AirBnB

Berlin

- Varinja Hartmann -

Applied Data Science Capstone

by IBM

For **week 1**, you will required to submit the following:

1. A description of the problem and a discussion of the background. (**15 marks**)
2. A description of the data and how it will be used to solve the problem. (**15 marks)**

A description of the problem and a discussion of the background

- Bedrooms
- Review scores
- Venues

→ Price!

Problem: Airbnb Prices – What determines the Airbnb price in Berlin?

- Imagine you are thinking about renting out your apartment on Airbnb from time to time. But you don't know what to ask for.
- What variable determines the most the price of rent of an Airbnb? Bedrooms, review scores, the neighborhood, etc.?

Price per Neighbouhood

- Can you predict the price for each Berlin neighborhood?
  - Visualize the data in a heatmap.
  - Predicting in which area I could get my highest Airbnb rate.
  - In the neighborhood you have selected what kind of things are there to see? (use Foursquare)
  - For example: My neighborhood is Steglitz-Zehlendorf. What would be the rent I could ask for 2 bedroom?

Week 2

Link to Notebook

- https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/16e59f11-ccff-4dea-9bb7-0edf32c3fad5/view?access_token=b60ad5ab434887f6de6c4de39f39ab0be892530ae1b1b6287705173bf39a6a1c

Capstone Project - The Battle of Neighborhoods (Week 2)

For the second week, the final deliverables of the project will be:

1. A link to your Notebook on your Github repository, showing your code. (**15 marks**)
2. A full report consisting of all of the following components (**15 marks**):
3. Introduction where you discuss the business problem and who would be interested in this project. → Week 1
4. Data where you describe the data that will be used to solve the problem and the source of the data. → Week 1
5. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
6. Results section where you discuss the results.
7. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
8. Conclusion section where you conclude the report.

A description of the data and how it will be used to solve the problem

Data Acquisition and Cleaning

- Airbnb Prices, Neighbourhoud Ratings from Inside Airbnb Dataset: http://insideairbnb.com/get-the-data.html
- Using: http://data.insideairbnb.com/germany/be/berlin/2019-07-11/visualisations/listings.csv
  - In total 24395 rows and 16 features
- Duplicate, highly similar features were dropped

See first line : # replacing NaN values with 0

df.fillna(0, inplace=True)

Data Wrangling: First Glance at the Data

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns


##getting the data and cleaning it up
df = pd.read_csv('http://data.insideairbnb.com/germany/be/berlin/2019-07-11/visualisations/listings.csv')
# replacing NaN values with 0
df.fillna(0, inplace=True)
print(df.shape)
df.head()
```

```
(24395, 16)
```

3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	1944	cafeheaven Pberg/Mitte/Wed for the summer 2019	2164	Lulah	Mitte	Brunnenstr. Nord	52.54425	13.39749	Private room	21	120	18	2018-11-11	0.25	1	364
1	2015	Berlin-Mitte Value! Quiet courtyard/very central	2217	Jan	Mitte	Brunnenstr. Süd	52.53454	13.40256	Entire home/apt	60	4	126	2019-07-04	3.18	4	0
2	3176	Fabulous Flat in great Location	3718	Britta	Pankow	Prenzlauer Berg Südwest	52.53500	13.41758	Entire home/apt	90	62	145	2019-06-27	1.18	1	279
3	3309	BerlinSpot Schöneberg near KaDeWe	4108	Jana	Tempelhof - Schöneberg	Schöneberg-Nord	52.49885	13.34906	Private room	28	7	27	2019-05-31	0.38	1	284
4	6883	Stylish East Side Loft in Center with AC & 2 b...	16149	Steffen	Friedrichshain-Kreuzberg	Frankfurter Allee Süd FK	52.51171	13.45477	Entire home/apt	125	3	124	2018-10-18	1.08	1	0

- The size of the dataset is 24395 rows and 16 columns
- In case there was NaN values I have replaced them with 0
- For evaluating the price
  - I will be using the categorical data neighbourhood_group and neighbourhood maybe also both latitude/longitude if the first are inconclusive.
  - Room type
- The column ‘Availability_365’ I will for now ignore as this is not important for my questions.
- Lets check if id and host-id are unique.

Data Wrangling

