Data Science Capstone Project: Using CDC data to analyze unhealthy behavior based on data from 500 cities

Vikram Seshadri

Introduction

Center for Disease Control collects census data on 500 major cities in the US every year. The dataset that they collected in 2018 which includes 2016 model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors, health outcomes, and preventive services.

CDC has identified obesity, less sleep (<7 hours a day), less physical activity, binge drinking, and smoking as unhealthy behaviors. These unhealthy behaviors are correctable by identifying the cause and motivating people to get into healthy habits. For example, obesity was identified as one of the unhealthy behaviors in the dataset. Obesity can be caused by eating unhealthy food, less physical activity, and lifestyle. All the three causes for obesity can be corrected by making good choices on food, signing up for gym/sports activity, and leading a good and healthy lifestyle. This project will focus on obesity, lack of sleep and physical activity for analysis.

Business Problem

The first problem would be to find the locations that are more seriously affected by unhealthy behaviors in the categories such as obesity, less sleep, and physical inactivity. Once these places are identified, it would be pertinent to find out the reason for such high numbers in unhealthy behavior. For example, is it possible that there are many fast-food restaurants in places where there is an obesity problem? Similarly, is there a lack of gym/outdoor recreational facilities such as parks/trials to enable physical activity.

Once the cause is found, there is an opportunity for the appropriate business startup in each of these locations to improve the quality of life long-term while being profitable. For example, setting up gymnasiums and sports complex in locations where there is a lack of physical activity is both profitable and improves the quality of life of people in that area. This CDC study can itself be an appropriate advertisement to motivate people to start healthy behaviors at a given unhealthy location.

A classification model can also be generated using the data to identify if a new location is healthy/unhealthy based on the surrounding venues. This model would be very useful to assess if businesses that can reduce unhealthy behaviors can be setup in a given location only based on the venue data.

Data

The dataset used for this problem will be obtained from the CDC website as a CSV file. The link to the dataset is provided below.
https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2018-relea/6vp6-wxuq

The dataset contains unhealthy behaviors, prevention, and health outcomes data for 500 cities in the United States. For this project, the data will be filtered to contain only the information on unhealthy behavior. The measure of the unhealthy behavior is in the form of Crude (CP) and Age-Adjusted Prevalence (AAP) along with their upper and lower confidence limits. The AAP and CP values are represented in percentage, denoting the percentage of population with specific unhealthy behavior. The dataset also contains the population count for each city, latitude and longitude information of each city, and detailed description of the measures and their meaning.

Latitude and longitude locations of these 500 cities along with Foursquare API are used to obtain the details of the venues, specifically the type of restaurants, gymnasium, yoga studio, and parks and recreations.

Depending on the details of the venues (or the lack thereof), an appropriate business model for each location can be recommended. Tools such as K-Means clustering and SVM can be used to find out the locations that will benefit from similar businesses that can improve the lifestyle.

Methodology

Data pre-processing

The data obtained from CDC as csv file is converted to a dataframe using Pandas. It has to be processed to remove information that are not necessary for this project. Several columns in this dataset such as state description, data source, and unique ID are not quite useful for the analysis and can be removed from the dataframe.

The geolocation data in the dataframe is in the format (latitude, longitude) as an object. It has to be converted to two separate columns containing latitude and longitude. The AAP and CP values for a given location are present in the dataframe as two separate rows. They have to be consolidated into one single row.

The unhealthy behaviors have to be isolated from the other two studies namely prevention and health outcome. There are five unhealthy behaviors that were observed in this study: Binge Drinking, Chain Smoking, Obesity, Lack of Sleep, and Lack of Physical Activity (LPA). For this study, only the latter three were considered. The other two were removed from the dataframe.

Further, these three unhealthy behaviors were separated into three individual dataframes containing each unhealthy behavior data. A stand-alone analysis was performed on the obesity data. Subsequently, the column name of these three dataframes for AAP_value were changed to AAP_obesity, AAP_LPA, and AAP_sleep. The three separate dataframes were once again merged together based on the location in order to perform joint analysis.

