Capstone Project
The Battle of Neighborhoods

for the IBM Data Science Professional Certificate

by Tobias Bertschinger

tobias.bertschinger@gmail.com

13 May 2020

Table of Contents

Introduction 3

Data 3

References 4

Introduction

The selected topic for this Capstone Projects analyses top visited tourist attractions around the world. The goal is to analyse whether the venues around those attractions share common features and whether it is possible to cluster attractions according to the nearby recommended venues.

Such findings may provide helpful insights for both tourism directors of the most visited attractions and to directs of attractions which are not currently visited by a comparable amount of people. Other interested parties could be business or restaurant owners as the analysis should provide insight into what visitors expect from well known and frequently visited tourist attractions at the venues such as cafés and restaurants nearby.

Data

The website Love Home Swap published an article called “The world’s 50 most visited tourist attractions” on 1 January, 2015 (Love Home Swap, 2015). Data for this list has been gathered from various sources such as the US National Park Services or websites from the attractions. Data driven travel blog Travel Stats Man has analysed the venues and published easily accessible data tables (Travel Stats Man, 2019).

Each entry includes the attraction name, the city and country of the attraction, annual visitors in millions of visitors and the rank in the top 50 list. As an example, the most visited attraction was The Strip in Las Vegas, USA, with an annual visitor count of 39.6 million.

In a first step, I will download this data and convert it into a pandas dataframe using python. The second step will involve downloading location data for all the attractions. For this, we will use the package geopy for pythons. As an example, for the most visited attraction mentioned before, we find latitude = 36.2859033 and longitude = -115.0071652. Checking this location on google maps reveals that the location is somewhat off The Strip. Therefore, we will either double check locations in the second part of the assignment or use the google maps API to solve this part of the problem.

After fetching the respective latitude and longitude data for each attraction, we can then use Forsquare’s Places API. Using the venues endpoint group and the explore endpoint, we will be able to fetch nearby locations.

We will then analyse recommended places nearby attractions according to their group (e.g. coffee shop) and calculate each group’s relative frequency. This will help us to sort the most common venues near each attraction.

Finally, we will be able to cluster attractions using k-Means clustering to analyse and visualize similiarity between the recommended venues near each attraction.

References

Love Home Swap. (2015, January 1). Retrieved from https://www.lovehomeswap.com/blog/latest-news/the-50-most-visited-tourist-attractions-in-the-world

Travel Stats Man. (2019, January 7). Retrieved from https://www.travelstatsman.com/07012019/the-worlds-most-visited-tourist-attractions/