Using Foursquare Data to Cluster U.S. Cities *

Michael Quigley

July 2021

*This report was used as the capstone project for the IBM Data Science Professional Certificate

1 Introduction

When a business decides to expand outside of it's home city, the choice of location for expansion can make the difference between success and failure. While a business may be successful in one city, there are numerous factors that could may another city less receptive to their services. Size, demographic differences and geographic factors can all effect which businesses are able to succeed in a given city. One strategy would be to expand into locations which are similar to a location where the business has already had success. Determining which cities are most similar is not a trivial task, but using Foursquare venue data and k-NN clustering, we can begin to extract useful insights for businesses looking to expand.

2 Data

The primary data source for this project is Foursquare. A request will be made to the Foursquare API for each city being compared. We will ideally be able to extract 100 venues for each city. The categories of these venues will then be used to calculate clusters.

In addition to venue data from Foursquare, data on relevant cities must also be collected. Data will be collected from the Wikipedia article `https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population` which contains data on the largest U.S. cities. The top 150 cities will be taken from this table and used for the model. Unfortunately the coordinates listed in this table are often inaccurate, so each individual city page will also be needed to extract accurate coordinates. After clustering is complete, more data extracted from these Wikipedia pages will be used to compare clusters.