

**Predicting Hotel Prices in Saudi Arabia by Using Machine Learning**

**Contents:**

1. **Introduction**
2. **Objectives**
3. **Business problem**
4. **Data description**
5. **Exploratory Data Analysis (EDA)**
6. **The approach used to analyze the data and extract the results**
7. **Results**
8. **Conclusion and future work**

**Introduction:**

As one of the largest and fastest-growing economic sectors in the world, the tourism sector tops the priorities of the Kingdom’s Vision 2030. The Kingdom of Saudi Arabia has adopted the National Tourism Strategy. This highlights the main axes of the sector’s aspirations that are in line with the objectives of the Saudi Vision, which include:

Increasing the tourism sector’s contribution to GDP from its current rate of 3% to more than 10% by 2030 and create thousands of jobs in the tourism sector by 2030. In addition to attracting domestic and international visits annually by 2030.

This sector includes various fields such as hotels, restaurants, development of tourist destinations, hospitality in general, and tour operators.

For this reason, and given the importance of the tourism sector in the Kingdom at the present time, we chose in our project to address hotel prices, which significantly affect the satisfaction and attraction of tourists.

**Objectives:**

* Analyzing the hotel's data and the available statistics to discover the extent of tourists' satisfaction with hotel prices in the Kingdom of Saudi Arabia and provide appropriate suggestions.
* Create models to predict hotel price and determine the best-used model.

**Business problem:**

Tourism is a new market in Saudi Arabia. Investment in the tourism sector in all its forms is one of the aspects of active investment, as it is one of the largest sectors that contribute to increasing the country's GDP. Therefore, the leakage of national income outside the country must be reduced by encouraging tourism investment within the country and removing obstacles that prevent to citizens and attract them internally instead of traveling abroad. Where the number of tourist trips outside the Kingdom during the summer of 2018, reached 9.4 million trips, a rate of 7.4% over the same period in 2017.

