Predicting City Quality Score Based on

Venue Frequency

Sunjin David Kim

August 23, 2019

1. Introduction

1.1 Background

Every year, UN News analysis the 125 most populous cities of United States to determine the best cities to live. These cities are given an overall score from 0-10, and a ranking from 1-125. In addition to original surveys conducted by UN News, the following data sources were used to determine the city scores.

- United States Census
- Gallup-Healthways Well-Being Index
- Federal Bureau of Investigation Uniform Crime Report
- Bureau of Labor Statistics

The following Indexes were used to determine the overall score of each city:

- Job Market Index
- Value Index
- Quality of Life Index
- Desirability Index

For more information concerning the methodology of the research, please go to:
https://realestate.usnews.com/places/methodology

1.2 Problem

In order to be competitive in the Realestate business, or to personally purchase home in a nice neighborhood before the prices of homes rise, it is important to use data to predict potential cities that will become popular in the future. Although the Rankings provided by USNEWS is very helpful in determining current hotspots, as potential home buyers, it is imperative to be ahead of the curve and purchase a home in a city before the popularity and prices rise. Therefore, using the data gathered from Foursquare, we will attempt to identify variables that correllate with city popularity represented by the City Scores, which could potentially help us find cities that will become popular in the future.

1.3 Interest

This analysis may be useful to the following groups:

1. Real-estate Agents looking for potential cities to work in.
2. Individuals looking for a good place to live.
3. City developers and officials seeking to improve their cities can also use this information to determine what types of venues to increase or decrease.

2. Data acquisition and cleaning

2.1 Data sources

I will be conducting an analysis the 5 top ranked and the 5 lowest ranked cities to live in the US according to the 2019 Best Places to Live Research conducted by US NEWS.

The Following Cities will be analyzed:
Rank #1: Austin, TX
Rank #2: Denver, CO
Rank #3: Colorado Springs, CO
Rank #4: Fayetteville, AR
Rank #5: Des Moines, IA

Rank #125: San Juan, PR
Rank #124: Bakersfield, CA
Rank #123: Stockton, CA
Rank #122: Shreveport, LA
Rank #121: Mobile, AL

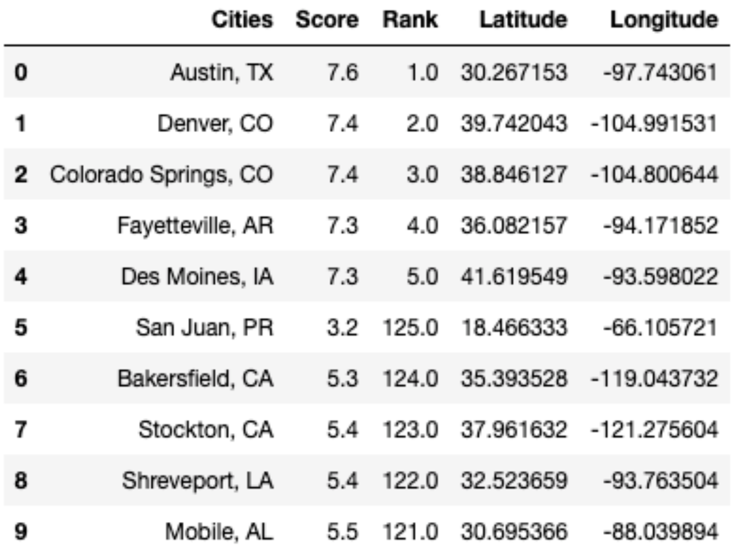
I will be obtaining the city scores from the research conducted by US News and the city coordinates from the web. It should be noted that the Lowest Ranked Cities, are not the worst cities to live in. Rather, they are the cities ranked 123-124th in the list of best places to live.

Venue Data

I will be using the Foursquare API to determine the frequency and ratio of venue types in each of the cities to determine whether there are any significant correlation between the venues found in a city and its overall score. If possible, we will use a machine learning algorithms to make predictive model that can can predict for us the city's overall score based on their venues
I will pick 4 random cities to test this model.

