London Airbnb

Yashvardhan Singh

Abstract

- Dataset: London Airbnb Data, investigating Airbnb activity in London, United Kingdom.
- Question: What is the best time, neighborhoods and prices to visit London by Airbnb.
- In this investigation we will know the data, clean, make some graphs in order to understand them better and finally use a linear regression algorithm to try to predict the value of a room given input data

Motivation

When you travel, it is very important to choose a good place to stay both by location, economy, or a balance between the two. My research focuses on knowing how London is in terms of costs, type of housing offered in Airbnb. and through this model offer a prediction of the value for the rent of a lodging data input variables.

As a particular data I would like to verify if start of academic session raises prices.

Dataset(s)

My dataset is London Airbnb Data-Investigating Airbnb activity in London, United Kingdom.

This dataset contains 6 .csv files with data since November 07th, 2018 and contain detailed listings data, review data and calendar data of current Airbnb listings in .

listings.csv is the principal file with a 82440 rows and calendar.csv with 85068 rows are the most important files in the dataset.

URL: https://www.kaggle.com/labdmitriy/airbnb#listings.csv

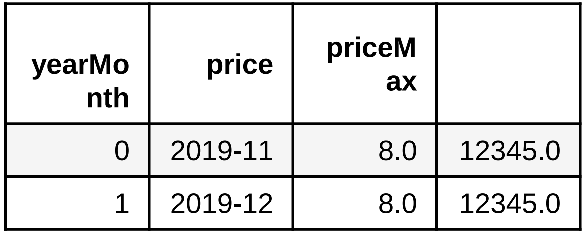
Data Preparation and Cleaning

It was necessary to change the price format because it included the $ sign and comma separation.

```
*********** Price from $0.00 TO $999.00
```

It was necessary to make a merge between the initial list and the neighborhoods to group them by communities.

It was necessary to take all dates to yyyy-mm format, to group by month



yearMonth	price	priceMax	
0	2019-11	8.0	12345.0
1	2019-12	8.0	12345.0

Research Question(s)

1. What is the best time, neighborhoods and prices to visit London by Airbnb?

2. What are the maximum, minimum and average prices per month in London?

Methods

- The method used was a first look at the 6 files of the dataset, choosing the most relevant (3). Know the fields and create a relationship between them.

- Once you have the data, you create graphs of lines, bars and rotate to obtain a better visualization.

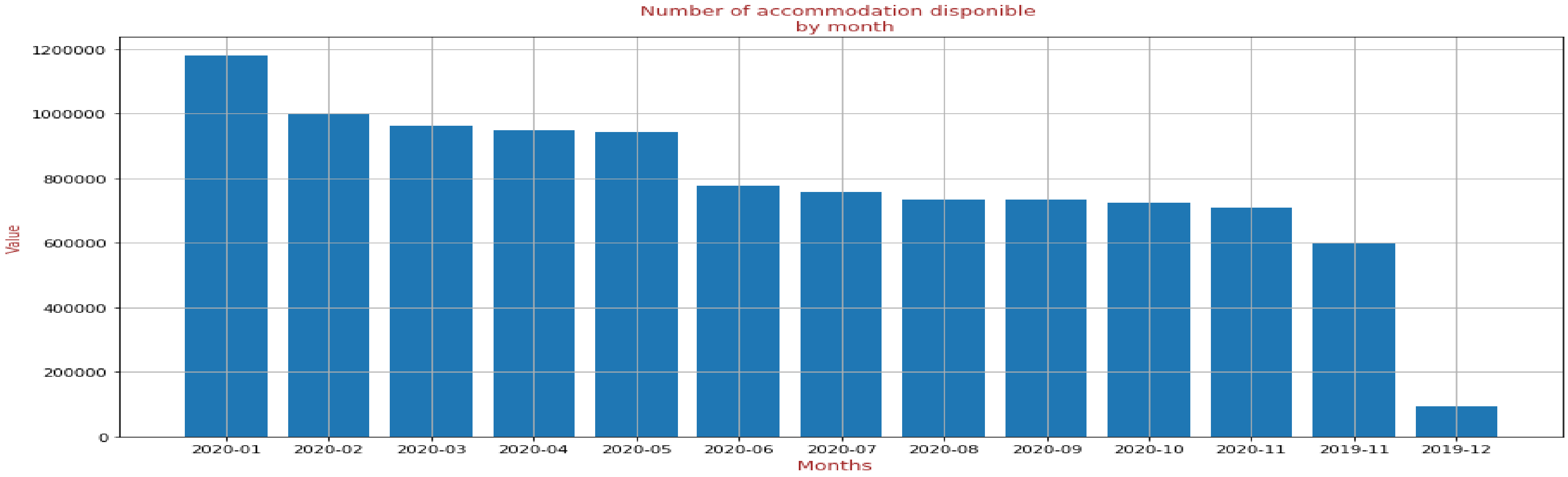
- Finally, a linear regression model is implemented with sklearn and the data was divided into training and testing using mean_squared_error.

