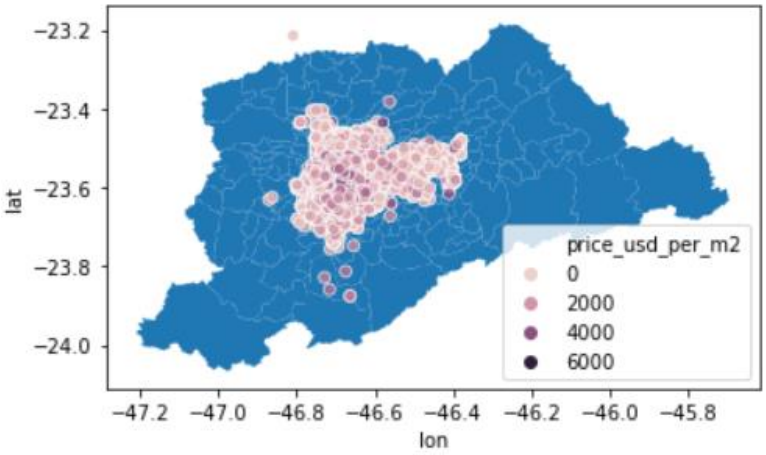
Prueba Data Science (uDA)

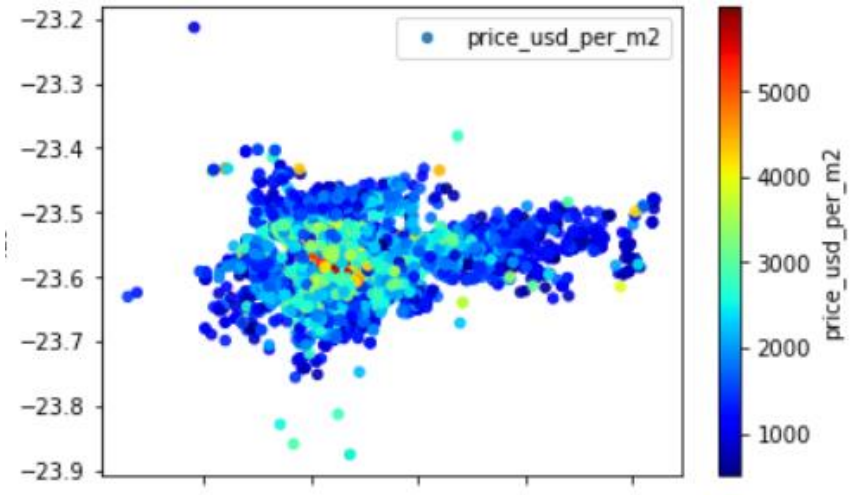
Primero, es abrir el archivo e inspeccionar el contenido: número de columnas, tipos de datos, relaciones entre ellos…

En este caso se comienza con la columna place_with_parent_names en el cual se encuentran contenidas las referencias al país, provincia, ciudad y barrio. A esto hay que eliminar las columnas vacías que se generan al utilizar el método *split*.

Por lo tanto, de todos los registros nos quedaremos con los referidos a la ciudad de Sao Paulo.