```
In [44]: #unique values count for each coloumn
         df.nunique()
```

```
Out[44]: id                                24395
         name                              23693
         host_id                           20442
         host_name                          6352
         neighbourhood_group                  12
         neighbourhood                       137
         latitude                          10195
         longitude                         13549
         room_type                             3
         price                               324
         minimum_nights                      104
         number_of_reviews                   337
         last_review                        1518
         reviews_per_month                   792
         calculated_host_listings_count       26
         availability_365                    366
         dtype: int64
```

In total we have 24395 data set. And we have 24395 Ids. Conclusion each ID is unique.
The other conclusion we can draw, some host have multiple venues.

Lets see further down the line if this is true. This would support the argument of some politicians that some people are trying to make money of an asset like living and these appartements are going to tourist not people who are living-working in Berlin.

Data Wrangling: df.describe

In [4]:
```
## Generate descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values. It also shows us the average room night in Berlin for the column price.
df.describe()
```

Out[4]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	2.439500e+04	2.439500e+04	24395.000000	24395.000000	24395.000000	24395.000000	24395.000000	24395.000000	24395.000000	24395.000000
mean	1.877282e+07	6.519884e+07	52.509952	13.405914	70.848166	6.960115	19.648576	0.896494	2.120476	71.265095
std	1.051497e+07	7.222803e+07	0.031244	0.058645	214.400730	24.789947	41.482242	1.435734	4.324981	111.062350
min	1.944000e+03	2.058000e+03	52.345800	13.097180	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000
25%	9.584722e+06	1.016728e+07	52.489115	13.375350	32.000000	2.000000	1.000000	0.050000	1.000000	0.000000
50%	1.957983e+07	3.540754e+07	52.509290	13.416470	50.000000	3.000000	5.000000	0.290000	1.000000	3.000000
75%	2.759866e+07	9.940649e+07	52.532770	13.439345	75.000000	4.000000	17.000000	1.080000	1.000000	110.000000
max	3.659933e+07	2.749825e+08	52.651670	13.757640	9000.000000	1000.000000	559.000000	43.960000	49.000000	365.000000

We use the method describe to obtain a statistical summary of the dataframe.
One of the first thinks we did was to drop all the NaN values and replace them with O.

The highest AirBnB is for 9000 Euro. The average is 70 Euro.

Now that the data is prepared we are starting to using the matplotlib to dive deeper and visualize our questions.

Data Development

```
4]: #Model Development - What correlates beside neighbourhoud_group the most with price --> availability_365.
    import matplotlib.pyplot as plt
    from sklearn.linear_model import LinearRegression
    df.corr()['price'].sort_values()
```

