Berlín Airbnb

Sebastian Toro

Abstract

Dataset: Berlin Airbnb Data, Investigating Airbnb activity in Berlin, Germany

Question: what is the best time, neighborhoods and prices to visit berlin 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.

GitHub project : https://github.com/sebastianToroTeam/PythonForDataScience

https://www.kaggle.com/brittabettendorf/berlin-airbnb-data#listings.csv

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 Berlin is in terms of costs, type of housing offered in Arbnb. 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 Oktoberfest raises prices

Dataset(s)

My datset is Berlin Airbnb Data-Investigating Airbnb activity in Berlin, Germany

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 Berlin.

listings.csv is the principal file with a 22552 rows and calendar_ sumarry.csv with 8231.480 rows are the most importan files in the dataset.

Url: https://www.kaggle.com/brittabettendorf/berlin-airbnb-data#listings.csv

Data Preparation and Cleaning

It was necessary to change the price format because it included the $ sign and comma separation. `######Price from  $1,000.00  max  $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	price2
2019-05	76.0
2019-05	76.0

Research Question(s)

1. what is the best time, neighborhoods and prices to visit berlin by Airbnb?
2. what are the maximum, minimum and average prices per month in Berlin?
3. what are the maximum, minimum and average prices per month in Berlin?

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. ***Fig1***

- 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.

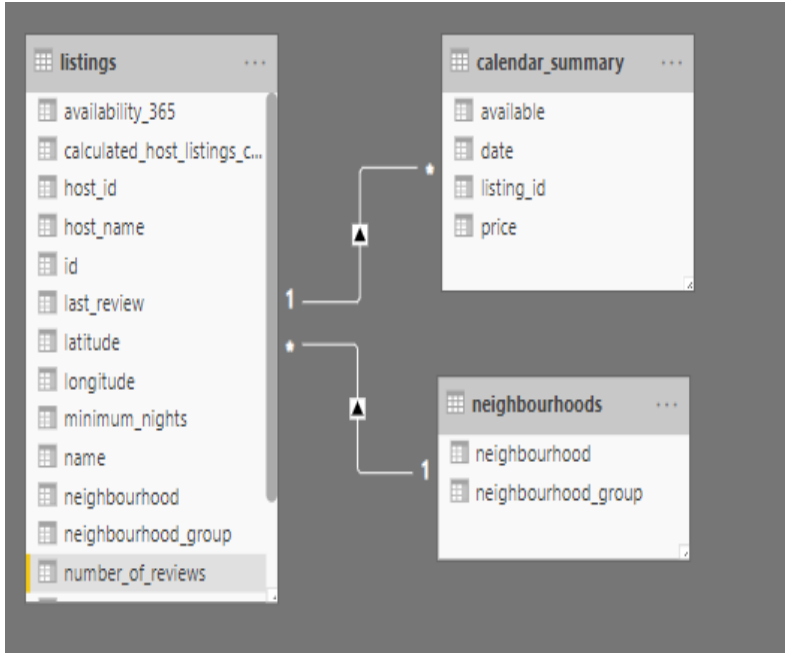
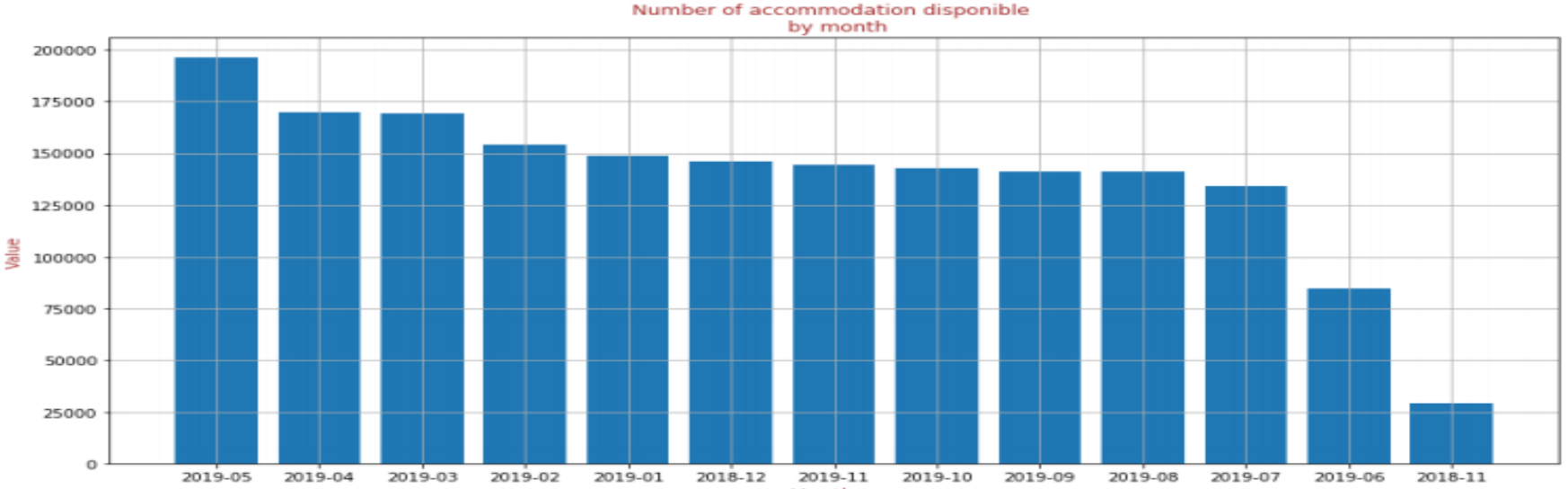


