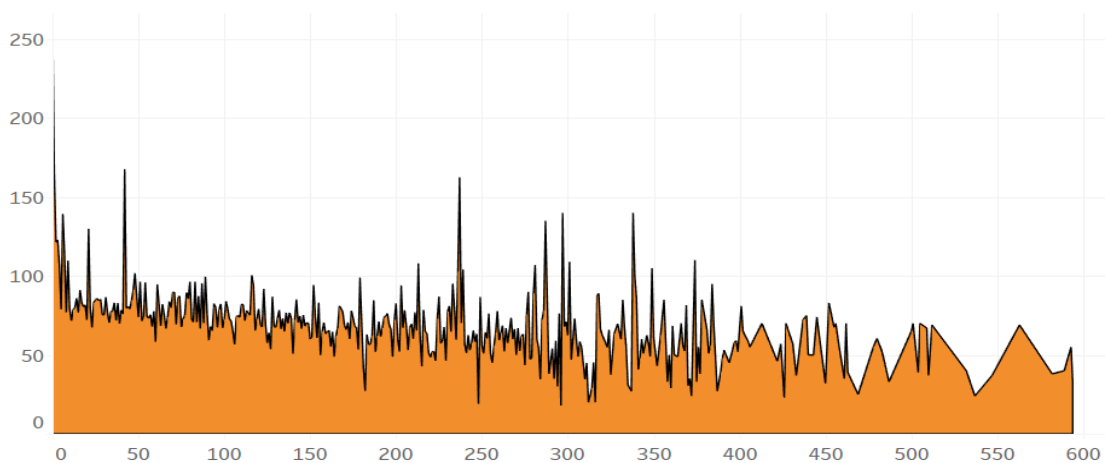


Figura 31. Precio promedio frente al número de Reviews

4.- Precio frente al tipo de propiedad: según las agrupaciones hechas en la parte de carga y transformación del dato, se puede apreciar que las más caras son los bungalows y casas de campo (grupo 3). A continuación, la categoría formada por los hoteles, hostales y similares que se ofertan en AirBnB (grupo 4). Las casas (grupo 2) en tercer lugar seguido del grupo de los apartamentos (grupo 1). Por último, el resto de tipos de propiedades que se alquilan en el que también se incluyen las habitaciones.

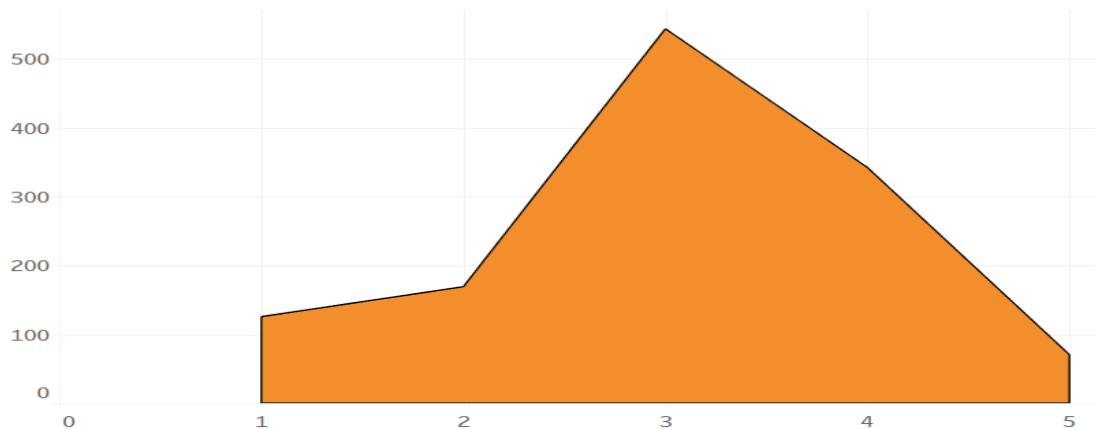


Figura 32. Precio promedio por el tipo de alojamiento

5.- Precio frente a la edad de la población: En esta agrupación de gráficas por edades, se puede observar que la distribución ideal de las edades de la población de una zona respecto a los precios de alquiler es una distribución estilo campana de Gaus. En dicha distribución habría pocos niños, una cantidad media de jóvenes, la mayoría de la población entre 30 y 59 años, y un descenso de la cantidad de gente con el aumento de la edad.

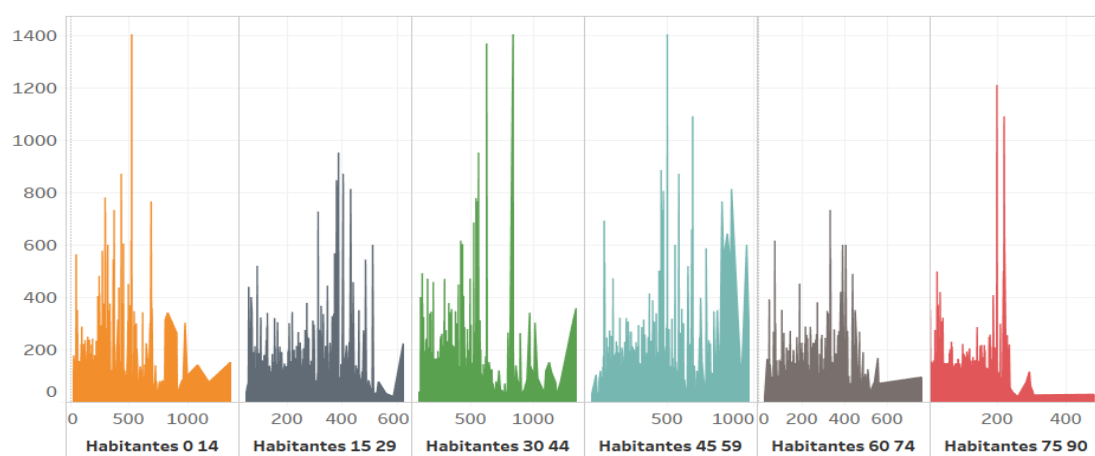


Figura 33. Precio promedio frente a la edad de la gente de la zona

6.- Precio de alquiler frente al tránsito: No es muy influyente el tipo de tránsito en el precio, sin embargo, a cuanto mayor tránsito menor es el precio medio de los alquileres.

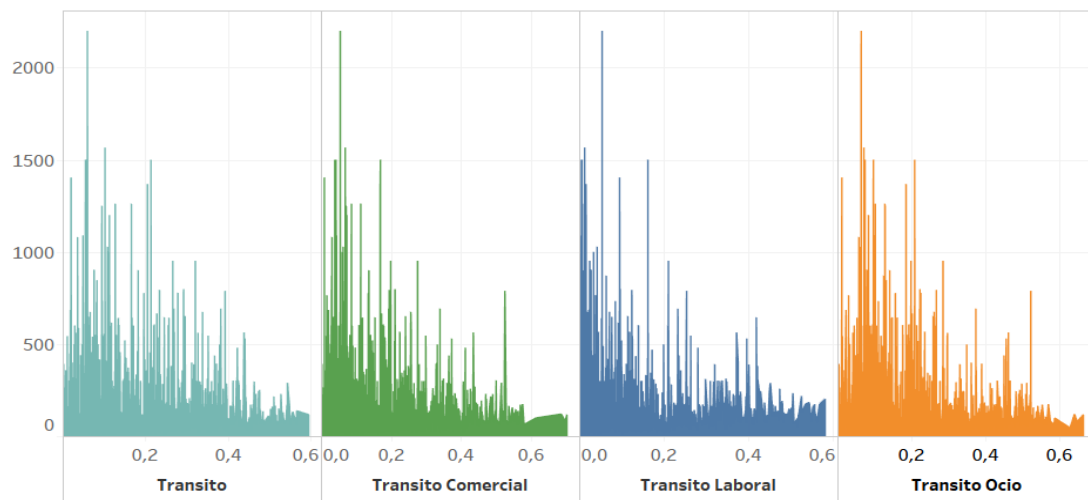


Figura 34. Precio promedio en función del tránsito de la zona

7.- Precio de alquiler frente a la valoración total: Las peores notas, tienen precios altos lo que puede ser que los clientes no están satisfechos con su estancia con el desembolso que han hecho. Sin embargo, las mejores notas son con precios más bajos proporcionando al cliente una satisfacción mayor según el desembolso.

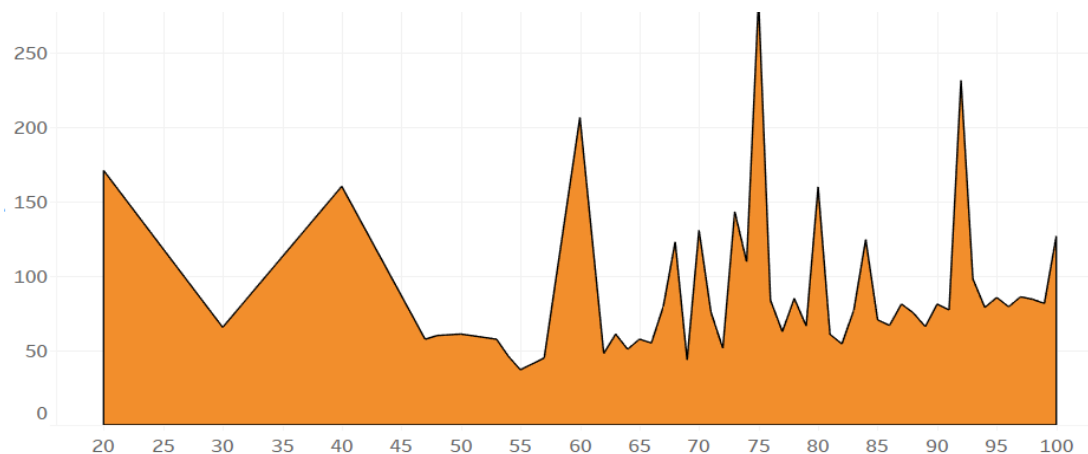


Figura 35. Precio promedio frente a la valoración del cliente

Estas son las gráficas de las cuales se han obtenido conclusiones claras de su repercusión en el precio. Para nuestra sorpresa, algunas gráficas como la que relaciona el precio de metro cuadrado con el precio de alquiler que se pensaba que iba a tener una gran importancia, no la tienen ya que no se acercan a ningún patrón. El resto de gráficas que

se han hecho para el estudio enfrentando el precio respecto a las variables: atractivo comercial de la zona, nacionalidades de la zona, precio del metro cuadrado de la zona, entre otras.