It is interesting to note that AAP values and CP values are highly correlated and it is sufficient to use AAP values for comparing unhealthy behaviors of different locations.

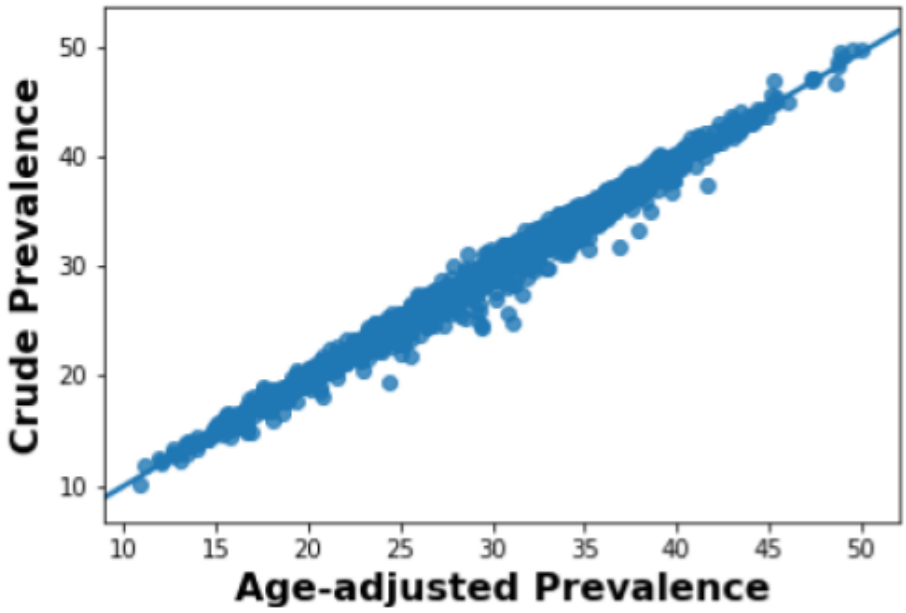

Figure 1: Age-adjusted and Crude prevalence are very linearly correlated.

Algorithms and Analysis

First, obesity data was analyzed separately using K-Means clustering similar to the Toronto dataset that was analyzed in one of the assignments. To perform this analysis, 15 locations with high AAP values and 15 locations with low AAP values were put together in the same dataframe. The venues data for these 30 locations were obtained using Foursquare API. In this case, only two clusters were used to see if the dataset would be naturally clustered into either healthy or unhealthy based on the venues in the given location.

In the second stage of analysis, all three unhealthy behaviors were put together in a single row for every location. Once again, 21 locations with all three AAP values greater than 35 and 23 locations with all three AAP values less than 29 were put together in one dataframe. In this case, three clusters were used with the hope that the locations would be naturally clustered based on their unhealthy behaviors with venues data. This analysis is unsupervised and is oblivious to the actual AAP values of the location.

Finally, supervised learning was used to train a model to predict whether a place has high or low prevalence of unhealthy behavior using Support Vector Machine algorithm. For this

model, 140 locations were used as training set and 112 locations were used as test set. To perform SVM, two variables were created called 'Problem Level' and 'SVM class'. Problem level is defined as the number of AAP values greater than 30 for a given location. For example, Brimingham, AL has AAP_obesity, AAP_LPA, and AAP_sleep greater than 30 and has a 'Problem Level' of 3. On the other hand, Bellevue, WA has all the three AAP values less than 30 and has a 'Problem Level' of 0. In the next stage, 'SVM class' was assigned a value of 1 for cities with 'Problem Level' < 2 and 0 for others. The following table shows examples for calculating Problem Level and SVM Class for different cities.

Table 1: Examples for calculating Problem Level and SVM Class using hypothetical cities.

	AAP_obesity	AAP_sleep	AAP_LPA	Problem Level	SVM Class
City 1	35	35	35	3	0
City 2	35	35	10	2	0
City 3	35	10	10	1	1
City 4	10	10	10	0	1

An SVM model was trained with the location and venue data and 'SVM class' and the test set for the same was used to predict the accuracy of the model. Similarly, an SVM model was trained with location and venue data and 'Problem Level'. These two models can be used to test any unknown location for its behaviors without having to do a survey in each and every location. To some extent, there is an underlying assumption that the rest of the locations in the US (or other parts of the world) have similar bearing on the extent of unhealthy behaviors.

