Factors of Cities Ranked as the Unhealthiest Cities

Yui Tamaki

March 14, 2021

1. Introduction

1.1 Business Problem

You are working for a health organization that wants to improve American's health and well-being. Staying healthy and fit has been the latest craze among many Americans thanks to many social media influencers promoting healthy eating and fitness routines but yet Americans still struggle to stay healthy. It seems as though there is a correlation between your health and the city that you live in. Your organization wants to find out what differentiates cities with the healthiest population to cities with the unhealthiest population so your organization can suggest cities what they can do to improve the city's health and well-being.

We will be comparing cities with the healthiest population per capita to cities with the unhealthiest population per capita to find the key differences that differentiates the cities. I can provide insights as to why a particular city is one of the unhealthiest cities in America by comparing the number of healthy food options and the number of fitness facilities. By providing these insights, the goal is to one day make the cities that were ranked as the unhealthiest cities to one of the healthiest cities in America.

2. Data Sources, Cleaning and Manipulation

2.1 Data Sources

We will be comparing cities that ranked as the healthiest cities in America to cities that ranked as the unhealthiest cities in America. The WalletHub has ranked 182 cities in America in their Healthiest and Unhealthiest Cities in America report. I scraped the entire table from the website and ranked the top 50 entries as the healthiest cities in America and the last 50 entries as the unhealthiest cities in America.

We will then look at Foursquare API's venues data for the 100 cities. The venues data that we are interested in are healthy food providers such as restaurants and markets. We are also interested in fitness facilities such as gyms and recreation centers.

2.2 Data Cleaning

The data that we scraped or downloaded from various sources include data that may not be relevant to this analysis so I have removed data that were irrelevant.

We first look at the scraped data from WalletHub. Data that we need from this data is the City data so I removed the following columns from the table: Overall Rank, Total Score, Health Care, Food, Fitness and Green Space.

	City
0	San Francisco, CA
1	Seattle, WA
2	Portland, OR
3	San Diego, CA
4	Honolulu, HI
...	...
177	Memphis, TN
178	Shreveport, LA
179	Gulfport, MS
180	Laredo, TX
181	Brownsville, TX

The second step of the data cleaning process is to look at the data downloaded from Foursquare API's venues data. In this data, I have identified all the venues that are related to health and fitness, which I have listed down below:

Health: Vegetarian / Vegan Restaurant, Doctor's Office, Pharmacy, Salad Place, Health Food Store, Supplement Shop, Medical Center, Farmers Market, Grocery Store, Supermarket, Organic Grocery, Fruit & Vegetable Store

Fitness: Gym / Fitness Center, Gym, Sports Club, Skating Rink, Tennis Court, Athletics & Sports, Basketball Court, Hockey Field, Baseball Field, Climbing Gym, Recreation Center, Bike Rental / Bike Share, Yoga Studio

After identifying venues that are related to health and fitness, I am now able to remove the venue data that does not fall into the health and fitness category. With this, the data cleaning process has completed so we now move on to the data manipulation phase.

2.3 Data Manipulation

Looking at all the cities that were ranked from the healthiest to the unhealthiest cities might be fun but we cannot really make any judgement as to why the cities were ranked in such a way. To find out what actually contributed to this ranking, we need to add a few data points.

Let us first add the latitude and longitude points of the cities. We can add these points by utilizing the Nominatim API that uses open street map data to find locations in the world with just the name and address of the location, also known as geocoding.

Here, we have the data for the top 50 healthiest cities with their latitude and longitude coordinates.

	City	Latitude	Longitude
0	San Francisco, CA	37.779026	-122.419906
1	Seattle, WA	47.603832	-122.330062
2	Portland, OR	45.520247	-122.674195
3	San Diego, CA	32.717420	-117.162773
4	Honolulu, HI	21.304547	-157.855676
5	Washington, DC	38.894992	-77.036558

The following dataset is for the top 50 unhealthiest cities with their latitude and longitude coordinates.