Después de los análisis gráficos respecto el precio, se analiza la satisfacción de los clientes mediante la valoración que ellos mismos ponen a los alojamientos, para sacar conclusiones de los aspectos más importantes.

1.- Valoración del cliente frente al atractivo comercial: Generalmente, con un alto atractivo comercial se obtienen mejores valoraciones, por lo que las zonas comerciales están mejor valoradas por los clientes.

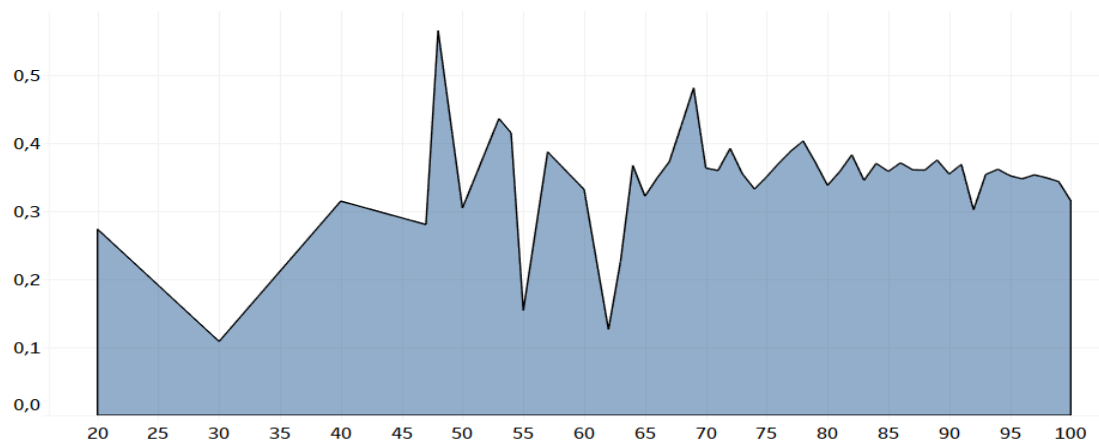


Figura 36. Atractivo comercial frente a la valoración de los clientes

2.- Valoración del cliente frente al número de camas: Los hospedajes que mejor valorados están por los clientes suelen tener dos camas, lo que proporciona una idea de que pueden ser parejas que viajen juntas, familias con un único hijo.

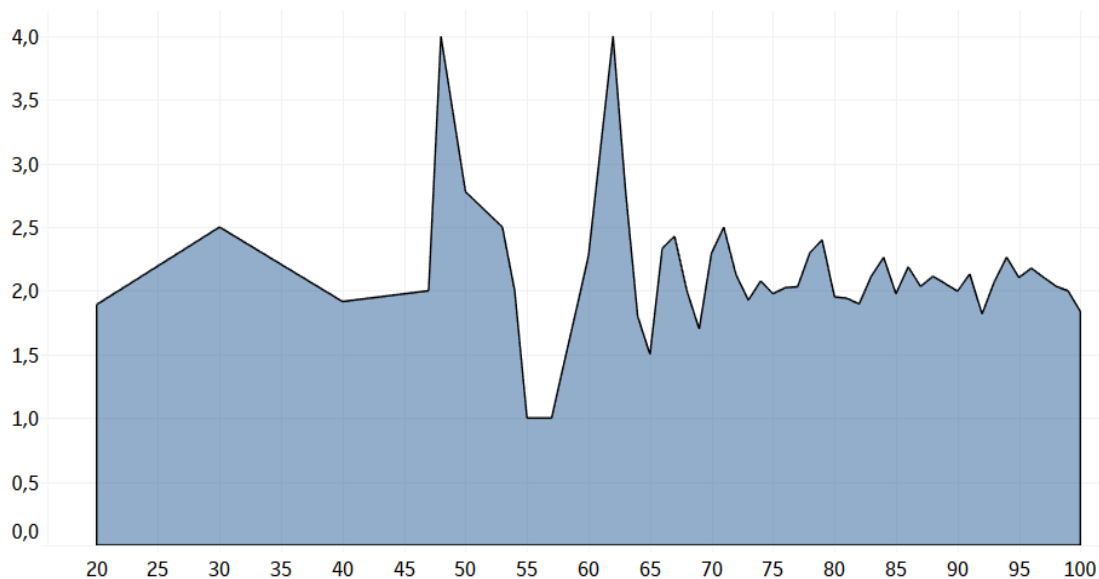


Figura 37. Número de camas frente a la valoración del cliente

3.- Valoración del cliente frente al porcentaje de extranjeros: El porcentaje de extranjeros se ha calculado mediante un campo calculado de Tableau, dividiendo los extranjeros entre los españoles y poniéndolo en porcentaje. Las mejores valoraciones tienen en común que generalmente hay un porcentaje de extranjeros entre el 20% y 25%.

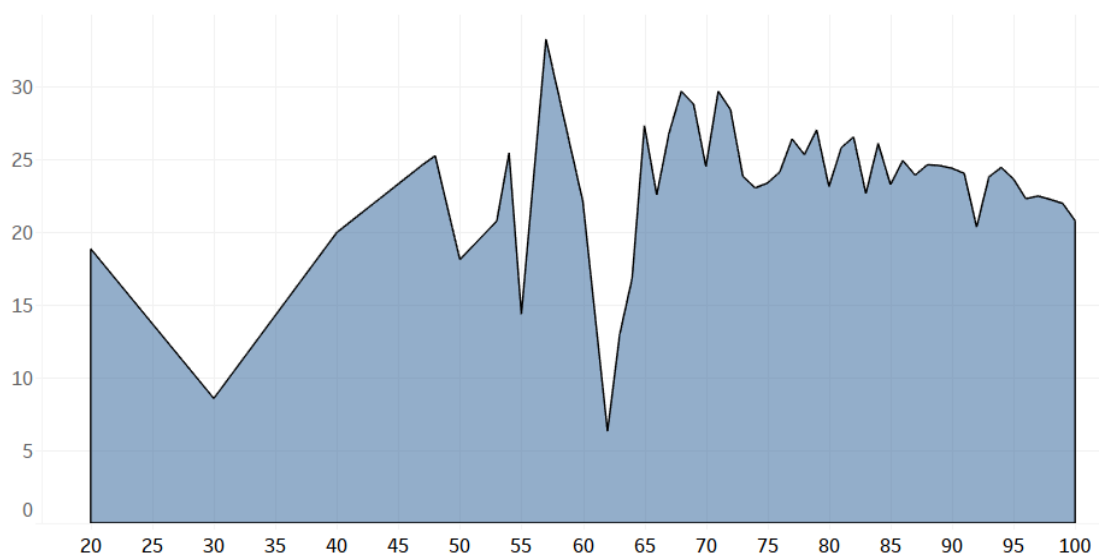


Figura 38. Porcentaje de extranjeros en la zona frente a la valoración de los clientes

4.- Valoración del cliente frente al tránsito: Al igual que en caso del precio de alquiler frente al tránsito, no afecta el tipo de tránsito pero las propiedades mejor valoradas comparten un tránsito aproximado de 0.3, mayoritariamente de ocio y comercial. Como

consecuencia, se entiende que son zonas cercanas a las de mayor tránsito y que dan opción a hacer actividades cerca.

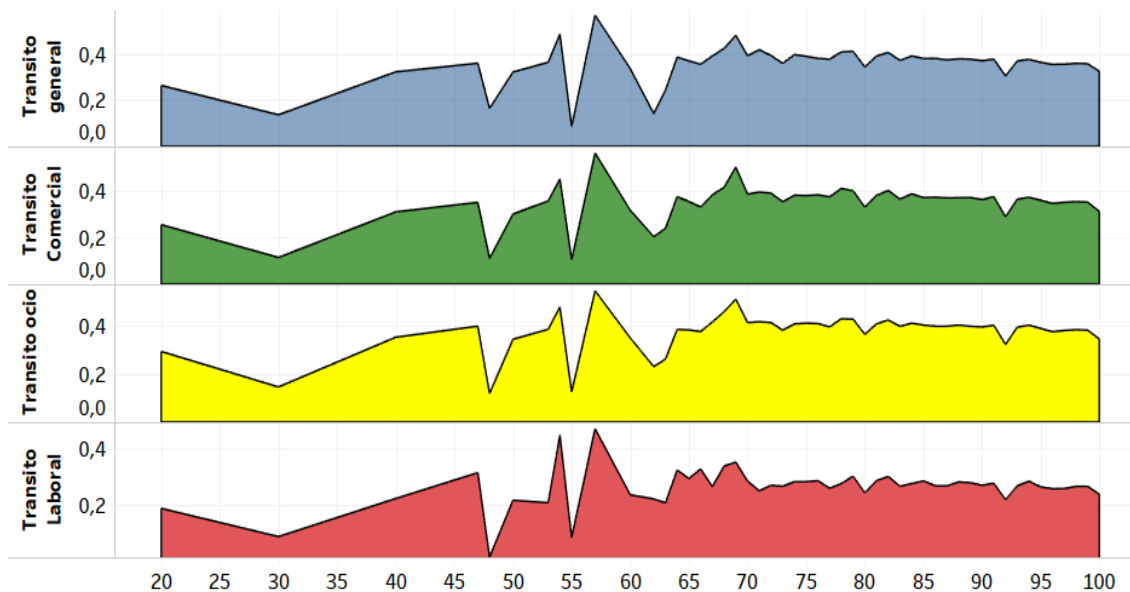


Figura 39. Tránsito de la zona frente a la valoración del cliente

5.- Valoración del cliente frente al turismo: Las zonas turísticas suelen valorar en los clientes, pero más que eso, se valora negativamente las zonas de alojamientos que tienen poco turismo. Esta variable por sí sola no va a determinar la valoración, pero junto con la anterior describen claramente que las zonas que más gustan son zonas con ambiente, que se pueda hacer planes, y se pueda hacer turismo.

