Checkpoint 1 - Grupo 07

Análisis Exploratorio:

El dataset provisto por la empresa Properati consta 460154 registros de propiedades, con 20 columnas detallando ubicación geográfica, precio, ambientes, entre otras características de cada una de ellas. A nuestro parecer los features que más se destacan son “place_l2” (posteriormente renombrada a “Barrio”), su “property_price” (Precio), “property_rooms” (Ambientes) y “property_type” (Tipo de Propiedad).

Una de las características más importantes a la hora de comprar una propiedad es su ubicación. En el data frame trabajado aparece de varias formas, una de ellas mediante coordenadas geográficas (según Latitud y Longitud) y otra es según varias divisiones de donde se encuentra (aunque estén marcadas en varias columnas, Ejemplo: Argentina, Capital Federal, Palermo, Palermo Chico). En todos estos casos el tipo de dato es objeto.

Otra feature a tener muy en cuenta es el precio de la propiedad, donde lógicamente, es un número (float) que al filtrar el dataset según las cuestiones solicitadas, solo trabajaremos los precios en dólares. Sin olvidarnos del tipo de propiedad, que toda persona que hoy en día quiere comprarse una, debe tener preferencias en si quiere un Departamento o una Casa, por ejemplo.

Y por último, pero no menos importante, la cantidad de ambientes que posee la propiedad. Es una característica fundamental a la hora de buscar propiedades a adquirir, ya que, como luego mostraremos, es una variable muy relacionada con el precio. Cuando una aumenta, en líneas generales, la otra también tiende a subir.

Preprocesamiento de Datos:

1) Cuando comenzamos a analizar el dataset y las variables del mismo, notamos que habían columnas que, en un principio, eran prescindibles asique nos metimos más en profundidad.

   Primero, vemos el porcentaje de nulos de las variables. Vimos que el 100% de los datos de las variables “place_l5” y “place_l6” son nulas. En conclusión, no son relevantes para el análisis y tomamos la decisión de eliminarlas.

   Luego, gracias a los filtrados que fueron solicitados, tenemos tres variables que en realidad son constantes, es decir, tienen un solo valor:

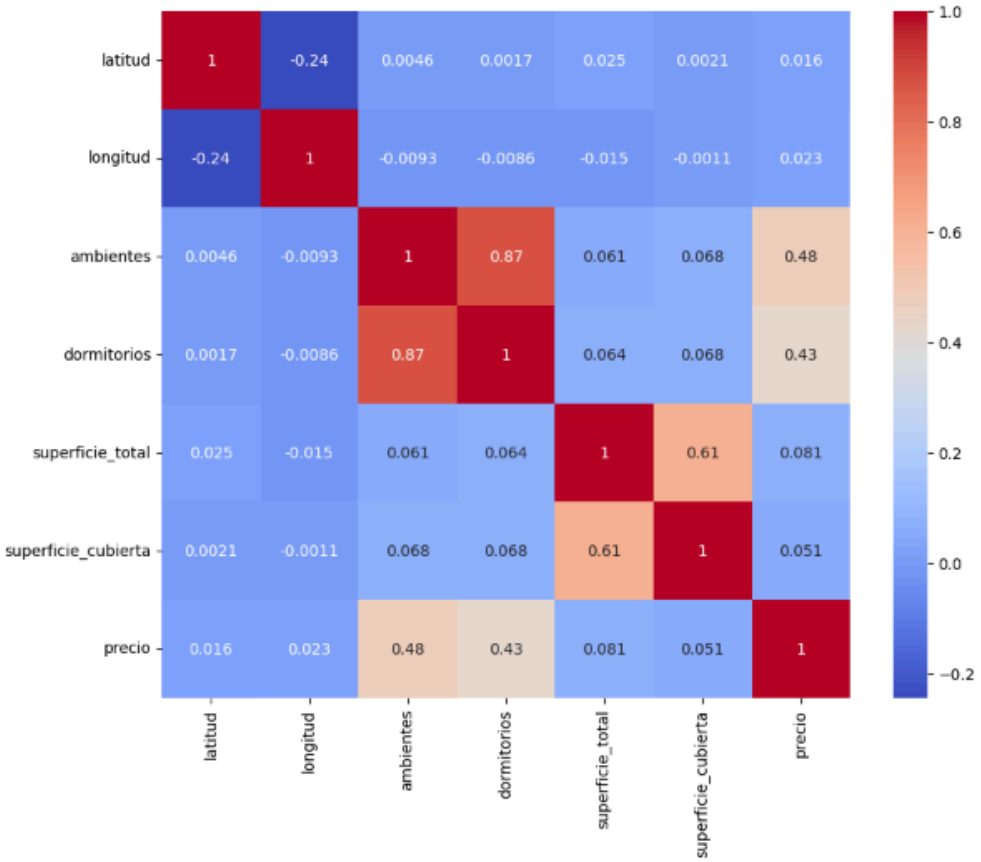
   - “place_l2” sólo contiene el valor “Capital Federal”