```
t[24]: longitude                       -0.034095
       minimum_nights                  -0.005389
       reviews_per_month               -0.003607
       number_of_reviews               -0.001224
       latitude                         0.004586
       id                               0.035210
       host_id                          0.040714
       calculated_host_listings_count   0.075612
       availability_365                 0.127587
       price                            1.000000
       Name: price, dtype: float64
```

Beside the neighbourhood_group the most relevant but far away for price is the criteria „availability_365“.

Exploratory data analysis

In order to set a price for my Airbnb apartment I need to analyze the dataframe in look at the different features that may / may not impact the AirBnb price.

- Which neighbourhood has the most Airbnb?
- Data distribution for neighbourhood_group and room_type.
- Price compared to neighbourhood_group and room_type.
- Reviews compared to
- What is the average of an Airbnb?
- How does the number of reviews affect the price for neighbourhood_group and room_type?

Top 3 Neighbourhoods in Berlin on Airbnb

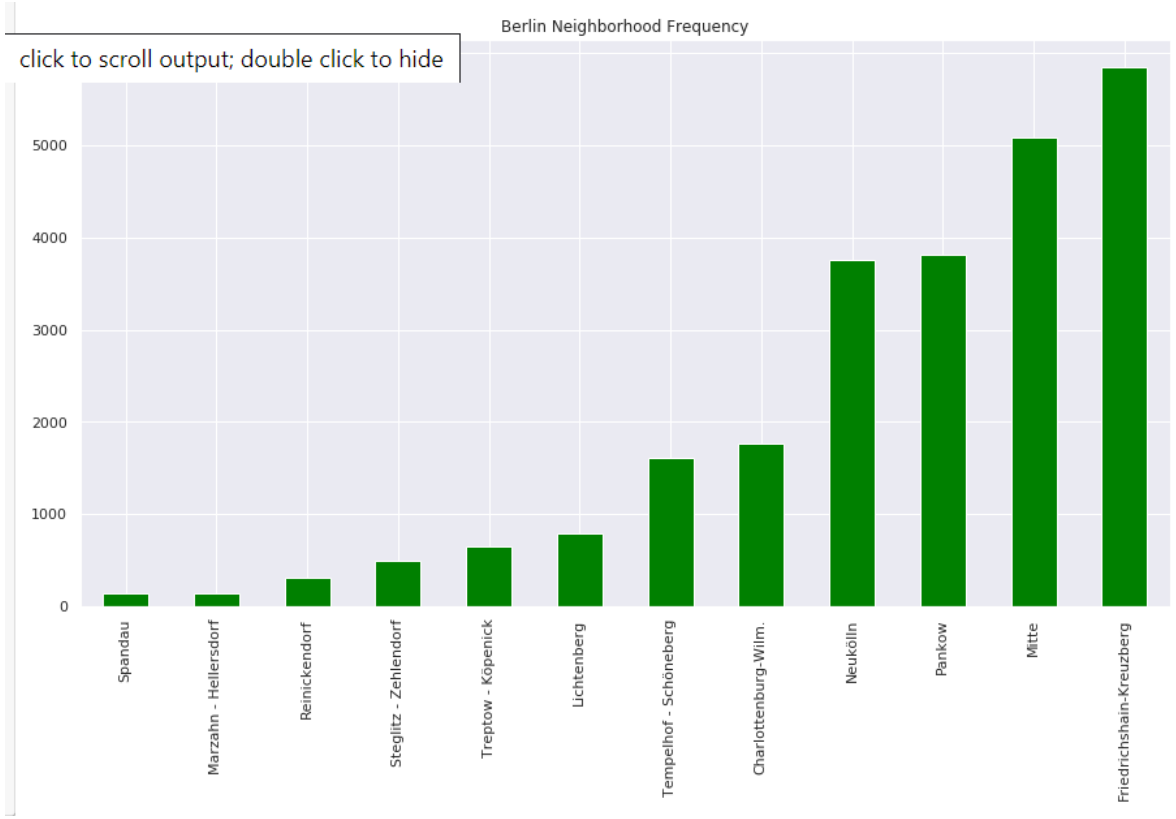
```
In [43]: # using Counter to analyze frequency of each listing based on neighborhood
         nh = Counter(df['neighbourhood_group'])
         nh

Out[43]: Counter({'Mitte': 5082,
                  'Pankow': 3818,
                  'Tempelhof - Schöneberg': 1610,
                  'Friedrichshain-Kreuzberg': 5854,
                  'Neukölln': 3753,
                  'Charlottenburg-Wilm.': 1766,
                  'Treptow - Köpenick': 647,
                  'Steglitz - Zehlendorf': 489,
                  'Reinickendorf': 304,
                  'Lichtenberg': 792,
                  'Marzahn - Hellersdorf': 142,
                  'Spandau': 138})
```

Which neighbourhoods are coming up the most often in Berlin?

1. Friedrichshain-Kreuzberg
2. Mitte
3. Pankow

Neighbourhoods on Airbnb in Berlin




```
In [40]:
##Let's see the average price of every listing available on SF
average_price = sum(df.price) / float(len(df.price))
average_price
print('The average price to rent a room in an AirBnb is', average_price)
```