Fig1

Findings

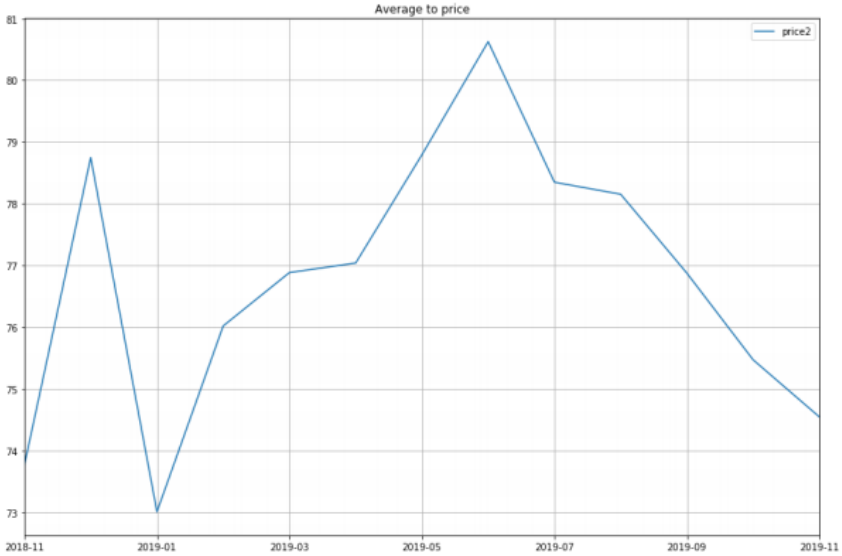
```
########## Years from  2018-11-07  TO  2019-11-08
********** Price from  0  TO  9000
```

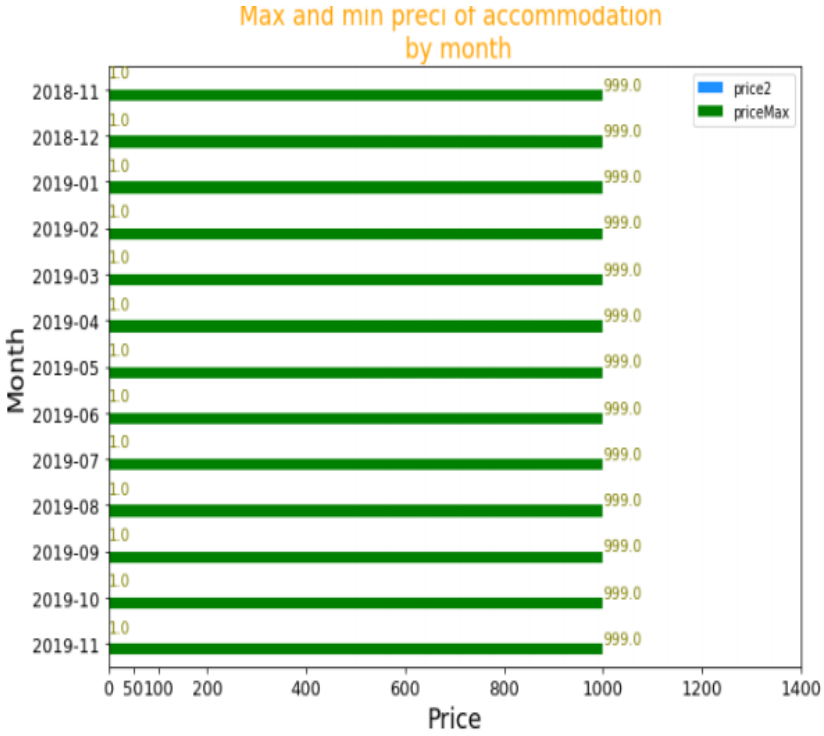
- First the calculation of the maximums and minimums in date and price
- The graph shows the number of accommodations registered per month.