- “operation” sólo contiene el valor “Venta”
- “property_currency” sólo contiene el valor “USD” (Dolares)

Por lo tanto, considerando que ya sabemos el valor constante de estas variables en todas las propiedades del dataset, concluimos que no aportan ninguna información adicional para que valga la pena dejarlas en el mismo y procedimos a eliminarlas.

También, en base a los nombres de las columnas, notamos que “start_date” y “created_on” pueden ser similares o hasta iguales. Así que creamos una función que itere en el dataframe y compare todos los valores de ambas las variables en todas las propiedades. En caso de que devuelva True, brindan la misma información, es decir, son iguales, y podemos eliminar una de ellas. Efectivamente, fue así. Por lo tanto la columna “created_on” es redundante y fue eliminada.

2) En un principio, se puede suponer alguna correlación entre variables que suene lógico. Por ejemplo, uno esperaría que cuanto más ambientes ó más metros cuadrados tenga una propiedad, más caro será su precio. Así que procedemos a buscar rápidamente esas posibles correlaciones lineales entre las variables cuantitativas mediante la correlación de Pearson.



Con el gráfico, observamos que la mayoría de las variables no están relacionadas entre sí y aportan mucha información al dataset, sin embargo existen algunas relaciones lineales:

- Ambientes y dormitorios: Una correlación evidente ya que un dormitorio es un ambiente, lo que la hace la correlación lineal más fuerte del gráfico.
- Superficie total y superficie cubierta: Otra correlación evidente ya que la superficie cubierta forma parte de la superficie total, pero a pesar de esto, no es una correlación lineal tan fuerte.
- Precio y ambientes: Una correlación lógica, que por transitividad se traslada también a la correlación entre el precio y los dormitorios. De estas tres relaciones destacadas es la más débil, lo cual no es raro en si, pero si es raro que no tengan una correlación lineal un poco más fuerte.

Por último, llama la atención que distintas variables no tengan una correlación lineal como el precio y la superficie, o por otra parte, la superficie y la cantidad de ambientes.

3) Por el momento, no percibimos la necesidad de crear una nueva columna o feature. Con las ya presentes, se logra analizar el dataset de una forma correcta. En caso de ser necesario, será una opción a tener en cuenta.

5) Aquí, hemos realizado un análisis creando un pequeño dataframe con la cantidad y sus porcentajes de los datos faltantes de cada columna.

	variable	porcentaje	cantidad
id	id	0.000000	0
fecha_inicio	fecha_inicio	0.000000	0
fecha_fin	fecha_fin	0.000000	0
latitud	latitud	3.950981	2979
longitud	longitud	3.950981	2979
barrio	barrio	0.432366	326
subdivision	subdivision	96.193583	72529
tipo	tipo	0.000000	0
ambientes	ambientes	1.119378	844
dormitorios	dormitorios	11.673895	8802
superficie_total	superficie_total	5.159220	3890
superficie_cubierta	superficie_cubierta	3.569013	2691
precio	precio	0.000000	0
nombre	nombre	0.000000	0

Como vemos en la imagen adjunta, hay un par de variables (columnas en el dataset original) que tienen un 0% de nulos en base al total, es decir, es una columna completa en cuanto a datos; como lo es el caso del “id”, de las fechas tanto de inicio como de fin, el precio, el nombre (una breve descripción acerca de la propiedad) y el tipo de propiedad. Luego hay casos donde tienen poco porcentaje de nulos, como lo es la cantidad de “ambientes” de la propiedad (1,1%) o el “barrio” donde se encuentran (0,4%).