Findings

- First the calculation of the maximums and minimums in date and price

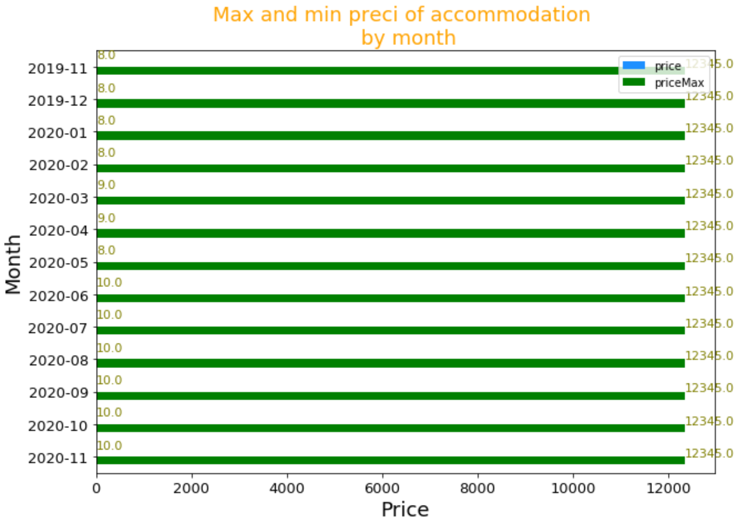
```
########### Years from 2019-11-05 TO 2020-11-04
```

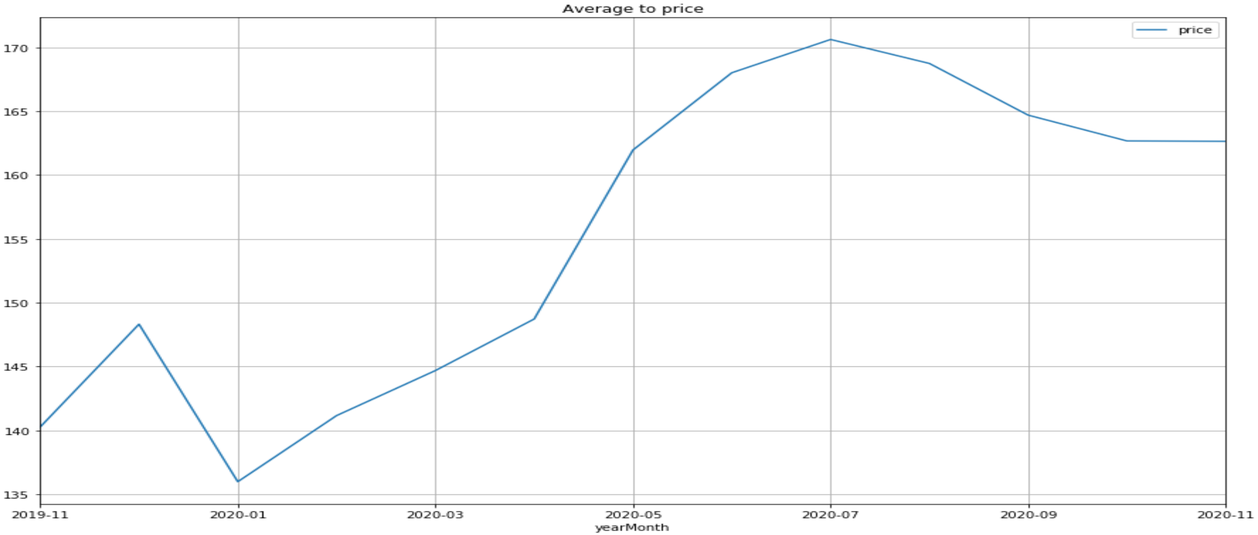
- The graph shows the number of accommodations registered per month



