Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Housing prices in Buenos Aires

21/03/2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

- **Exploratory Data Analysis** is a process that helps analysts to make sense of their data by looking for patterns, anomalies, and outliers. This helps in better understanding of the data before any modelling begins.
- The dataset used in the EDA process is Buenos Aires-Sheet1 which contains 16 variables and 8606 observations.
- The variables included, operations, property type, latitude and longitude, price of the properties in local and foreign currency, the surface are covered by the property etc.
- We seek to analyze the factors affecting price of properties in Buenos Aires.

Problem Statement

We seek to investigate;

- Is there a relationship between size of a property and property price?
- Is there a relationship between neighborhoods and property price?
- Is there a relationship between location and price?
- What factors are important in determining the price of a property?

Approach

The approach in analyzing this data included;

- Cleaning the data.
- Perform exploratory data analysis (EDA).
- Feature selection.
- Build the model.
- Test models.

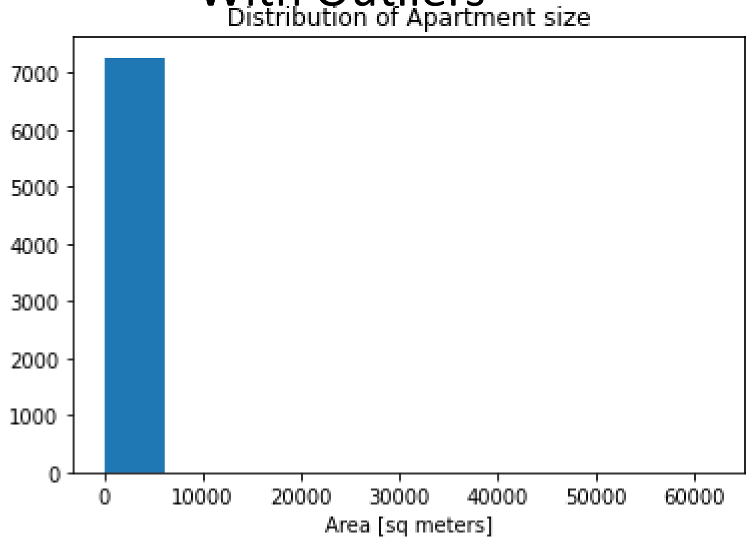
Data Cleaning and EDA

- The EDA processes done in this dataset include;
- The first thing I did was to check each data type of columns and check for missing values. I then dropped the ‘expense’ and ‘floor’ columns because they each had 89% and 85% missing data respectively.
- I then split the ‘place_with_parent_names’ column to create the neighborhood and state column. I also split the ‘lat-lon’ column and created the lat and lon columns to represent latitude and longitude respectively and changed their data types to float. I then dropped the ‘place_with_parent_names’ and ‘lat-lon’ columns.
- I dropped three columns ‘operation’ and ’currency’, for having low cardinality and also dropped 'properati_url‘ column fro having a very high cardinality and wouldn’t be useful in the analysis of the data.

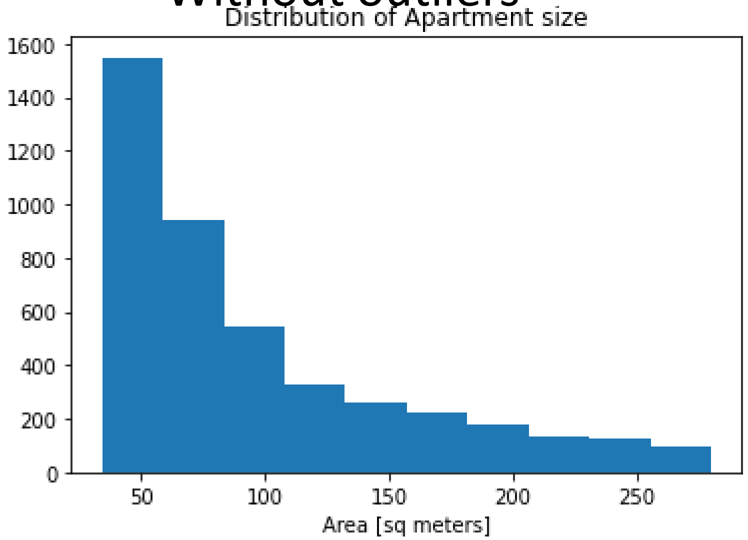
Continuation of EDA

- After plotting the histogram of the surface area covered, it could be observed that there was outliers present in the dataset. The outliers present below the 0.1 quantile and above the 0.9 quantile were removed.