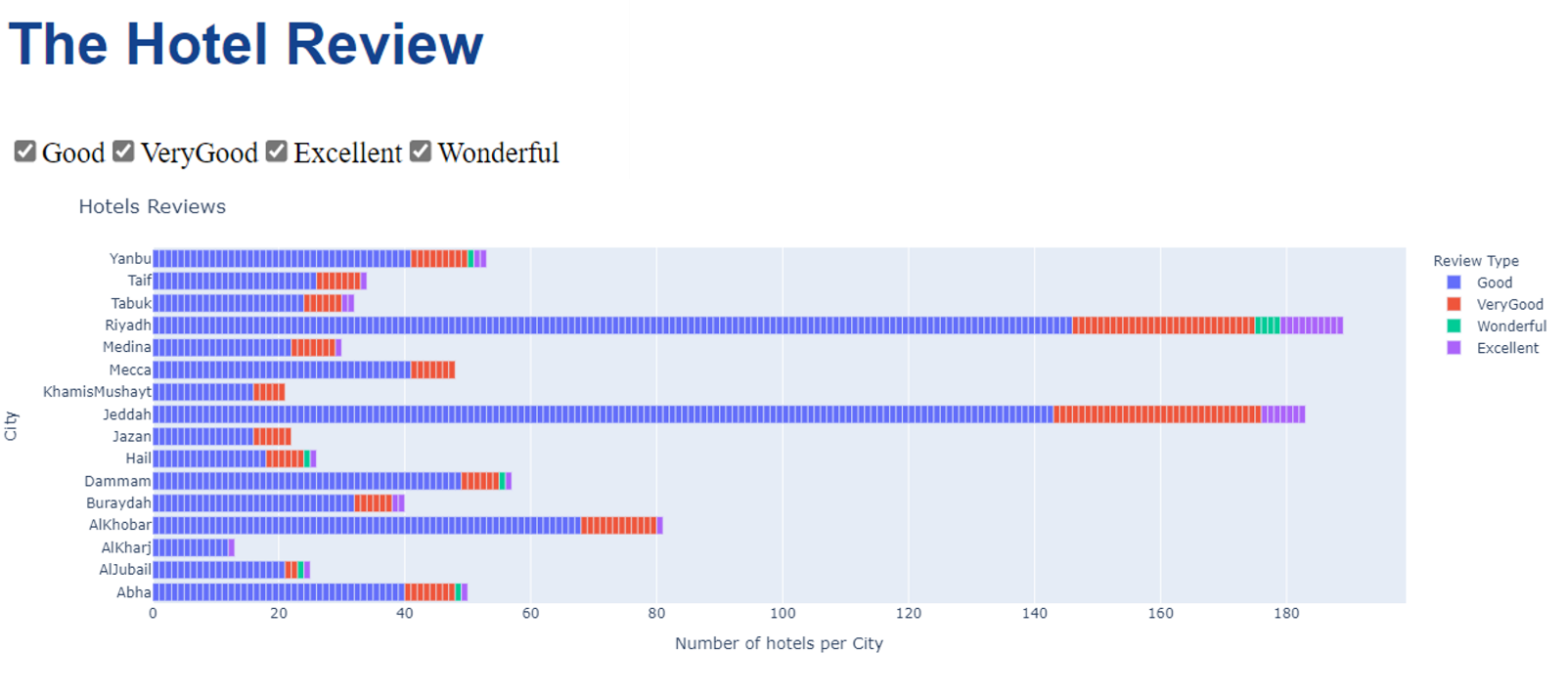
**Data description**

The data was downloaded from the Kaggle website https://www.kaggle.com/ghidaqahtan/booking-hotels. It pertains to the tourism sector and describes many aspects related to hotels in particular. Also, it provides accurate statistics for hotels in different regions of the Kingdom of Saudi Arabia, as it contains 1026 rows and 9 columns giving (hotel\_name, city, review\_badge, review\_title, review\_number, price, stars, taxes\_and\_charges, hotel\_type).

**Exploratory Data Analysis (EDA)**

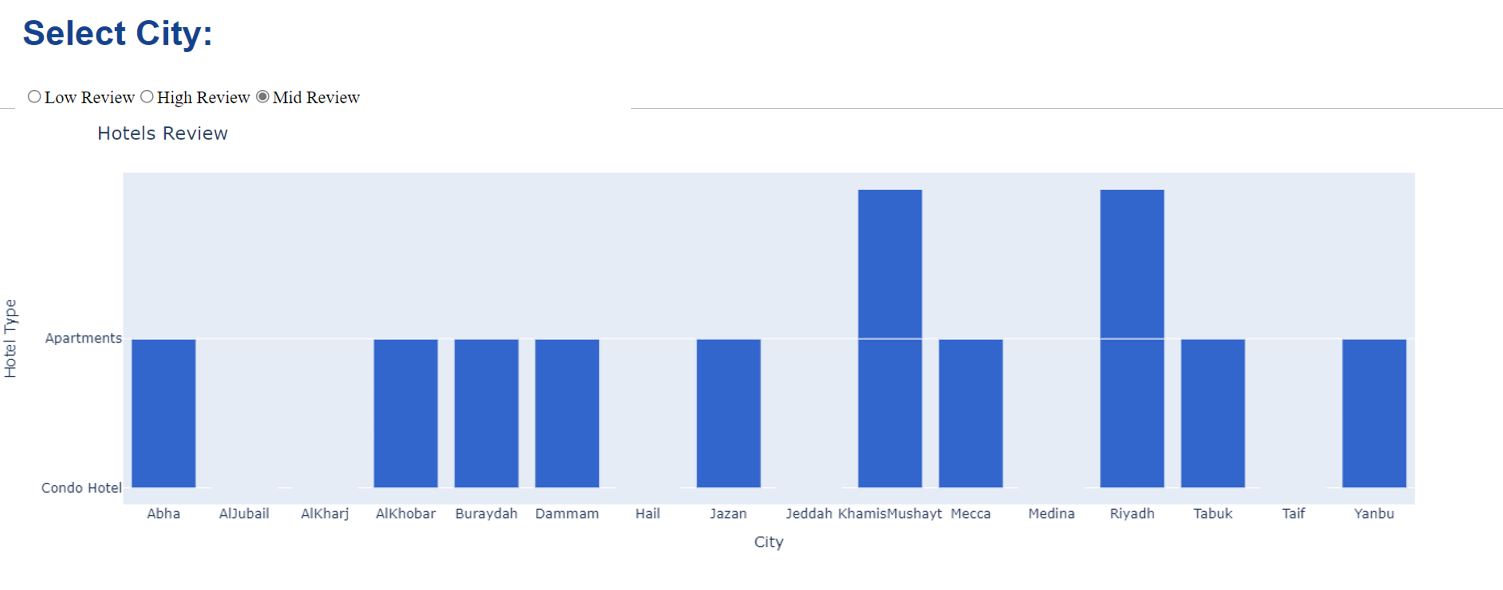
This part, is about creating visualization to help us explore the data and determine the directions of research and further investigation.