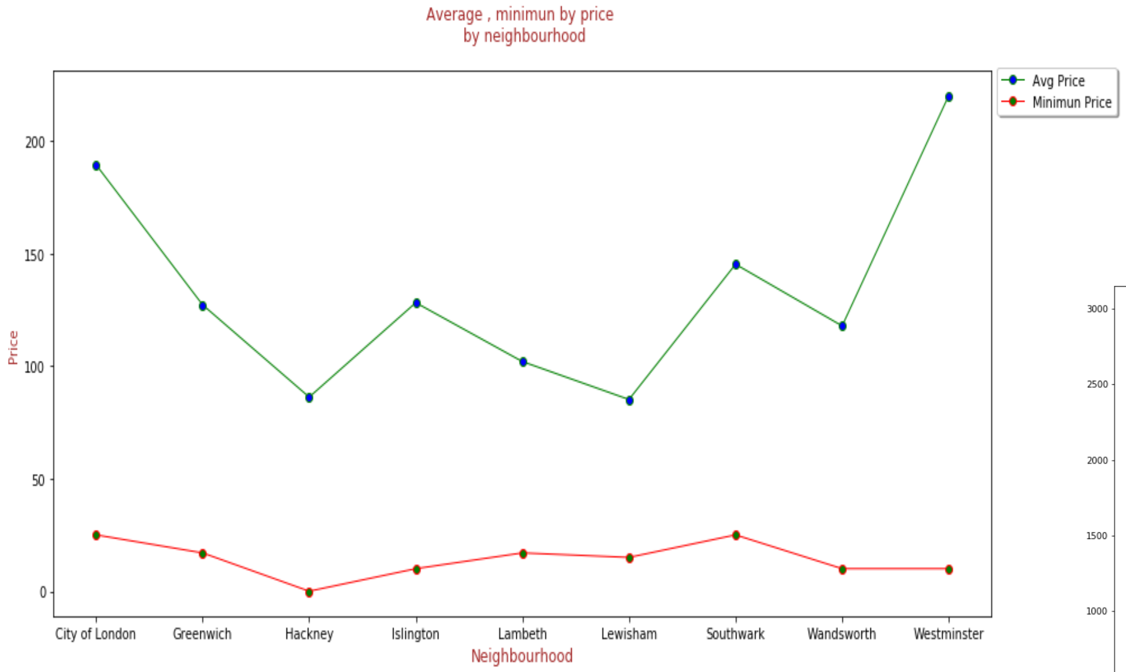
Findings

- The graphs shows of the maximums and minimums and average in date and price registered per month. Is not clear by the difference but the number is 10 and 12345