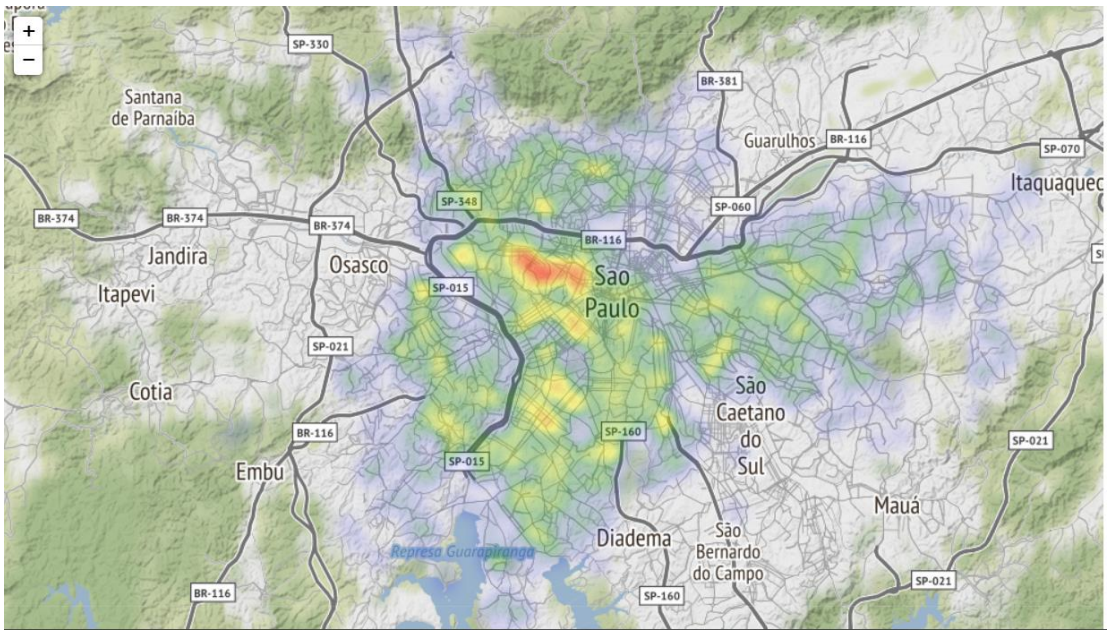
Para poder visualizar de forma rápida y entender el entramado de la ciudad, los barrios y los precios se ponen a disposición los siguientes gráficos espaciales.



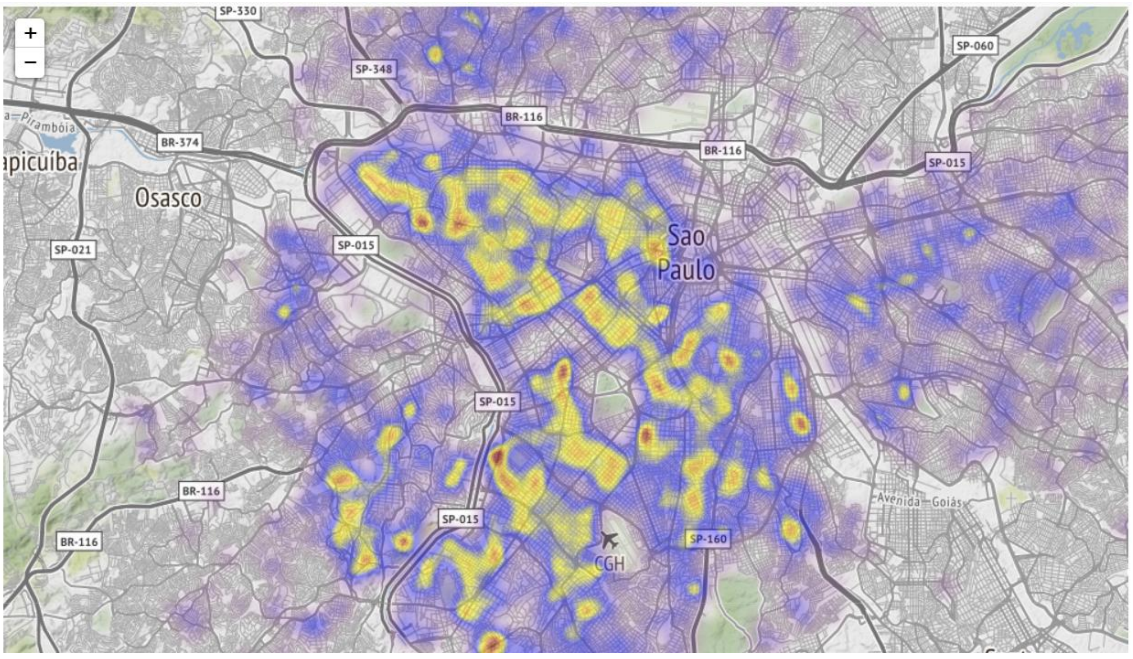


En este gráfico se pueden sacar algunas ideas como que las ofertas que hay son escasas en la periferia de la ciudad de Sao Paulo. Además los precios más altos por metro cuadrado se encuentran en el centro de la ciudad y en el extremo inferior.