* The hotel reviews



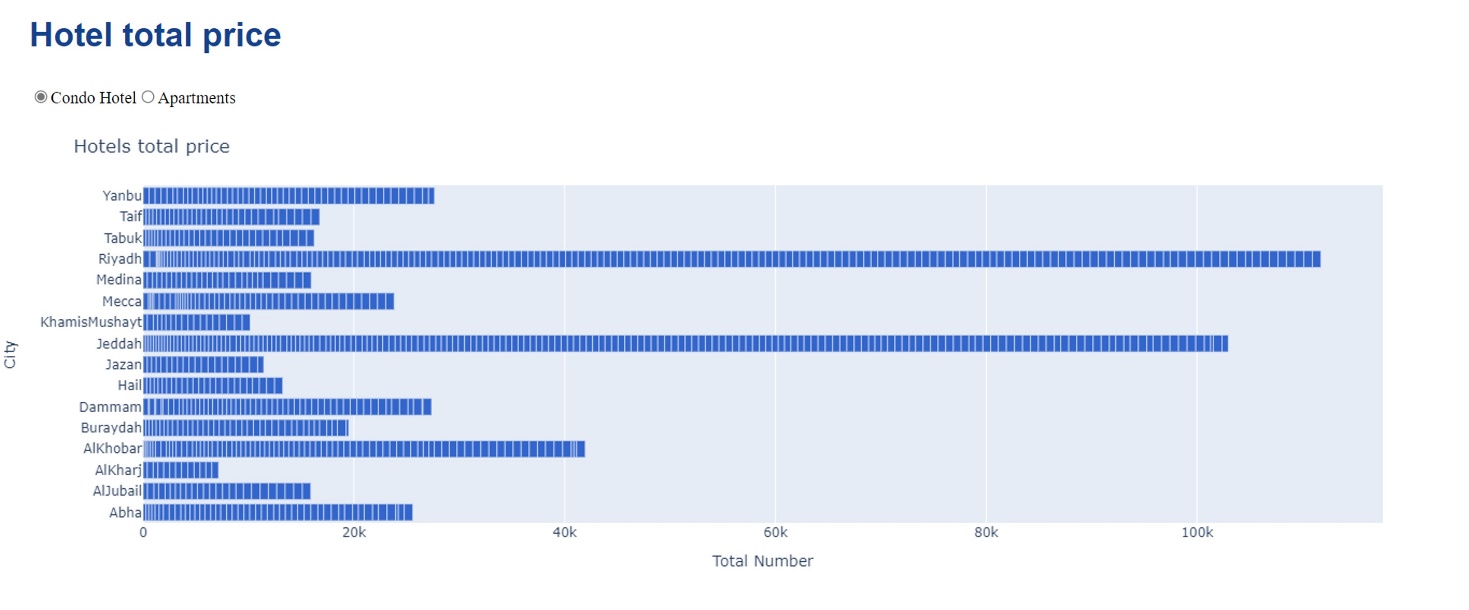
The data displays 4 types of hotel review types (very good, good, wonderful, excellent). The highest number of hotels which get excellent and wonderful reviews were in Riyadh. Al-Jubail, Taif, Medina, Hail, Dammam, Buraydah and Abha come after Al-Riyadh in getting excellent review type. Also, Al-Riyadh has the highest number of hotels which get wonderful reviews type. Jeddah and Riyadh have the highest number of hotels which received a good and very good review.