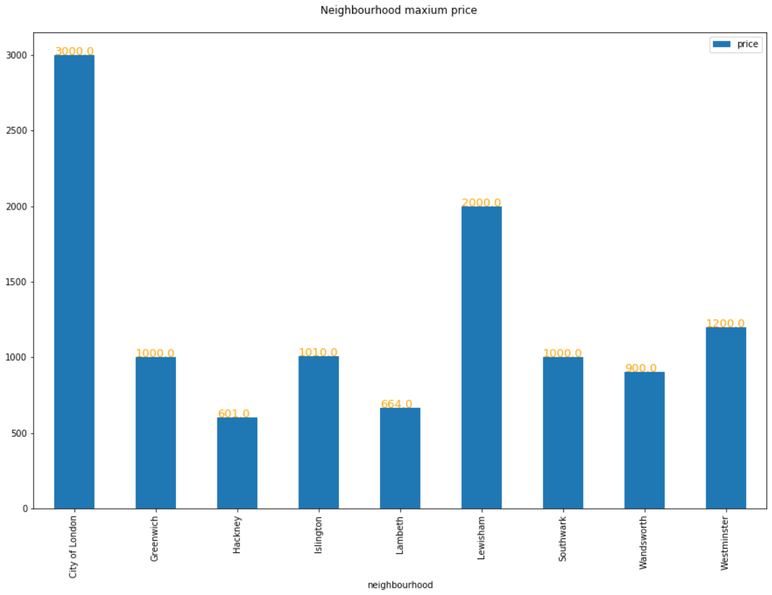




Findings

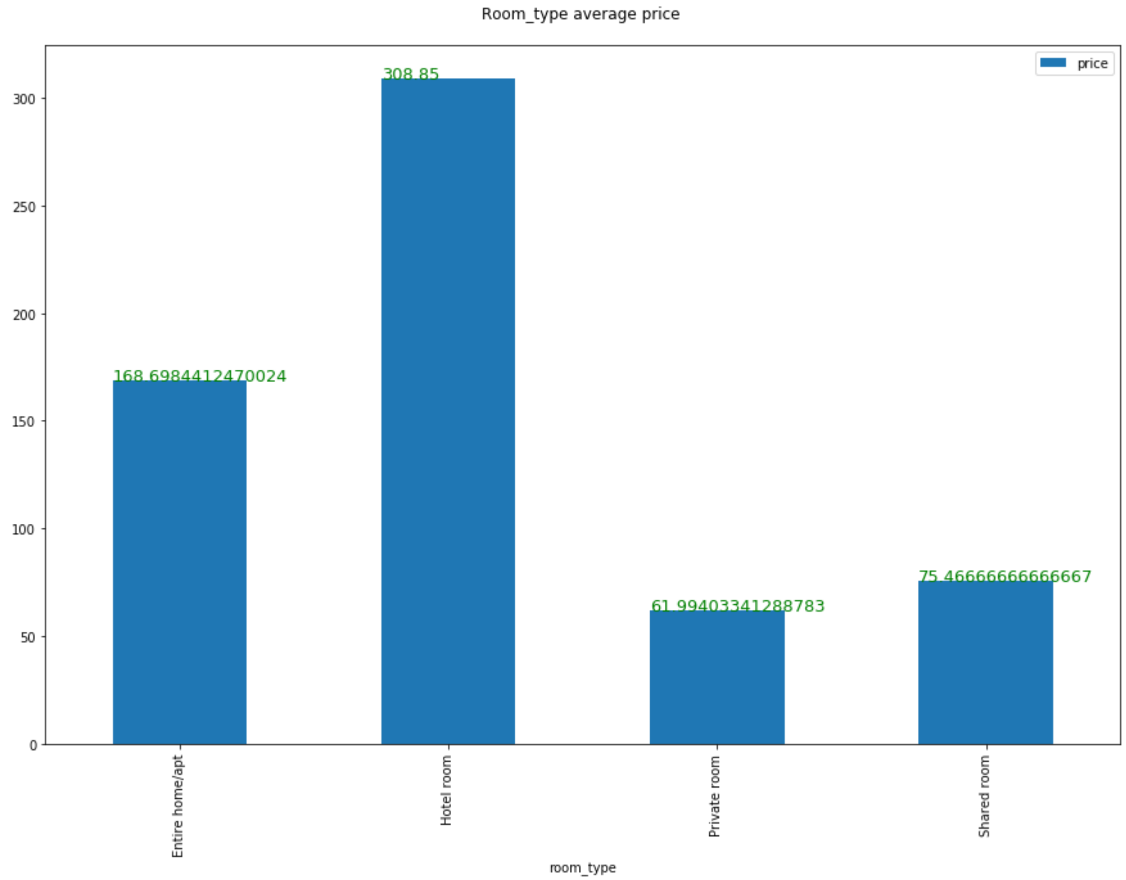


- This graph show the average, minimum and maximum by price and neighborhood group



