

**IBM Data Science Professional Certificate**

**Applied Data Science Capstone**

**Final project report**

**Faddy Kamel**

**07/04/2020**

## 1. Introduction

In this project I will attempt to find a relationship between the most common types of restaurants (including coffee shops, bars, etc) in a city and the city's obesity rate. If the impact on the obesity rate of a new restaurant opening can be predicted, this problem can be tackled more effectively. Since it is known in the medical community that nutrition is at least 50% responsible for the overall health of people, finding a correlation here can help deploy health improvement efforts more effectively.

The main stakeholders of this study would be doctors, health officials, fitness companies, and lawmakers. Doctors & health officials will have more information to predict the trend in the population they are treating and therefore can be more accurately proactive in preventing obesity. Lawmakers can use that information perhaps to pass legislation limiting the density of unhealthy restaurants or creating community initiatives to promote healthier lifestyles. Finally, fitness companies can leverage a population's obesity rate and willingness to reduce it to generate revenue.

## 2. Data

The obesity data will come from governing.com, which itself took it from the CDC. Unfortunately, the original CDC database does not seem to be accessible. The dataset contains a list of 192 American metropolitan areas (city & state) and the following data for each:

1. Normal weight (% of population)
2. Overweight (% of population)
3. Obese (% of population)
4. No physical activity (% of population)

The geographical coordinates for each city in the above dataset will come from <https://simplemaps.com/data/us-cities>. This dataset contains data for 28,889 incorporated US cities & towns taken from the United States Census Bureau, so it is reasonable to expect to find data for all 192 cities of interest.

The restaurant data will come from the Foursquare API. To make the calls, I will use the GPS coordinates of each city, and I will retrieve:

- Venue name
- Venue location
- Venue category

To make the search results more relevant, I will limit the queries to venues that fall under the "Food" category adding the "category\_id" parameter in the URL. I will use the query results to identify the most common types of food venues in each city.

### 3. Methodology

#### 3.1. Data formatting

To make the obesity dataset easily usable, it was formatted in 3 main ways. First, the “No physical activity” column was dropped. Second, the “Area” column was converted from a “City-ST Metro Area” template to a “City, ST” template using simple string operations. Finally, the data in the 3 other columns was converted from string to float, to allow for statistical analysis & visualization.

The GPS coordinates dataset also needed slight formatting. First, all columns except the ones listed below were dropped. The “city” and “state\_id” columns were then merged into an “Area” column with the template “City, ST”, then dropped.

- city
- state\_id
- lat
- lng

This gives us an obesity dataset and a GPS coordinates dataset with a common column, “Area”, containing data in the same format in both datasets.

That common column was used to merge both datasets and dropping any city that was not in the obesity dataset. The final dataset’s parameters are the following:

