Rental Prices Analysis

Author’s name

Institutional Affiliation

Project Introduction

Rental price is one of the major reasons for increase in vacancy rates among Real Estates agencies in London. There are many factors that influence rental prices ranging from house attributes, geographical location and other surrounding features. However, these attributes differ in the extent at which they affect the prices. This project aims to find out the factors that influence rental prices.

Data was collected and stored in Excel files which were then read into pandas data frame. Descriptive charts and graphs were plotted to understand the data in detail. The data was the preprocessed by scaling to prepare it for Machine Learning models. The train data was trained on five different models and their performances evaluated. This whole process was carried out using Jupyter Notebook, an interactive computing platform that runs on browsers.

Project Objectives

The objective of this project was to;

* Load and inspect the data,
* Clean the data,
* Analyze and visualize the data
* Data preprocessing,
* Predict the house prices using machine learning models and evaluate the models.

Explanatory Data Analysis

**Loading and Data Inspection**

There were three datasets for the months of May, June and July inform of comma delimited (csv) files. They were loaded into “pandas” data frame. Most of the datasets were non numeric. There were only two numeric features, number of bathrooms and bedrooms.

**Data Cleaning**

The raw data was messy which required some wrangling to clean them. A month column was added to each dataset before merging the datasets row wise. The price had a Euro sign and “per month” unit string of words and they were stripped off. The area variable had value with an extra string characters “km” which were also stripped. The two variables, area and price were converted to numeric. The null values in all the numeric variables were filled by the variable mean. The non-numeric variables had no null values. There also existed duplicated rows in the dataset that were dropped keeping the first one.

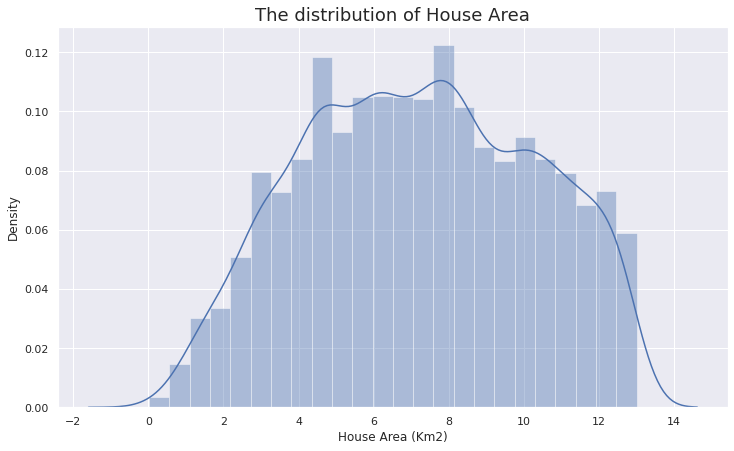
**Data Visualization**

A bar chart was plotted to compare the average house prices for houses that were furnished, unfurnished and those that were furnished at tenant choice.



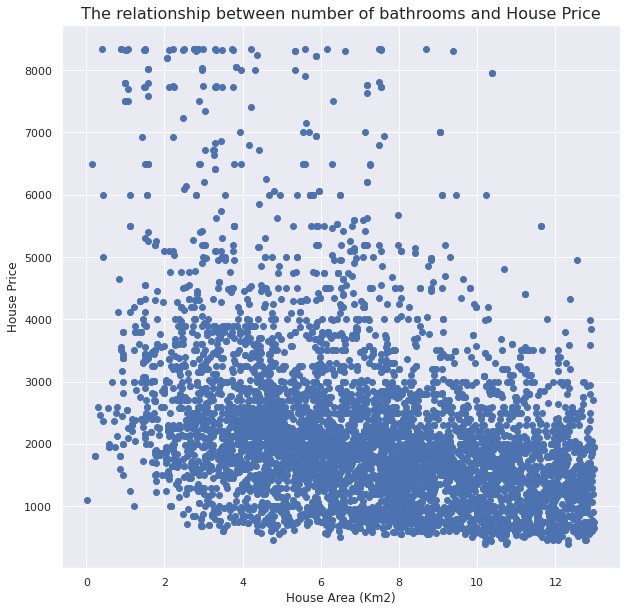
All the prices had an average price of above 2000 Euros per month. The houses that were furnished under at tenant’s choice were the most expensive while the unfurnished ones were the least expensive.

Thereafter a distribution plot was plotted to show the distribution of house floor areas in square units.



Most houses had an area of between five and nine square units. There were some outliers in the in the house area like 2 square units which is impossible which were removed.