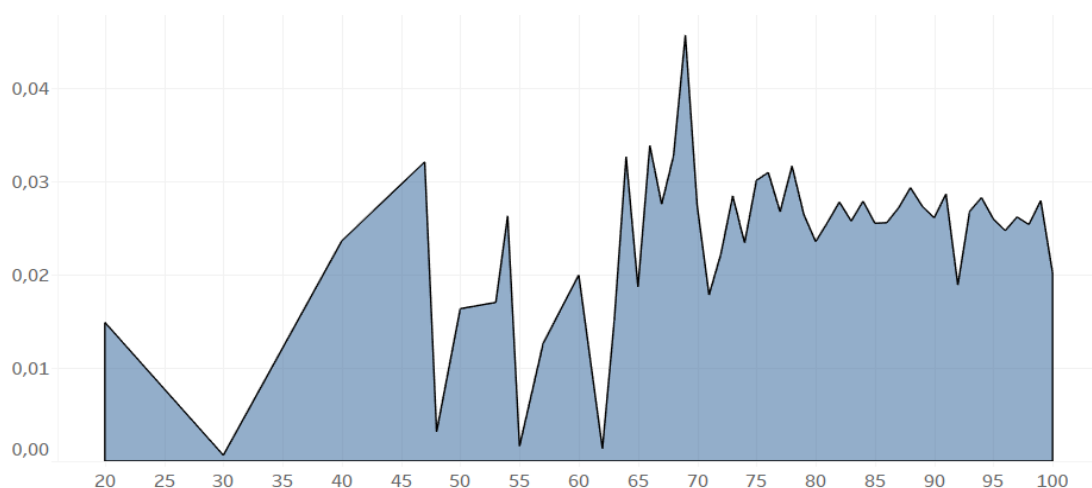


Figura 40. Turismo de la zona frente a la valoración

6.- Valoración del cliente frente al precio de alquiler: La gente generalmente se queda más satisfecha con una buena relación calidad - precio, ya que sus expectativas no son tan altas y al finalizar la experiencia están muy satisfechos.

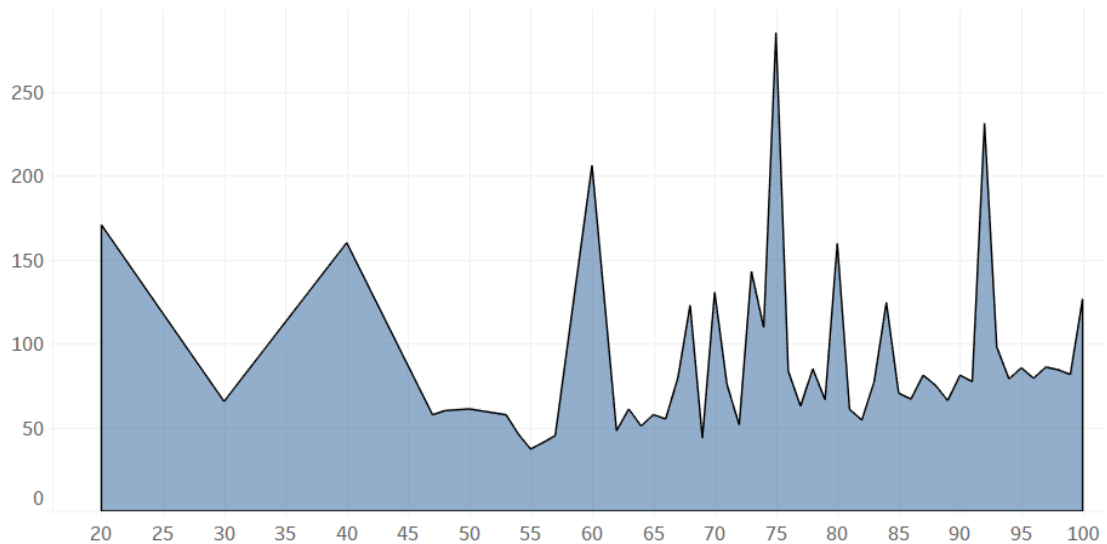


Figura 41. Precio frente a la valoración del cliente

7.- Valoración del cliente frente al número de reviews: La cantidad de reviews que tiene el alojamiento es acorde a la valoración del mismo, cuantas más opiniones mayor valoración. Los clientes se detienen a poner una opinión de satisfacción que de descontento.

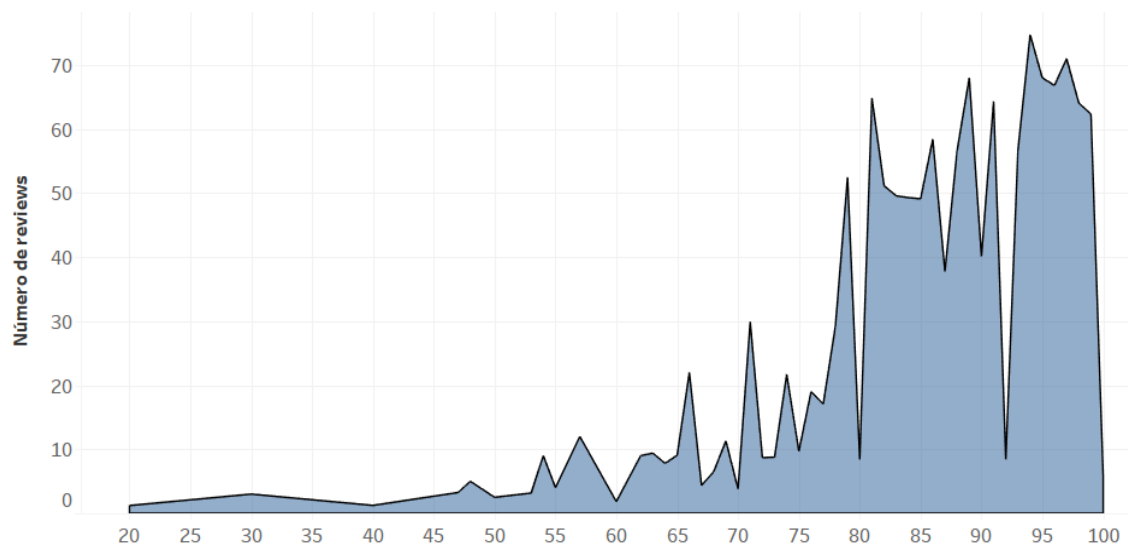


Figura 42. Cantidad de reviews frente a la valoración de los clientes

8.- Valoración del cliente frente al número de habitantes de la zona: Las zonas de grandes aglomeraciones de personas no son bien valoradas. Ayuda a describir la zona ideal pero no aporta gran relevancia a simple vista.

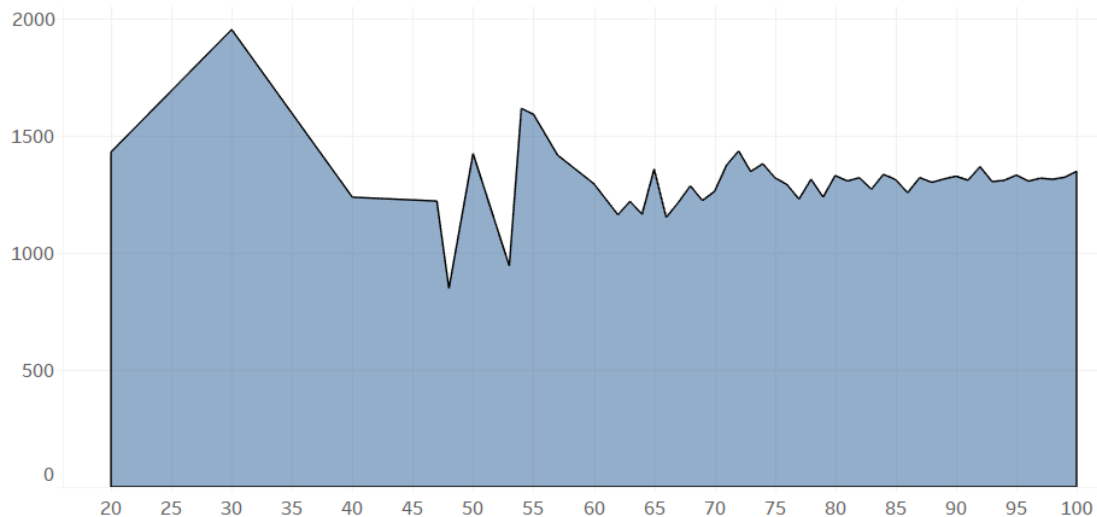


Figura 43. Número de habitantes frente a la valoración de los clientes

Además de las anteriores gráficas se han hecho más enfrentando otras variables, como el porcentaje de mujeres y hombres que hay en la zona, la riqueza de la zona, precio del metro cuadrado, entre otras pero visualmente no aportan valor al entendimiento.

Analizando los datos, haciendo una consulta buscando en las descripciones de las casas las palabras final o champions, se ha visto los alojamientos con precios desorbitados que se han dado en la plataforma. Los precios van desde 5000 hasta 27 pero analizando las ofertas las más baratas que son pocas se cree que estas ofertas no fueron coincidentes temporalmente con la fecha de la final de la champions.

Descripcion,	€	
Piso final Champions league 2019,	5.000	▲
floor for champions,	3.500	
Se alquila para final de la champion league,	3.000	
Piso fin de champion , cerca del wanda.,	3.000	
Apartamento para la final de champions en Madrid,	3.000	
Great apartment to enjoy the champions league!,	2.950	
Alquiler piso para final de la Champions league,	2.600	
Final champions,	2.200	
Champions league final,	2.050	
Piso final champions league,	2.000	
Luxury apartment Chamberí final champions league,	2.000	
~10 minutes on foot from the wanda stadium~,	2.000	
Mi Casa durante la final de la Champions,	1.890	
Habitacion matrimonio juveniles final champions,	1.700	
Piso luminoso a 12 minutos del wanda metropolitano,	1.500	
Piso grande para la final de la LDC (+8 personas),	1.500	
2 rooms near wanda,	1.500	
Habitacion para la final de la champions,	1.450	
Habitaciones para 3 para la final de champions,	1.400	
Three rooms flat - Champions final Madrid,	1.200	
Apartment at 8 minutes to wanda metropolitano,	1.200	
Room available near to wanda metropolitano,	1.150	▼

Figura 44. Tabla con los apartamentos para la final de la champions

4.- VISUALIZACIÓN DE DATOS

Abarcando otros aspectos interesantes como es la visualización de los datos para una empresa, es algo bastante importante se mire desde el punto de vista que se mire. Desde el punto de vista del cliente, tener una plataforma en la que visualizar lo mejor que se pueda el producto que se contrata proporcionando una buena imagen de la marca, que le proporciona toda la información posible en una interfaz amigable en la que se sienta cómodo y haciendo hincapié indirecto en las zonas que a ti te interesa que el cliente se fije. Desde el punto de vista de la empresa es importante conocer la máxima información acerca de lo que ofreces, de tus clientes y del funcionamiento de la empresa. Por ello, se han desarrollado tres dashboards, ejemplificando la importancia que tiene poder entender la información mediante un vistazo rápido. El primero de los dashboards dirigido a los clientes cuando buscan un alojamiento y los otros dos dashboards dirigidos internamente para la propia empresa. La visualización se ha hecho con la herramienta por excelencia del sector Tableau Desktop.