Results

Basic statistics on unhealthy behavior

The histogram of the AAP of obesity, sleep, and LPA are shown in Figure 2. The basic descriptive statistics for the three AAP values for all the cities is provided in the table below. Scipy library was used for the normality test of the data. While the p-values of sleep and LPA were less than 0.05, obesity had a p-value of 0.25 suggesting that obesity data was non-normal. The average value of AAP of lack of sleep was much higher than LPA and obesity.

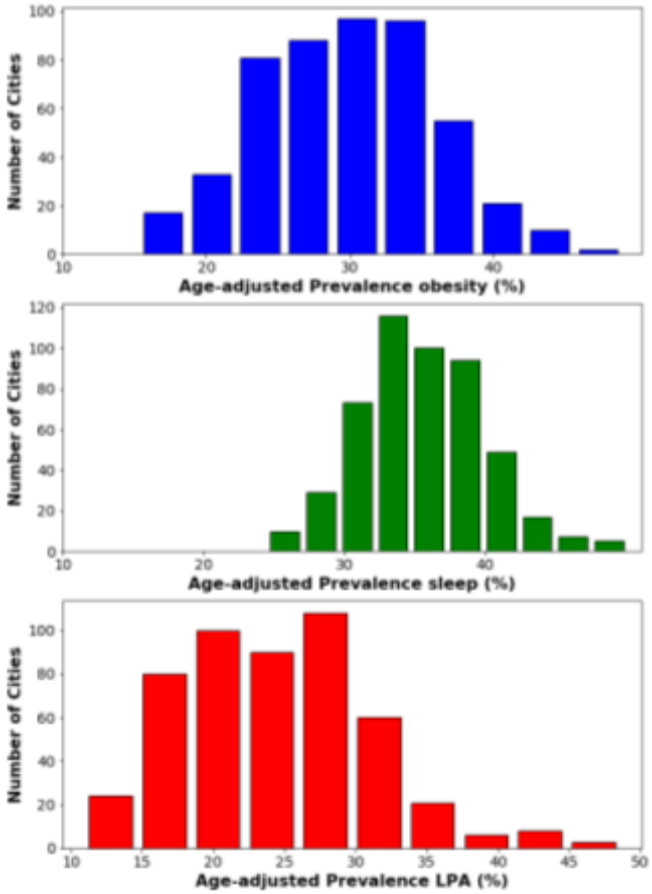

Figure 2: Histogram of AAP values and the number of cities.

Table 2: Descriptive statistics of AAP values for all the cities.

	Age-Adjusted Prevalence (%)		
	Obesity	Sleep	LPA
Mean	29.8	35.5	24.3
Standard deviation	6.1	4.4	6.6
Minimum	15.3	24.5	10.9
25 percentile	25.3	32.4	19.1
50 percentile	29.8	35.4	24.3
75 percentile	33.9	38.5	28.8
Maximum	49.0	50.1	48.6
P-value Normality test	0.25	0.007	2.35e-5

Obesity in United States

In the first part, we focus only on the obesity issue. The dataframe for obesity looks like the table shown below.

	Location	AAP_value	AAP_LCL	AAP_HCL	PopulationCount	MeasureId	lat	long	CP_value	CP_LCL	CP_HCL
0	Gary,IN	49.0	48.7	49.3	80294	OBESITY	41.590478	-87.347291	49.1	48.8	49.4
1	Detroit,MI	47.4	47.3	47.5	713777	OBESITY	42.384702	-83.105318	47.0	46.9	47.1
2	Flint,MI	45.4	45.2	45.6	102434	OBESITY	43.023634	-83.692064	45.0	44.8	45.2
3	Jackson,MS	44.5	44.3	44.8	173514	OBESITY	32.316272	-90.212453	43.6	43.3	43.9
4	Brownsville,TX	44.4	44.0	44.7	175023	OBESITY	25.998198	-97.456634	43.8	43.5	44.2