El número de ofertas se reparte de la siguiente forma, concentrándose sobre todo en el núcleo central norte:



Aun así, utilizando mapas de calor en cuanto al precio del metro cuadrado, se pueden obtener ciertos puntos en los que el precio es donde es el más alto de la ciudad.



Hay que remarcar que estos gráficos están realizados a partir de los datos que contenían la latitud y la longitud, es decir, un 11,6% del total de datos de Sao Paulo, pero ayudan a comprender la distribución.

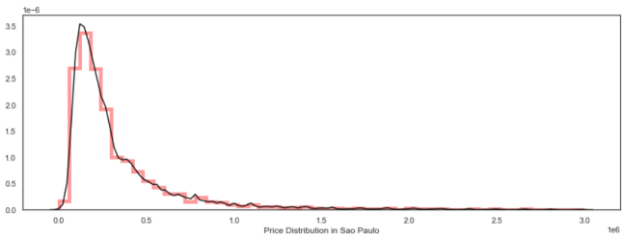
Para la visualización del resto de datos y el uso de los modelos de aprendizaje automático se usarán la totalidad de datos con las siguientes condiciones:

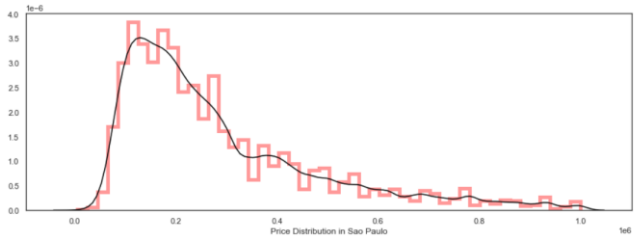
Con respecto a los datos faltantes:

- La columna de gastos se considerará que es 0 cuando no se indique nada. Los datos se dejan en moneda BRS. No se realiza conversión debido a que su valor no necesita estar en la misma moneda al usar los modelos de predicción.
- Al igual que esto, se considerará como planta 0 cuando no se indica nada en la columna de floor.
- Para el número de habitaciones no especificadas, se utilizará la mediana ya que se trata de un valor discreto. En este caso el valor será de 2.
- Se eliminan los registros que no contengan precio al ser el objeto de predicción.
- Para el entrenamiento del modelo, además, se eliminan las siguientes variables entre los cuales también se encuentran datos faltantes: "lat-lon", "lat", "lon", "operation", "surface_total_in_m2", "properati_url", "country", "province", "city", "price_usd_per_m2", "place_name".

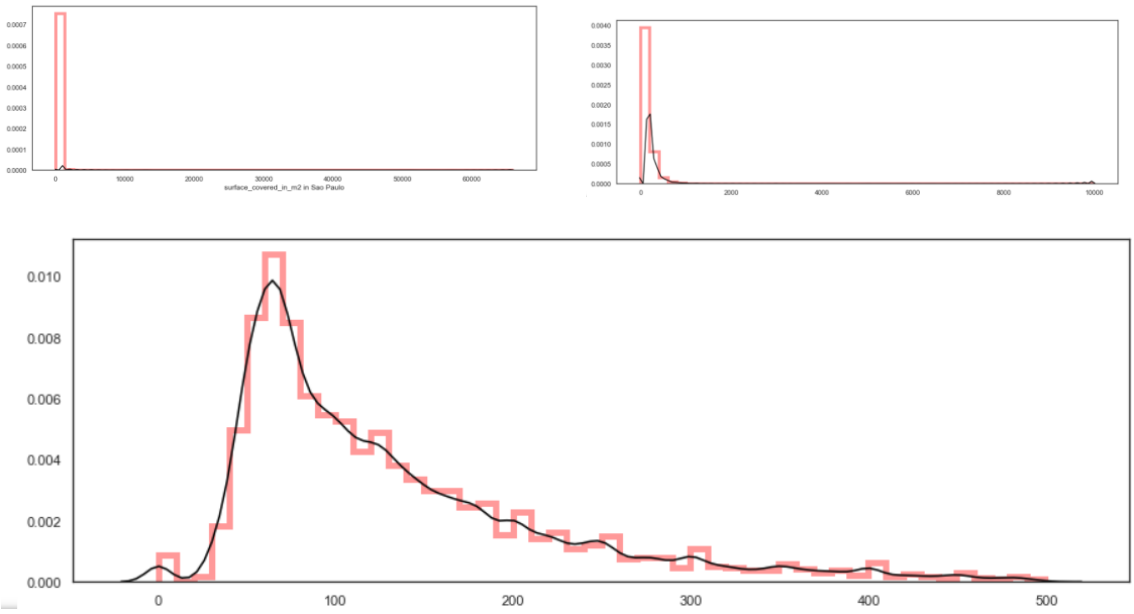
Tras aplicar esta limpieza, toca imponer cuáles serán los rangos numéricos en los que se mueva el modelo, para que de esta forma los valores extremos/errores no repercutan fuertemente en el modelo.

