**Data Science Project Protocol**

*Author(s):*

# Introduction

Here you have to give some known facts about the field you will work on.

Try to focus on the problems that are most common and then state the goals of the project.

* You can try to answer to the following question:
* Which questions do we want to answer?
* What is known about the problem?
* How we define the outcome(s)?
* What is known to influence the outcome?
* Do we have any possible new knowledge that has not been in use before?

This part must be between half to one and half page.

[vladi]

* These are the works that were used as part of the background and domain research
  + <https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6>
  + <https://towardsdatascience.com/going-dutch-how-i-used-data-science-and-machine-learning-to-find-an-apartment-in-amsterdam-part-def30d6799e4>
  + [https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a](https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a" \t "_blank)
  + [http://users.ece.northwestern.edu/~xws864/eecs349/doc/FinalReport.pdf](http://users.ece.northwestern.edu/~xws864/eecs349/doc/FinalReport.pdf" \t "_blank)
  + [https://medium.com/asif-anwar/utrecht-apartment-hunting-da9883131ff6](https://medium.com/asif-anwar/utrecht-apartment-hunting-da9883131ff6" \t "_blank)
* The question we would like to answer is: what would be the occupation of the particular apartment in a particle week, given set of features( related to apartment description, price and reviews)

Airbnb is a home-sharing platform that allows home-owners and renters (‘hosts’) to put their properties (‘listings’) online, so that guests can pay to stay in them. Hosts are expected to set their own prices for their listings. Although Airbnb and other sites provide some general guidance, there are currently no free and accurate services which help hosts price their properties using a wide range of data points.

Paid third party pricing software is available, but generally you are required to put in your own expected average nightly price (‘base price’), and the algorithm will vary the daily price around that base price on each day depending on day of the week, seasonality, how far away the date is, and other factors.

Airbnb pricing is important to get right, particularly in big cities like London where there is lots of competition and even small differences in prices can make a big difference. It is also a difficult thing to do correctly — price too high and no one will book. Price too low and you’ll be missing out on a lot of potential income.

This project aims to solve this problem, by using machine learning to predict level of occupation of a property in San-Francisco based on property’s features.

# Methodology (Project design)

## Data

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

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + 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?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will 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)

The dataset used for this project comes from [Insideairbnb.com](https://insideairbnb.com/" \t "_blank), an anti-Airbnb lobby group that scrapes Airbnb listings, reviews and calendar data from multiple cities around the world. The dataset

includes scrapping of San-Fransisco Airbnb listings starting 2015 and ending 2020, and includes information on about 25,000 listings that were active during this period.

Perodic scrapping includes:

Listings.csv – holds listings description including reviews statistics and ratings

Calendar.csv – hold future calendar of the listings (along with per-day availability and price)

The data is quite messy and has some limitations. The major one is that it only includes the advertised price (sometimes called the ‘sticker’ price). The sticker price is the overall nightly price that is advertised to potential guests, rather than the actual average amount paid per night by previous guests.

Scrapping includes the future calendar of the properties along with the current property, host and reviews details.

To improve the model precision, property calendar was considered from the scrap date (usually first day of the month, till the next scrapping day), in any case property’s calendars are at most one month ahead.

Subjects inclusions:

# 1. Take all the links from isideairbnb

# 2. Filter only sf (san-fransisco) calendars (in Linux shell)

# grep san-fra insideairbnb.txt |grep calen > sf.calendars

# 3. Clean it from html (in Linux shell)

# cat sf.calendars | cut -d \" -f2 > sf.calendars.cleaned

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss the limitations of the model, when it can be used and when not.