The dataset contains the locations with 15 high and 15 low AAP-value locations. Foursquare was used to obtain venue details for the categories below:

1. College_Gym = '4bf58dd8d48988d1b2941735'
2. Food = '4d4b7105d754a06374d81259'
3. Outdoor_and_rec = '4d4b7105d754a06377d81259'
4. Medical_centers = '4bf58dd8d48988d104941735'

A radius of 20 km and limit of 100 venues were used for the Foursquare API request. The resulting dataframe 'obesity_venues' contain 1380 rows with columns containing location details, venue details and category.

```
obesity_venues.shape
```

```
(1380, 7)
```

```
obesity_venues.head()
```

	Location	Location Lat	Location Long	Venue	Venue Category	Venue Lat	Venue Long
0	Gary,IN	41.590478	-87.347291	Starbucks	Coffee Shop	41.545950	-87.509370
1	Gary,IN	41.590478	-87.347291	Chicago Skyway	Bridge	41.719509	-87.545325
2	Gary,IN	41.590478	-87.347291	Albanese Confectionery	Candy Store	41.470291	-87.269804
3	Gary,IN	41.590478	-87.347291	Starbucks	Coffee Shop	41.600668	-87.558552
4	Gary,IN	41.590478	-87.347291	Panera Bread	Bakery	41.565249	-87.508483

```
print('There are {} uniques categories.'.format(len(obesity_venues['Venue Category'].unique())))
```

```
There are 172 uniques categories.
```

For performing K-Means clustering on this data, one-hot encoding was performed on the venues followed by grouping of data based on location and averaging the values by venue category. The number of clusters was chosen to be two to see if the locations would be naturally clustered into healthy and unhealthy locations depending on the type of venue. The location details and the AAP values were removed and the K-Means was performed only on the venues. After performing K-Means on the data, the location data was merged back with the venue data and the resulting dataframe is shown below.

	Location	AAP_value	PopulationCount	lat	long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	
0	Gary,IN	49.0	80294	41.590478	-87.347291	1	Coffee Shop	Park	Hospital	Fast Food	Gym / Fitness	Food Court	G
1	Detroit,MI	47.4	713777	42.384702	-83.105318	1	Gym / Fitness	Coffee Shop	Park	Hospital	Waterfront	Bridge	
2	Flint,MI	45.4	102434	43.023634	-83.692064	0	Coffee Shop	Fast Food	Hospital	Gym / Fitness	Lake	Park	Ca
3	Jackson,MS	44.5	173514	32.316272	-90.212453	0	Fast Food	Hospital	Coffee Shop	Gym	Café	Gym / Fitness	
4	Brownsville,TX	44.4	175023	25.998198	-97.456634	0	Fast Food	Coffee Shop	Mexican	Playground	Grocery Store	Sporting Goods	
5	Camden,NJ	44.1	77344	39.936191	-75.107296	1	Convenience Store	Park	Hospital	Plaza	Coffee Shop	Sandwiches	
6	Macon,GA	44.0	91351	32.832042	-83.649582	0	Fast Food	American	Seafood	Fried Chicken	Grocery Store	Medical	
7	Kansas City,KS	43.8	145786	39.123462	-94.744192	1	BBQ	Hospital	Coffee Shop	Gym	Food Court	Grocery Store	
8	Dayton,OH	43.5	141527	39.779768	-84.199793	0	Coffee Shop	Park	Fast Food	Hospital	Gym	Plaza	