4.1. Dashboard 1 - ¡Escoge tu destino!

En este caso este dashboard centrado en el cliente que está buscando su próximo alojamiento en Madrid en la página web, su objetivo es que el cliente tenga una visión lo más representativa posible de lo que va a contratar, desglosando lo que incluye cada alojamiento, sus características y su localización.

Antes de comenzar con el storytelling, se han escogido tonos de color suaves de fondo para que el usuario se sienta cómodo y se pueden destacar las zonas en las que interesa que el usuario se fije más. Iniciando con el storytelling de la navegación por el dashboard, consiste en desde el principio incitar al cliente al cliente a contratarnos, como por ejemplo con el título en grande para que contrate uno de nuestros alojamientos. Tras ello, se muestran una serie de filtros en los cuales el usuario va a introducir las características del alojamiento que le interesa contratar. Lo cual va a filtrar el resto de la visualización. Después de los filtros, se muestran una serie de características generales de lo que se está filtrando para más información del cliente y

que esté seguro del filtrado efectuado o lo modifique. A continuación, se resalta mediante colores distintos un top recomendados según la valoración de las casas en el cual puedes elegir tú la cantidad de recomendaciones modificando el parámetro “N.º de recomendaciones”, esto son las colaboraciones pagadas con particulares para colocar sus alojamientos en los lugares más vistosos de las plataformas de alojamientos. Una vez mostradas las recomendaciones, se destacan en un fondo más claro los resultados del filtrado con su precio listas para ser seleccionadas y un mapa donde se localizan los distintos alojamientos.

Por último, se ha creado dos gráficas coloridas para llamar la atención del cliente donde se muestra información individual de cada alojamiento cuando sea seleccionado en cualquier lugar del dashboard, ya sea en las recomendaciones, en el mapa o en el listado de los resultados.

Este dashboard se ha publicado en Tableau Online y se ha puesto en producción con la actualización de los extractos de los datos periódicamente. Se puede observar en el siguiente [enlace](#).

4.2. Dashboard 2 - Analítica del grueso de datos

Este dashboard está enfocado en exponer dónde está el grueso de los alojamientos de Madrid junto con sus características frente al precio, para facilitar la identificación de anomalías ya que los algoritmos no son siempre perfectos el objetivo de este es proporcionar una visión sencilla de las características de cada apartamento por distrito y barrio y sus respectivos precios incluyendo tasas de limpieza, invitados extra y fianzas.

Al estar cerca de realizar unas predicciones se ha diseñado un dashboard más analítico, escogiendo colores que ayuden a identificar los datos. Respecto a la navegación del dashboard, se ha proporcionado libertad respecto a la navegación por distritos y filtrar por su número de alquileres, filtrado de precios y filtrado por tipo de propiedad. A continuación, se visualizan los gráficos con los datos según se filtre resaltando el tipo de propiedades y las tasas de limpieza, invitados y la fianza junto al precio de cada piso dividido por barrios. Dichos barrios están en gráfico

de barras y se puede interactuar con ellos al seleccionar cada una de estas barras subdivide la información en barrios por si el usuario desea visualizar la información con más detalle. Además a medida que se va filtrando, el número de alquileres que se va modificando junto a un gráfico de burbujas agrupadas que da una visión del porcentaje por tipo de propiedad aplicados los filtros actuales. Asimismo, este gráfico también actúa como para el dashboard. Por último, se dispone de un mapa de Madrid del cual se visualiza por colores la cantidad de alquileres de la que dispone cada distrito que como se ha mencionado anteriormente puede ser filtrada, cuando se selecciona cada distrito se muestra el número de alquileres de los que dispone y el porcentaje del total que supone este.

Este dashboard está publicado en Tableau Online y puede ser visualizado con más detalle en el siguiente [enlace](#).

4.3. Dashboard 3 - Impacto de datos censales de Madrid en el precio

Por último, este dashboard está enfocado en plasmar cómo interactúan algunas variables censales de Madrid como las clases sociales, la riqueza o el precio de las viviendas con el precio para proporcionar una rápida visualización de cómo se relacionan las variables del censo de Madrid con una de nuestras variables objetivo, el precio de los alquileres de AirBnB.

De igual manera que el Dashboard 2, es un dashboard analítico en el cual se ha dado preferencia a una rápida visualización del dato frente a una visualización más ilustrativa. En este caso la navegación es similar, permitiendo así filtrar por precios, tipo de propiedad y distritos, respecto a las visualizaciones mostradas en el dashboard se muestran gráficas que principalmente muestran cómo interactúan los datos censales de Madrid frente al precio promedio. En la primera gráfica, se muestra en una visualización de gráfico de líneas como cuando aumenta o disminuye el número de extranjeros por distrito el precio tiende a actuar de forma contraria a este patrón. A continuación, se muestra un gráfico de mariposa mostrando cómo impacta en el precio la disponibilidad de un servicio mostrando la diferencia que ha supuesto disponer de este. Seguidamente se aporta una gráfica sobre cómo interactúa el precio promedio según va aumentando el número de habitaciones. Por último, se muestran tres gráficas con formato de gráfico

combinado de barras y líneas, mostrando la interacción de la riqueza, clase social y el precio por metro cuadrado de las viviendas frente al precio promedio de los alquileres de Airbnb.

Al igual que los dos últimos dashboards este también ha sido publicado en Tableau Online y puede visualizarse en el siguiente [enlace](#) o si se quieren visualizar los dos últimos dashboards mencionados anteriormente se ha creado un storytelling [aquí](#).

5.- PREDICCIONES

Llegados a este punto, en el cual se ha hecho un proceso ETL para el tratamiento y transformación de los datos, un análisis geográfico y más profundo de los datos, se pretende explotar al máximo el dato para este tipo de plataformas de alojamientos turísticos. En consecuencia, se han realizado predicciones como son sobre el precio de un alojamiento, para que la oferta sea lo más acorde posible a las pretensiones de los clientes siendo una manera de fijar precios inteligentes. Además de esta predicción se predice la satisfacción que tendrá un alojamiento, lo que nos puede proporcionar información para que apartamentos debemos mantener o no en la plataforma siempre buscando lo mejor para el cliente y la empresa. Estas predicciones son solo dos ejemplos de las múltiples maneras que se pueden explotar los datos en este gremio.

En este punto tenemos unida la información de las secciones censales de cada uno de los alojamientos con los alojamientos, pero poder comenzar con estas predicciones se hizo la unión de los demás datasets utilizando el id del alojamiento para que los datos se unan correctamente. Se han añadido los datasets `calculada_listing_detailed`, `gsd_madrid`, `calculada_promedio_estancias` y `calculada_servicios_airbnb` mediante una query uniendo los campos con un join que junta las tablas por los ids de los alojamientos.

5.1 PREDICCIÓN DEL PRECIO

En esta predicción se pretende predecir un precio justo para los distintos alojamientos utilizando la variable `price` como etiqueta, que es el precio que tiene cada alojamiento para su alquiler por noche. Para ello, se utiliza la plataforma Azure Machine Learning.

Antes de realizar entrenamiento de algoritmos y predicciones, inicialmente se eliminan los posibles picos de todas las variables discretas mediante un bloque llamado Clip Values asegurándonos así de que se mantengan los rangos correctos, ya que están en el dataset como un valor entero. A continuación, se eliminan variables como el id del dataset para eliminar el ruido en los algoritmos intentando maximizar su rendimiento. Tras ello, se han categorizado variables como la política de cancelación o el tipo de propiedad y normalizado los datos numéricos para una mejor entrada en los modelos.

En relación con las anomalías, disponemos de dos algoritmos a elección: One-Class Support Vector Algorithm o Análisis de componentes principales, para escoger cual opera mejor se han evaluado ambos para escoger el que obtenga los mejores resultados.

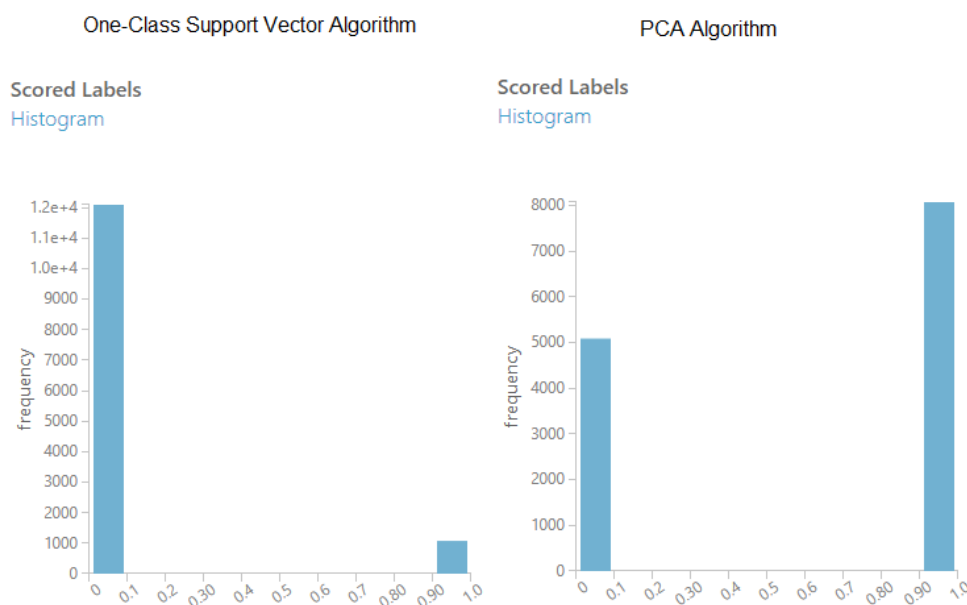


Figura 45. Comparativa de los algoritmos de detección de anomalías

El algoritmo PCA opera bastante peor ya que, detecta más anomalías que valores comunes así que por descarte se utiliza el otro para la detección de anomalías. A continuación, se eliminan los outliers (alrededor de 1000 registros), ejecutando una transformación SQL que selecciona los registros que no son outliers mediante las columnas Score del modelo de anomalías, que dicha columna después será eliminada para evitar bajar la eficiencia del resto de algoritmos.