Relacionado a este último, vemos una cantidad mayor, aunque siga siendo baja, cantidad de nulos en las coordenadas geográficas, “latitud” y “longitud” (casi un 4%). Antes de ello, observamos que la cantidad de registros que contienen “latitud”, “longitud” y “barrio” nulos, simultáneamente, suman 132. Esta magnitud comparada con el total de registros (75399) de nuestro data frame actual, la consideramos despreciable y esas propiedades serán eliminadas. Luego, para la imputación de los datos faltantes hemos decidido usar un geocodificador mediante un dataset de barrios en CABA propiciado por el gobierno. De esta forma, hemos iterado en todo nuestro dataset a trabajar (el de properati) para ver los “barrios” de cada propiedad e imputarles sus respectivas latitudes y longitudes. Hemos tomado esta convención de “creerle” al barrio donde pertenecen las propiedades por sobre a sus coordenadas por 2 grandes motivos: el 1ero ya que al realizar un gráfico del mapa de CABA, nos percatamos que muchas coordenadas son erróneas e incluso otras están por fuera del rango de CABA. Y el 2do motivo, es que tomando las coordenadas para imputar sus respectivos barrios hubiéramos necesitado un geocodificador inverso y esto hubiera sido un poco más costoso en implementación. Por ende, procedimos a eliminar las propiedades con “barrio” nulo, además de imputar todas las coordenadas a los demás.

Siguiendo con la lista de las variables, observamos que las columnas “superficie total” y “superficie cubierta” tienen porcentajes distintos de nulos en el dataset, marcando casi 5,2% y 3,6% respectivamente. Gracias a que estas columnas tienen una correlación bastante alta, para estos datos faltantes lo que decidimos hacer es realizar un método multivariado de imputación con Regresión Lineal para que se estudien e imputen ambas variables entre sí. De todas formas, antes de seguir realizamos un chequeo de que la “superficie cubierta” no debe ser mayor que la “superficie total” así que en esos casos donde sí sucedía, calculamos el promedio de “superficie cubierta” en todas las propiedades que tengan X cantidad de “superficie total” y fuimos imputando esos promedios de “superficie cubierta” a cada propiedad que tengan esa cantidad X de “superficie total”. Al final de este algoritmo, todas las propiedades deberían cumplir que su “superficie total” es mayor a su “superficie cubierta” o en su defecto, los empatamos para que todos los valores sean coherentes. Posdata: luego de este algoritmo hemos chequeado y eliminado todas las propiedades que tengan una

“superficie total” menor a 20. Ya que, creemos que ninguna propiedad “real” puede tener menos de 20 mts^2.

Y por último, las columnas “ambientes” y “dormitorios” también tienen porcentajes distintos de nulos en el dataset, marcando 1,1% y 11,6% respectivamente. Estas columnas, también tienen una correlación bastante alta, pero en su defecto, aquí no hemos utilizado el método de Regresión Lineal debido a que, generalmente, imputaba valores muy cercanos entre sí o incluso idénticos. Creemos que no todas las propiedades “reales” tienen la misma cantidad de “ambientes” y “dormitorios”. Por lo tanto, realizamos el mismo algoritmo que con las superficies. Calculamos el promedio de “dormitorios” en todas las propiedades que tengan X cantidad de “ambientes” y fuimos imputando esos promedios de “dormitorios” a cada propiedad que tengan esa cantidad X de “ambientes”. Al final de este algoritmo, todas las propiedades deberían cumplir que su “ambientes” es mayor a su “dormitorios” o en su defecto, serán valores iguales pero no tantos como si hubiéramos imputado mediante la Regresión Lineal.