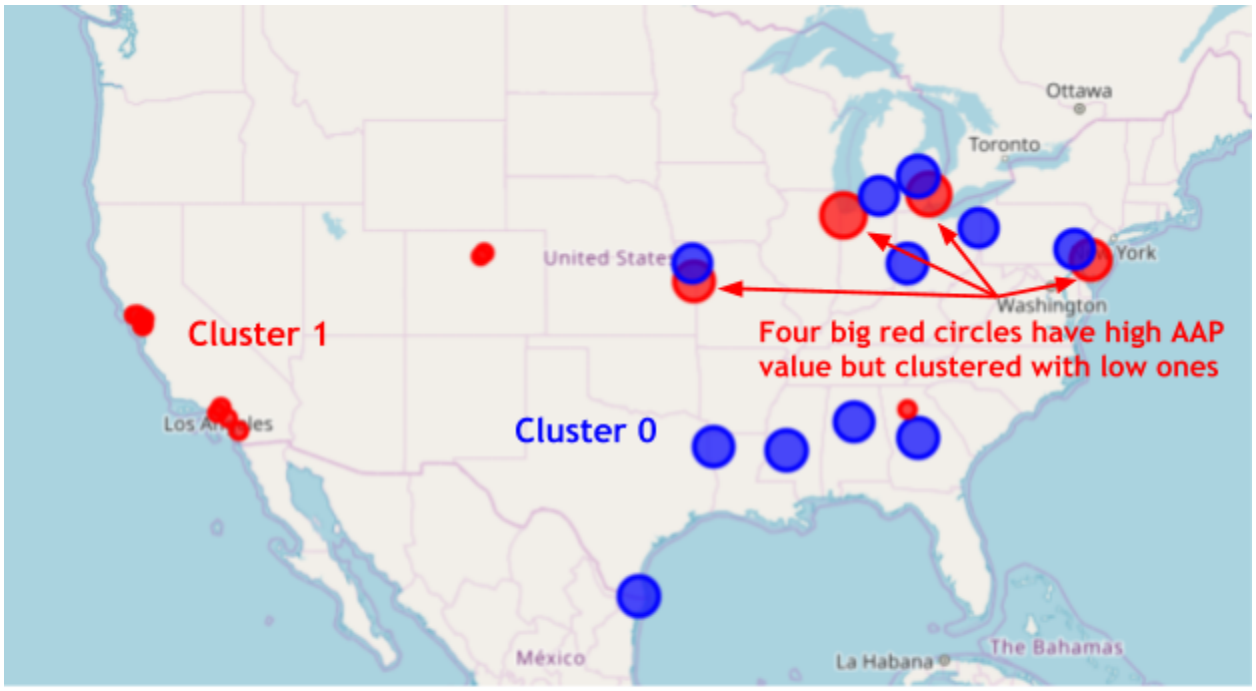

Figure 3: The two clusters from K-Means clustering are shown in red and blue. The size of the circle is proportional to the AAP of obesity. From the results, it can be observed that all the locations with high AAP values were clustered together.

Figure 3 shows the clustering of data by the K-Means algorithm. The radius of the marker in the figure is proportional to the obesity AAP value of the location. It can be seen that all the places that had high obesity were clustered together and are represented in blue. The second cluster is represented in red and it mostly contains location with low obesity AAP value. There are 4 locations that were present in cluster 1 with high obesity AAP value. It has to be noted that K-Means clustering can produce varying results depending on the choice of starting points.

Unhealthy Behaviors in United States

In this analysis, all three unhealthy behaviors were taken together for every location. The first five rows of dataframe for this analysis looks as below.

	Location	PopulationCount	AAP_obesity	AAP_LPA	AAP_sleep	lat	long
0	Birmingham,AL	212237	42.6	35.8	41.8	33.527566	-86.798817
1	Hoover,AL	81619	28.9	20.2	32.8	33.376760	-86.805194
2	Huntsville,AL	180105	33.3	25.9	38.3	34.698969	-86.638704
3	Mobile,AL	195111	39.7	29.3	39.7	30.677625	-88.118448
4	Montgomery,AL	205764	36.9	30.0	38.1	32.347265	-86.267706

Another dataframe was created to contain location with all the AAP values greater than 35 and all AAP values less than 29. This construction helped in understanding if the locations form a natural cluster based on their unhealthy behaviors. There were 44 cities in total for this analysis of which 21 cities had high AAP values and 23 cities had low AAP values. Three clusters were used to describe the data and the K-Means analysis was performed similar to that of obesity data. For the Foursquare API request, additionally Spiritual centers ('4bf58dd8d48988d131941735') and Massage_studio ('52f2ab2ebcbc57f1066b8b3c') were added to category ID list.