```
The average price to rent a room in an AirBnb is 70.84816560770649
```

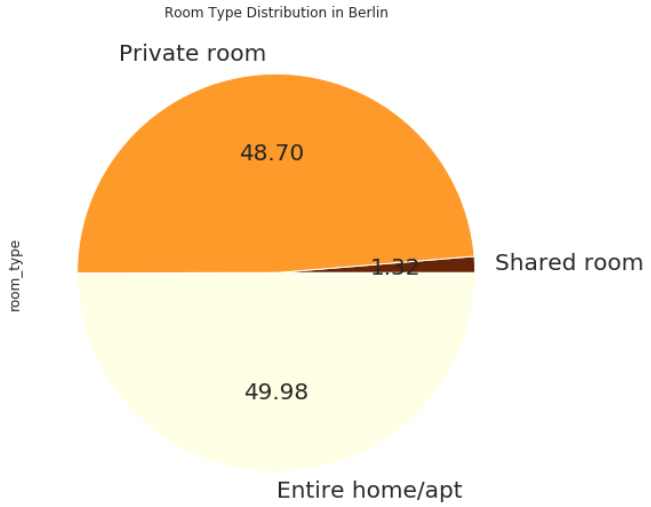
The average price is 70 Euro. This matches our Data Wrangling result.

Data distrbution in Berlin of room types

```
In [41]: room = df.room_type
         r = Counter(room)

         room_df = pd.DataFrame.from_dict(r, orient='index').sort_values(by=0)
         room_df.columns = ['room_type']
         room_df.plot.pie(y = 'room_type',
                          colormap = 'YlOrBr_r',
                          figsize=(8,8),
                          fontsize = 20, autopct = '%.2f',
                          legend = False,
                          title = 'Room Type Distribution in Berlin')
```

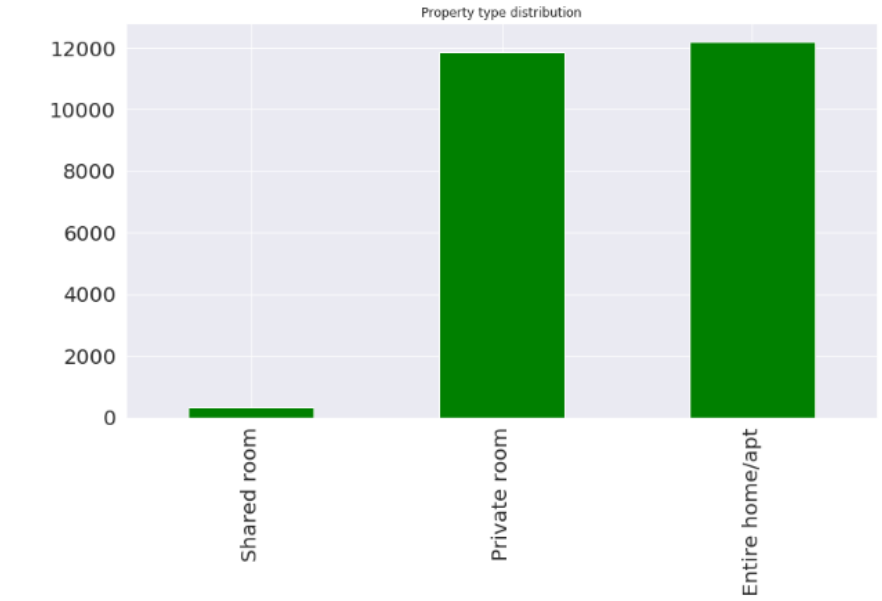
```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x7f439a3ff6d8>
```




```
: property = df.room_type
  p = Counter(room)