2.2 Data cleaning

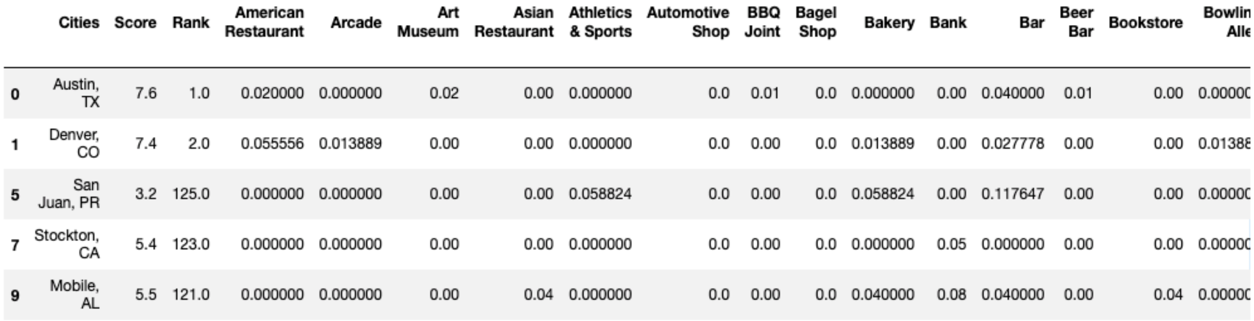
We began by creating a dataframe with the 5 top scoring cities and the 5 lowest scoring cities. We included their name, score, rank, latitude, and longitude.



	Cities	Score	Rank	Latitude	Longitude
0	Austin, TX	7.6	1.0	30.267153	-97.743061
1	Denver, CO	7.4	2.0	39.742043	-104.991531
2	Colorado Springs, CO	7.4	3.0	38.846127	-104.800644
3	Fayetteville, AR	7.3	4.0	36.082157	-94.171852
4	Des Moines, IA	7.3	5.0	41.619549	-93.598022
5	San Juan, PR	3.2	125.0	18.466333	-66.105721
6	Bakersfield, CA	5.3	124.0	35.393528	-119.043732
7	Stockton, CA	5.4	123.0	37.961632	-121.275604
8	Shreveport, LA	5.4	122.0	32.523659	-93.763504
9	Mobile, AL	5.5	121.0	30.695366	-88.039894

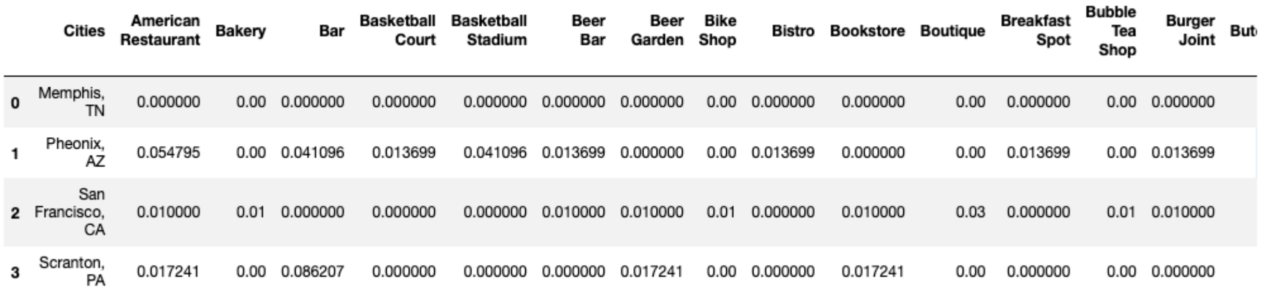
Next, using the Foursquare API, we gathered venue information of each city, setting the limit to 1000 and radius to 2000. Several cities had to be dropped from the data due to a lack of venue information. Cities with less than 10 venues were dropped, which includes Bakersfield, Colorado Springs, Des Moines, Fayetteville, and Shreveport.

The frequency of each venue category for each city was calculated using the information gathered from Foursquares. A new dataframe was created with the city name, score, rank, and the venue frequency. The rank of the city was later removed as it was redundant and a relative number based on comparison of other cities.



	Cities	Score	Rank	American Restaurant	Arcade	Art Museum	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Beer Bar	Bookstore	Bowlin Alle
0	Austin, TX	7.6	1.0	0.020000	0.000000	0.02	0.00	0.000000	0.0	0.01	0.0	0.000000	0.00	0.040000	0.01	0.00	0.00000
1	Denver, CO	7.4	2.0	0.055556	0.013889	0.00	0.00	0.000000	0.0	0.00	0.0	0.013889	0.00	0.027778	0.00	0.00	0.01388
5	San Juan, PR	3.2	125.0	0.000000	0.000000	0.00	0.00	0.058824	0.0	0.00	0.0	0.058824	0.00	0.117647	0.00	0.00	0.00000
7	Stockton, CA	5.4	123.0	0.000000	0.000000	0.00	0.00	0.000000	0.0	0.00	0.0	0.000000	0.05	0.000000	0.00	0.00	0.00000
9	Mobile, AL	5.5	121.0	0.000000	0.000000	0.00	0.04	0.000000	0.0	0.00	0.0	0.040000	0.08	0.040000	0.00	0.04	0.00000

This process was repeated later with the test cities that were picked at random. (San Francisco, Phoenix, Memphis, and Scranton)