* Hotel’s review



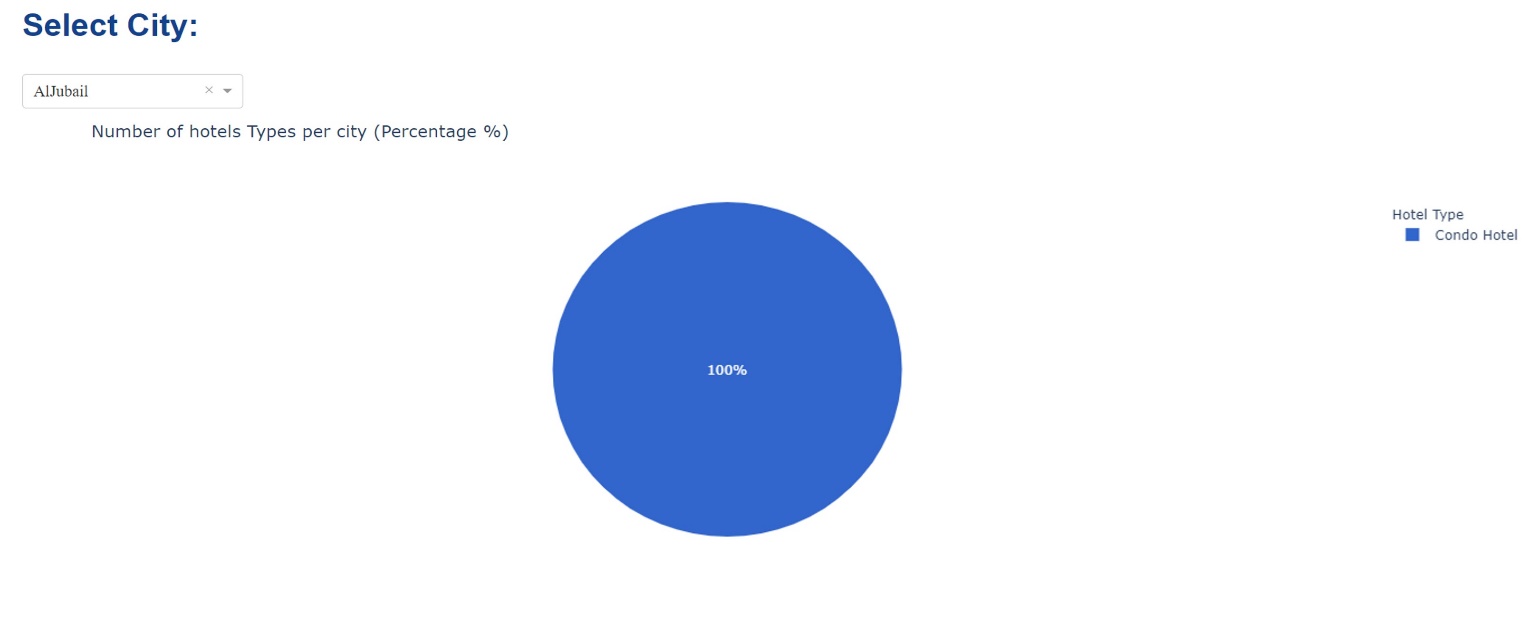
This plot displays the reviews (Low, High, Mid) for each Condo Hotel and Apartments. The upper section of the plot represents the Apartments, and the below section represents the Condo Hotel. For the Mid Review, Riyadh and KhamisMushayt have mid reviews apartments. Dammam, Yanbu, Riyadh, Abha, AlKhobar, Jazan, Tabuk, and KhamisMushayt have mid reviews Condo Hotels.

* Hotel total price



In this plot, we noticed for condo hotels type Riyadh and Jeddah have the highest total price, and KhamisMushayt and Al-Kharj have the lowest total price.