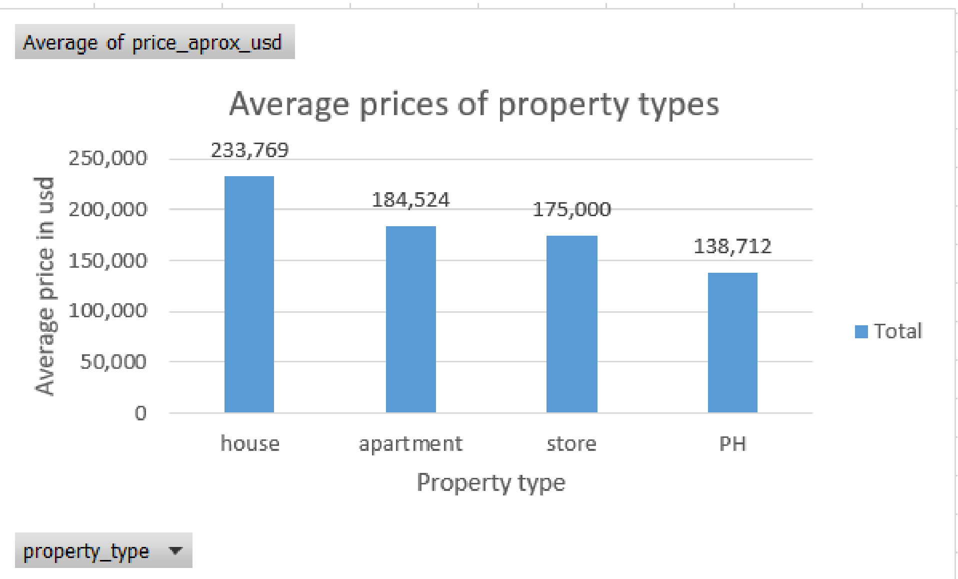
  - With Outliers

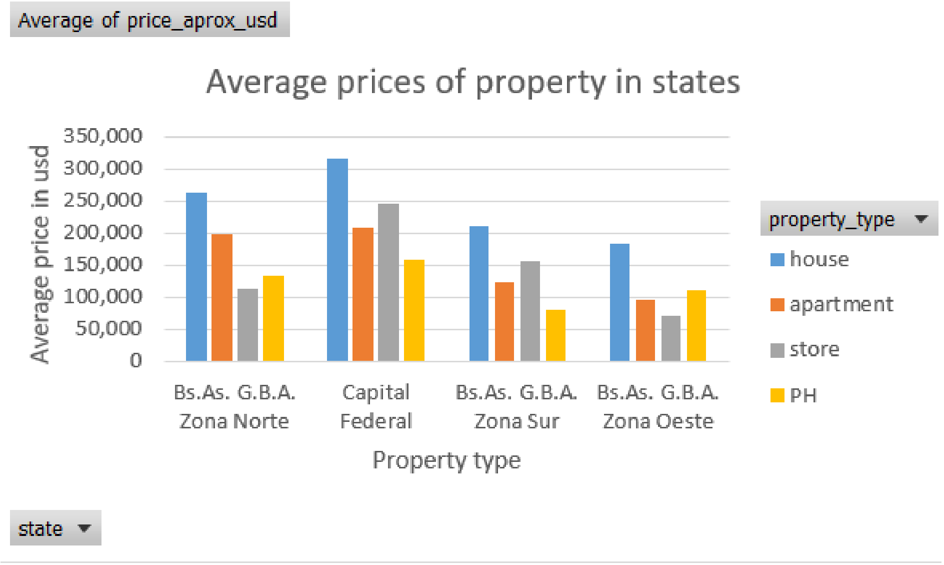


Without outliers

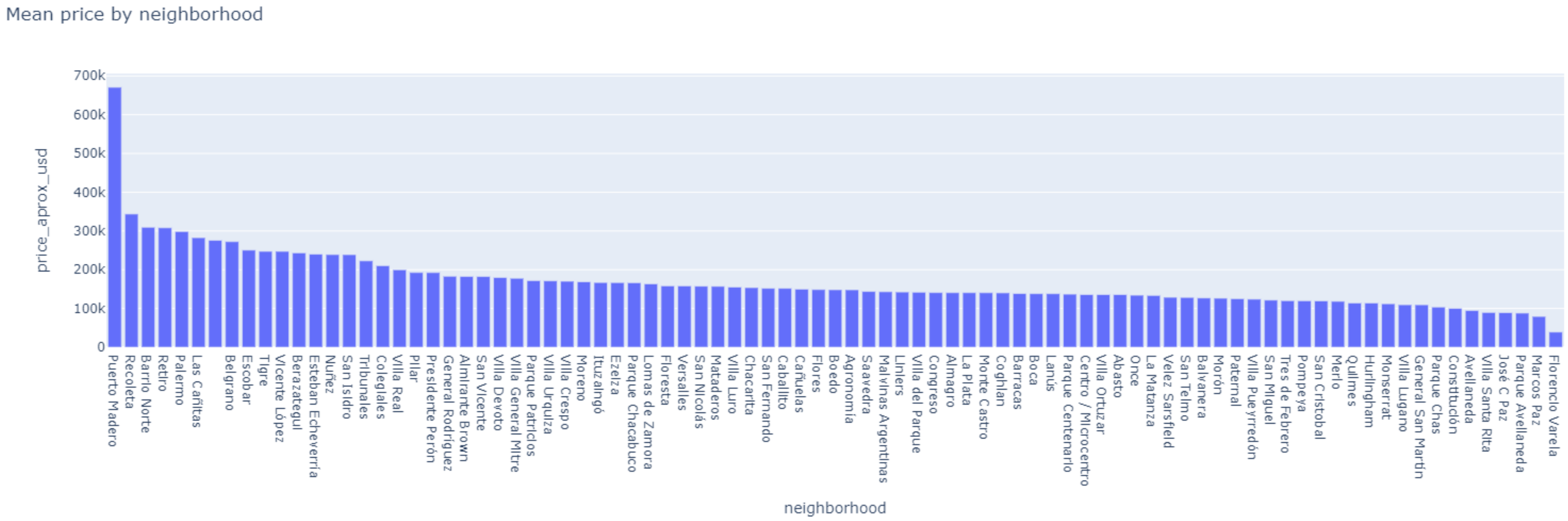


