C4-NYC-AIRBNB

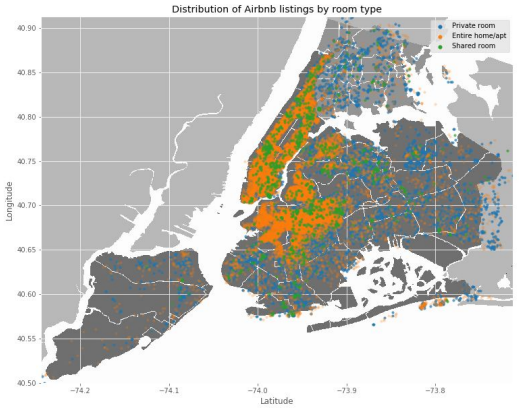
UNIVERSITY OF TARTU

Introduction

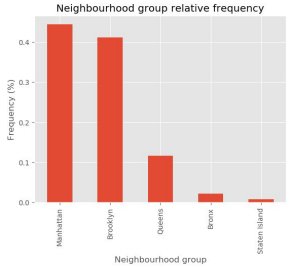
Our objective is to analyze Airbnb listings in NYC in 2019 in search of correlations and patterns, and to develop a model capable of giving predictions for listings based on their location and type. Another possible model would be capable of predicting the room type or location based on price and other parameters.
In the end, we seek to create a script for processing the data, analyzing it, creating the models and displaying the results.

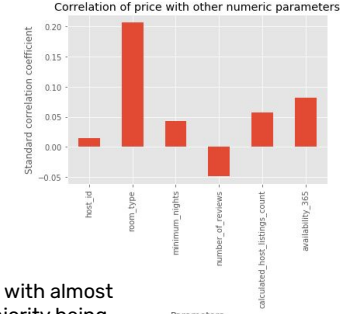
Data Analysis

All of our data is from Kaggle, which was gathered directly from Airbnb beforehand.
We used Python and various libraries for data processing and analysis, including cleaning up and filtering out data; one-hot encoding the data and creating subsets for different kinds of visualizations.

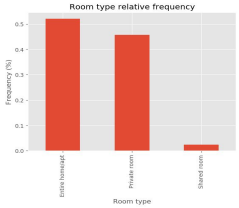


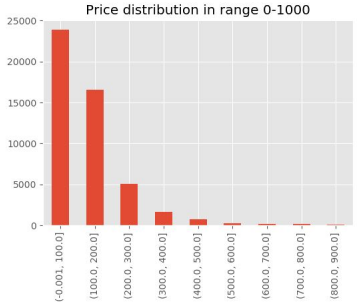
We discovered that the price of a listing is most heavily dependent on the type of the room, as well as its' availability throughout the year.
The most expensive apartments are located in Manhattan, mostly being the type of "Entire home/apt".





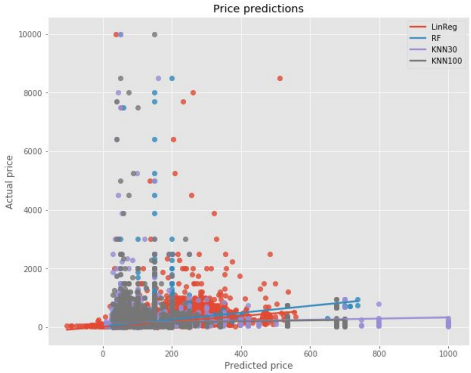
The data was quite unbalanced, though, with almost no listings for Shared rooms, and the majority being located in Manhattan and Brooklyn.
Besides, the majority of prices for the rooms were In the range 0-200 $/night, with some rare cases reaching 10,000.





Models

We developed four primary models using linear regression, random forest and k-nearest neighbors algorithms.
The models we set up predicted prices mostly in the range 0-200 $/night, being very inaccurate, with some surprising outliers even below 0$. Despite this, the models can be still used if the price range is known to be around 100 or 200 $ per night.



Conclusions

We have reached all of our goals and objectives. The analysis was a success and we have created multiple helpful visualizations.
While our models lack accuracy, they still can be used in some situations.
We have developed a script that handles all of the steps from loading in the data to creating and saving the visualizations, as well as creating and training the models, which can be tweaked and improved to yield better results in the future.

Anton Slavin, Elen Liivapuu

December 2020