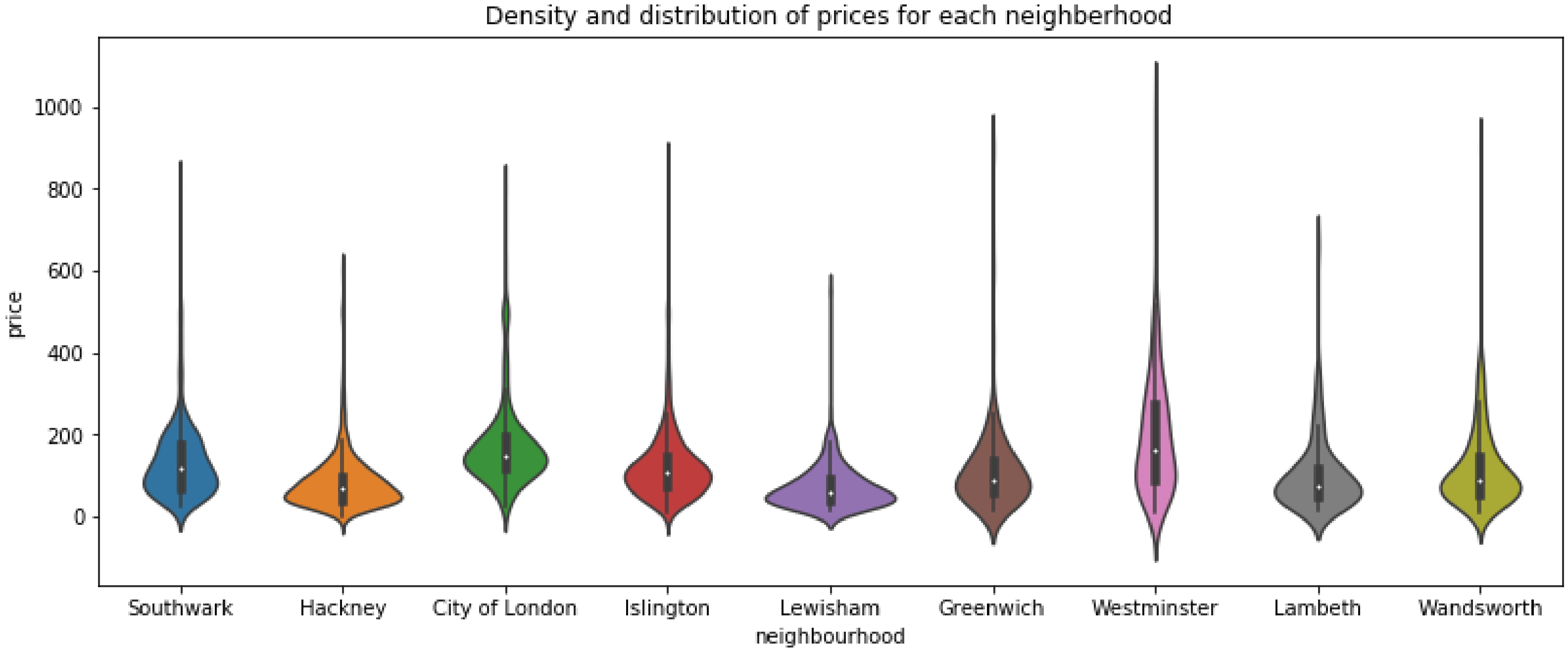
Findings

- This graph show the average price by room type



Findings

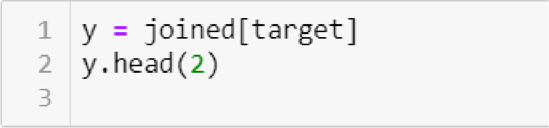
- This graph show the density and distribution of prices of Airbnb in London neighborhood.