Como ya se apreció en el análisis de los datos, se vieron distintas agrupaciones de precios que se formaban por tipo de propiedad y barrios. Para ello, se utilizará el algoritmo de agrupación por clusters K-Means. El cual da como resultado la creación de la variable "Assignments" que cuyo valor indica al cluster que pertenece. Para así analizar la influencia que tiene en los distintos algoritmos empleados. Ejecutando el bloque Permutation Feature Importance obtenemos la importancia de dichas variables utilizadas en el modelo. En la siguiente comparativa se muestran los distintos algoritmos y las variables que más relevancia tienen para cada uno de ellos. Los algoritmos

evaluados especificando de izquierda a derecha son: Boosted decision tree, decision tree, regresión lineal, regresión lineal bayesiana.









Feature	Score	Feature	Score	Feature	Score	Feature	Score
							
Assignments	0.414058	Assignments	0.705552	Assignments	1.713865	Assignments	0.302557
calculated_host_listings_count	0.21343	neighbourhood	0.109147	extranje_europeos	0.503327	calculated_host_listings_count	0.053154
cumulated_price	0.208745	minimum_nights	0.074322	extranjeros_americanos	0.396466	availability_60	0.052678
host_response_rate	0.111172	has_pets	0.049963	maximum_nights	0.273577	extranjeros	0.046495
host_listings_count	0.10683	number_of_reviews	0.040272	neighbourhood	0.270242	availability_30	0.035006
number_host_verifications	0.087171	bedrooms	0.029359	extranje_asiatcos	0.243263	habitantes_hombres	0.02667
poi_bus_b3	0.047018	host_total_listings	0.02606	availability_60	0.162023	neighbourhood	0.012119
has_pets	0.041785	host_response_rate	0.025721	habitant_30_44	0.127531	habitant_15_29	0.010186
minimum_nights	0.027297	number_host_verifications	0.025267	transito_laboral	0.107036	extranje_asiatcos	0.009901
availability_30	0.02266	cumulated_price	0.020673	poi_bus	0.083505	availability_90	0.009664
number_of_reviews	0.020033	host_listings_count	0.020441	extranjeros_africanos	0.072118	habitant_30_44	0.008785
neighbourhood	0.016127	room_type	0.019856	poi_bus_b3	0.052872	review_scores_rating	0.008129
cancellation_policy	0.015525	host_identity_verified	0.018894	habitantes_hombres	0.050637	room_type	0.00763
availability_90	0.015417	has_aid_kit	0.01805	has_cooking_complements	0.047916	has_pets	0.007163
has_aid_kit	0.015285	is_family_friendly	0.015612	has_cook_devices	0.038206	numero_habitantes	0.006076
extra_people	0.014097	cleaning_fee	0.014632	numero_habitantes	0.036481	host_total_listings	0.005779
cleaning_fee	0.014035	review_scores_value	0.014182	is_family_friendly	0.029452	host_listings_count	0.005779
maximum_nights	0.01224	poi_bus_b3	0.011965	poi_metro_	0.027741	transito_ocio	0.005121
		extra_people	0.011937				

Figura 46. Comparativa de la importancia de las variables

Como se puede observar en todos los casos la variable assignments siempre se encuentra en la que mayor importancia obtiene por lo que se puede intuir que implementar una clusterización mejora el análisis. En relación con las demás variables, se puede destacar algunos de los servicios de los que dispone cada alquiler, cierta importancia en este aspecto junto al número de alquileres de los que dispone cada anfitrión.

Por último, se analiza qué algoritmo de regresión proporciona los mejores resultados respecto a la predicción de los precios, creando sus respectivas mallas de hiperparametros, utilizando el coeficiente de determinación como métrica para la regresión, y haciendo una división de los datos train/test de un 70% - 30% de los datos respectivamente. Entre los distintos algoritmos probados para predecir el precio del alquiler por noche, que han sido regresión lineal, regresión lineal bayesiana, XG Boost y árbol de decisiones, el mejor resultado fue el siguiente:

Metrics

Mean Absolute Error	0.230258
Root Mean Squared Error	0.548668
Relative Absolute Error	0.787076
Relative Squared Error	0.408496
Coefficient of Determination	0.591504

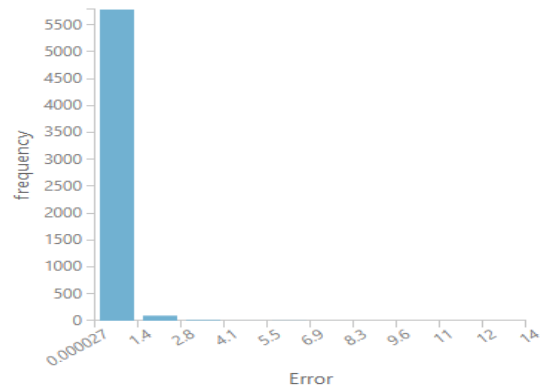


Figura 47. Resultados del algoritmo XGBoost

El algoritmo que mejor interactúa con los datos con una gran diferencia es el XG Boost obteniendo un alto coeficiente de determinación y una media de error absoluto baja. En cambio con un tratado más pobre de los datos, sin variables categorizadas ni una realización de clustering se obtiene una media de error bastante más alta y un coeficiente de determinación bastante más pobre.

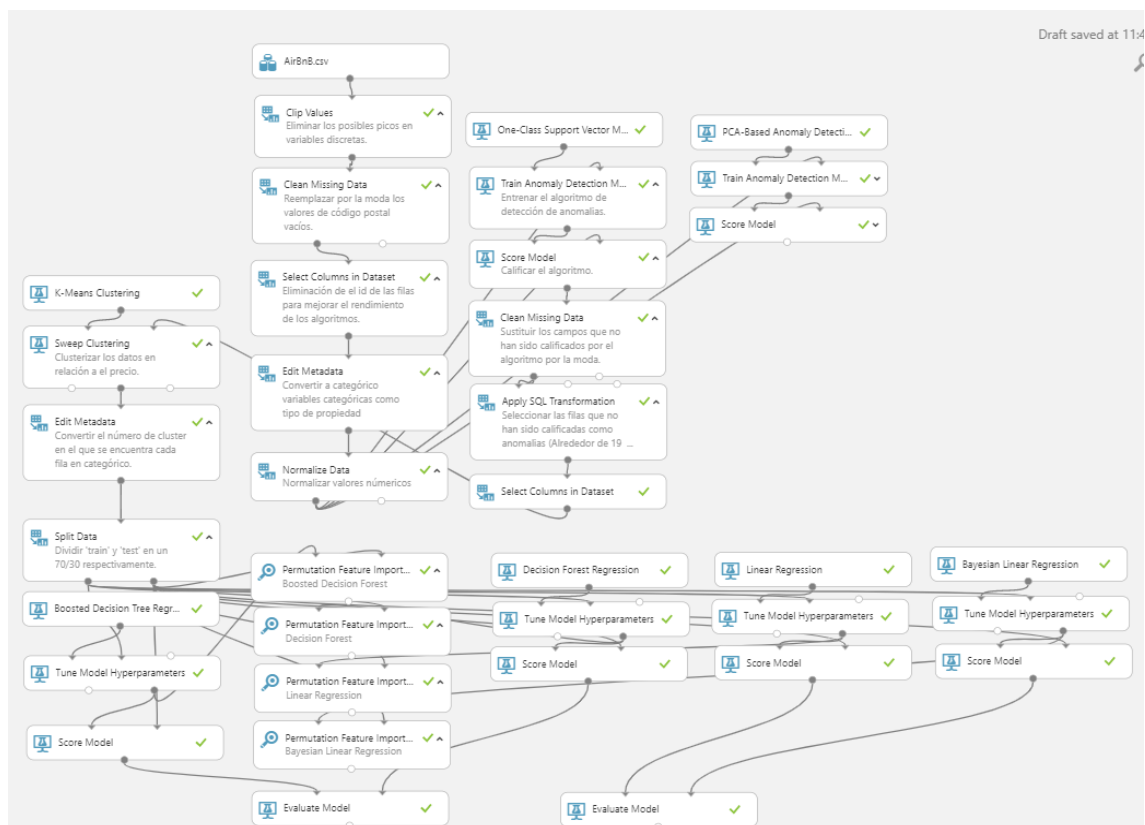


Figura 48. Experimento resultante de la predicción del precio

Para más información del experimento pulsa aquí

<https://gallery.cortanaintelligence.com/Experiment/Predicci-n-del-precio-AirBnB>

6.2 PREDICCIÓN DE LA SATISFACCIÓN DEL CLIENTE

En cuanto al caso de la predicción de la satisfacción de los clientes, esta satisfacción corresponde a la variable `review_score_accuracy` la cual tiene valor de entre 0 y 100, siendo 100 la mejor nota posible. Para ello, se han aplicado algunos algoritmos de machine learning mediante el Azure Machine Learning. Una vez tenemos el dataset previamente tratado, hace falta aplicar algunos procesos como son algunos cálculos de métricas, tratamiento de missing values, normalizar variables, es decir, ingeniería de las características. El proceso que se ha seguido han sido varios experimentos distintos hasta conseguir el modelo que mejor predice la valoración que le dan los clientes en para cada uno de los alojamientos.