  room_df = pd.DataFrame.from_dict(p, orient='index').sort_values(by=0)
  room_df.columns = ['room_type']
  room_df.plot.bar(y= 'room_type',
                   color = 'Green',
                   fontsize = 20,
                   legend = False,
                   figsize= (13, 7),
                   title = "Property type distribution")
```

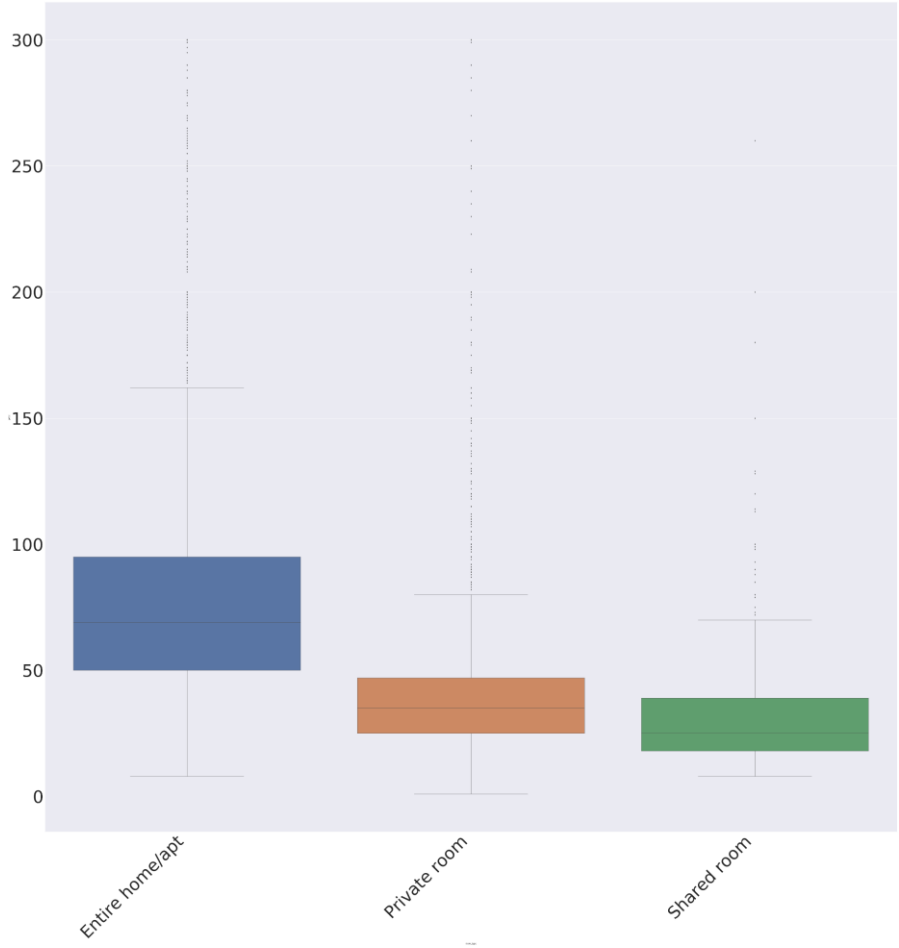
```
½2]: <matplotlib.axes._subplots.AxesSubplot at 0x7f439a3abc18>
```



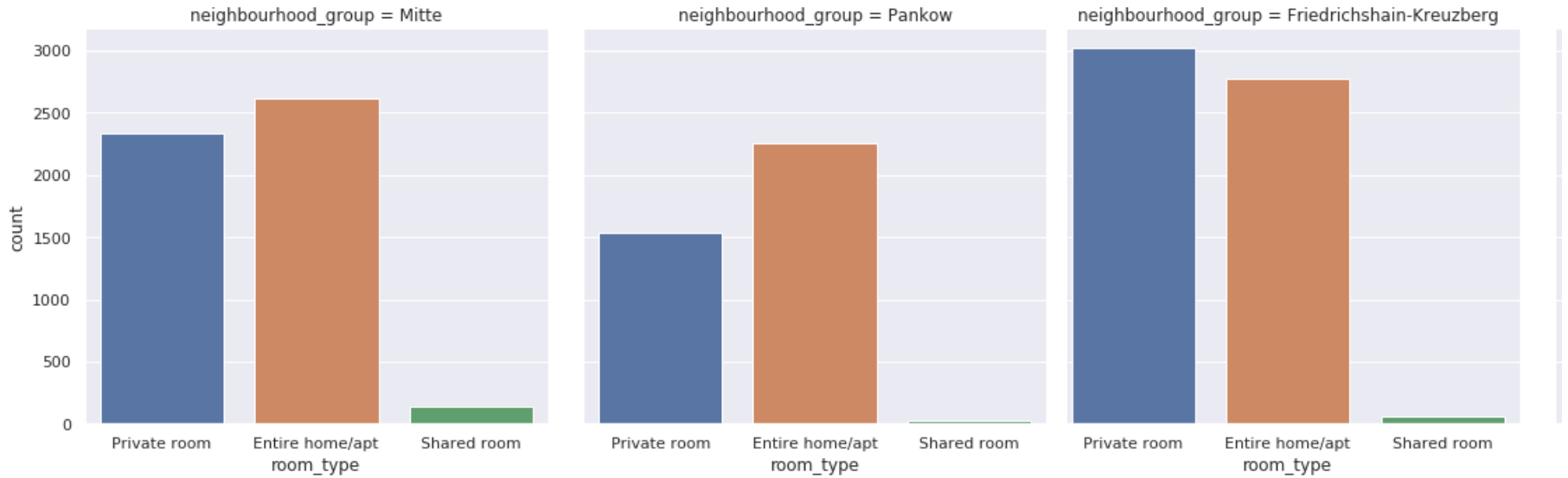
Like the Top 3 neighbourhoods suggested shared room is negletable.
Most people on Airbnb are renting out a) entire home or b) private room.
If you look at the chart of total rooms approx 12000 are in private rooms and entire homes.

Room-type and Price

- As expected an entire home /apt is more expensive that a private room.
