**Data Science Project Protocol  
Naama Jan**

**Introduction:**

For this project, I decided to choose the Airbnb Berlin dataset.  The dataset can be found on Kaggle:  
<https://www.kaggle.com/brittabettendorf/berlin-airbnb-data>

[Airbnb](https://www.airbnb.com/) is an online marketplace where people can offer and rent out their houses and apartments (properties). Airbnb became extremely successful for tourists and business people, who are looking for an overnight accommodation.

The datasets are based on properties list with description (features) and information whether the properties are being rented in each day from November 2018 until November 2019.

The project goal is predicting the probability of the properties to be rented in the next 3 months. Defining the outcome variable as booked if there are at least 70 rented days in the 3 months following every month.

The outcome variable is the "booked up" state of the following 3 months for each month starting in March until August (the first prediction starts in March since I'm using November-February for training. August is the last month since the last prediction is from August until November). I define the booked-up variable as a binary variable where it's set to 1 if at least 70 out of 90 days are booked and 0 otherwise.

I plan to examine the occupancy of the properties, according to the variables in the model. In general, there are variables that I expect to have high significant influence on the outcome. For example, price, weather, location (neighborhood), the distance to the city center, the size of the property, room type, reviews (especially review scores rating) and etc.

~~After creating a model which predict the next quarter occupancy of the property, I plan to check whether a host can optimize the price (I will check whether an increase /decrease in price will affect the occupancy of the properties).~~

**Data**

The main datasets are based on properties list with description (features) and information whether a property is being rented in each day from November 2018 until November 2019 (every record describes property in a day).

As mentioned in the introduction, our target is predicting whether a property will be rented in the next 3 months (defining the booked\_up\_target base on occupancy of the property).

These datasets are taken from [Kaggle](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data) and ensembled from 6 different csv files:

* [Calendar Summary](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=calendar_summary.csv)
* [Listings Detailed](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=listings.csv)
* [Listings Summary](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=listings_summary.csv)
* [Neighborhoods](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=neighbourhoods.csv)
* [Reviews Detailed](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=reviews.csv)
* [Reviews Summary](https://www.kaggle.com/brittabettendorf/berlin-airbnb-data?select=reviews_summary.csv)

For enrichening the main data sources, I used [climatestotravel](https://www.climatestotravel.com/climate/germany/berlin) website for getting average weather data. Assuming that weather info can influence on tourism in Berlin which can directly influence the demand for properties.

Importing the following datasets from this website:

* Sunshine hours
* Average precipitation
* Average temperatures

The data sections in this project are divided to the following steps:

1. The first step in the project is collecting the datasets from Kaggle to SQL server and creating a flat table (by relevant joins between the tables). Most of the features are created at this stage.
2. There are enrich features that requires analysis of text. These feature have been added to flat table using “1- Addition to Flat file.ipynb” notebook.
3. In “[2 - Berlin\_Airbnb\_EDA.ipynb](https://github.com/NaamaJan/airbnb-berlin-availability/blob/main/2%20-%20Berlin_Airbnb_EDA.ipynb)” notebook, I perform an analysis of the detailed Berlin listings data. I am using visualizing and analyzing data to extract insights from the variables in the data.

In the beginning, I used descriptive statistics to explore the data, which can help describe the data set's basic features and obtain a summary of the data. Then, I perform Data Visualization analysis to provide an accessible way to see and understand trends, outliers, relationships, variability, patterns in data and to notice if there is a problem with data quality.

I have been paying attention to the correlations and the differences between the variables in the data set and describe the target variable "booked\_up\_target", his distribution and its relationships with the variables.

This section also helped me create new variables or perform variable transformations.

1. In [3- Berlin\_Airbnb\_Data\_Cleansing.ipynb](https://github.com/NaamaJan/airbnb-berlin-availability/blob/main/3-%20Berlin_Airbnb_Data_Cleansing.ipynb) notebook includes reduce the bias by filling NA, fixing outliers and co.

2 - Berlin\_Airbnb\_EDA

Here you have to describe how you plan to manipulate the data. For this you have to answer to the following questions:

* On which time frames will your project be based on?
  + Time-frame for training
* איך מגדירים כל שורה, מה הקריטריונים להוספה ולהורדה-פרק זמן וlisting id
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data? outliers
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How will you deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)