Portland vs. Memphis: Comparing America's Healthiest and Fattest Cities

Daylen Mackey

June 27th, 2019

1. Introduction

1.1 Background

Obesity rates in the United States are continuously climbing. It's no secret that an unhealthy diet and lack of exercise contribute to the problem, but there are still many unknowns.

Wallethub compared data from the *U.S. Census Bureau, Bureau of Labor Statistics, Centers for Disease Control and Prevention, County Health Rankings, United States Department of Agriculture Economic Research Service, Child and Adolescent Health Measurement Initiative, Gallup-Sharecare, and Trust for America's Health, and ranked the 100 'Fattest' cities in the country.*

1.2 Problem

I'd like to take a look at two cities on the opposite end of the list and compare the number of fast food restaurants in the two cities, as well as the number of gyms. If significant differences are found, to what extent?

Since the foursquare data must be used, could there be significant differences in venue frequency between these two cities? If there are significant differences, should governments implement restrictions? Could this location data influence future business placement?

2. Data acquisition and cleaning

2.1 Data Sources

The data that inspired me to take on this topic was the initial article by wallethub. 100 American cities were ranked based on their "Fatness" measurement (fatness measurement explained here). I wanted to compare two cities on opposite ends of that list and find any significant differences

in venues. I choose Portland and Memhpis because they were on opposite ends of the scale (97 and 3 respectively), and they had similar population sizes. While this isn't the only condition that should be accounted for, it seemed like a good starting point.

A list of neighbourhoods for the two cities were scraped from Wikipedia pages and fed into the Nominatim geolocator library to find their longitudes and latitudes. Now with a data frame of all the neighbourhoods and their coordinates, venues were collected from the Foursquare Developer API.

2.2 Data Cleaning

2.21 Wikipedia Scraping

Data retrieved from Wikipedia needed to be cleaned to it would fit register correctly in the Nominatim database.

The Portland list of neighbourhoods had many instances where brackets were used to indicate what else what in the neighbourhood ie. *Northwest District (includes Uptown, Nob Hill, Alphabet Historic District).* To solve this, I wrote a block of code to split the string if a '(' was detected, and only keep the first half – leaving only the neighbourhood name.

The Memphis neighbourhood list had many "slashes" in the list to indicated when a neighbourhood went by multiple names. Because Not all neighbourhoods may have coordinates in the GEOPY library, rows that had a "slashed" name, were broken into two rows: one row with the name on the left of the slash, one with the name on the right of the slash. I chose this approach because it reduces the likelihood of me getting an 'NaN' coordinate value, and I can always filter out duplicates later.

Once the neighbour lists were collected, I made a new column for each data frame and concatenate strings for the GEOPY library to read (example below)

City	**Before**	**After**
Portland	'Forest Park'	'Forest Park, Portland, Oregon'
Memphis	'Edge District'	'Edge District, Memphis, Tennesee'

2.22 Foursquare Data

For the Foursquare queries, I used a 2.5km radius because I wanted to capture as many venues as possible without exceeding my foursquare account request limitations. The 2.5 km radius

sometimes overlapped with other neighbourhoods, so foursquare returned duplicate venues ie. 4 rows with the information for the same McDonalds. This was solved by using the simple drop_duplicates method.

2.3 Feature Selection

After data cleaning, Portland and Memphis had dataframe shapes of (2650,4) and (2002,4) respectively. Because this study is focusing on the quantity of venues under specific categories, the only columns kept were the venue category, and their location.

Feature Selection Justification

Fast Food	Grocery Stores	Fitness Centres	Discount Stores
• Larger numbers imply increased consumption of fast food • Lower numbers imply reduced consumption of fast food	• Grocery stores grant access to healthier food options • Larger number of grocery stores implies more customers • Smaller number implies fewer customers	• Provide outlet for physical activity • Larger numbers imply increased membership numbers • Smaller numbers imply reduced membership numbers	• People who consume the most subsidized foods have a 37 percent greater risk of obesity than those who consume the least • Larger number may imply lower-income regions • Larger number implies increased consumption of subsidized foods

3 Methodology and Exploratory Data Analysis

3.1 Calculation of Target Variables

The four categories considered were Grocery Stores, Gym/Fitness Centres, Fast Food Restaurants, and Discount Stores. Four new data frames were made consisting only of those categories so they could be separately counted, visualized, and plotted on folium maps.