	City	Latitude	Longitude
0	Indianapolis, IN	39.916401	-86.051957
1	Lewiston, ME	44.100351	-70.214776
2	Little Rock, AR	34.746481	-92.289595
3	Arlington, TX	32.701939	-97.105624
4	El Paso, TX	31.775415	-106.464634
5	Chattanooga, TN	35.045722	-85.309488

Now that we have the latitude and longitude coordinates for all 100 cities, we can utilize the Foursquare API to retrieve the venue data for the cities. By providing the city name, latitude and longitude, Foursquare API can pull up all the venues in the specified city. The following dataset is the venue data for the top 50 healthiest cities. Foursquare API was able to retrieve 2484 venue information for the 50 cities.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	San Francisco, CA	37.779026	-122.419906	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
1	San Francisco, CA	37.779026	-122.419906	Herbst Theater	37.779548	-122.420953	Concert Hall
2	San Francisco, CA	37.779026	-122.419906	War Memorial Opera House	37.778601	-122.420816	Opera House
3	San Francisco, CA	37.779026	-122.419906	San Francisco Ballet	37.778580	-122.420798	Dance Studio
4	San Francisco, CA	37.779026	-122.419906	Main Stage Of Davies Symphony Hall	37.777703	-122.420476	Performing Arts Venue

Here is the venues dataset for the top 50 unhealthiest cities with 1263 venue information.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Indianapolis, IN	39.916401	-86.051957	Sahm Golf Course	39.916499	-86.052166	Golf Course
1	Indianapolis, IN	39.916401	-86.051957	Sahm Park	39.915540	-86.056717	Park
2	Indianapolis, IN	39.916401	-86.051957	Sahm Aquatic Center	39.916161	-86.053200	Water Park
3	Indianapolis, IN	39.916401	-86.051957	Sahm Park Playground	39.916161	-86.053200	Playground
4	Indianapolis, IN	39.916401	-86.051957	Sahm Park Basketball	39.919759	-86.053497	Basketball Court

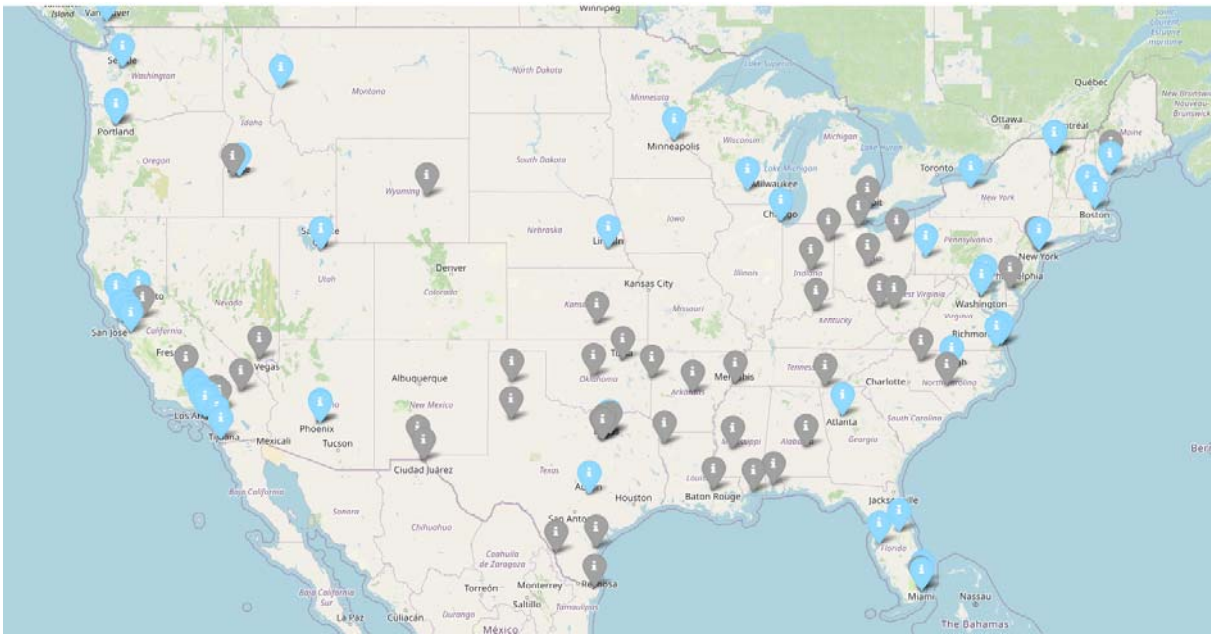
Now that we have gathered enough information and prepared the dataset for both healthiest and unhealthiest cities dataset, we can now go ahead in to the analysis phase.

3. Analysis and Methodology

During the analysis phase, we look at graphs and maps that compares the healthiest and unhealthiest cities. We first compare the spread of the cities on a map of America. We then explore the venue categories using a bar graph and then lastly look at a clustered map of the venues.