- Normal weight
- Overweight
- Obese
- Latitude
- Longitude

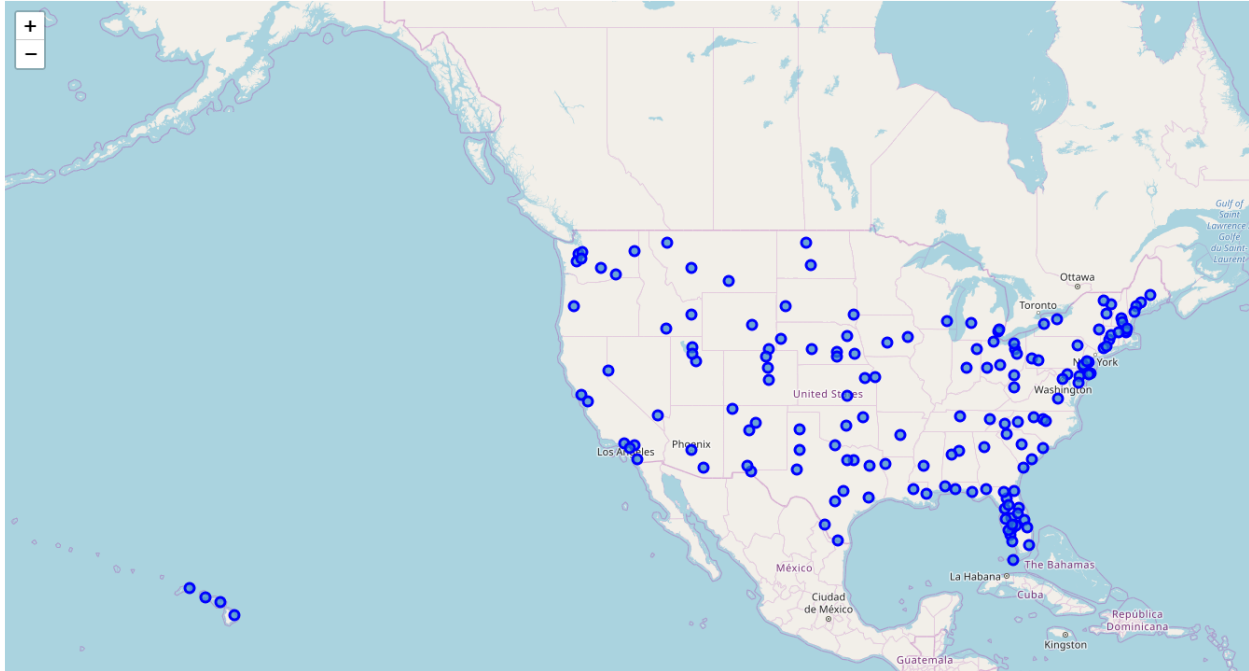
#### 3.2. Obesity dataset analysis

The first step of the project was to look more into the obesity data itself to have an idea of what the distribution is between cities, and within each city.

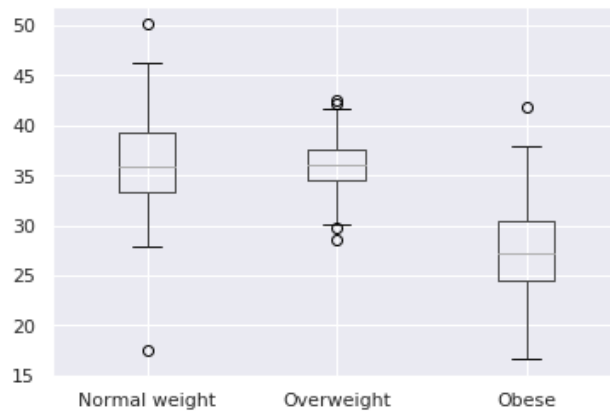
##### 3.2.1. Data visualization

To visualize the data, I used a Folium map and a boxplot.

The reason behind the Folium map was to make sure the dataset covers the entire country. 5 states are not represented: Alaska, Minnesota, Illinois, Missouri, and Kentucky. As these states are not very populated, the dataset is acceptable. Otherwise, the data is coherent with the population distribution, meaning that a lot of data is available for major areas such as Seattle, WA, Los Angeles, CA, and New York, NY. However, it is noteworthy that simply from looking at the map, Florida is disproportionately over-represented, with 22 cities.



The boxplot shows whether there were any outliers in the dataset and confirm the normal distribution. We see that the “Normal weight” and “Obese” columns contain outliers. Wauchula, FL is the outlier with very high obesity, and Fort Collins, CO is the outlier with a very high percentage of population being of normal weight.



### 3.2.2. Statistical analysis

Using the `df.describe` method on the final dataset, we obtain the following. The results obtained here quantify what is seen in the boxplot but do not teach us anything new. Note that we only have 160 cities in the final dataset, which means 31 cities from the obesity dataset are not in the GPS coordinates dataset. This could be due to the fact that the GPS coordinates dataset does not include unincorporated areas.

	Normal weight	Overweight	Obese	Latitude	Longitude
count	160.000000	160.000000	160.000000	160.000000	160.000000
mean	36.328754	36.235626	27.441875	37.057594	-92.798208
std	4.621789	2.518319	4.378614	6.404878	18.273823
min	17.600000	28.500000	16.600000	19.688600	-159.352000
25%	33.275000	34.599998	24.475000	32.324475	-103.936575
50%	35.800001	36.099998	27.550000	38.927000	-86.465050
75%	39.525000	37.700001	30.425000	41.983225	-80.343850
max	50.200001	42.500000	41.799999	48.237400	-68.790600

### 3.3. Restaurant dataset analysis

As it seems from Foursquare documentation that the default number of venues per query is 50, I put a limit of 100 results per city to have a dataset large enough to conduct the study on. The query returned 13,568 results between the 160 cities.