3.2 Statistical Significance

To determine the statistical significance when comparing categories from different populations, I decided to use a one tailed, Z Score calculation for two population proportions with an alpha of .05, this would let us know if any of the categories were statically different.

Fitness Centres

I hypothesized that the number of fitness centres in the two cities would not be statistically different. This prediction was made on the assumptions that simply having fitness facilities or purchasing a membership does not dictate overall fitness.

Fast Food Restaurants

I hypothesized the number of fast food restaurants would be higher in Memphis. This assumption was mostly founded on rumours that the state of Tennessee has more fast foods.

Grocery Stores

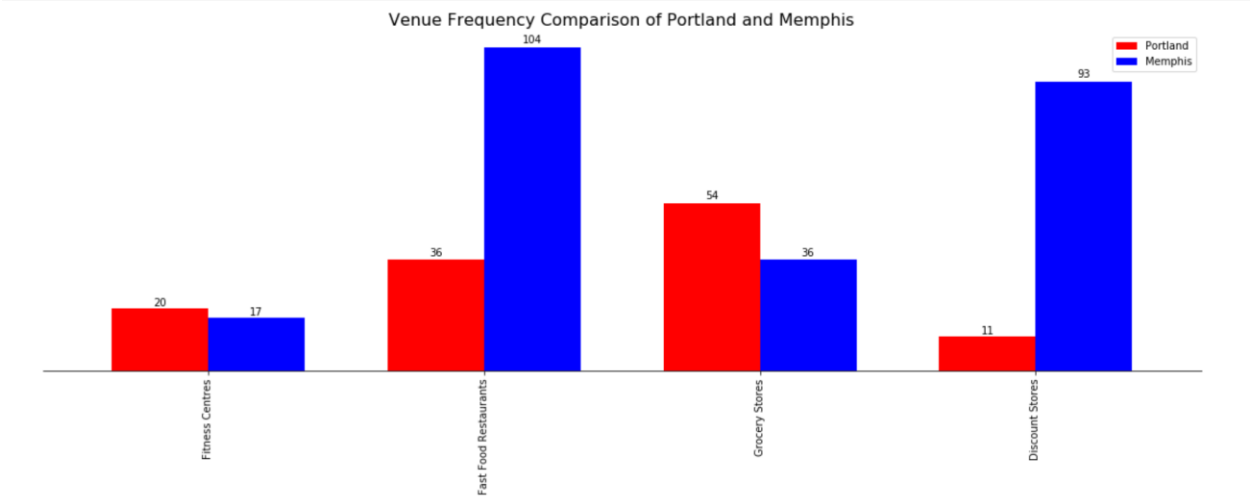
I hypothesized the number of grocery stores would be higher in Portland. This was made under the assumption that the number of Grocery Stores would directly correlate with population health.

Discount Stores

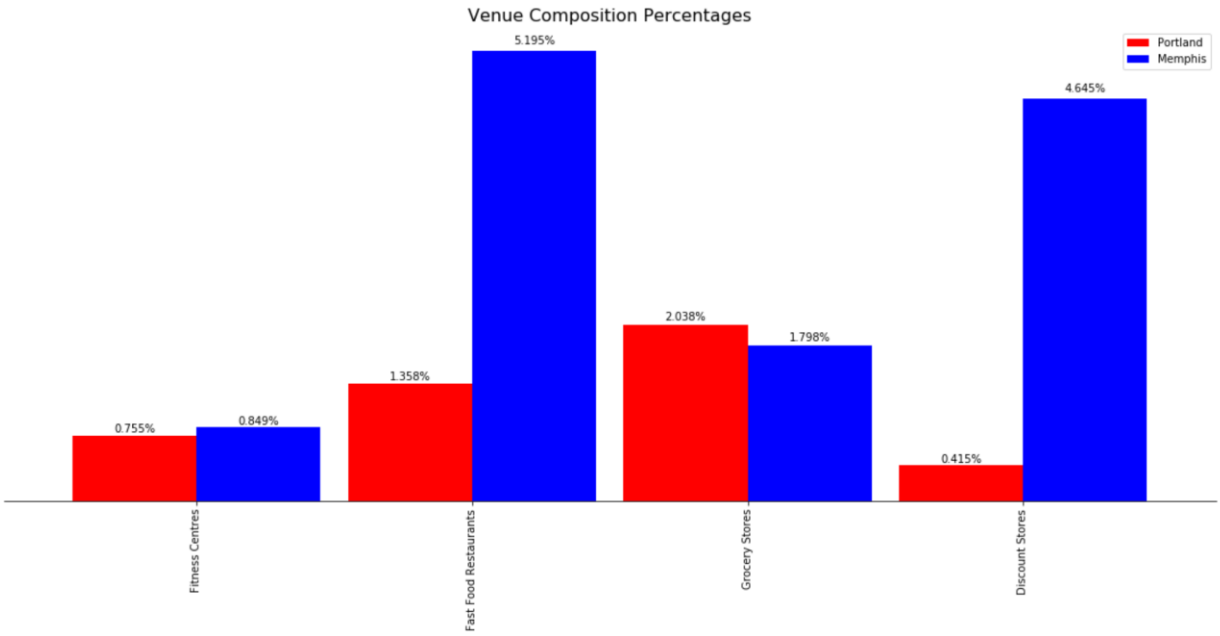
I hypothesized Memhpis would have a higher number of discount stores. This hypothesis was made on the knowledge that Portland has fewer low income areas, and likely, fewer discount stores