3.1 Location Differences between Healthy and Unhealthy Cities

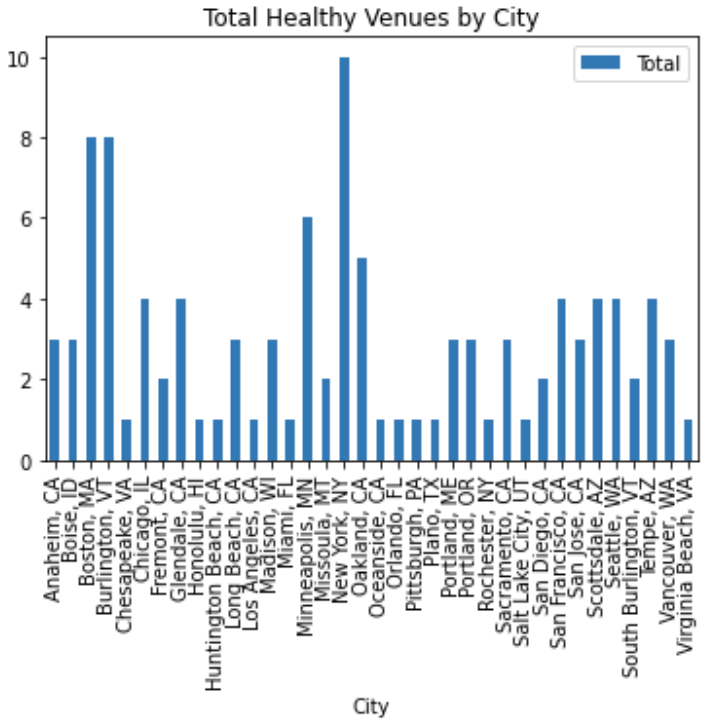
We added latitude and longitude coordinates to the scraped city data earlier in the data cleaning and manipulation phase. We use this dataset to map the cities onto a map of America. The top 50 healthiest cities are marked with a light blue pin and the top 50 unhealthiest cities are marked with a light gray pin.



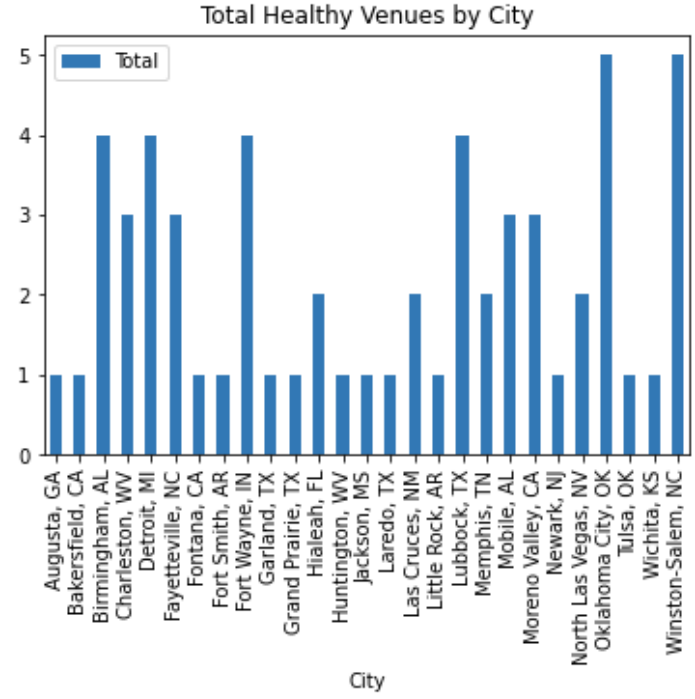
From the map, you can see that the healthiest cities are closer to the coastal region and the unhealthiest cities are in the central region of America.

3.2 Bar Graph

Next, we will look at the venue categories from Foursquare API. The venue categories for each city contained venues that were not related to health and fitness so during the data cleaning phase, we removed all venues that did not pertain to health and fitness. We created a bar graph that shows the total health and fitness venues each city provides. First, we are going to look at the bar graph of the top 50 healthiest cities in America.



We see in the above bar graph that the maximum total number of health and fitness venues is 10 and about a quarter of the cities have at least 4 or more venues that are related to health and fitness. Let us now look at the top 50 unhealthiest city's bar graph.