Una vez llegado a este párrafo de texto, habríamos completado todo el dataset con valores coherentes habiendo analizado cada columna en particular, imputando mediante, promedios, medianas, predecidos con la Regresión Lineal o utilizando un geolocalizador.

4) Sí, se han encontrado muchos valores atípicos (Outliers) a lo largo de nuestro análisis del data set. Los pasaremos a explicar en el siguiente orden: Coordenadas, Ambientes y Dormitorios, Superficies (Totales y Cubiertas) y Precio.

   Para el caso de las coordenadas, “latitud” y “longitud”, (como hemos mencionado), al realizar un gráfico del mapa de CABA, nos percatamos que muchas coordenadas son erróneas e incluso otras están por fuera del rango de CABA. Las que son erróneas las hemos modificado (en el punto de imputación) y las que están por fuera del rango de CABA las hemos analizado como Outliers. A las propiedades con esta característica, hemos tomado la decisión de buscar a mano en el dataset propiciado por el gobierno, los centroides de cada coordenada de su respectivo barrio. De esta forma, imputamos esos centroides en la “latitud” y “longitud” de la propiedad que “estaba por fuera de CABA” respetando y guiándonos por su “barrio”, de la misma manera que hemos hecho en el punto de imputación.

   Para el caso de “ambientes” y “dormitorios”, hemos graficado tanto boxplots para ambas variables como un scatter plot (Dispersión) para observar todas las propiedades de nuestro dataset. Con ellos, a simple vista se veían claros Outliers de casos donde los “ambientes” eran muy altos pero los “dormitorios” eran muy bajos. Lo “ideal” sería que el gráfico de

Dispersión se vea similar a una recta diagonal (matemáticamente hablando, y=x) donde tendría sentido que se relacionen de esta forma, ya que la cantidad de “dormitorios” debe ser menor o igual que la cantidad de “ambientes” en cada propiedad. Para estos Outliers hemos decidido trabajarlos a mano buscándolos en el dataset y analizando su descripción, que en varios ocasiones, nos brindaba información clave para esas propiedades. Por ejemplo, varias descripciones (variable nombre en nuestro dataset) mencionan la cantidad de “ambientes” de la propiedad que se está vendiendo. Y en los casos que no era así, hemos decidido imputar medianas o promedios analizando ambas variables entre sí. Es decir, si tengo una propiedad de 4 ambientes, ¿Cuál es el promedio de dormitorios en el dataset? ¿Y la mediana?. Siguiendo esta convención hemos solventado muchos Outliers (sin sentido) en esta sección, mejorando la precisión y legibilidad de los gráficos.

Para el caso de “superficie total” y “superficie cubierta”, los hemos trabajado de forma similar a los “ambientes” y “dormitorios” ya que también tienen una correlación positiva entre ellas de la misma forma. Hemos graficado tanto boxplots para ambas variables como un scatter plot (Dispersión) para observar todas las propiedades de nuestro dataset. Con ellos, a simple vista se veían claros Outliers de casos donde las “superficie totales” eran muy altas pero las “superficie cubiertas” eran muy bajas, incluso a veces, viceversa. Lo “ideal” sería que el gráfico de Dispersión se vea similar a una recta diagonal (matemáticamente hablando, y=x) donde tendría sentido que se relacionen de esta forma, ya que la cantidad de “superficie cubierta” debe ser menor o igual que la cantidad de “superficie total” en cada propiedad. Para estos Outliers hemos decidido trabajarlos a mano buscándolos en el dataset y analizando sus “ambientes” y “dormitorios”, que llegado a este punto, como ya las habíamos tratado antes, ahora nos brindan buena información para esas propiedades. Por ejemplo, varias descripciones (variable nombre en nuestro dataset) mencionan la cantidad de “ambientes” de la propiedad que se está vendiendo. Y en los casos que no era así, hemos decidido imputar medianas o promedios de superficies analizando ambas variables entre sí y también los “ambientes” y “dormitorios”. Es decir, si tengo una propiedad de 4 ambientes, ¿Cuál es el promedio de “superficie total” en el dataset? ¿Y la mediana? O por otra parte, si tengo una propiedad de 4 ambientes, ¿Cuál es el promedio de “superficie cubierta” en el dataset? ¿Y la mediana?. Siguiendo esta convención hemos solventado muchos Outliers (sin sentido) en esta sección, mejorando la precisión y legibilidad de los gráficos.