- Precio en dólares USD: como se puede comprobar en la gráfica (en escala de máximo 3 millones y 1 millón), la gran parte de los precios se encuentran entre los cien mil y trescientos mil dólares. Por lo tanto, viendo los valores del cuartil menor y del mayor y la distribución gráfica se decide el siguiente rango, también guiado por el sentido común:
    - Inmuebles de $60.000 a $1.000.000





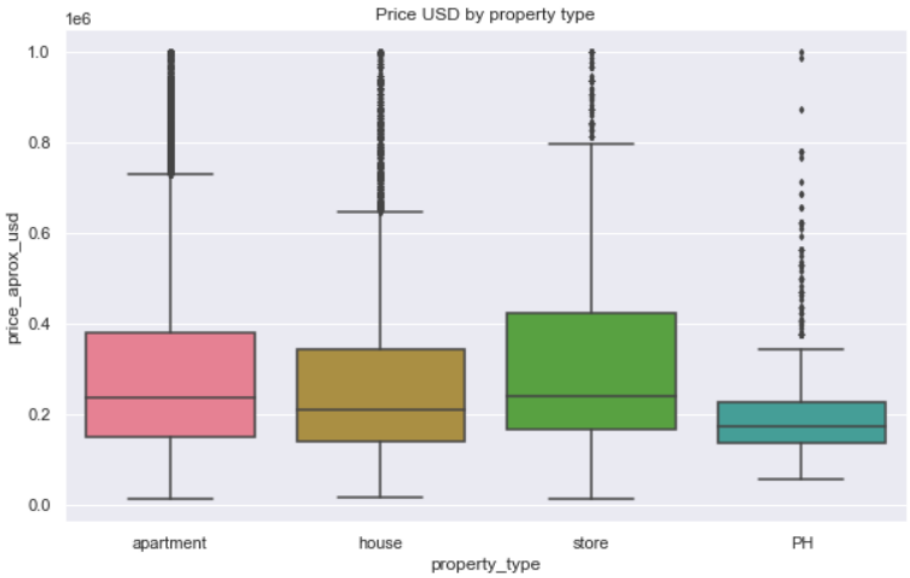
- Superficie en m^2: al igual que el anterior, se revisa gráficamente la distribución. Viendo que hay valores muy extremos, se reduce el rango a:
    - Superficie desde 40 m^2 a 500 m^2.