Mean Average prices

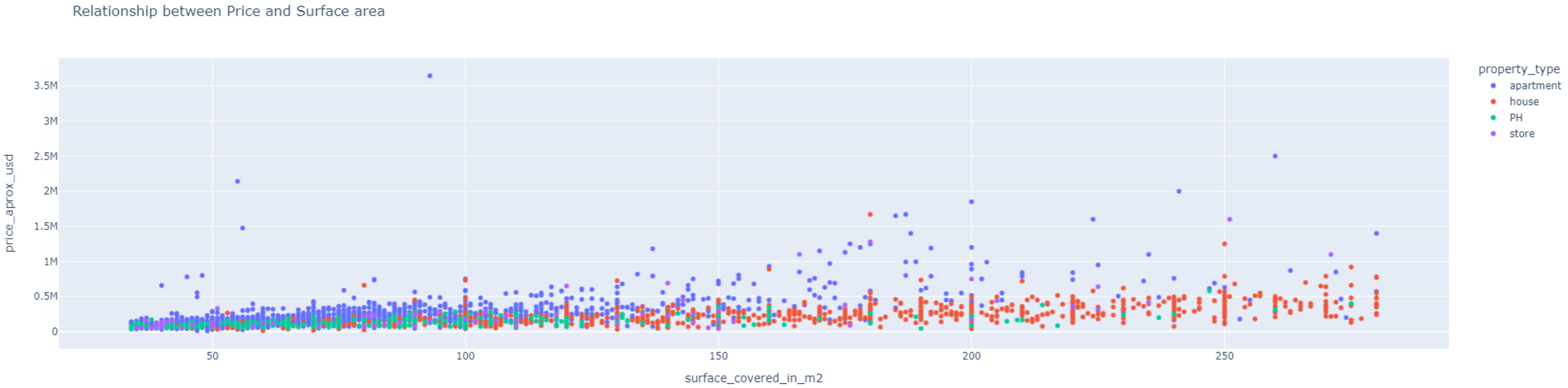




Is there a relationship between neighborhoods and property price?