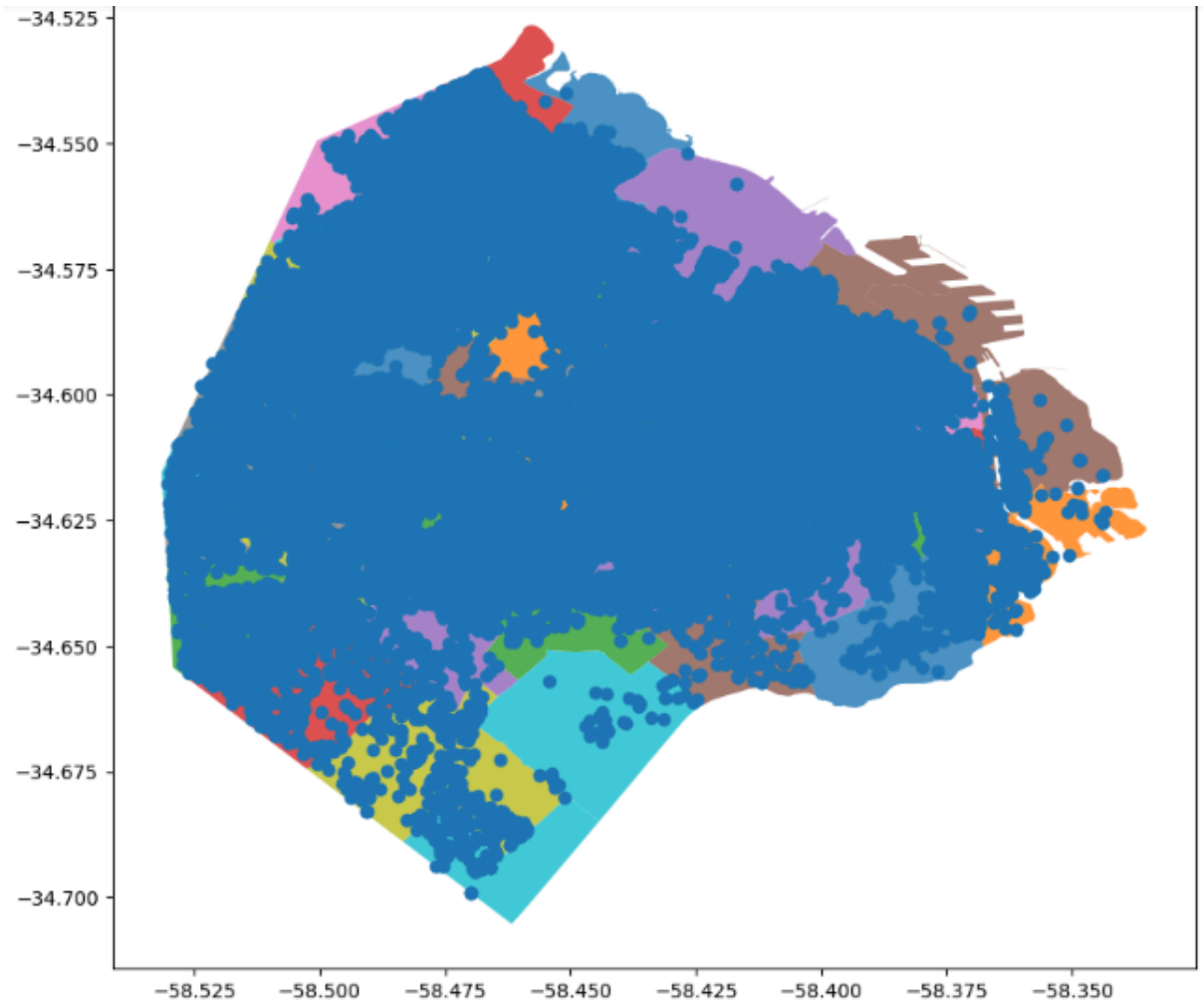
Luego, en el caso de “precio” decidimos trabajar la detección de Outliers con un método univariado, ya que los gráficos que hemos realizado de esta columna, son casi ilegibles por la cantidad de Outliers tan extremos en él. Por ejemplo, hay una propiedad en “Caballito” de 2 “ambientes” que su “precio” es de 21 millones de USD. Por casos así, los gráficos, como