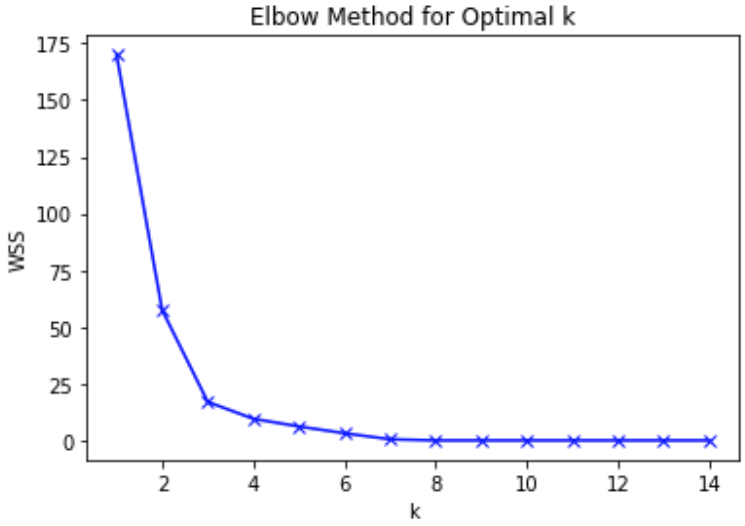


You can see in the above bar graph that the highest number of health and fitness venue for the unhealthiest cities is 5 and only 6 venues have at least 4 or more health and fitness venues in their cities.

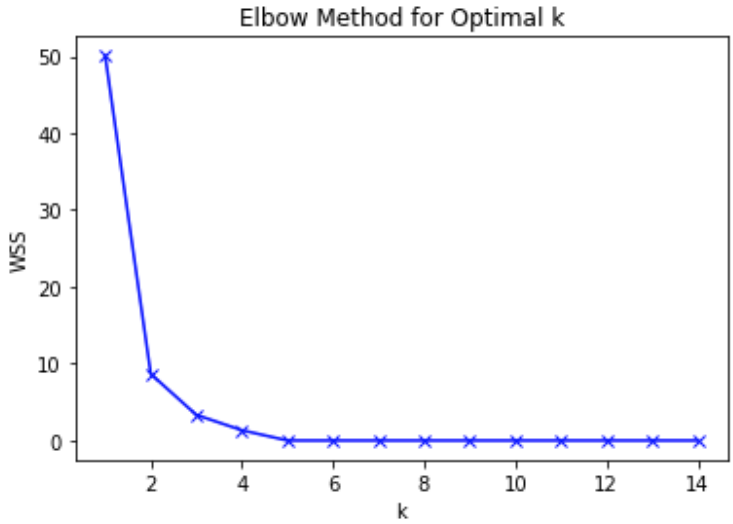
From the two bar graphs, we can see that the healthiest cities have a higher maximum total health and fitness venues compared to the unhealthiest cities, which suggests that the number of health and fitness venues that the city provides can potentially affect the overall health of their city's population.

3.3 Clustering

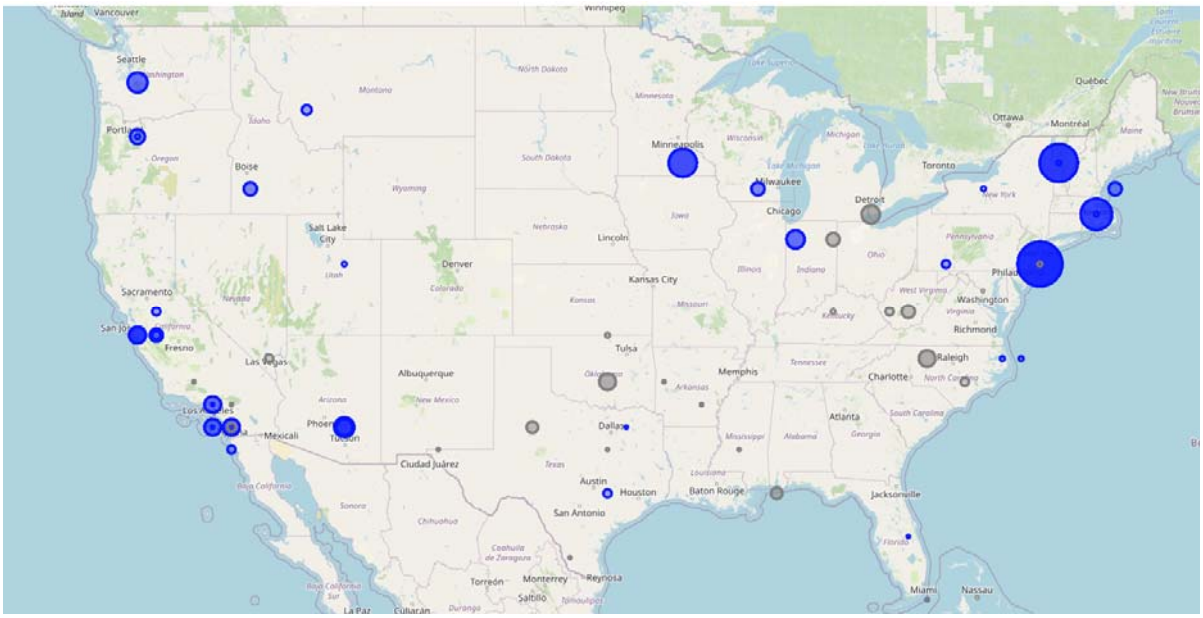
Lastly, we look at a clustered map. We cluster all the health and fitness venues based on their total numbers for both the healthiest city dataset and the unhealthiest city dataset. The healthiest city's cluster is in blue and the unhealthiest city's cluster is in gray. In order to create a clustered map, we need to determine the optimal number of clusters for the particular dataset. To achieve this, I used the K-Means clustering technique using the elbow method. The following graph is for the healthiest city dataset. From the "elbow" point of the graph, we can determine that the optimal number of clusters for this dataset is 3.