- You can even see the median between an entire home is higher than between a shared room

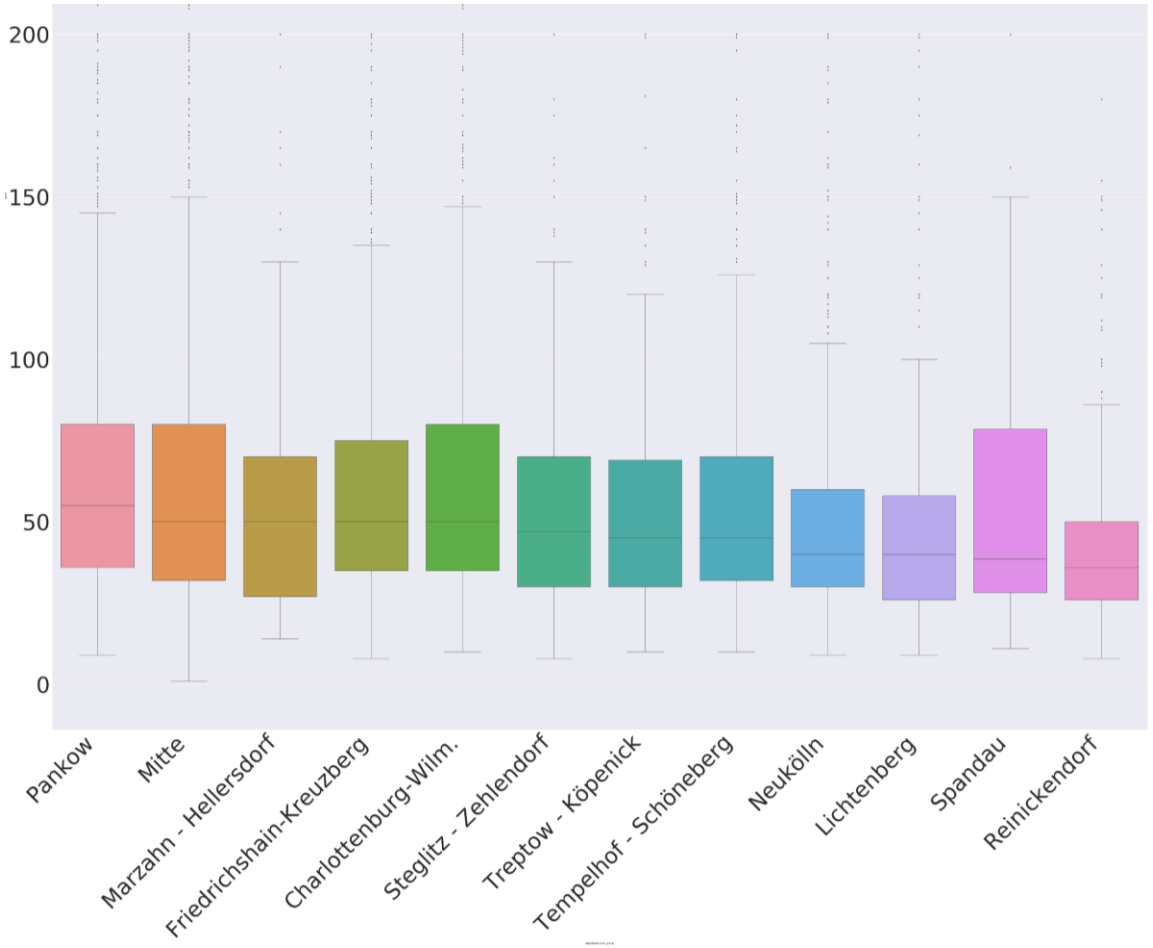


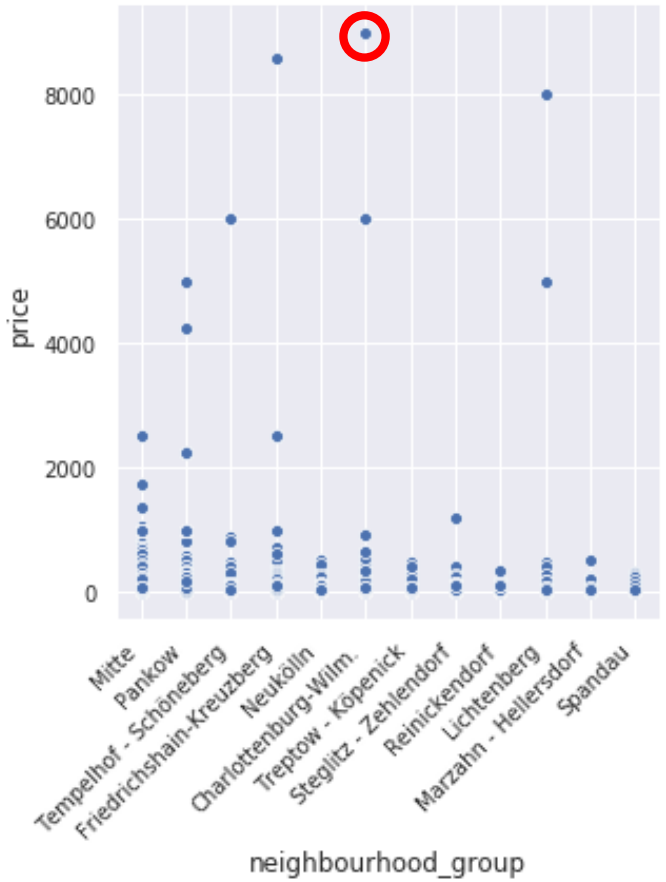
Data distribution for neighbourhood_group and room_type.



The room type private room and appartement are the most common once. Shared room is negletable.
Except in Friedrichshain-Kreuzberg Appartement is stronger than private room.
Lets look at the big picture for Berlin

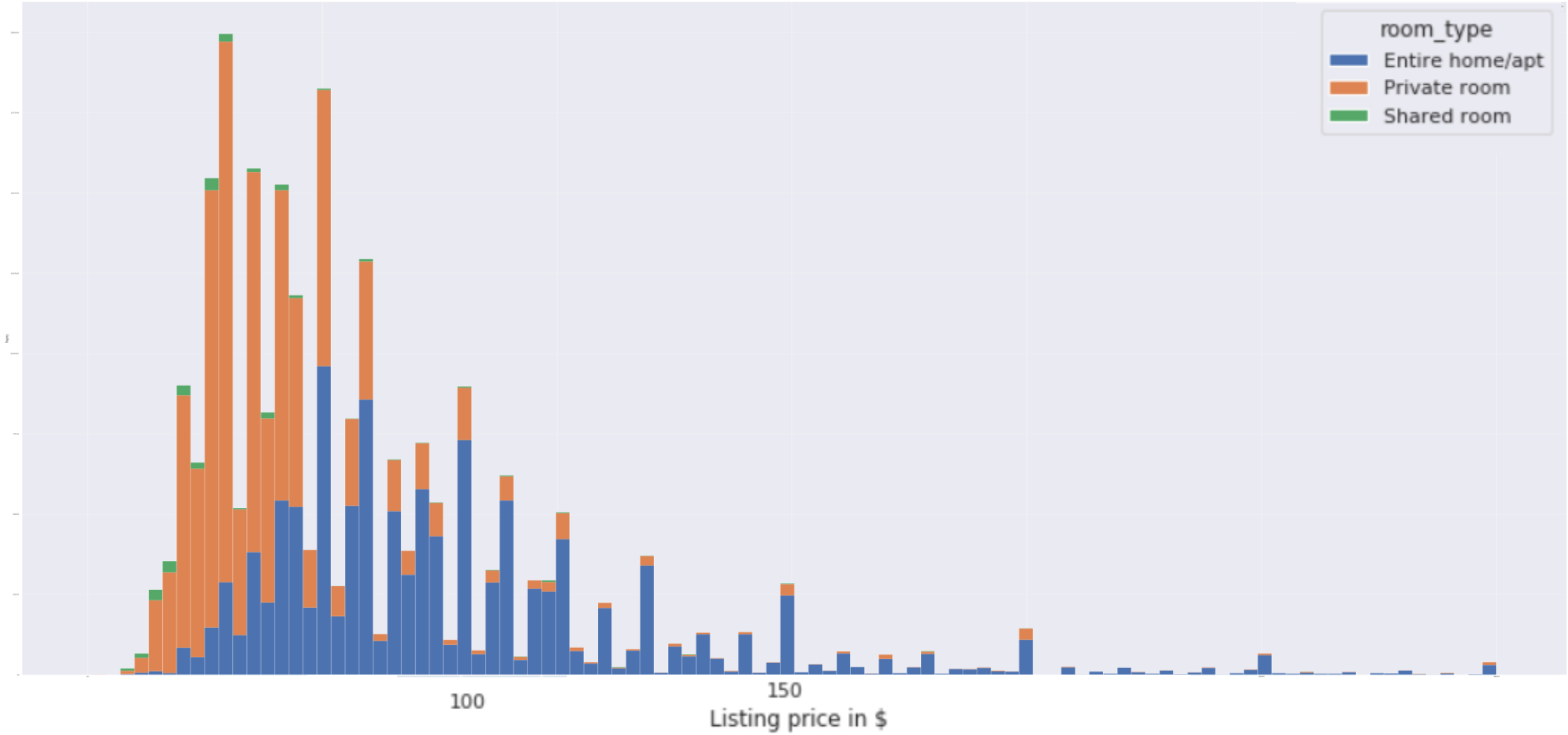
Price compared between neighbourhoods





The most expensive Airbnb Listing is in Charlottenburg-Wilmersdorf- Looking at the boxplost we can also see that thee average between Reinickendorf between Pankow is different. Meaning an Airbnb listing in Pankow, Mitte is in average higher than in Spandau or Reinickendorf.

Listing Histogramm