el boxplot, dejan mucho que desear visualmente hablando y hay un enmascaramiento de otros posibles Outliers que no se ven debido al desplazamiento que generan esos casos extremos. Por lo tanto, utilizamos el método “Z-Score Modificado” para detectar todos los Outliers mediante su métrica y su regla de oro (umbral en 3,5). De esta forma, según este método, hay aproximadamente 6800 Outliers. La decisión que hemos tomado para ellos ha sido, analizando las variables “ambientes”, “tipo” y “superficie total” de cada propiedad, calcular un promedio de precios para imputar en cada caso que sea Outliers. Mismo algoritmo que realizamos para Superficies. Es decir, si tengo un departamento de 4 ambientes con 170 mts^2, ¿Cuál es el promedio de precios con esas características en el dataset?

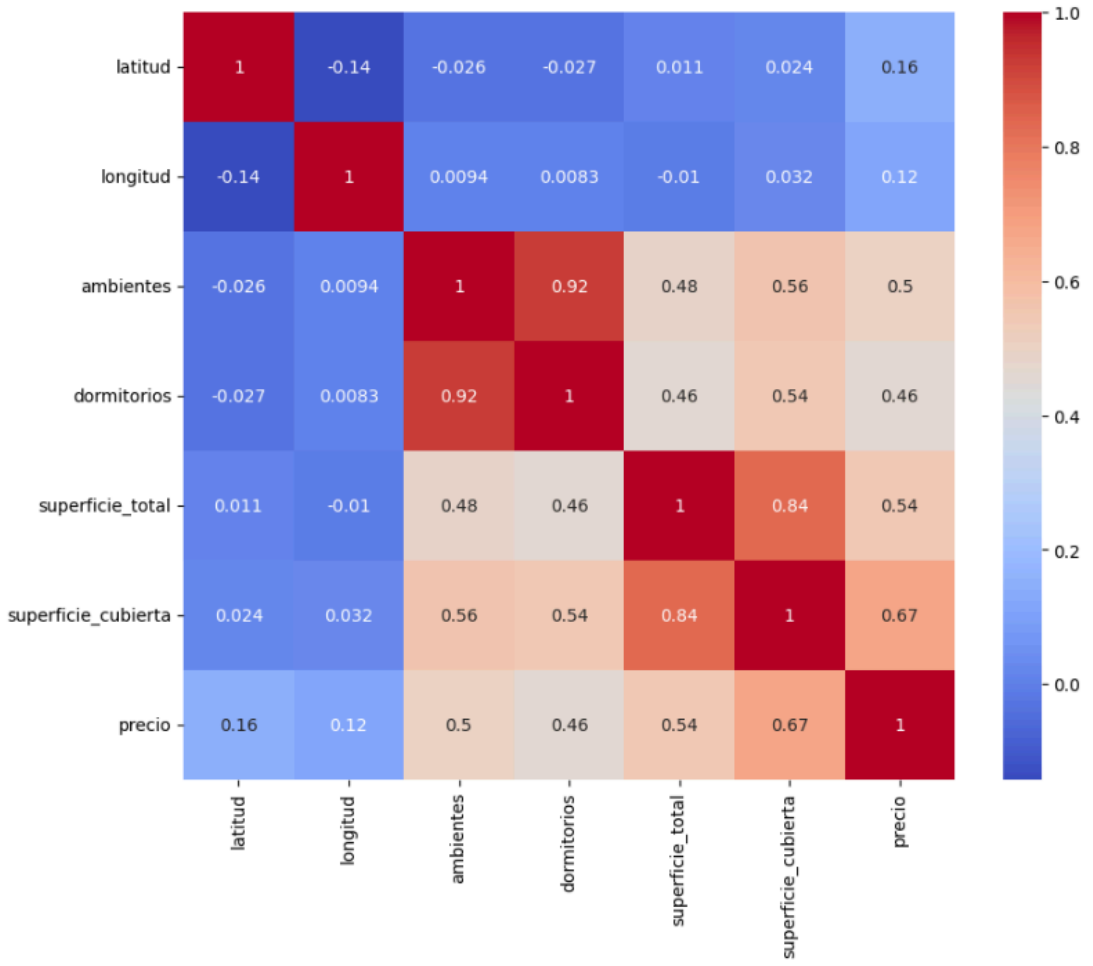
Y por último, como es solicitado, también hemos trabajado la detección de Outliers de “precio” de forma multivariada con “superficie total”, mediante el método “Distancia de Mahalanobis”. De todas formas no hemos llegado a terminar de cerrar esta parte y tomar la decisión sobre alguna convención a tomar. Es lo único faltante de este punto.