En primer lugar, se tiene un experimento cuyo objetivo es dejar un dataset listo para estudiar con el modelo a aplicar, por lo que se han hecho algunos procesos. Para comenzar, se ha calculado el porcentaje de hombres y mujeres que hay en cada sección censal respecto de la cantidad total de personas de cada sección censal ya que nos proporciona una medida global para cada sección censal. Además del cálculo del porcentaje de personas por sexo, se ha calculado como métrica el porcentaje de personas por cada grupo de edades, separadas de 15 en 15 años hasta mayores de 60 años y el porcentaje de casados, viudos, solteros, divorciados, separados. Por último, en cuanto a las métricas se ha calculado el porcentaje de españoles, extranjeros, europeos, africanos, asiáticos y americanos. Todas estas métricas se han calculado con el fin de que el modelo entienda cómo es cada barrio.

En segundo lugar, normalizar las variables numéricas continuas mediante el método ZScore para un mejor funcionamiento de los algoritmos mediante el bloque Normalize Data. A las variables que no se tienen que tener en cuenta como tal las hago categóricas como son el id del alojamiento, el barrio y la zona de Madrid, pero como la zona de Madrid sí que interesa conocerla, hago one hot encoding de la variable mediante los bloques Edit Metadata y Convert to Indicator Values, dejando para cada una de las zonas de Madrid una variable que sea 1 si es la que corresponde o 0 en caso de que no corresponda a esa sección censal.

En tercer lugar, se tratan los missing values que se tienen en algunas columnas relacionadas con el host como son, desde cuando es host, cuanto tarda en responder, cantidad de alojamientos que dispone. Para tratarlos se utiliza la media de las demás casas para que no influyan en el análisis.

Seguidamente se quiere sacar del dataset a los outliers para lo cual Azure Machine Learning proporciona dos modelos que son PCA y SVM, probando con PCA como resultado elimina demasiadas filas y por lo tanto, para nuestro análisis no funcionaba. En cambio, SVM funciona bien, elimina 10 registros y es el que se ha utilizado. En relación con los hiper parámetros del mismo, mediante el bloque Tune Model Hyperparameters se consiguen los óptimos y se guarda el modelo con nombre PFM - Modelo SVM para outliers. También se guarda un dataset con los datos globales sin los outliers llamado PFM - Dataset.

Por último en este experimento, se aplica un modelo de agrupación mediante clusters K-Means para poder estudiar la predicción de la satisfacción de los clientes además de globalmente para cada cluster. De igual manera que con los outliers se tunean los hiper parámetros para que saque el modelo más óptimo pero esta vez mediante el bloque Sweep Clustering, el cual como resultado ha dado cinco grupos diferenciados, que se han guardado en distintos datasets mediante SQL, el modelo K-Means también se ha guardado. En la siguiente página se puede observar el Experimento 1. Para más información pulse aquí <https://gallery.cortanaintelligence.com/Experiment/PFM-Tratamiento-aplicando-clusters-y-sin-outliers>

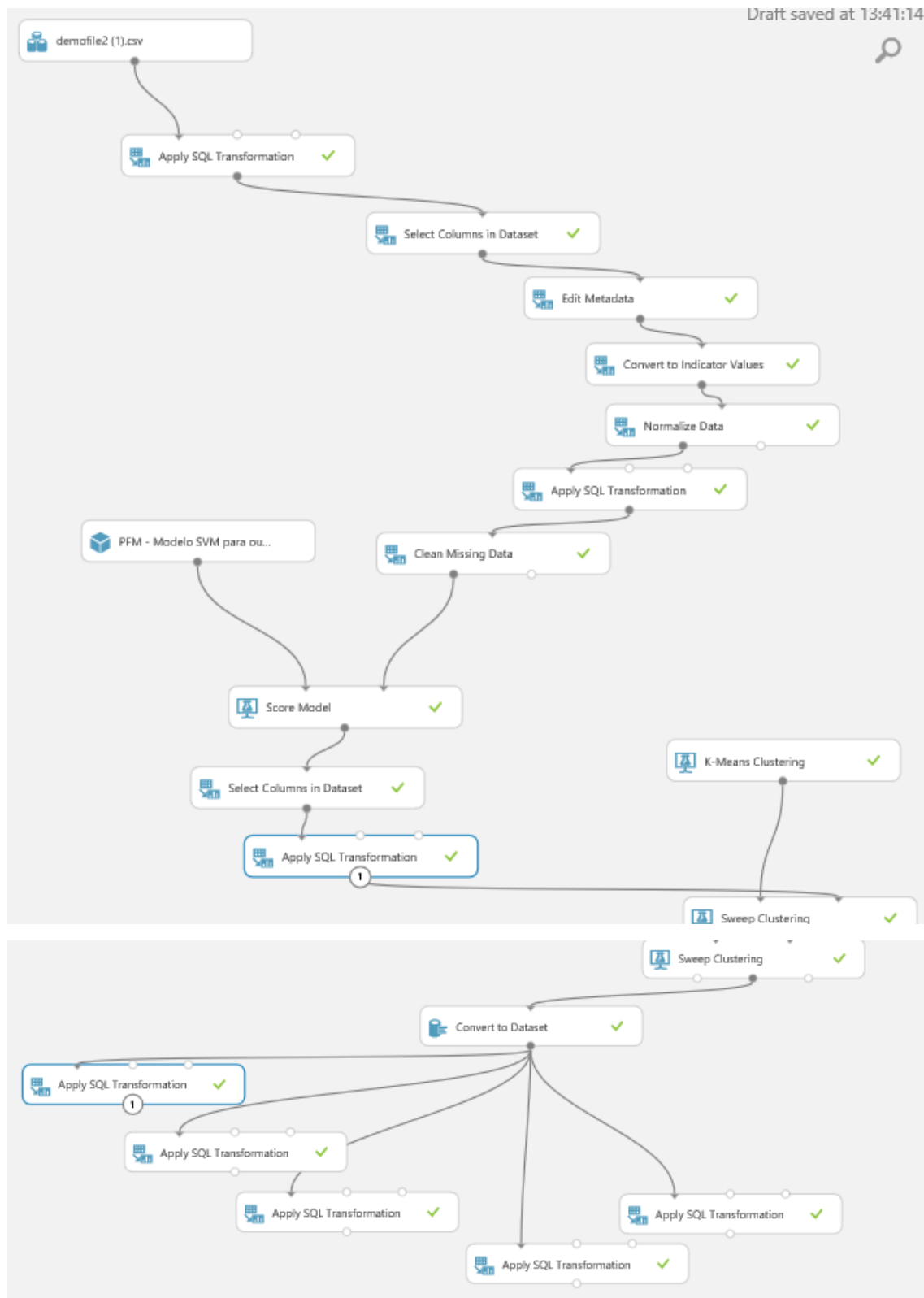


Figura 49. Experimento 1 de la predicción de satisfacción en Azure ML

Después de tratar los datos para dejarlos listos para introducirlo en un modelo de regresión capaz de predecir como de satisfactorio será un alojamiento para los clientes. Para dicho modelo se tiene como objetivo conseguir un coeficiente de determinación r^2

lo más cercano posible a 1, que sería la predicción absolutamente perfecta. Para la elección entre los distintos modelos de regresión se ha probado con la regresión lineal dando resultados malos, con el algoritmo de Naive Bayes el cual no se puede aplicar porque supone que cada una de las variables es totalmente independiente de las demás cosa que en nuestro dataset no se cumple. Por tanto, se ha optado por los árboles de decisión con XGBoost mediante el bloque Boosted decision tree regression. El proceso que se ha seguido para este experimento ha sido el siguiente.

En primer lugar, se carga el dataset correspondiente, ya sea el que corresponde a un clúster o el dataset global sin outliers. Dichos datasets almacenados anteriormente. Tras ello, se eliminan columnas que no interesan como Score Labels proveniente de la aplicación de los anteriores modelos.

Después, se separa el dataset en train y test antes de la irrupción del modelo de regresión en el proceso. Esta división se hace en una proporción 80 - 20 respectivamente.

Para finalizar, se añade el modelo al proceso y se entrena sus hiperparámetros para que escoja el modelo más óptimo para nuestros datos con el bloque Tune Model Hyperparameters. Una vez entrenado el modelo se valora mediante un bloque Score Model y un bloque Evaluate Model para sacar las distintas conclusiones, quedando el experimento de la siguiente manera.

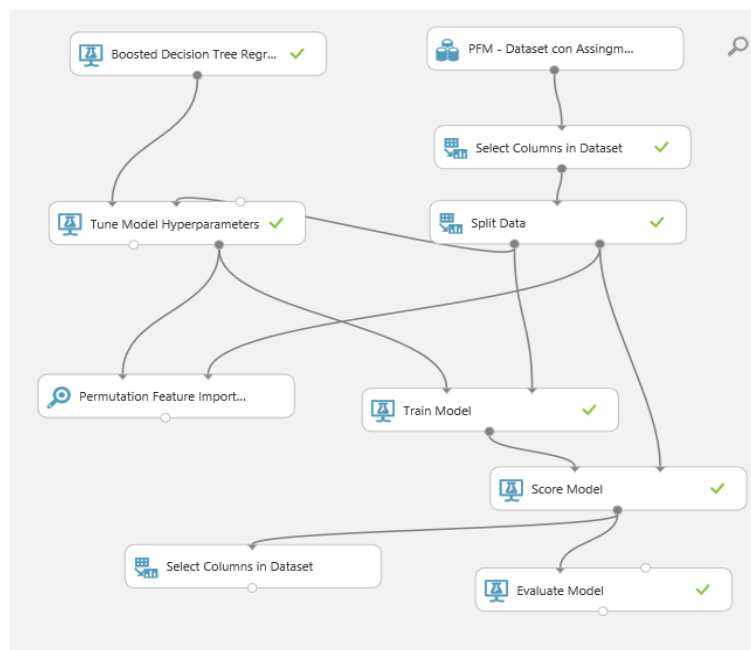


Figura 50. Experimento 2 de la predicción de la satisfacción en Azure ML

Analizando los resultados obtenidos se han extraído varias conclusiones con respecto a aspectos como la elección del modelo, los distintos análisis, la relevancia de las distintas variables y el propio dataset. En relación con los distintos modelos aplicados, el mejor modelo para nuestro problema son los árboles de decisión sin hacer el one hot encoding de las zonas de Madrid y con la variable que determina a que cluster pertenece, con los cuales se ha obtenido los siguientes resultados:

Metrics	
Mean Absolute Error	1.081534
Root Mean Squared Error	1.923547
Relative Absolute Error	0.233298
Relative Squared Error	0.061816
Coefficient of Determination	0.938184

Figura 51. Resultados de la aplicación del modelo

La aplicación de un modelo para cada uno de los clusters en nuestro problema es una mala práctica debido a que los datos que se tienen no son una cantidad muy grande entonces las distintas clasificaciones se quedan un poco cortas en cuanto información para crear modelos de regresión con buen coeficiente de determinación. Otro aspecto que se pensaba que iba a mejorar los resultados es el one hot encoding de las distintas zonas de Madrid, sin embargo empeora un poco el modelo al añadirlo. Quitando la variable que almacena a que cluster pertenecen y haciendo el one hot encoding los resultados son los siguientes:

Metrics	
Mean Absolute Error	2.253528
Root Mean Squared Error	4.064798
Relative Absolute Error	0.484102
Relative Squared Error	0.261702
Coefficient of Determination	0.738298

Figura 52. Resultados de la aplicación del modelo

En relación con el estudio de la relevancia de cada una de las variables en el modelo final para así perfeccionar el modelo, se ha hecho mediante el bloque Permutation Feature Importance. Las variables que más impactan en la satisfacción de los clientes es

principalmente al cluster que pertenece, la valoración de si era lo que esperaban, la valoración de la limpieza, valoración de la comunicación, el número de reviews y las valoraciones del checking con el host.

Por último, manejando los datos se apreció un claro evento relacionado con el precio de las casas, ya que no era acorde el precio que tenían muchas de ellas frente a la realidad, y es que en 2019 la final de la Champions de fútbol masculino se celebró en Madrid por lo que se ofertaron muchos alojamientos a precios desorbitados por la altísima demanda que había. Para más información del experimento pulsa aquí <https://gallery.cortanaintelligence.com/Experiment/PFM-Prediccion-5-Global>

6.- ALTERNATIVA DE ARQUITECTURA CLOUD

En este apartado se quiere diseñar la arquitectura para centralizar los datos de una plataforma de alquiler turístico en Madrid mediante un Data Lake, debe se escalable a una plataforma global. Para ello, la plataforma cuenta con datos recogidos de diversas fuentes relacionales de cada alojamiento:

1. Calendar: Amplio histórico por días de cada apartamento, con su precio, disponibilidad, mínimo de noches y máximo.
2. Listings: Tabla con las características básicas de cada uno de los apartamentos
3. Listings detailed: Extendida de la anterior en la cual se añade toda la información recogida de cada uno de los apartamentos.
4. Neighbourhood: relación de las zonas que hay en cada barrio
5. Reviews: fecha de cuando se pone una review y a que apartamento
6. Reviews detailed: Es un extendido de la tabla anterior que añade quien hace la review, su nombre y el comentario.

Además, la plataforma de alojamiento va a hacer uso de los datos del catastro de Madrid, conociendo más información relacionada con la localización de cada uno de los alojamientos en cartera. Esta arquitectura a diseñar debe de satisfacer una serie de requisitos que son:

- Creación del Data Lake con los datos centralizados
- Diseñar un Data Warehouse
- Aplicación de modelos matemáticos para hacer predicciones
- Creación de visualizaciones y dashboards
- Control de acceso para los usuarios
- Implementación de una API que podrá ser explotada de manera segura por una aplicación de desarrollo

Solución propuesta

En cuanto a la solución propuesta, se ha tenido en cuenta los cinco pilares de AWS, seguridad, confiabilidad, rendimiento, optimización de los costes y excelencia operacional.

Un aspecto importante para la elección de la región donde se trabaja es que la empresa de alojamiento turístico sigue el Reglamento General de Protección de Datos (RGPD), por lo tanto, la región a elegir además de tener todos los servicios que se van a utilizar debe cumplir esta regulación. La elección final teniendo en cuenta la localización, los servicios necesarios y las regulaciones, ha sido la región eu-west-1, correspondiente a Irlanda (Europa).

Inicialmente, se conocen las fuentes de datos de las cuales se saca la información:

1. Listings: Tabla con las características básicas de cada uno de los apartamentos. Almacenada inicialmente en AmazonRDS.
2. Listings detailed: Extendida de la anterior en la cual se añade toda la información recogida de cada uno de los apartamentos. Almacenada inicialmente en AmazonRDS.
3. Neighbourhood: relación de las zonas que hay en cada barrio. Almacenada inicialmente en AmazonRDS.
4. Reviews: fecha de cuando se pone una review y a que apartamento. Almacenada inicialmente en AmazonRDS.
5. Reviews detailed: Es un extendido de la tabla anterior que añade quien hace la review, su nombre y el comentario. Almacenada inicialmente en AmazonRDS.
6. GSD_Madrid: Información muy amplia por cada sección censal de Madrid, desde habitantes, nacionalidades, cantidad de casas hasta la riqueza o el turismo. Almacenada inicialmente en DynamoDB.

Una vez estudiadas las fuentes de datos y donde están situadas, las vamos a centralizar en un Data Lake con el servicio de almacenaje S3 de AWS. Dado lo cual, algunas de ellas van a sufrir cambios. Haciendo uso del servicio Data Pipeline se hacen los procesos de transformación ETL (que serían los mismos que los descritos en el apartado ETL) y

movimiento de los datos de las fuentes que están en Amazon RDS a S3, y los datos del catastro de Madrid, que están en Dynamo DB, a S3. Sin embargo, a la información del catastro de Madrid se podría añadir un disparador o reloj y una función Lambda que actualicen la información periódicamente cada trimestre y la vuelquen en S3. Con toda la información centralizada el Data Lake en S3, para conseguir una optimización mayor de costos se utiliza el servicio Glacier S3 que tiene un coste de almacenaje menor que S3 para almacenar los históricos o datos que no se consultan pero que hay que guardar como los distintos alquileres muy antiguos de la tabla calendar. Además de este almacenaje, se almacenan las tablas que se van a explotar en Dynamo DB reduciendo así los costes en producción de la arquitectura. Esta decisión tiene que ver con el tamaño de los objetos que actualmente no es muy grande teniendo en cuenta que se están utilizando datos únicamente de Madrid y su acceso de lectura tampoco es muy grande, lo que encarecería el uso de DynamoDB. Tanto el nuevo número de acceso de lectura como el nuevo tamaño de almacenaje habría que tenerlo en cuenta si se escalase para una plataforma global del mundo y nuestros clientes.

Después de crear el Data Lake, utilizando el servicio Amazon RedShift se crea un datawarehouse que contiene información de distintas fuentes de datos, quedando las siguientes dimensiones: `tratada_calendar`, `cálculo_de_días_ocupados`, `calculada_listing_detailed` y `calculada_servicios_airbnb`, siendo estos datos ya procesados que van a ser explotados.

En paralelo, partiendo de la información de S3, se pretende la aplicación de un modelo matemático a un determinado dataset para la predicción tanto del precio como de la satisfacción de un alojamiento. Con el fin de conseguir dicho dataset, se utiliza el servicio AWS Glue, que va a tratar la información, transformarla para dejar únicamente la información necesaria para hacer la predicción y la variable a predecir. Tras esto, mediante Amazon Sagemaker se utilizan modelos matemáticos para la predicción.

Por otro lado, en la visualización el servicio que se utiliza es Amazon Quicksight con el cual se crean los distintos dashboards que van a permitir una rápida visualización de los datos. Este servicio se nutre de las nuevas tablas que están en Dynamo DB, que son las que se quieren explotar.

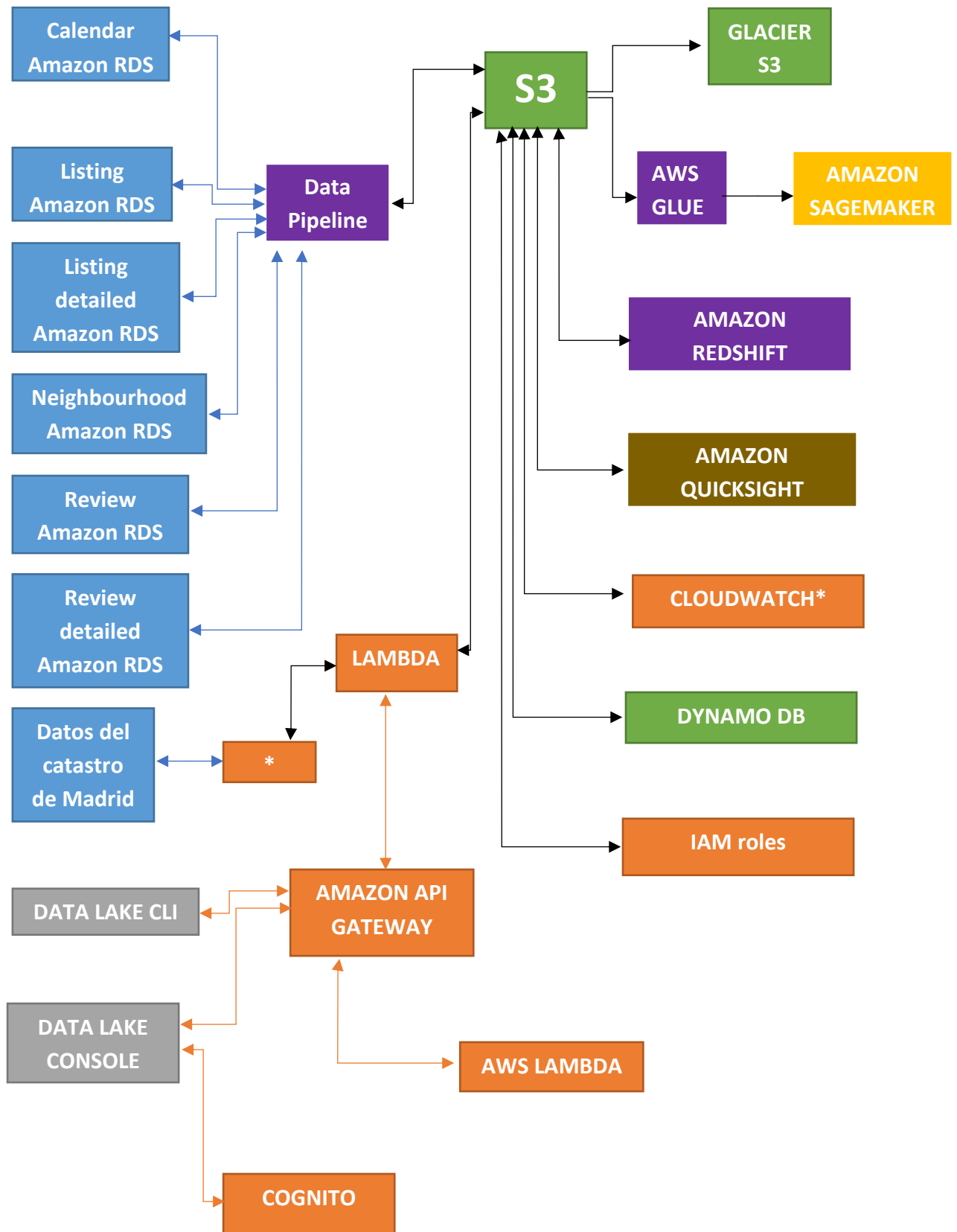
Otro de los requisitos iniciales propuestos para hacer el proyecto de arquitectura cloud es la implementación de una API que será explotada por una aplicación de desarrollo, lo que se implementa mediante el servicio Amazon API Gateway, que va a permitir a los desarrolladores la creación, mantenimiento y monitoreo entre otras cosas de la API a cualquier escala.

