**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.

**Audience:**

The audience for this project is for other data scientist curious on how to compare cities but also for people to see how similar the cities of Austin, Seattle, and Chicago. It will give an in-depth comparison on how these cities are different or alike but also give them hopefully an eye opener on each of these cities.

**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. The Categories are Diabetes, Cancer, Cardiovascular, Stroke, Rehab, Vaccine, Diarrhea, Obesity

Climate:

This data frame was retrieved from a website called infoplese. It has the average weather temperature, rainfall and 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.

**Methodology:**

To compare the three cities, I wanted to have viable categories to compare them. If I choose something like How many Pumpkins each city grows, it wouldn’t really tell me anything unique about the city other than it’s pumpkin growing stats. So I decided which five categories really tell me about a city and it’s people. I came up with Health, Cost of Living, Venues, Weather and Census. Health tells us what the people of that city’s lifestyle/ diseases are most common, Cost of Living tells us how expensive a city is, Venues tells us what the people like to do for fun, weather tells us how the living conditions in a city and census tells us the population of a city and if it somewhere people desire to live or if it’s a ghost town. With these five data sets, I will try to see which 2 cities are more alike from Austin, Chicago and Seattle. How I analyzed the data was by hierarchical clustering and dendrograms for each five categories to see which city was most alike in these categories.

**Data Collection:**

Most of the data sets were already readily available to use. The only data set that I had to collect was the Venue data set. How I collected the Venue Data set was by using the explore foursquare API for each city. I gathered the 30 most popular venues around the official government building in each city. I used a foursquare API to collect the data into a json dictionary, then I used a for loop to iterate through the dictionary to create a data frame with three column, Name of Venue, Category and Location. Below you will see how I did this:

![A screenshot of a cell phone

Description automatically generated]()

**Data Preparation:**

**Venue**

After making the data frame, I made it into a csv and called it into another script to prepare it more. I wanted the count of each Main Category. So I created another column in the data frame called Main Category. In this column, I would assign each row whether it belonged to the Entertainment, Food & Beverages, Lodging and Other Category. Once I did that, I used a value\_counts() function to get the counts of each group and then a pivot() function to get a data frame of the counts for each category. Below I have displayed how I did this:

![A screenshot of a cell phone

Description automatically generated]()

I did this for all three cities and concatenated the dataframes into one.

Census:

For Census dataset, I had to clean up the header, as there were so many sub headers. I used all the header as they had useful information. The main headers were Population by sex, 2010, Population by sex, 2000, Population by age, 2010, Population by age, 2000, Percent of population by age, 2010, Percent of population by age, 2000, Local to national ratio2, 2010, Local to national ratio2, 2000 and Numeric change in population by age, 2000 to 2010Percent change in population by age, 2000 to 2010. There were many subcategories like gender and age range as well. So I decided to delete all the header as they were not important when it came to the analysis and be unnecessary work of renaming 55 columns. I then filtered out for the three cities and the then removed all non integer values.

Weather:

For the Weather dataset I filtered out for the three cities and removed all column that are not integers. The columns I used were Jan. (Average temperature for 30 years between the months of January, February and March), Apr. (Average temperature for 30 years between the months of April, May and June), Jul. (Average temperature for 30 years between the months of July, August and September), Oct. (Average temperature for 30 years between the months of October, November and December), Average annual (Average annual Precipitation in inches), Average annual.1 ((Average annual Precipitation in days), Averageannual(in.)3 (Averageannual(in.)3 in snowfall.

Cost of Living:

The Cost of Living data set is based of New York. New York in this dataset is 100 in each category; Cost of Living Index, Rent Index, Cost of Living Plus Rent Index, Groceries Index, Restaurant Price Index and Local Purchasing Power Index. This means that if a city has 120 in a category, they are 20% higher than New York. For this data set I filtered out the cities and removed all columns with non integers.

Health Searches:

For the Health Searches data set, which are the searches of a particular disease or remedy on an index of 0 (being the lowest) to 100 . This data set has cancer, cardiovascular, vaccine, obesity, diarreha, stroke, rehab, depression and diabetes data from 2004 to 2017. . For this data set I filtered out the cities and removed all columns with non integers.

Exploratory Analysis:

Once I got all the datasets ready, I had to normalize each data set, to make sure the scale of the number is correct. After normalizing the data, I created 5 dendrograms to tell me which two cities were similar, I also used Algomerative Clustering to map the cities to show the clusters:

Example of normalization:

![A screenshot of a cell phone

Description automatically generated]()

Weather:

0 is Austin, 1 is Chicago, and 2 is Seattle

![A screenshot of a social media post

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

Census:

0 is Austin, 1 is Chicago, and 2 is Seattle

![Chart, box and whisker chart

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

Health:

0 is Chicago, 1 is Austin, and 2 is Seattle

![A screenshot of a social media post

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

Cost:

0 is Seattle, 1 is Chicago, and 2 is Austin

![Chart, box and whisker chart

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

Venue:

0 is Austin, 1 is Chicago, and 2 is Seattle

![Chart, box and whisker chart

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

**Results**

|  |  |
| --- | --- |
| Cities | Similarity Count |
| Austin - Seattle | 0 |
| Austin - Chicago | 1 |
| Chicago - Seattle | 4 |

From our exploratory analysis, we can see that Chicago and Seattle have the most common with 4 categories and Austin and Chicago have 1 category in common, while Seattle and Austin do not have any.

**Discussion**

In out explanatory analysis, we can see that Chicago and Seattle are the cities that have the most common. When it came to Census, Venues, Cost and Weather, Chicago and Seattle were similar, and for health it was Austin and Chicago.

**Conclusion**

In conclusion, from our analysis and data we can say that Chicago and Seattle are the most similar. This necessarily doesn’t mean that they are more similar than Austin and Seattle or Austin and Chicago, it just means that from the data we used Chicago and Seattle are the two cities out of Austin, Chicago and Seattle that are similar.