Most Private rooms are below 75 USD. With the increase in price entire homes are more shown

Foursqaure API – for the hot spot #1

1. Friedrichshain-Kreuzberg (5854)

- Most of the Airbnb listings are in the middle of Berlin – right in the citycentre.
- All top 3 neighbourhoods parts (if you know Berlin) are in the city centre.
- Looking at the results of Foursquare the are primarly food related.

Conclusion

- In this data capstone analysis project, I looked at 1 Airbnb's data set of Berlin.
- Summarized findings:
- Most of the listings are clustered in the neighborhood Friedrichshain-Kreuzberg and the center of the city.
    1. Friedrichshain-Kreuzberg (5854)
    2. Mitte (5082)
    3. Pankow(3818) are the top 3 neighborhoods with Airbnbs.
- Average price of all AirBnB listings is 70 Euro.
- Prices very differ widely in regards to neighborhood, property and room types. We can as a rule of thumb say neighborhood matter. I will be able to get a higher rent in Pankow, Mitte versus an offering in Reinickendorf.
- It comes to a draw between renting an apartment entirely and renting a private room. The difference is very small.
- Listings with prices less than 300- get the most reviews, meaning that they are booked most often. To be more precise most of the listings are below 150 USD – hence a strong correlations.
- Reviews have a meaning – but you should not stress on it. As the price is influenced more on availability →Offer and demand

A link to your Notebook:

- The following link shows you my notebook and code:
- https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/16e59f11-ccff-4dea-9bb7-0edf32c3fad5/view?access_token=b60ad5ab434887f6de6c4de39f39ab0be892530ae1b1b6287705173bf39a6a1c