Finalmente, el control de acceso a los usuarios se realiza mediante el servicio Cognito, que proporciona la autenticación y autorización para la administración de los usuarios mediante inicios de sesión. Las identificaciones IAM con permisos específicos se incluyen por si en algún caso se requieren permisos específicos para algún usuario.

En la siguiente página se encuentra el diseño con los servicios utilizados y las conexiones entre los mismos.



6.1. Diseño de la arquitectura



7.- CONCLUSIONES

Conclusiones globales:

- La calidad de los datos iniciales ya era buena a pesar de lo cual ha facilitado mucho el desarrollo de la ETL y el análisis descriptivo, pese.
- Respaldando la predicción a la hora de descartar anomalías se utilizaban los dashboards del grueso de los datos.
- En general tiene mucho sentido usar este tipo de tecnologías, para conocer más a tus cliente y conocer mucho mejor lo que tu ofreces consiguiendo beneficiarse ambos, tanto la plataforma como el cliente.
- Las características de los barrios que más atraían a la gente y mejor valorados todas tienen en común que son zonas que están bastante concurridas pero sin grandes aglomeraciones, pero cercanas a zonas comerciales y de ocio.
- El trabajo en equipo nos ha ayudado a abordar mucho mejor el problema con un buen reparto de tareas con revisión periódica conjunta y marcandonos objetivos de fechas.
- Podría ser que la plataforma que provee los datos los de sesgados con sus mejores alojamientos pero que no representan al 100% la realidad.
- Ha sido una pena no tener datos geográficos de otras ciudades para tener otra visión no únicamente los de Madrid, que han tenido bastante impacto en los resultados de la predicción obtenidos.
- En cuanto a los resultados de la predicción del precio, inicialmente teníamos un coeficiente de determinación de alrededor de un 0.41 y una media de error bastante alta pero utilizando técnicas de visualización para determinar el comportamiento de los datos y creando nuevas métricas se ha conseguido impulsar esta predicción hasta un coeficiente de determinación de 0.59 y bajando la media de error hasta 0.23.

- En relación con la predicción de la satisfacción de los usuarios, ha mejorado bastante desde los resultados iniciales con la aportación de métricas, cálculos de porcentajes, agrupaciones, llegando a un valor de determinación de 0.93 con un error medio de 0.108 puntos en una valoración sobre 10.
- Creemos que se pueden mejorar algunos aspectos como una ingesta de datos automática que descargue directamente de insideairbnb.com y los introduzca a la ETL.
- La utilización de las predicciones para conocer hacia donde va la tendencia de cada alojamiento, en el caso de la satisfacción las acciones que se podrían tomar como un apartamento que atrae malas valoraciones sacarlo de la cartera provocando que no se manche la imagen de marca e intentando ofrecer lo mejor para el cliente.

Conclusiones de carga y transformación de datos:

- Es importante que los archivos csv estén estandarizados, ya que supone un trabajo extra el arreglarlos para añadirlos a nuestra base de datos.
- La creación de un servidor global, ha facilitado todo el tratamiento de los datos
- Transladar el datalake a postgresSQL ha facilitado su lectura en cuanto a la velocidad.
- Hay que tener en cuenta las ip públicas van cambiando y que las variables del `kettle.properties` sean iguales en todos los miembros que tienen acceso para facilitar las conexiones a la base de datos.
- Ha sido muy importante utilizar procesos SQL para tratar la transformación de los datos siendo más efectivo muchas veces que mediante una transformación en Pentaho
- Además utilizar la herramienta QGIS con la extensión de PostGIS ha agilizado las uniones de datos de AirBnB y los datos censales de Madrid.

- La carga de los datos a PostgreSQL es relativamente rápida, excepto por los datos de la tabla calendar que tenía alrededor de 7 millones de filas.
- Respecto a los datos sobre las estancias se agruparon en promedios para tener una mejor perspectiva sobre cada alquiler y su comportamiento y facilitar la creación del csv a la hora de unir los datos.
- Para conseguir sacar los servicios que inicialmente estaban desestructurados, inicialmente tratarlos con pentaho fue más difícil así que se atacó con consultas SQL a esta columna para obtener determinados datos de esta y posteriormente tratarlos con pentaho que al final algunas de estas métricas obtenidas resultaron tener valor en las predicciones tanto de precios como de valoraciones de alquiler.
- Columnas que inicialmente estaban desestructuradas y aparentemente no tenían gran valor resultaron ser de ayuda en el análisis y predicciones de los datos al ser tratadas y categorizadas.

Conclusiones análisis descriptivo:

- La idea inicial era hacer el análisis en geomarketing pero se quedaba un poco obsoleto y se decidió mejorarlo en Tableau
- QGis fue importante para la unión espacial de los datos mediante una joint por localización, haciendo posible juntar las tablas de Madrid con los alojamientos.
- La gran mayoría de las casas se concentraban en el centro de la ciudad y alrededor del 90% de estas son apartamentos.
- Hay una gran cantidad de casas que tienen un precio desproporcionado debido a que se ofertaron en mayo del 2019 para acoger a los aficionados ingleses que venían a ver la final de la Champions de fútbol a Madrid.
- Los precios de las casas más caras se concentran en el centro de Madrid, por su localización y en el barrio de San Blas, debido a la cercanía con el estadio Wanda Metropolitano.

- La importancia de los servicios que incluye cada alojamiento cambia según el precio de la casa, en las casas más baratas y en las más caras incide en el precio de mayor manera que en las casas de un rango de precios medio.
- La mayoría de alojamientos ofrecidos en los datos de AirBnB tienen una valoración demasiado buena lo que puede ser apostado para publicitar la compañía.
- La correlación de los precios frente a las camas y el número de habitaciones es bastante importante.
- Las mejores valoraciones cuentan con un gran número de reviews, lo que muestra una gran satisfacción de la clientela que las alquila, seguramente dejando una opinión positiva.
- La tendencia de los precios está relacionada con la cantidad de habitantes nacionales que hay en la zona.

Conclusiones de visualización de datos:

- La visualización de los datos nos ha ayudado a tener un mayor entendimiento de los datos y confirmar algunas de las tendencias que se pensaba inicialmente.
- La clusterización de los datos ha sido importante debido a que para cada cluster tiene unas peculiaridades distintas.

Conclusiones de las predicciones:

- La unión de los datos en un dataset facilitó bastante el trabajo de cada uno de las predicciones.
- A la hora de realizar las predicciones se eliminaron las estancias que duraban más de dos meses las cuales se consideran anomalías.
- La categorización de las variables mediante el bloque de Azure ML, aumentó en gran medida el resultado de la predicción.

- La repercusión de la ingeniería de las características destaca sobre la elección del modelo dentro de los modelos que daban buen resultado.

Conclusiones arquitectura cloud:

- Este apartado está destinado a la escalabilidad de la plataforma. Hacerlo únicamente para los datos de Madrid no sería necesario.
- Tendría mucho más sentido si fuese una plataforma global de muchas más ciudades del mundo o capitales importantes.

8.- BIBLIOGRAFÍA Y ANEXOS

8.1. Anexo de ejecución

Para la ejecución correcta del trabajo realizado se debe seguir el siguiente orden:

1. En primer lugar, se ejecutaría el job global del apartado ETL, este paso crea las tablas, las trata y junta las relacionadas con los alojamientos.
2. Seguidamente, mediante QGis hay que cargar la capa de GSD_Madrid y cambiar el sistema de coordenadas predeterminado EPSG:4326 al EPSG:23030
3. Hacer la unión mediante la localización con QGis como se explica en el apartado de Análisis y entendimiento del dato.
4. Tras ello, ya dispones del dataset enriquecido con toda la información lista para hacer las visualizaciones y predicciones.

8.2. Enlaces bibliográficos

- 1.- ¿Qué es Airbnb y cómo funciona?
- 2.- Big Data y Data Science. El éxito de Airbnb “Nuestra Obsesión por los Datos”
- 3.- <https://www.startupranking.com/airbnb/competitors>
- 4.- Inside Airbnb. Adding data to the debate.
- 5.- <https://www.airbnb.es/help/article/1168/cómo-puedo-activar-o-desactivar-los-precios-inteligentes>
- 6.- https://www.qgistutorials.com/es/docs/3/performing_spatial_joins.html
- 7.- Normalize Data - ML Studio (classic) - Azure
- 8.- Operadores de comparación (Transact-SQL) - SQL Server
- 9.- Sweep Clustering - ML Studio (classic) - Azure
- 10.- Evaluate Model - ML Studio (classic) - Azure