Visualizaciones:

Como primera visualización gráfica a traer al informe, elegimos el gráfico de puntos en CABA, luego de ser depurado y con los Outliers de coordenadas geográficas corregidas. Este gráfico muestra la relación del “barrio” con “latitud” y “longitud” mediante los puntos azules. Como observamos, la visualización es coherente ya que ningún punto está por fuera de CABA, cosa que en una versión anterior de este mismo gráfico sí sucedía. También es muy notable la numerosa cantidad de propiedades en el dataset debido a la superposición de todos los puntos en el mapa.



Como segunda visualización gráfica a traer al informe, elegimos el Heatmap, el gráfico de correlaciones lineales entre las variables cuantitativas mediante la correlación de Pearson. Sin embargo, esta es una nueva versión más actualizada que la anterior mostrada, ya que este último lo hemos hecho luego de analizar y tratar todos los datos faltantes y Outliers. Como podemos observar, en general, las correlaciones han mejorado entre todas. Antes habíamos mencionado que sonaría “lógico” o uno supondría que la “superficie total” y “superficie cubierta” estarían correlacionadas a “ambientes” y “dormitorios” pero NO sucedió. En cambio, habiendo depurado el dataset de errores, nulos y outliers, ahora se vé que las correlaciones entre estas cuatro columnas del dataset, han mejorado bastante; incluso hasta aproximarse a 0,50. También hablando de a pares, las correlaciones entre “ambientes” y “dormitorios” se han reforzado llegando hasta 0,92. De la misma forma, el caso de “superficie total” con “superficie cubierta”, donde su correlación escaló hasta 0.84. En el caso del precio, mejoró todas sus correlaciones excepto con “latitud“ y “longitud”, por obvias razones. Pero las demás, ahora precio si tiene un apoyo sobre las otras columnas, llegando también a valores aproximados a 0,50.



Estado de Avance:

1) Análisis Exploratorio y Preprocesamiento de Datos:

Porcentaje de Avance: 90%/100%
Tareas en curso: Análisis de Valores Atípicos (Outliers en Precios)
Tareas planificadas: Hemos tenido un inconveniente a la hora de imputar los promedios de precios, en análisis univariado de esa misma columna.
Impedimentos:

d) Valores Atípicos: estamos terminando con el análisis de outliers en la variable precio de forma univariada y multivariada (con superficie como es pedido).

2) Agrupamiento:

Porcentaje de Avance: 0%/100%
Tareas en curso:
Tareas planificadas:
Impedimentos:

Tiempo Dedicado:

Integrante	Tarea	Prom Hs. Semana
Matias Agustin Ferrero Cipolla	Exploración Inicial Visualización de Datos Datos Faltantes Valores Atípicos	9
Franco Ricciardo Calderaro	Exploración Inicial Visualización de Datos Datos Faltantes Valores Atípicos Armado de Reporte	9
Carlos Alejandro Orqueda	Exploración Inicial Visualización de Datos Datos Faltantes Valores Atípicos	8
Sebastian Kraglievich	Exploración Inicial Visualización de Datos Datos Faltantes Valores Atípicos	8