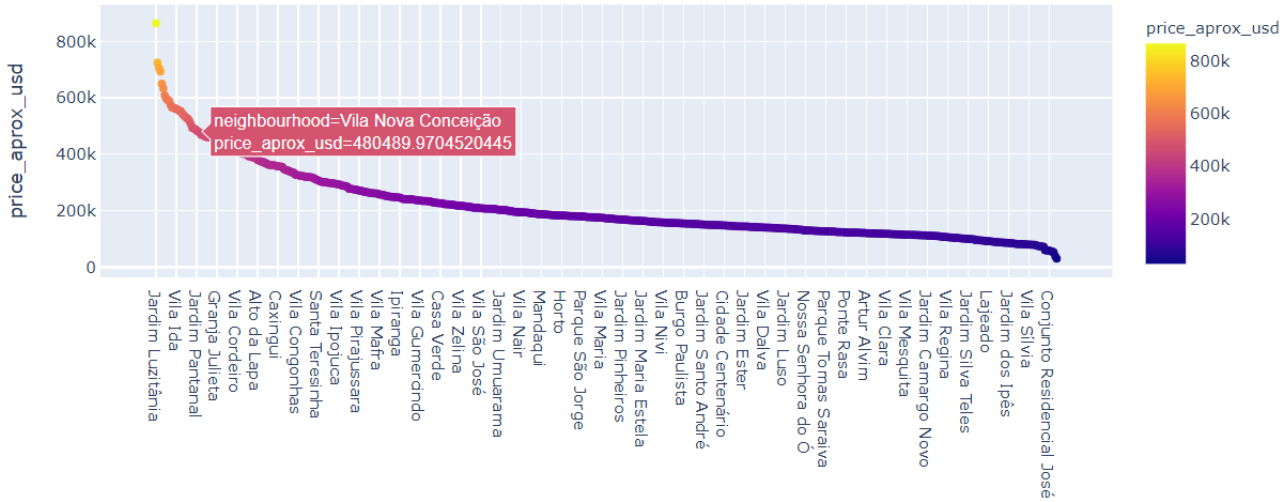
Tras toda la limpieza y filtrado nos quedamos finalmente con un 68% de los datos totales.

A partir de esto, con las columnas restantes se pueden obtener ciertas métricas y estadísticas. A nivel visual se pueden consultar con los siguientes gráficos:

- Precio del inmueble dependiendo del tipo. Por orden sería las tiendas, apartamentos, casas y propiedad horizontal.



- Precio medio por barrio. Se crea un gráfico interactivo en el cual poder visualizar las diferencias entre barrios con respecto al precio de los inmuebles. Esto también nos indica la importancia de esta variable para el cálculo del precio.



Finalmente, tras haber hecho toda la tarea de limpieza y análisis toca crear un modelo de regresión que pueda predecir mediante las variables numéricas disponibles.

La variable categórica property_type se convierte en variable dummy, mientras que la variable, también categórica, neighbourhood se deja como dummy para

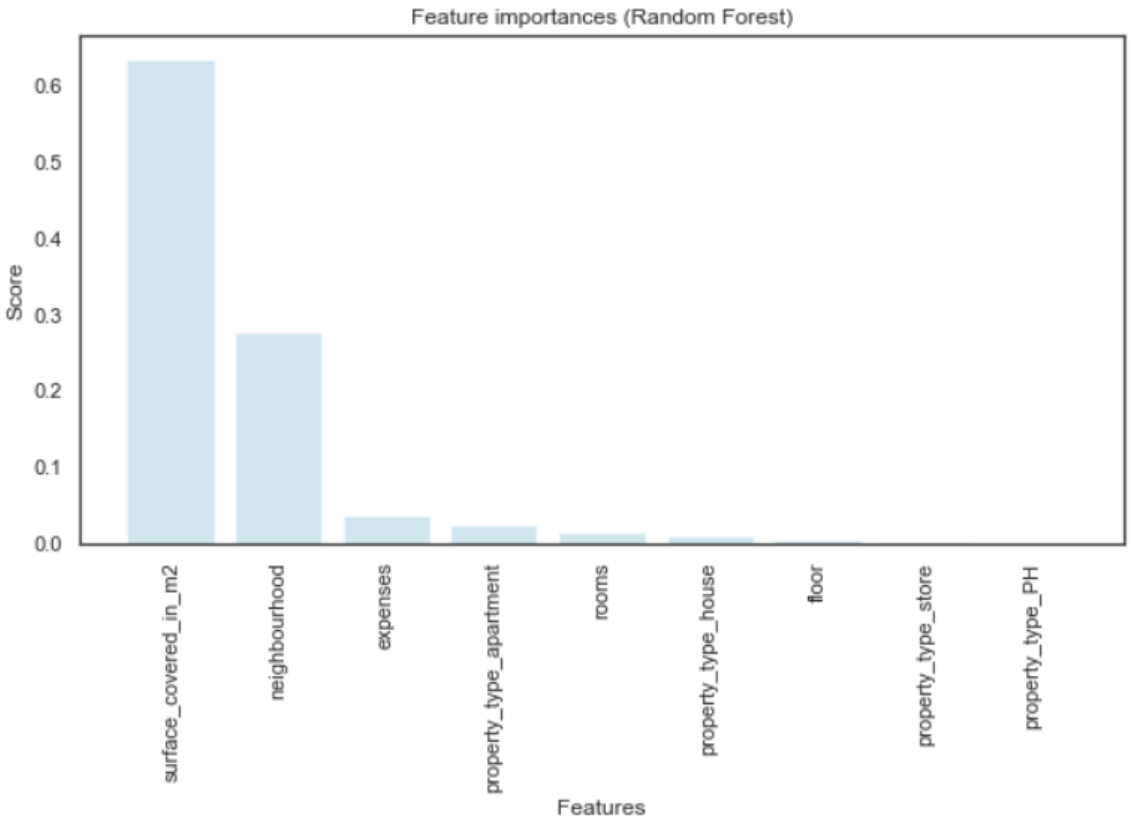
la regresión lineal mientras que para Random Forest y Gradient boosting se utiliza como número ordinal dependiendo del precio medio del barrio debido al coste computacional.