We now look at the unhealthy city dataset and apply the same technique. From the graph, we can conclude that the optimal number of clusters for the unhealthy city dataset is 2. The elbow method allows us to determine the optimal number of clusters by looking at the graph where the inertia (y-axis) starts to decrease in a linear fashion.



We have now determined the optimal number of clusters for both datasets. Based on the total number of venues, each city is assigned with a cluster number. Once all the cities are assigned a cluster number, we map the clusters on the map. The blue clusters are the healthiest city dataset and the gray clusters are the unhealthiest city dataset. Visually, we can determine that the blue clusters are much bigger than the gray clusters. The bigger the clusters, the more health and fitness venues are available for that city.



4. Results and Discussion

We collected data from WalletHub's Healthiest and Unhealthiest Cities in America report by scraping the table from the website and from Foursquare API's venue data. The data that we

scraped from WalletHub had a total of 182 cities but we were only looking at the top 100 cities in which we labeled the top 50 cities as the healthiest cities and the bottom 50 cities as the unhealthiest cities. The data contained several columns/data that were not relevant to this analysis so we cleaned the data by removing all unnecessary data. We then retrieved data from the Foursquare API's venue data. For this analysis, we were focusing on health and fitness venues so we identified all the venues that are health and fitness related. After identifying which venues we needed for this analysis, we removed all other venues. This left us with 108 venues for the top 50 healthiest cities and 59 venues for the top 50 unhealthiest cities to work with.

Once the data was ready for analysis, we mapped all 100 cities on the map of America. The map provided some visual understanding of how the healthiest and unhealthiest cities were spread across America. We saw that most of the healthiest cities were located in the coastal areas and most of the unhealthiest cities were located in the central region of America. However, the map is not enough to determine what constitutes the cities as a healthy city or an unhealthy city.

We then looked at the data by using a bar graph. The graph was based on the total number of health and fitness venues of each city. The maximum total of the healthiest city dataset was 10 whereas the maximum total of the unhealthiest city dataset was 5 which we start to see the factors that differentiates the two datasets apart.

To further analyze the two datasets, we clustered the data into appropriate groups and mapped the result on to the map of America. We used the elbow method of the k-means clustering technique to determine the optimal number of clusters for each dataset. The optimal number of clusters for the top 50 healthiest city dataset was 3 and 2 for the top 50 unhealthiest city dataset. We mapped the clusters on to the map, which provided a better insight as to what factors differentiates the two datasets.

After analyzing the two datasets, we can see that the number of health and fitness venues that a city provides to their people does have an impact on the overall health of the community and thus differentiates those cities that have a higher rate of healthier individuals than those cities that have a lower rate of healthier individuals.

From this analysis, each individual city can utilize this information to provide their community with adequate access to health and fitness venues so their community can see an improvement in their overall health.

5. Conclusion

The objective of this project was to find out what constituted the differences between cities that are healthy and unhealthy. The cities and venues were chosen based on WalletHub's Healthiest and Unhealthiest Cities in America report and Foursquare API's venues data. By analyzing the dataset, we can conclude that the number of venues that a city provides to their community plays a major role in the community's overall health. Based on this finding, we can suggest cities that were ranked as one of the unhealthiest cities to make efforts to providing more health and fitness venues for their community.