	Cities	American Restaurant	Bakery	Bar	Basketball Court	Basketball Stadium	Beer Bar	Beer Garden	Bike Shop	Bistro	Bookstore	Boutique	Breakfast Spot	Bubble Tea Shop	Burger Joint	But
0	Memphis, TN	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.00	0.000000	
1	Pheonix, AZ	0.054795	0.00	0.041096	0.013699	0.041096	0.013699	0.000000	0.00	0.013699	0.000000	0.00	0.013699	0.00	0.013699	
2	San Francisco, CA	0.010000	0.01	0.000000	0.000000	0.000000	0.010000	0.010000	0.01	0.000000	0.010000	0.03	0.000000	0.01	0.010000	
3	Scranton, PA	0.017241	0.00	0.086207	0.000000	0.000000	0.000000	0.017241	0.00	0.000000	0.017241	0.00	0.000000	0.00	0.000000	

2.3 Feature selection

After data cleaning, data from 234 venues in 5 cities were left.

For the test, 255 venues from 4 cities were used.

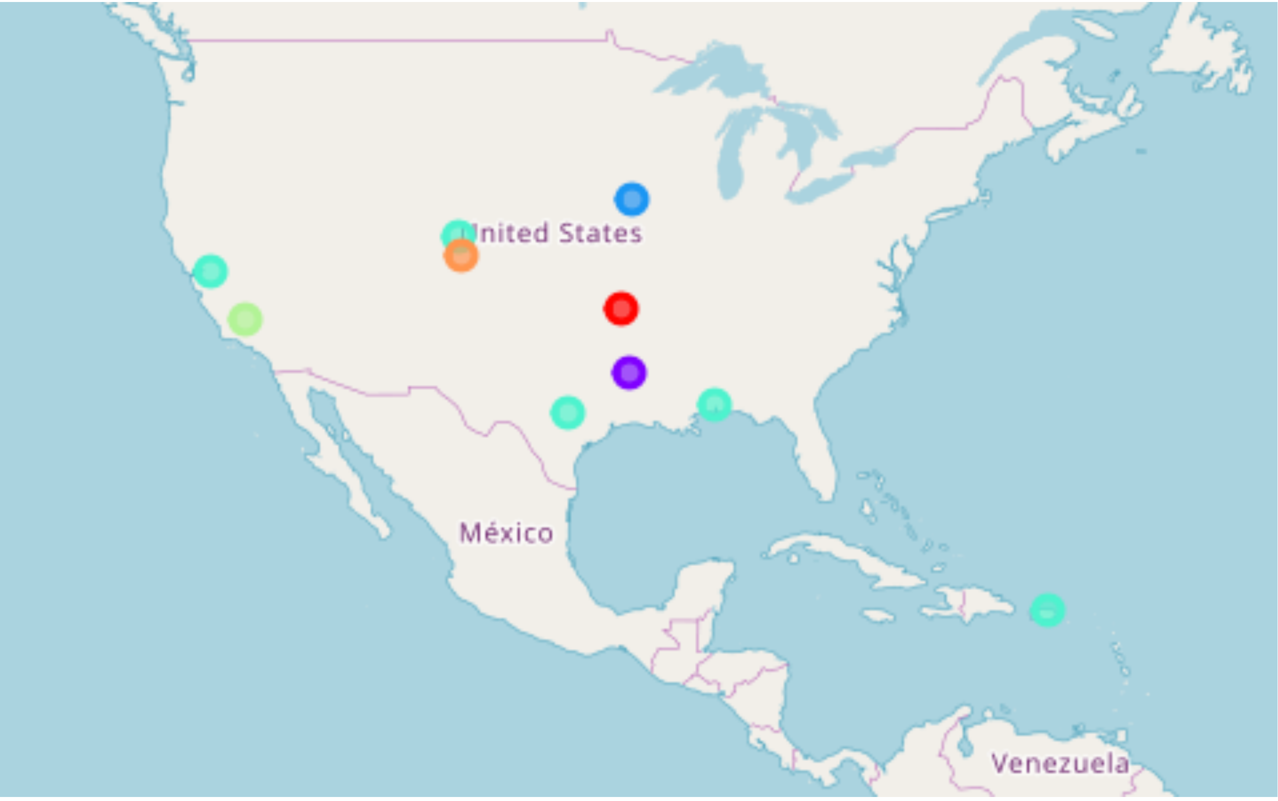
For future research, these number should be adjusted to get a better result.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 K-Cluster Analysis

To explore the data gained from Foursquare, we conducted a K-cluster analysis to see if the venue frequencies by themselves would create clusters that more or less aligned with their city scores. The initial analysis with the 10 cities revealed that there were cities with not enough data that was skewing the clusters. The initial cluster visualized onto a map can be found below.



Upon eliminating the cities that did not have enough venues and reducing the k to 2, the cluster formed according to city scores, with Austin and Denver in once cluster and San Juan, Stockton, and Mobile in the other.

