Project 1:
Data: Feature extraction, and visualization

,Luka Avbreht (s191963)

September 30, 2019

0.1 Section 1: A description of your data set

1. What is the problem of interest? (describe what your data is about in general terms, i.e. to someone who knows nothing of machine learning)

Our data is a dataset of all AirBnB apartments and other locations available in New York City in summer of 2019. AirBnB is a very popular booking application where people can list their homes, or book a stay in someone's else's home. In our data set we have all kinds of info, from names of properties offered, to coordinates and review stats of each home. Our data set is freely available on the link [1].

2. Who made the data and why? Did they, or somebody else, work with the data and report results? If so, what were their results?

Data set was published by AirBnB for public use. Their main inspiration is as follows.

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

As far as we know, no one did anything significant with this dataset.

3. What is the primary machine learning modelling aim? (Is it primarily a classification, a regression, a clustering, an association mining, or an anomaly detection problem?)

Whit this data set it is possible to do a project with main aim to almost any of this classes. We will primary use it for clustering (apartments with similar audience, similar properties such as proximity to some important points...), an association mining and some regression (TODO TODO), but there is also a huge possibility for anomaly detection network.

4. Which attributes are relevant when carrying out a classification, a regression, a clustering, an association mining, and an anomaly detection? Specifically: Which attribute do you wish to explain in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task?

This is the list of all attributes:

- **id** listing ID
- **name** name of the listing
- **host_id** host ID
- **host_name** name of the host
- **neighbourhood_group** location

- **neighbourhood** area
- **latitude** latitude coordinates
- **longitude** longitude coordinates
- **room_type** listing space type
- **price** price in dollars
- **minimum_nights** amount of nights minimum
- **number_of_reviews** number of reviews
- **last_revieW** latest review
- **reviews_per_month** number of reviews per month
- **calculated_host_listings_count** amount of listing per host
- **availability_365** number of days when listing is available for booking

5. Are there any data issues? Either directly reported in the accompanying dataset description or apparent by inspection of the data? (such as missing values or incorrect/corrupted values)

We can se some minor issues with our data.

- We don't know the exact date on which the data was picked (prices are the main concern since they change over the year)
- Some of the properties are outdated, they are not listed on AirBnB anymore. We are missing some data that may come in handy. (such as which users are so called super hosts,
- TODO

But overall we think our data set is very clean and offers enough info to carry out all required tasks.

0.2 The responsibility assignment

Table 1: Responsibility assignment

Section 1	Luka Avbreht	45%
Section 2		
Section 3		

Bibliography

[1] AirBnB open data set, (accessed on 20.9.2019)
https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/version/3?