**Capstone Project: Which City is more Similar**

**By Firdhous Puraliyil**

Introduction

Cities have unique traits that define them and give them ‘names’. For example, Austin is known for it’s uniqueness and weirdness, which how is Austin, TX got the saying ‘Keep Austin Weird’. This saying got me thinking, I wonder which cities are similar and different from each other. I wonder if Seattle is weird or if it has a more of a Blue Collar Vibe like a working city as Pittsburgh, or if it had the lax atmosphere as a Miami. This led to me come up with my capstone project idea, comparing cities to see which ones are alike.

First, I had to choose how many cities I would compare and which cities I would compare. I decided to go with three cities, so I could decisively see which city is similar. The next step was choosing the cities. I choose Austin, because how unique it was. Based on the first selection, I wanted to choose a city that people say are somewhat like Austin, which was Seattle. Lastly, I choose Chicago because I thought it was a city that was not as similar as the other two from word of mouth.

The next step I had to choose is what categories would I want to compare these cities with. I wanted to have data that regarded the people of these cities as well data that didn’t regard the people of these cities, so the 5 categories I decided to look into was Cost of Living, Popular Venues, Weather, Population and Health. With these five data sets I will see which city is like each other and which is the outlier among the three.

**Data:**

*Health searches by US Metropolitan Area, 2005-2017*

<https://www.kaggle.com/GoogleNewsLab/health-searches-us-county>

# *Climate of 100 Selected U.S. Cities*

# <https://www.infoplease.com/math-science/weather/climate-of-100-selected-us-cities>

# *Venue:*

# Foursquare API

# *Cost of Living:*

# <https://www.kaggle.com/debdutta/cost-of-living-index-by-country>

# *Census:*

# <https://www.census.gov/library/publications/2012/dec/c2010sr-01.html>

# Data Meaning:

# Health:

This data frame shows us the health searches from 2004 to 2017. It scales from 0, being the lowest searches for that disease to 100 which is the most of searches of that disease. This dataset is created through google API and will allow us to understand the health searches of the people from the three cities, giving us an idea of what disease prevails in that area.

Climate:

This data frame was retrieved from a website called infoplese. It has the average weather temperature, rainfall snowfall for at least 30 years. This will give us an idea on how similar the weather is between these three cities and which two cities are the most similar.

# Venue:

# Using the explore API, I will be retrieving the most popular venue around the city’s mayor/capitol building, so see how similar they are. I got the data using the API call and retrieved the venue’ name, category and location. From there I used a python script to categorize them in 4 main categories, Entertainment, Food & Beverages, Lodging and Other. After that I concatenated all three data frames to one to give me the counts of each main category for each city

# Cost of Living:

# This data set is found off Kaggle and it is the cost of living compared to New York. New York is 100 in all the fields, so if something is 120 in cost of living it is 20 % greater than New York. What I will use this data set for is to compare the cost of living among the three cities to New York. The fields are Cost of Living Index, Rent Index, Cost of Living Plus Rent Index, Groceries Index, Restaurant Price Index and Local Purchasing Power Index.

# Census:

# The Census Data contains Census data from 2010 and 2000. It has various categories such as population by sex, for both years, by age, local to national ratio, numeric and percent change in population etc. This will be used to compare how the population has changed over time within these three cities.