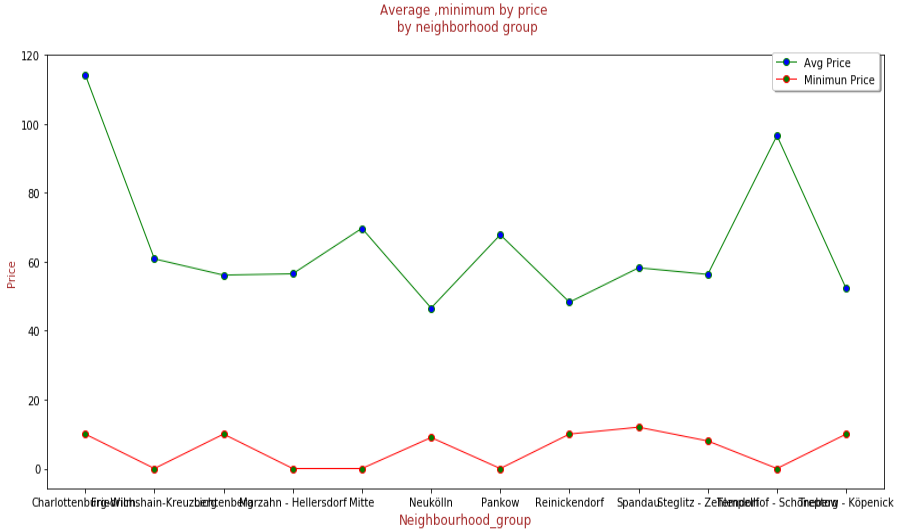
Findings

- 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 1 and 999

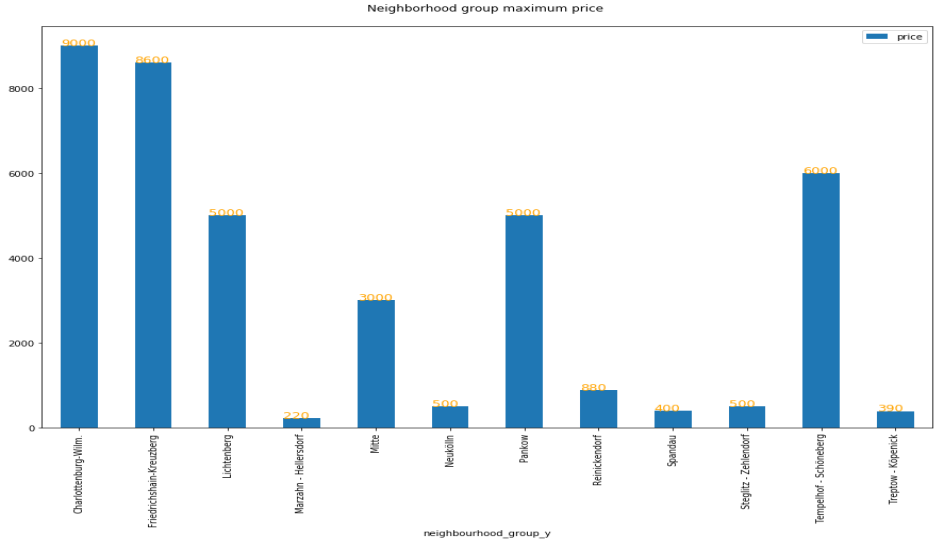




Findings

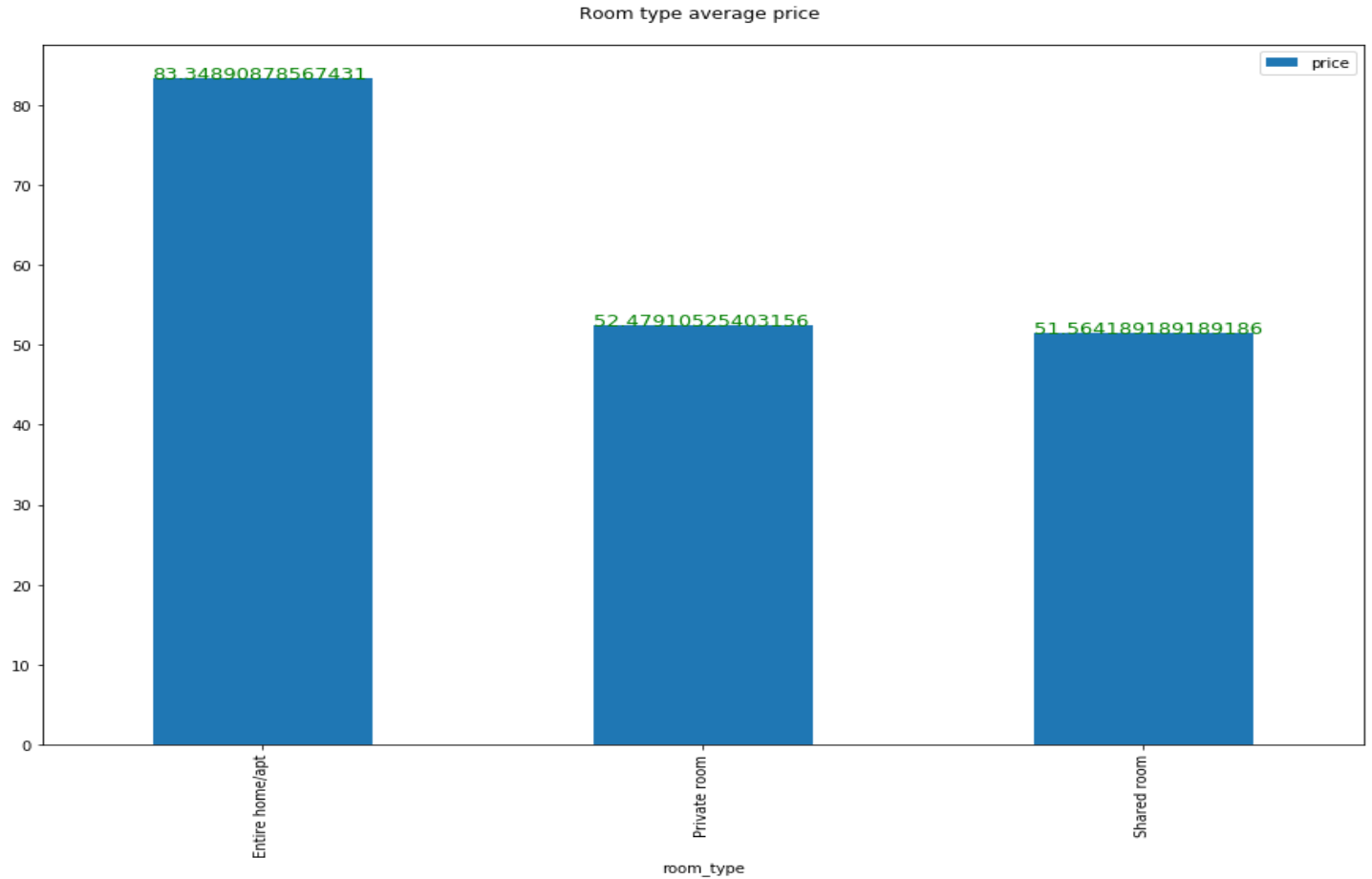


- This graph show the average, minimum and maximum by price and neighborhood group



Findings

- This graph show the average price by room type



Findings

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

	latitude	longitude	minimum_nights	number_of_reviews	reviews_per_month	availability_365
0	52.534537	13.402557	4	118	3.76	141
1	52.548513	13.404553	2	6	1.42	0

```
y = listing[target]
y.head(5)
```

	price
0	60
1	17
2	90
3	28
4	42

Findings

- This is the description of the test dataset, with an average of 80.32. After that, the mean square error test is applied with a result of 80.06 which shows that it is a good model.

```
y_test.describe()
```

	price
count	6134.000000
mean	59.499022
std	80.326896
min	0.000000
25%	30.000000
50%	47.000000
75%	70.000000
max	5000.000000

```
RMSE = sqrt(mean_squared_error(y_true = y_test, y_pred = y_prediction))
###std ->80.326896
print(RMSE)
```

```
80.06893531413805
```

Limitations

The data is only since 2018-11-07.
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 2019 the number of accommodations is almost similar.
- The average cost is 73 to 76 euros but for 2019-06 it was 80.625369, which shows that it is a time of great movement by the beginning of summer.
- The 3 most expensive communities are :

Charlottenburg-Wilm.	$9000
Friedrichshain-Kreuzberg	$8600
Tempelhof - Schöneberg	$6000

Conclusions

- The average price by room type is

room_type	price
Entire home/apt	83.348909
Private room	52.479105
Shared room	51.564189

-
- 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 festival, OktoberFest there is no big difference with respect to the other times of the year in prices and availability, as is the summer

Acknowledgements

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

Also investigate the stations in berlin, the most common celebrations and the number of visitors and the most attractive places in the city.

References

- https://stackoverflow.com/questions/31468176/setting-values-on-a-copy-of-a-slice-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/pandas-docs/stable/reference/api/pandas.to_numeric.html
- https://robertmitchellv.com/blog-bar-chart-annotations-pandas-mpl.html