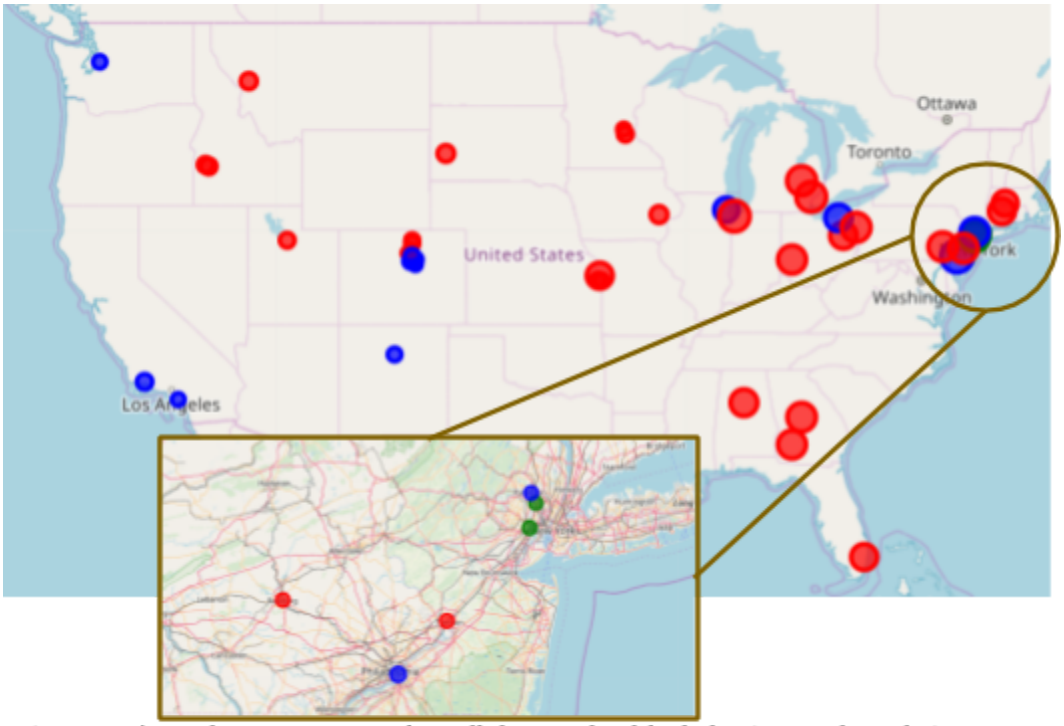

Figure 4: Three clusters were used on all three unhealthy behaviors and no obvious pattern was detected in the clustering.

From the K-Means analysis, there was no obvious pattern to the clusters formed with all the unhealthy behaviors put together. Two cities in New Jersey namely Newark and Passaic formed one of the clusters as they had identical top five venues. In terms of unhealthy behavior, 7 out of the remaining 19 unhealthy locations were present in the wrong cluster whereas 5 out of 23 healthy locations were misplaced in cluster with unhealthy locations. Nevertheless, it was interesting to note that the locations that were clustered into unhealthy behaviors had Fast Food in the top ten venues.

Modeling Unhealthy Behaviors using SVM

As described in the 'Data' section of the report, the locations were categorized based on the Problem Level and SVM class. For performing the Support Vector Machine analysis, the training and testing datasets needed to be merged together. This was done to ensure that the number of venue categories are the same in the training and testing dataset. Further, for training the SVM model, 35 locations from each Problem Level was used totalling 140 locations. A total of 112 locations were used for testing the model. The location data was removed from the dataframe to perform the SVM analysis.

Table 3: Total number of training and test datasets used for SVM model.

Problem Level	SVM Class	Total no. of locations	Training dataset	Testing dataset
3	**0**	80	35	35
2	**0**	166	35	35
1	**1**	212	35	35
0	**1**	42	35	7

In the first case, SVM model was trained with 'SVM Class' as the output variable for the model. For this SVM model RadialBasis Function was chosen as the kernel with parameter C=1.0 and gamma as 'auto'. It is evident from the confusion matrix in Figure 5 that the model was able to accurately predict the SVM class. Further, the model accuracy was able to provide the same level of prediction for C = 0.01 to 100.