#### 3.3.1. Statistical analysis

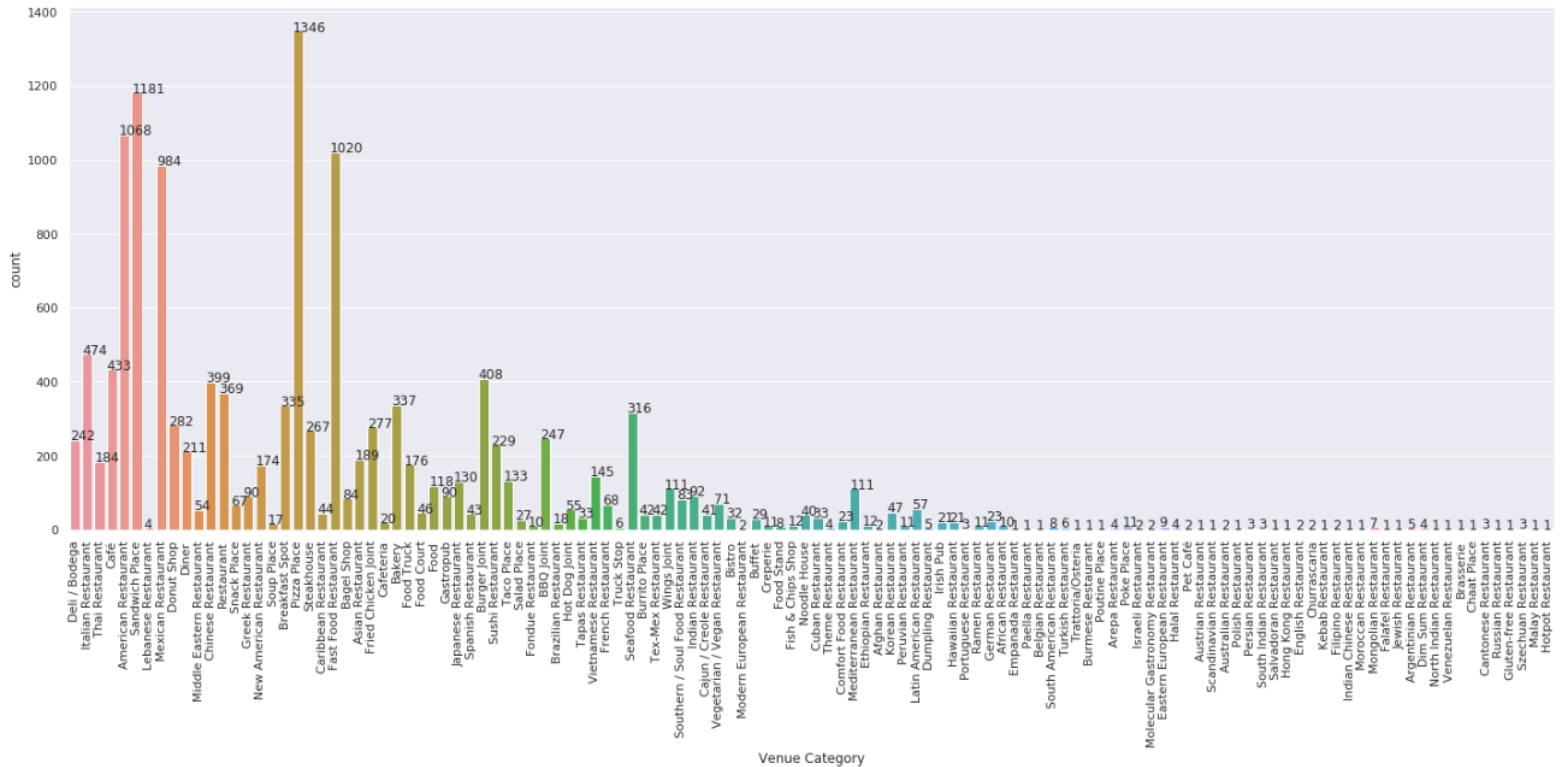
Starting with the statistical analysis can give us an indication whether we need a visualization tool, and if so which one to use.

The describe method shows us that there are 120 unique venue categories, and the most common one is "Pizza Place". This large number of categories makes this dataset much easier to understand visually.

	Area	Area Latitude	Area Longitude	Venue	Venue Category
count	13568	13568.000000	13568.000000	13568	13568
unique	160	NaN	NaN	9036	120
top	Little Rock, AR	NaN	NaN	SUBWAY	Pizza Place
freq	100	NaN	NaN	327	1346

#### 3.3.2. Data visualization

Using a Seaborn countplot, it is very easy to see which categories of food venues are the most common. After Pizza Place, Sandwich Place is the 2<sup>nd</sup> most common category with 1181 venues, followed by American Restaurant, with 1068.



### 3.4. Finding the restaurant-obesity relationship

The first step here is to determine what the most common venue category and the frequency of the category per city are. To do that, the JSON results are imported into a dataframe that is analyzed, and a new dataframe is created with the following data: Area, Most common venue category, and Venue frequency.

Once that is complete, all data is compiled into a single dataframe that will be used for the analysis, which has the following format.