* Number of hotel types for every city



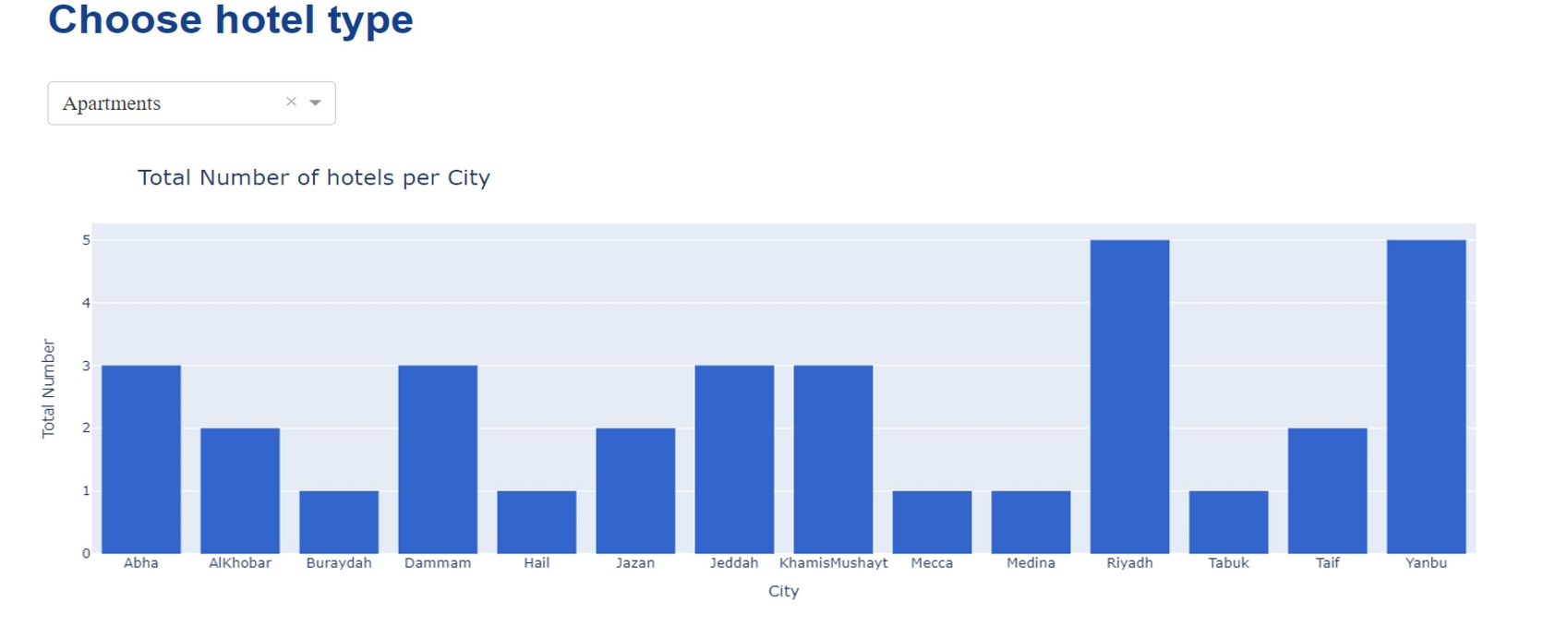
This plot displays the total number of hotels for every city based on hotel type. Some cities have only one hotel type. Al-Jubail have 100% of condo hotel.

Chart, pie chart

Description automatically generated

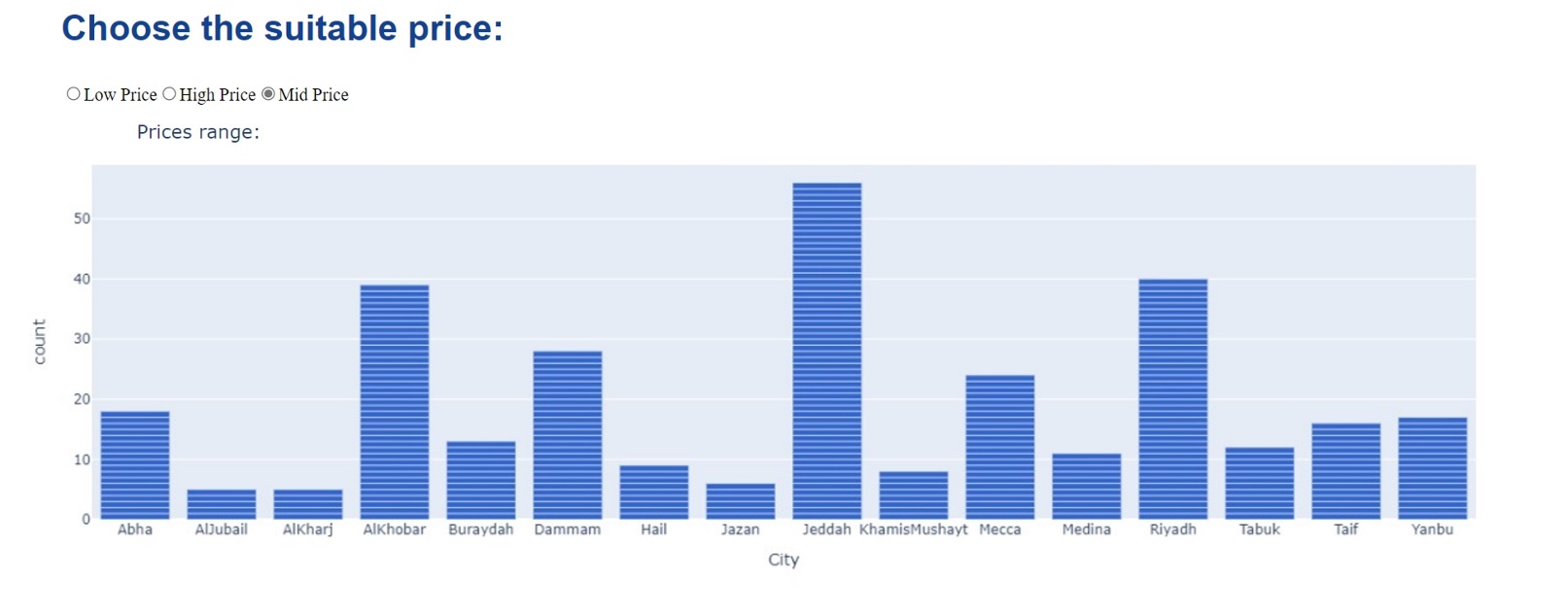
The other cities have both condo hotels and apartments and we noticed that condo hotel is higher for the other cities. For example, in Yanbu, condo hotel. is 90.6%, and apartments 9.43 %