Although the SVM model may have overfit, the sample size of the testing dataset was reasonably large (almost as large as the training set itself). For such a large test set, some misidentifications would be expected whereas the model was able to make a clear distinction based on the venue data.

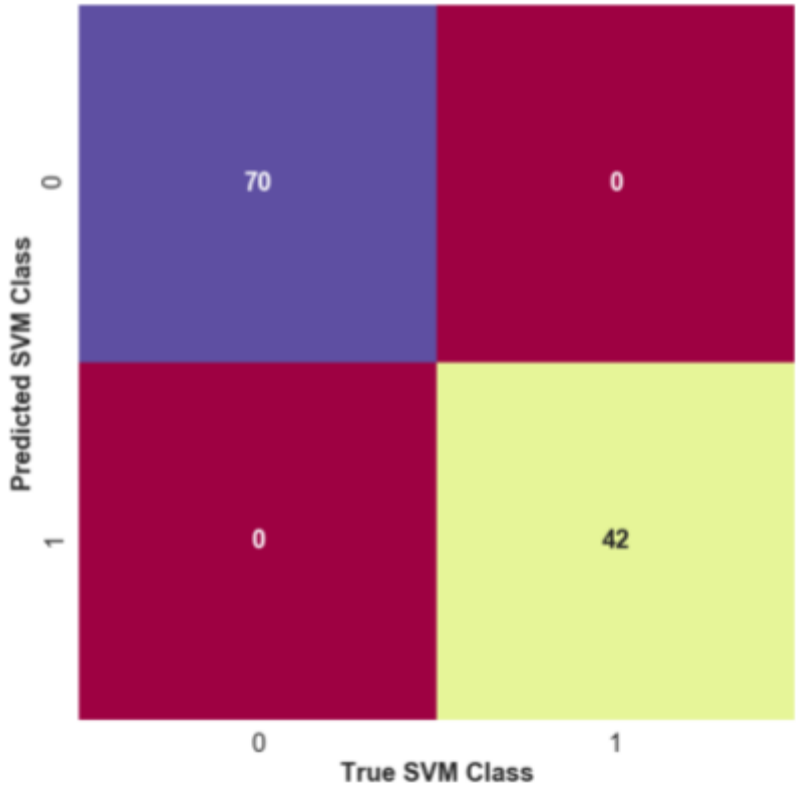

Figure 5: Comparison of True and Predicted SVM Class using SVM algorithm.

Modeling Problem Level instead of SVM class

Instead of using SVM Class in the previous study, Problem Level can be used to understand the classification in further granularity. The parameters used for the SVM analysis was the same as that of the 'SVM Class' analysis.

Figure 6 shows that when locations were analyzed using Problem Level as the output variable, there were significant number of misclassifications. It is interesting to note that these misclassifications were very localized. For example, there were 28 places that had problem level of 3 and were predicted as 2 and 12 places vice versa. Similarly 2 locations that had problem level of 0 was classified as 1 and 12 vice versa. In this model, a total of 54 locations were misclassified out of 112 locations.

For this model, the accuracy and f1 scores were 0.52 and 0.51 respectively. Given the nature of the classification of 'Problem Level', it is clear why 'SVM Class' had such a high accuracy. Further, it has to be noted the C parameter in the SVM model did not affect this prediction as well.