Area	Normal weight	Overweight	Obese	Latitude	Longitude	Most common venue category	Venue frequency
City, ST	float	float	float	float	float	String	float

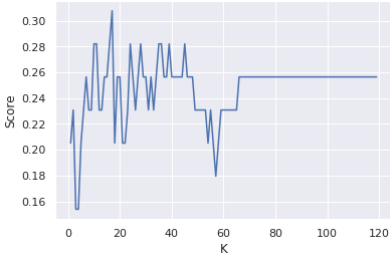
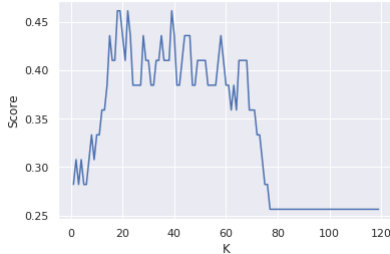
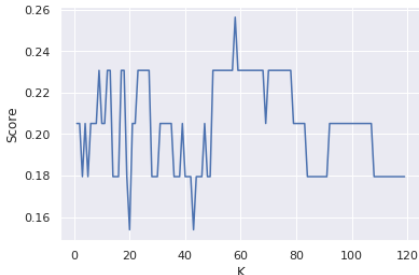
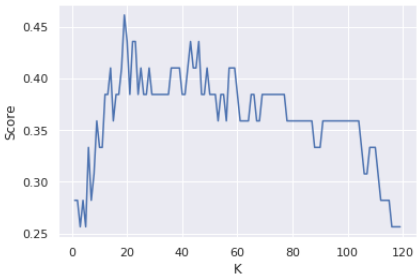
Before starting the analysis, the data was normalized so all percentage columns would range from 0 to 100. Since we are interested in the relationship between a numerical, continuous value and a categorical one, the best algorithm to use is a classifier. The input variables are the Obesity rate & Venue frequency, and the output variable is the Most common venue category.

Many algorithms were tested, including K-Nearest neighbors, Naïve Bayes Gaussian and Decision Tree, all from the scikit-learn library. Regression algorithms and Naïve Bayes Categorical were not suitable since the predicted variable is categorical, and the input variables are numerical. Being unsatisfied with the performance of these algorithms, I added Latitude and Longitude to the input variables.

## 4. Results

### 4.1. Finding the best algorithm

The results below show that the best algorithm to solve this problem with this dataset is the K Nearest Neighbors, uniform weights.

Algorithm	Score without GPS coordinates as input variables	Score with GPS coordinates as input variables
<b>K-Nearest neighbors, uniform weights</b>	K: 17 Score: 0.307692 	K: 18 Score: 0.461538 
<b>K-Nearest neighbors, distance weights</b>	K: 58 Score: 0.25641 	K: 19 Score: 0.461538 
<b>Naïve Bayes Gaussian</b>	0.3076923076923077	0.38461538461538464
<b>Decision Tree</b>	0.2564102564102564	0.28205128205128205

### 4.2. K-Nearest neighbors' performance

Here are the 8 unique categories of most common restaurants in the selected cities, with Jaccard index and F1 score for each.

Unique values in y_true	array(['American Restaurant', 'Fast Food Restaurant', 'Café', 'Mexican Restaurant', 'Sandwich Place', 'Bakery', 'Vietnamese Restaurant', 'Pizza Place'], dtype=object)
Jaccard index	array([0. , 0. , 0. , 0.28571429, 0.46153846, 0.44444444, 0.22222222, 0. ])
F-1 score	array([0., 0. , 0. , 0.44444444, 0.63157895, 0.61538462, 0.36363636, 0.])

## **5. Discussion**

The results above indicate that it is difficult to establish a direct relationship between the most common type of restaurant in a city, and that city's obesity rate. Including the geographical location of a city yields marginally better results for all algorithms except KNN, which saw its score increase by 50%. Out of 8 restaurant categories, 4 are impossible for the algorithm to predict. The algorithm is acceptable at predicting Sandwich places and Bakeries, poor for Mexican restaurants and Vietnamese restaurants, and completely fails when it comes to American restaurants, Fast food restaurants, Cafés and Pizza places.

It is interesting to see Fast food restaurants and Pizza places be among the least correctly predicted categories, as they could be considered as the 2 unhealthiest from the list.

## **6. Conclusion**

The purpose of this work was to evaluate the relationship between the obesity rate in a city and the most common type of restaurant in that city. The results show that the relationship is weak across all types of restaurants, and moderate for Sandwich places and Bakeries.

For health officials and city mayors to have more relevant conclusions on how to decrease the obesity rates in their cities, future work could include the percentage of population per city not physically active and/or the 2<sup>nd</sup> most common restaurant category.