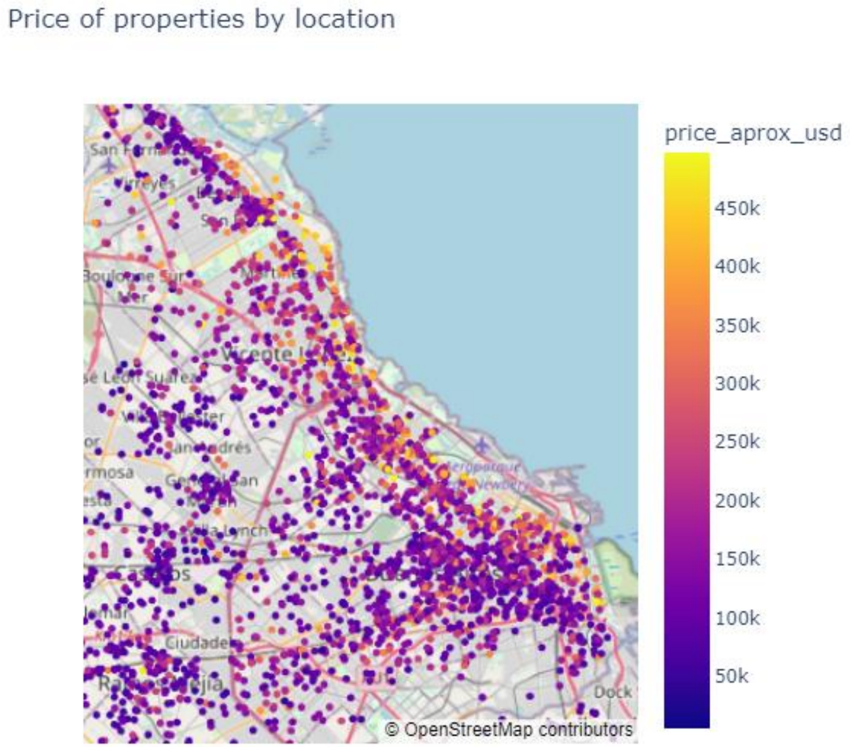

Mean price by neighborhood

Is there a relationship between size of a property and price?



- It can be noted from the above graph that most houses are bigger in size compared to the other property types.
- It can also be noted that as the size of the property increases there is a slight increase in the property price.

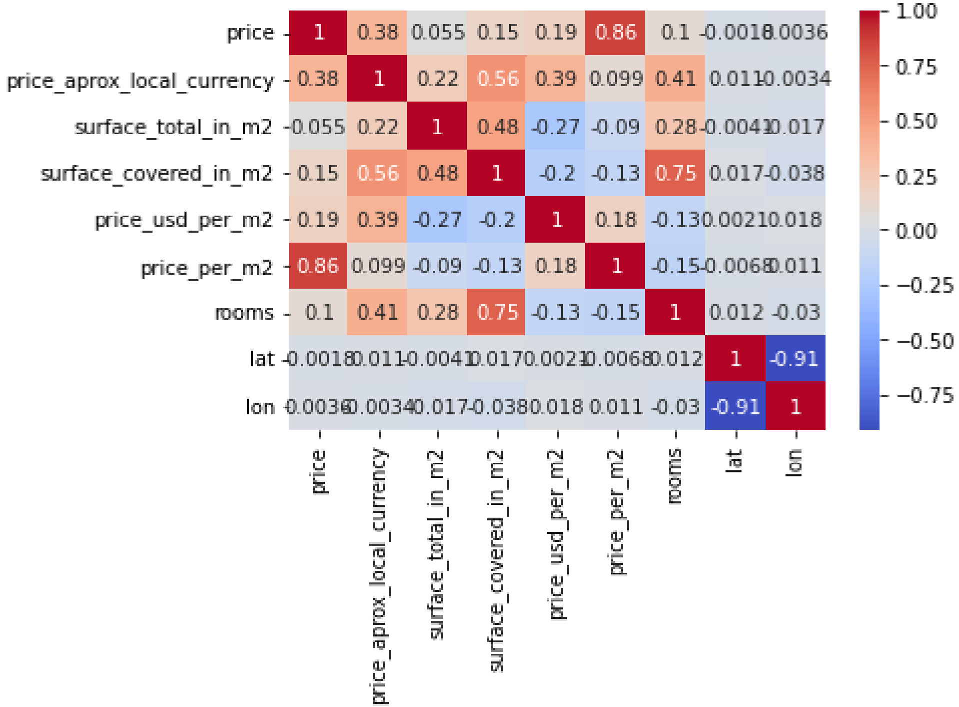
Is there a relationship between location and price?