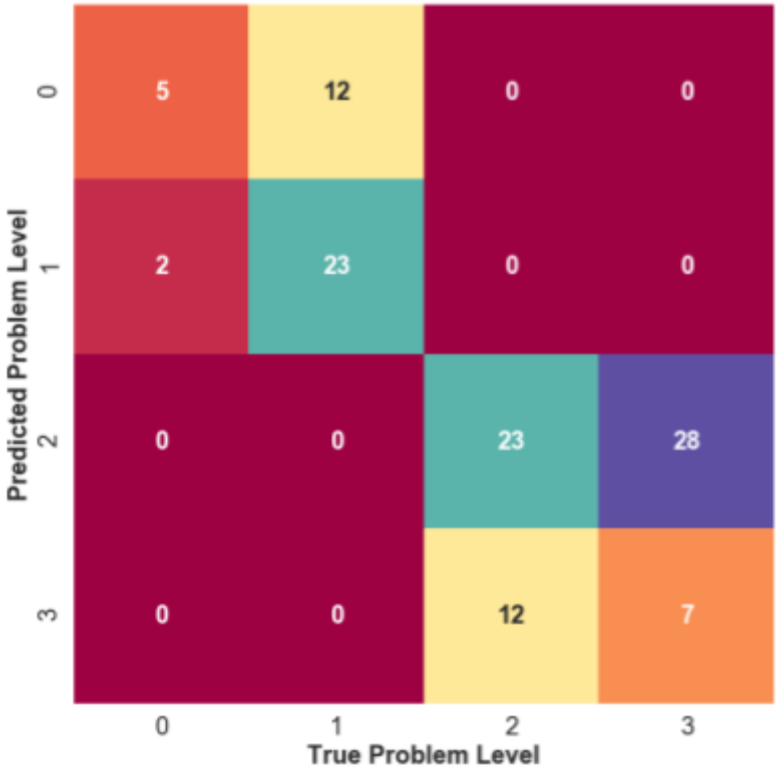

Figure 6: Confusion matrix for the predicted and true problem levels. There are significant number of misclassifications.

Discussions

From the data analysis thus far, a few interesting observations can be made. First, there are a combination of factors that make a place unhealthy. The number of columns in the dataframes containing venues category and their frequencies exceed 150 in each of the analysis. Therefore, it is hard to pinpoint one specific factor that can be the cause for unhealthy behaviors. In the obesity and all unhealthy behavior K-Mean clustering, a heuristic observation suggests that a place would be unhealthy if it contained Fast Food. However, there are places where Fast Food is the most common venue and that effect is offset by other places such as Gym and sports centers that the location is still classified as healthy. Most of the places that had low AAP values in general did not contain Fast Food in their top venues.

Interestingly, SVM analysis was able to classify places in general as healthy/unhealthy using SVM Class with very high accuracy. There were significant number of misclassifications when Problem Level was used. Even in this case the misclassifications were only within Problem Level 0 and 1, and Level 2 and 3. This shows that the algorithm will perform very well to classify whether a place is seriously problematic or not. From a business point of view, making this clear distinction in SVM Class is very important.

In these seriously affected places, opening new venues such as gyms, parks, fitness centers, or trails to the extent that fast food and eateries are not the most sought after venues can improve the location health significantly. These efforts can be both private or joint venture with the public sector. For the cities that were not included in the study, the SVM model will be able to provide a confident classification of the unhealthiness measure based on the venue details.

Conclusions and Future Work

The CDC data was downloaded and pre-processed to obtain dataframe with desired rows and columns. Foursquare API was used to obtain popular venues in a given location with category ID as input. K-Means clustering and Support Vector Machines were used to perform unsupervised and supervised learning of the dataset respectively. K-Means provided a good way to analyze existing data and the trained SVM model is a good tool to classify new data from new locations with very high confidence. From the studies conducted in this project, some of the top venues that will benefit from having fitness centers and/or nutrition/diet specialists are:

1. Detroit MI
2. Flint MI
3. Birmingham AL
4. Gary IN
5. Macon GA

Further studies can be conducted on this dataset to identify the sets of venues that cause a location to be classified as healthy/unhealthy. Dimensionality reduction can be used to obtain linear combinations of the features that classifies the location into one or the other category. There are also two other unhealthy behaviors namely binge drinking and chain smoking, and other categories such as prevention and health outcome in the original dataset that were out of the scope of this project. Analyzing these data will also provide some rich knowledge on ways to improve the overall health of the country using data science. It would be interesting to invoke the socio-economic factors for these locations as well to obtain more clarify on the unhealthy behavior.