Predicting population size based on abundance of different venues

Franziska Metge

May 24, 2020

1 Introduction

In this Capstone project I want to investigate if the population size of a city can be estimated by the amount of different venues within a 5km radius around the city center. Furthermore, I want to determine which type of venue is the best predictor for population size and what type of model is the most accurate. Therefore, I will implement different models introduced in the Machine Learning Course using an Ipython Notebook and the python specific libraries such as numpy, pandas, and sklearn.

This analysis could reveal crucial correlations between venues and population size. It could identify outlier cities. Cities with a higher density of venues to population size could possibly be interesting for tourism agencies, while cities with a lower density of venues to population size could possibly be interesting for business developers.

2 Data

I will use the population data made publicly available by the United Nations (Demographic Yearbook – 2018[1]). I will use numpy and pandas libraries to clean the data. Secondly, I will use the geopy library to acquire the coordinates for all cities. I will separate cities into different classes/groups based on their population size, i.e. < 0.5 Mio, $0.5 - 1$ Mio, $1 - 5$ Mio, $5 - 10$ Mio, $10 - 20$ Mio, and > 20 Mio. I will use the foursquare app to look for all venues within a 5km radius of the city's center. In order to deal with the limited amount of requests that can be made to foursquare, I will only use a subset of all cities. The data will be stored in a table with one row per city containing the cities name, country, coordinates, population size, population size category, number of venues from different categories in one-hot encoding.
Finally, this data set will be split into a training and test data set to train different models and evaluate their accuracy.

3 Methods

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.

4 Results

Results section where you discuss the results.

[1] https://unstats.un.org/unsd/demographic-social/products/dyb/dyb_2018/

5 Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

6 Conclusion

Conclusion section where you conclude the report.