Price of properties by location

- It can clearly be observed that properties along the beach are more expensive that properties towards the mainland.

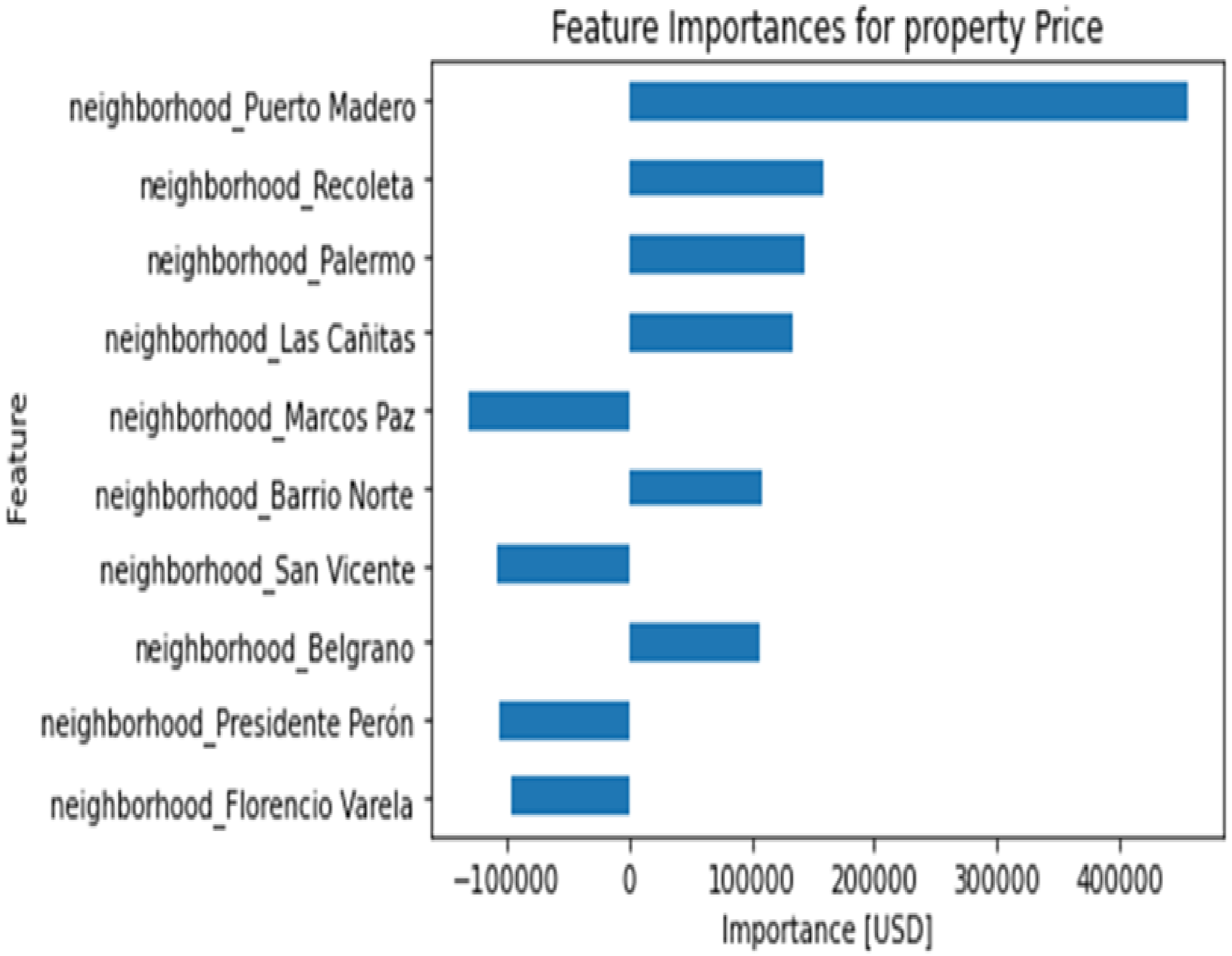
Feature selection and modeling

- A heat map was used to plot the correlation between different variables and select features used to build the predictive model