* Total number of hotels per city



Yanbu and Riyadh have the largest number of apartments hotels. In terms of the lowest number of hotels, Tabuk, Medina, Mecca, Hail and Buraydah has the lowest number of apartments.

* Suitable Price



The plot displays the prices range (Low, High, Mid), as it shown Jeddah, Riyadh, and AlKhobar have the most mid prices.

**The approach used to analyze the data and extract the results:**

By using python and machine learning we studied and analyzed hotel prices and we compared them between the regions of the Kingdom of Saudi Arabia based on the type of hotels and hotels price. This was done by:

* **Cleaning (wrangling) the data:**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Clean data will boost overall productivity and assist you to make better decisions by allowing you to use the best information possible. This was done by:

1. Created total price column so that we can collect price column and taxes\_and\_charges column to get the final amount paid by the hotel guests.
2. Handling missing values by make a data frame for every city to properly replace the null values. We replaced the null values in columns by using the mode, mean and median according to what is appropriate for each subset of the data frame.

* **Machine learning models**

Machine learning methods assist individuals in exploring, analyzing, and deducing meaning from large data. We followed these steps in our analysis:

1. **split the data:**

In machine learning, it is a common practice to split your data into two different sets. These two sets are the training set and the testing set which works well with large datasets. We used the most split ratio of 80:20.

1. **Scaling the data:**

We scaled our data because it is a useful practice as a preprocessing step before using machine learning models in order to standardize the input dataset's range of functionality.

1. **Random Forest Regressor:**

Random forest is a learning algorithm that works on the principle of combining many decision trees to decide the final outcome. In this step we made a comparison between the real and predicted values then we calculated some functions as (MSE), (MAE) and (RSME).

1. **XGBoost Regression:**

To solve Machine Learning tasks by using the technique of the boosting tree. Also, we calculated the cost functions (MSE), (MAE) and (RSME).

1. **Grid Search for Regression**

we have defined the hyper parameters of the model which we want to pass to through GridSearchCV for getting the best parameters. Also, we have calculated grid best\_ score and grid best\_params. Finally, we printed the tree.

**Results**

* The cost functions for RandomForestRegressor model are:

MSE = 506.86165429834256

MAE = 8.791834254143646

RMSE = 22.51358821463923

* The cost functions for XGBoost Regression model are:

MSE = 656.7611973158267

MAE = 11.53004531438838

RMSE = 25.627352522565157

**Conclusion and future work**

For location that does not have apartment hotels, our model can give this a recommendation on whether to open it in these cities or not based on the market trends and historical data. Also, we noticed based on all regression models we used in this project for hotel data the best cost functions results were for the RandomForestRegressor model.