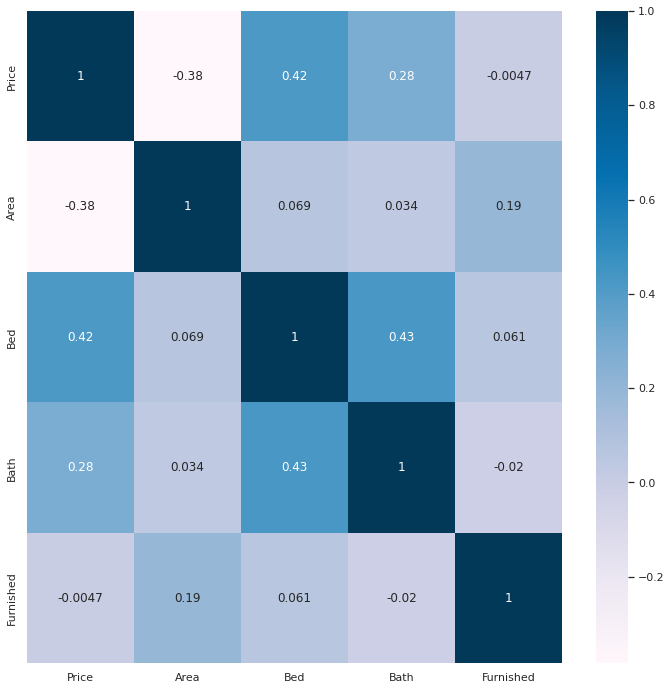
The relationship between price and area were visualized using scatter plot.



There were some houses with large area that were cheaper than some houses with small area showing that there is no relationship between area and price.

**Correlation**

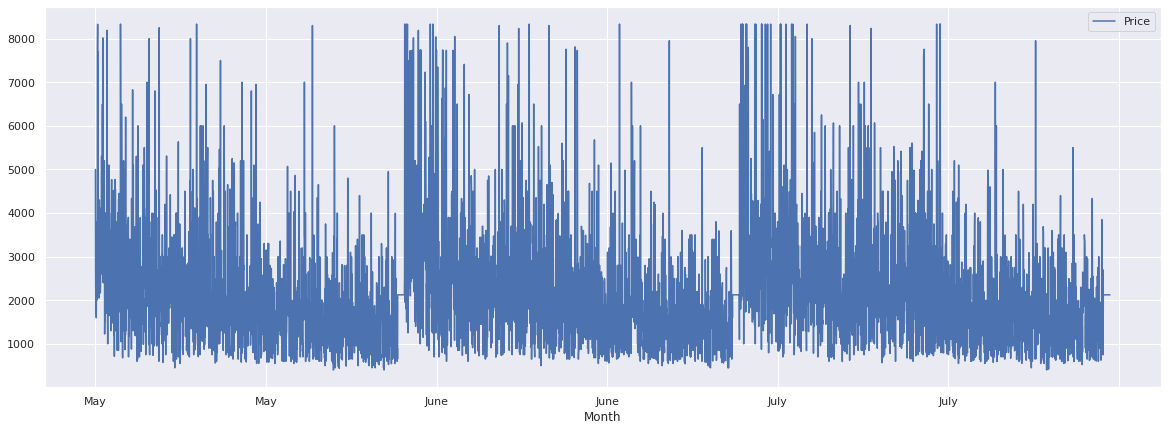
The correlation between the numeric variables was computed and the results were plotted on heatmap.



The correlation was compared against the price variable, bed and area had the highest absolute correlation with the Price variable.

**Trend**

The trend in prices from May to July was visualized in a line chart. There was no clear



Data Preprocessing

The price variable was the dependent variable and the other variables were the independent variable. The independent datasets were then scaled by minmax scaler in the range of 0 to 1. Both the independent and dependent datasets were split into train and test split in the ratio of 80% to 20%.

Data Modelling

Since the target feature (house price) was a continuous data, only regression models had to be used. The regression models were Linear regression, Ridge regression, Random Forest regression, Support Vector Machine regression and Ada Boost regression. The models were fitted using the train data and later evaluated using test data. The R-squared score was used to compute the models’ accuracies. The first model, Linear regression had a low accuracy of about 38%. The other regression models were then introduced. The best model overall was the Random Forest Regressor which had an accuracy score of approximately 67% and the worst model was the Support Vector Machine regression which an accuracy score of less than 1%.

Conclusion

The main focus for this research was to find factors that affected rental prices. The house area, number of bedrooms and bathrooms for each house, house status, and the prices were analyzed.

From the research, the line chart did not show a clear trend for the rentals price from May to July, however, the bar chart showed that there was a gradual upward trend. Tenants preferred houses with a moderate size, very big and small houses drove away more customers according to the distribution plot. Furthermore, custom furnished rental houses were more expensive than both the unfurnished and furnished ones. The model’s accuracies were very low.

Investors are recommended to invest in real estates since its values appreciates over time. Additionally, the real estate owners are recommended to build and invest in moderate house size since they attract most customers. There were other factors that influence rental prices which are proximity of social amenities like hospital, schools and banks from the houses (Sirmans, G & Benjamin J.D, 2019). Availability of smart appliances in the rentals also attracts more tenants. The Random Forest regression accuracy can be improved by adding more features and more data rows.

References

Sirmans, G & Benjamin, J.D (2019). Journal of Real Estate Research: Determinants of Market Rent. Pg 357 - 380. https://www.researchgate.net/publication/5142421\_Determinants\_of\_Market\_Rent/citation/