3.1.2 Correlation

Using Pandas .corr() function, the correlation of the venue frequency and the city scores were calculated. Upon inspection, 7 venues were found to be positive correlated with the city scores with a correlation score greater than 0.83. (Café, Clothing Store, Dessert Shop, Restaurant, Salad Place, and Theater)

3.2 Predictive Modeling

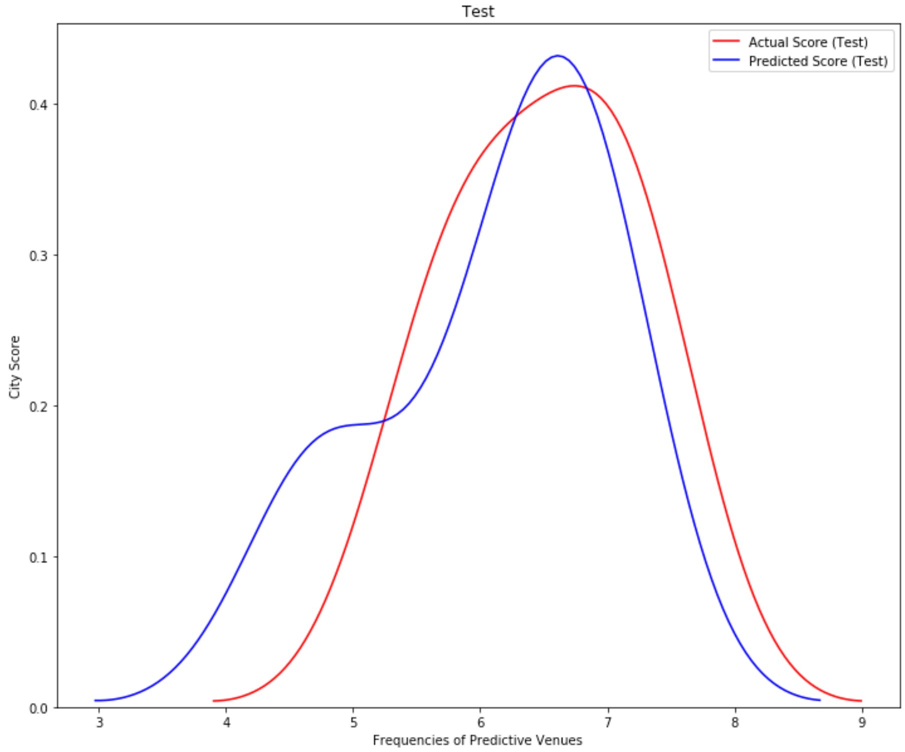
3.2.1 Multi Linear Regression Model

A multi linear regression model as created using the data from the remaining 5 cities. The city score was chosen for the dependent variable and the frequency of the 7 positively correlated venues were set as the independent variable. The model was tested using the venue frequency data gathered for the 4 test cities ((San Francisco, Phoenix, Memphis, and Scranton), and the predicted score was compared to the actual scores found on US NEWS.

4. Results

The results of the Regression Model test was the following.

```
Residual sum of squares: 0.53
Variance score: -0.25
```



There was an error in the modeling that mislabeled the city scores, which will be fixed at a later time.

5. Discussion

First I would like to discuss the lack of data for 5 of the original cities. It seems that the data provided by the Foursquare API is not comprehensive, especially in regions where the app is not frequently used, therefore, the data that remained are based on cities where people tend to use Foursquare and for venues that are often uploaded by those using Foursquare. Furthermore, we must recognize that 234 venues in 5 cities is not ideal. With a greater database, we may end up with other venue types that have a positive correlation with city scores.

Despite the shortcomings in data as well as analysis ability, interesting results were found from this research. The negative Variance Score would indicate that there is some kind of error that needs to be dealt with, which is beyond the scope of my ability or this research. However, the model visualization reveals that the model is able to predict the score of a city fairly accurately. The extra curve in the prediction seems to indicate that there was some overfitting of the model. In future studies, each variable can be tested separately, and different combinations of variables could be used to improve the model.

1. Conclusions

In this study, I analyzed the relationship between venue frequency and the overall city score based on the study conducted by US News. Using the correlation function and the K-cluster analysis, I identified that there was a positive correlation between the city score and the venue frequency of 7 venue categories (Café, Clothing Store, Dessert Shop, Restaurant, Salad Place, and Theater). I created a Multi Linear Regression Model and tested my theory that city scores can be predicted using venue category frequency. Although, many improvements can be made, the model was fairly successful in predicting the scores of randomly picked cities. Such prediction models can be used to assist real-estate agents and individuals looking to purchase a home in deciding which city to invest in. It can also help city planners in picking the next venue to invite into their city.