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

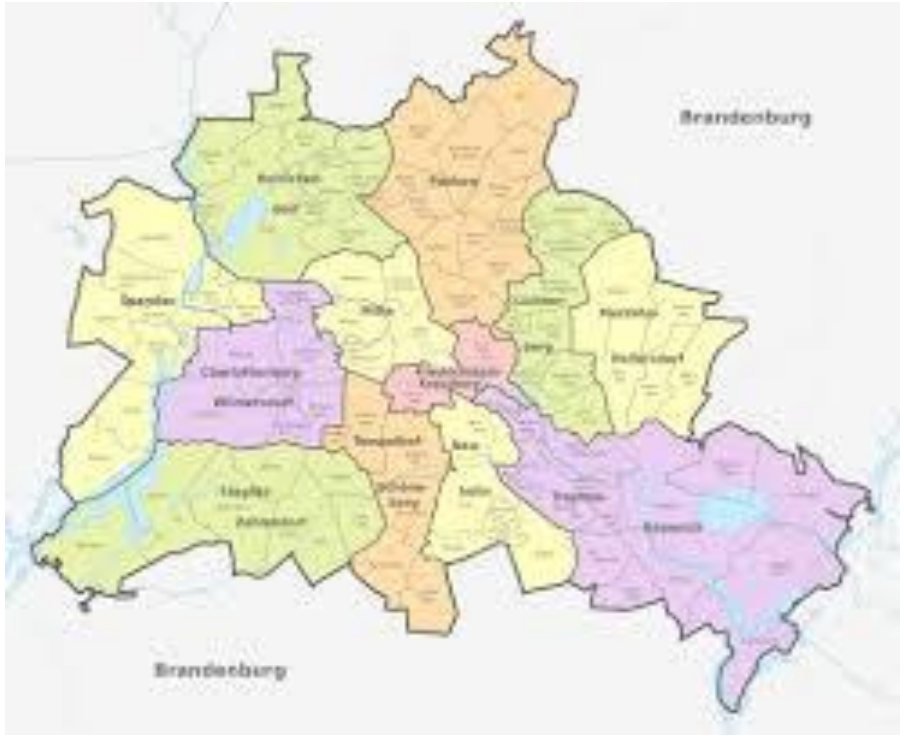
- Bedrooms
- Review scores
- Venues

→ 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?

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
- Cleaned data contains 24 features
- https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/05736402-c2de-4e09-bc50-0f5962bea4a0/view?access_token=99c4e5e8fb78d961c59eaf40106ebbbc52936f00a587953719b7c3c6864745dc

First Glance at the Data

```
:
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('http://data.insideairbnb.com/germany/be/berlin/2019-07-11/visualisations/listings.csv')
print(df.shape)
df.head()
```

```
(24395, 16)
```

19]:

	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

- Lets check if id and host-id are unique.
- 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.

Week 2

Still in progress till the 30th. 10.2019

Please dont count this for week 1 – as not needed.

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.
4. Data where you describe the data that will be used to solve the problem and the source of the data.
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 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

Business Problem:

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.

- Data distribution for neighbourhood_group and room_type.
- Price compared to neighbourhood_group and room_type.
- How do reviews per month affect the Price?
- How does the number of reviews affect the price for neighbourhood_group and room_type?
- Linear Relation between sights (foursquare) and price – What can I see in the neighborhood where I found the best price-Airbnb rate?