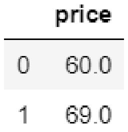
Findings

- These are the fields selected to do the regression lines and the target is Price field

latitude	longitude	minimum_nights	number_of_reviews	availability_365	
0	51.50205	-0.10015	1	91	353
1	51.49865	-0.10284	3	146	0



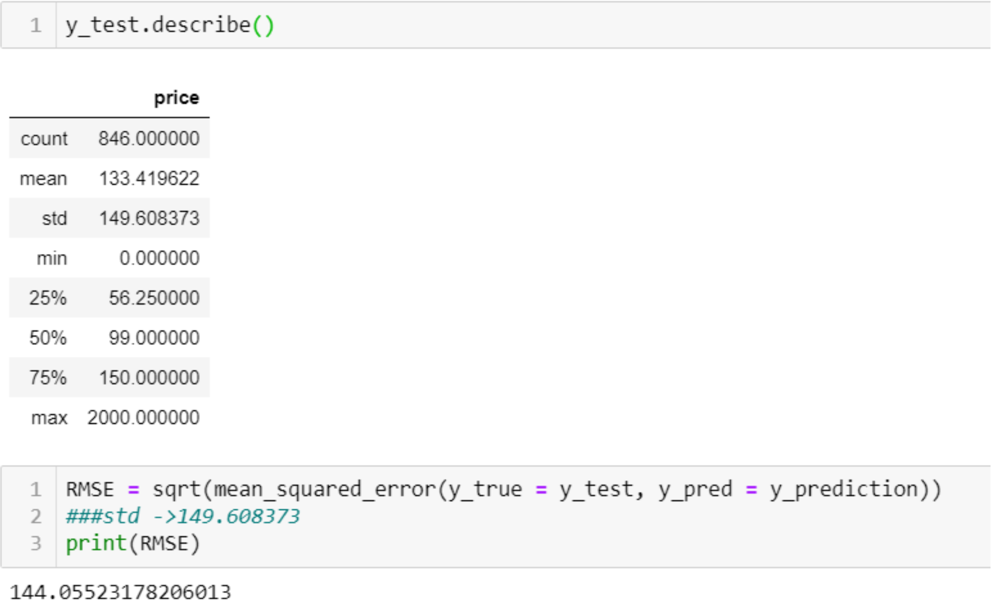
```
y = joined[target]
y.head(2)

```



	price
0	60.0
1	69.0

Findings

- This is the description of the test dataset, with an average of 149.608373. After that, the mean square error test is applied with a result of 144.0552317 which shows that it is a good model.



```
y_test.describe()
```

	price
count	846.000000
mean	133.419622
std	149.608373
min	0.000000
25%	56.250000
50%	99.000000
75%	150.000000
max	2000.000000

```
RMSE = sqrt(mean_squared_error(y_true = y_test, y_pred = y_prediction))
###std ->149.608373
print(RMSE)
```

```
144.05523178206013
```

Limitations

- The data is only since 2019-11-05.
- It would be good to be able to cross the data with other cities or apply the model to others.
- I would like to show the average prices by neighborhoods as is done in gapminder.

Conclusions

- There are accommodations for all prices
- In May is where more accommodations are available.
- From July to November 2020 the number of accommodations is almost similar.
- The average cost is 135 to 160 dollars but for 2020-07 it was 171, which shows that it is a time of great movement by the beginning of the academic session.

Conclusion

- On average it is just as expensive to rent a private room as a shared room

- The regression model applied proved to be very attached to the test dataset average

- In the future the dataset can be used to further explore the data using the location of the apartments and the observations left by users.

- For the Session, Academic session there is a difference with respect to the other times of the year in prices and availability.

Acknowledgements

use the London Airbnb Data dataset, which I found in the repository of https://www.kaggle.com.

Also investigate the stations in London, the most major events and the number of visitors in the city.

References

- https://stackoverflow.com/questions/31468176/setting-values-on-a-copy-of-aslice-from-a-dataframe?rq=1
- https://github.com/scentellegher/code_snippets/blob/master/pandas_groupby_unstack/Plot_groupby_multiple_columns_unstack.ipynb
- https://pandas.pydata.org/pandasdocs/stable/reference/api/pandas.to_numeric.html
- https://robertmitchellv.com/blog-bar-chart-annotations-pandas-mpl.html