Se crea un conjunto X e Y tanto de entrenamiento como de test para luego calcular la eficacia del modelo.

Tras usar los tres algoritmos de regresión con los mejores parámetros posibles para cada modelo, se obtienen los siguientes datos:

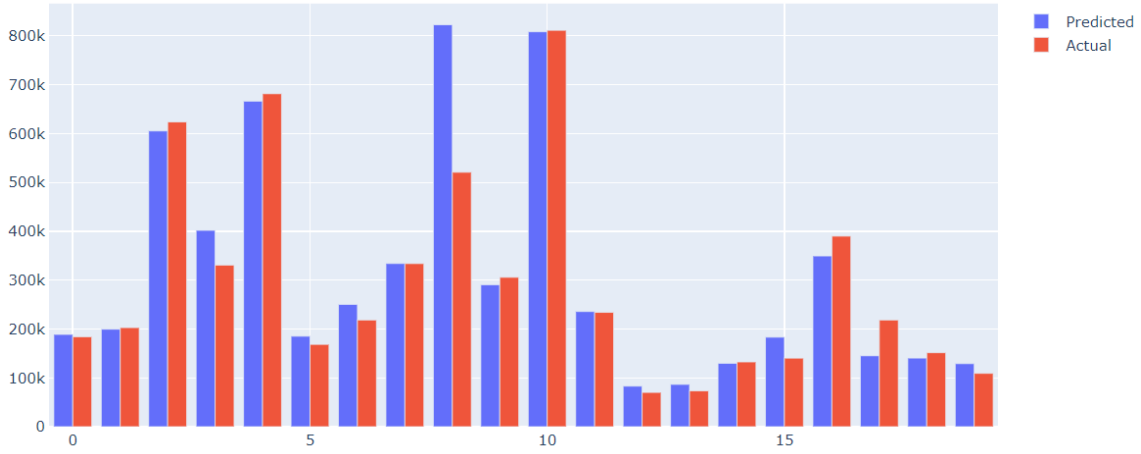
	algorithm	CV error	CV std	training error	test error	training_r2_score	test_r2_score
0	Linear Regression	1.009602e+10	1.863563e+08	1.001261e+10	1.013833e+10	0.736655	0.736350
1	Random Forest Regressor	5.862760e+09	1.126650e+08	2.575151e+09	5.912975e+09	0.932270	0.846231
2	XGBoost	6.081523e+09	1.034349e+08	4.777608e+09	6.149966e+09	0.874343	0.840068

Por lo tanto, por cualquiera de los dos criterios (R^2 y error cuadrático) nos quedamos con el Random Forest Regressor. El resultado es R^2 de 0.846 que es bastante bueno contando que se tienen muy pocas variables explicativas y de gran repercusión.



Finalmente se predicen los resultados en los conjuntos de test y se calcula el margen entre la valoración predicha y el precio de mercado. Guardamos este resultado en un nuevo dataframe que contiene la información completa, así

como la predicción y la diferencia en los que el resultado es que el inmueble está infravalorado.



Esto se guarda en una Excel para poder distribuir fácilmente en forma de tabla a los departamentos interesados.