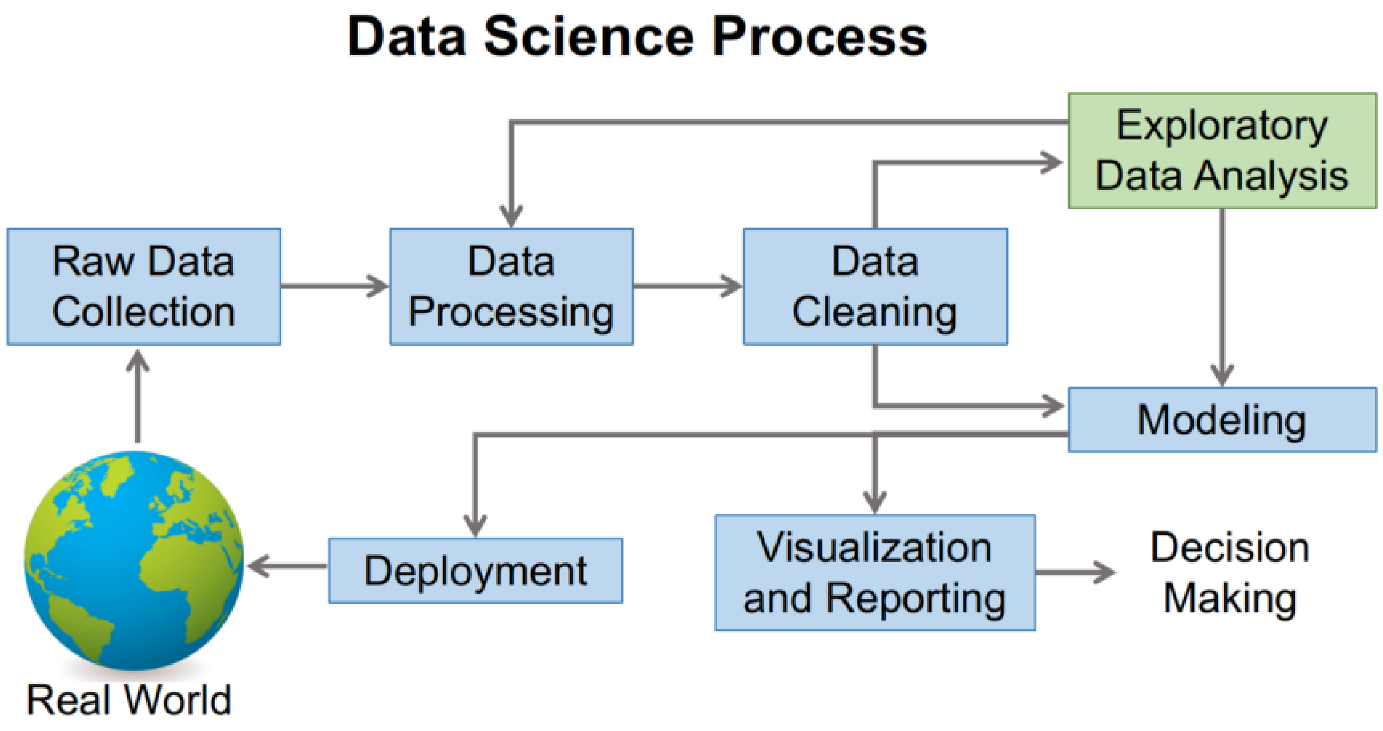


- Since we seek to predict **price_aprox_usd** we will drop all the column that contain price leakage for example; **price**, **price_aprox_local_currency, price_per_m2, price_usd_per_m2.**
- **Rooms** and **surface_total_in_m2** was dropped because it could cause multicollinearity.

- To build a predictive model, we used the features **surface_covered_in_m2**, **lat, lon**, and **neighborhood** to predict the target variable **price_aprox_usd**.
- This data set was split into training and test sets, categorical values were encoded using OneHotEncoder and a model was built with the resulting data..
- Evaluation of the dataset was done using data containing the features used.
- The bar graph on the right reflects the 10 most common features that are most important when determining a property price. Neighborhoods is by far the most important of these, with some neighborhoods making a property very expensive and others having a negative impact on prices, making them very cheap.



EDA Summary



EDA is a very important step in data analysis.

The EDA process can help identify obvious errors, find interesting relations among variables, detect outliers or anomalous events, and better understand patterns within the dataset.

Exploratory data analysis is the process of analyzing and better understanding a dataset before making any assumptions.

Recommendations

- When looking for a property in Buenos Aires, one ought to consider the neighborhoods and the location and size of the property to determine the price of the property.

Thank You

Data Glacier

Your Deep Learning Partner