4. Results

	Fitness Centres	Fast Food Restaurants	Grocery Stores	Discount Stores
Portland	20	36	54	11
Memphis	17	104	36	93



	Fitness Centres	Fast Food Restaurants	Grocery Stores	Discount Stores
Portland	0.754717	1.358491	2.037736	0.415094
Memphis	0.849151	5.194805	1.798202	4.645355



After adjusting results for the number of venues returned,

Fitness Centres: The value of z is -0.359. The value of p is .35942. The result is not significant at $p < .05$.

Fast Food Restaurants The value of z is -7.583. The value of p is < .00001. The result is significant at $p < .05$.

Grocery Stores: The value of z is 0.5873. The value of p is .2776. The result is not significant at $p < .05$.

Discount Stores: The value of z is -9.6631. The value of p is < .00001. The result is significant at $p < .05$.

5 Discussion

In this study, I analyzed the frequencies of venue categories in Memphis and Portland. The four categories in question were Fitness Centres, Fast Food Restaurants, Grocery Stores, and Discount Stores.

Fitness Centres

The number of Fitness Centres were statistically similar with percentage of total venues at .755% and .849% for Portland and Memphis respectively. This similarity means that it does not account for the differences in obesity rates between the two cities.

Potential explanations for this is that the number of fitness centres is not an accurate predictor of population health. In order for gyms to remain running, they need people to buy memberships. Membership holders are not required to train hard, use the equipment properly, or put their membership to use at all. Therefore, the idea that the number of fitness centres is not an accurate predictor of population health seems reasonable.

Fast Food Restaurants

The number of fast food restaurants were statistically different. This implies that there is a correlation between the number of fast food restaurants in a city, and that city's obesity rate. It is worth noting that while this seems like a obvious implication, we cannot draw conclusions from this study. There may be exceptions to this rule – there may be cities with large number of fast food restaurants with lower than average obesity rates.

Grocery Stores

The number of Grocery Stores were statistically similar with percentage of total venues at 2.038% and 1.798% for Portland and Memphis respectively. A potential explanation for this is that grocery stores are simply a necessity – regardless of population health.

Discount Stores

The number of discount stores were statistically different with percentage of total venues at .415% and 4.645% for Portland and Memphis respectively. This implies that there is a correlation between the number of discount stores in a city, and that city's obesity rate. While we can't directly draw conclusions, there are few things we can infer. Memphis may have more lower-income regions, hence the increase in discount stores. This increased number may be indicative of disposable income that can be used on fitness, or earnings dedicated to food.

Potential Improvements for Future Studies

A full data set could not be acquired due to the limitations of foursquare. Not all venues are registered on foursquare, and due to the free account limitations, I was not able to access all available venues without exceeding the API's results restriction (search radius was reduced to 2.5 kilometers, and results were limited at 10000).

6 Conclusion

While statistically significant differences were found in 2 categories ("Fast Food Restaurants" and "Discount Stores"), much more could be learned from other categories (which may serve as superior predictors of population health)

Future Improvements:

- Using neighbourhoods to query venues is inefficient and may leave gaps. This could be improved by using a grid system to set equidistant queries within city border lines
- The Foursquare API limits the number of return results with a free account potentially reducing true results – upgrading may improve data quality