Data Science Capstone Project:

Using CDC data to analyze unhealthy behavior based on data from 500 cities

Vikram Seshadri

Introduction to the problem

CDC collects health data

500 cities

Categories:

- Unhealthy behaviors
- Health Outcomes
- Prevention

Unhealthy Behaviors

Obesity

Lack of Sleep (< 7 hrs a day)

Lack of Physical Activity (LPA)

Analyzed

Binge drinking

Chain Smoking

Not Analyzed

Potential Causes of obesity, lack of sleep and LPA

Food habits (Fast Food vs. Healthy diet)

Lack of venues that provide physical activities

Lifestyle choices

Lack of recreational places

Business plan

Analyze the causes for unhealthy behavior

Observe the general trends in locations that are healthy

Build a model to find the healthiness measure of new locations

Build venues that improve healthy behaviors (e.g.) Gym, Trail, Healthy food places etc...

Dataframe pre-processing

Unnecessary columns were removed

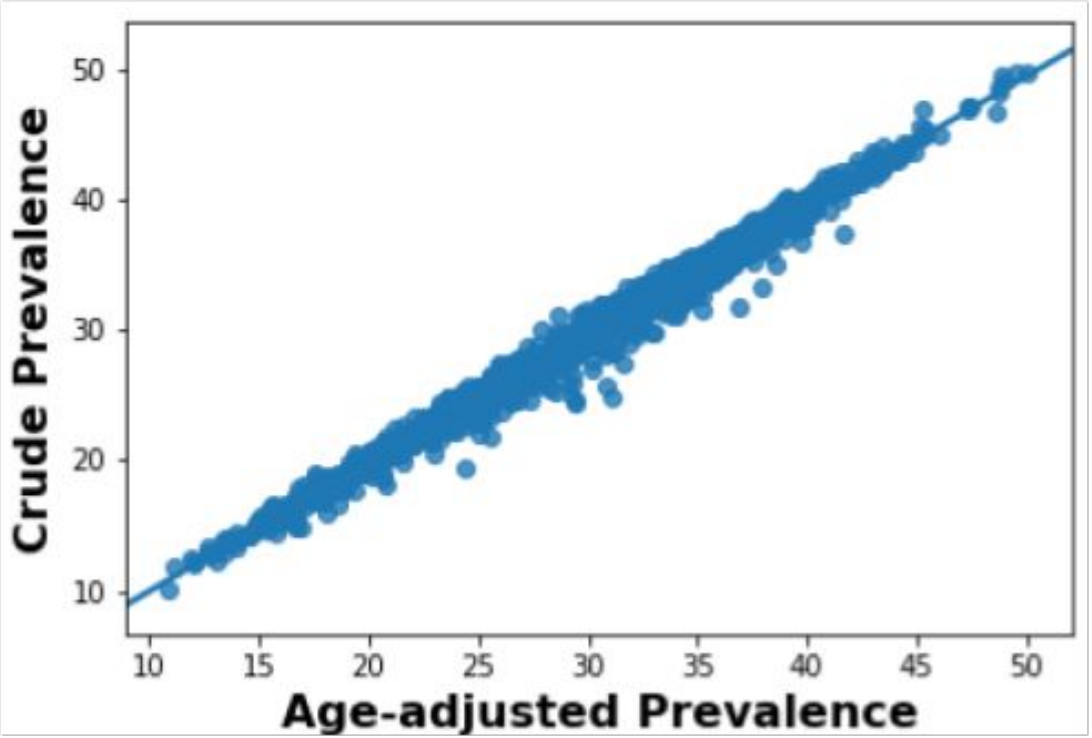
Columns with Latitude and Longitude based from Geolocation data

Consolidating Crude and Age-adjusted prevalence for every location.

Remove Prevention, Health Outcome data

Remove Binge drinking and Chain smoking data

Age-adjusted and Crude Prevalence are linearly correlated



Sufficient to use Age-adjusted prevalence (AAP) values for further analysis

Algorithms and Analyses

Venue information from Foursquare

Basic statistics with Scipy

Unsupervised Learning: K-Means Clustering

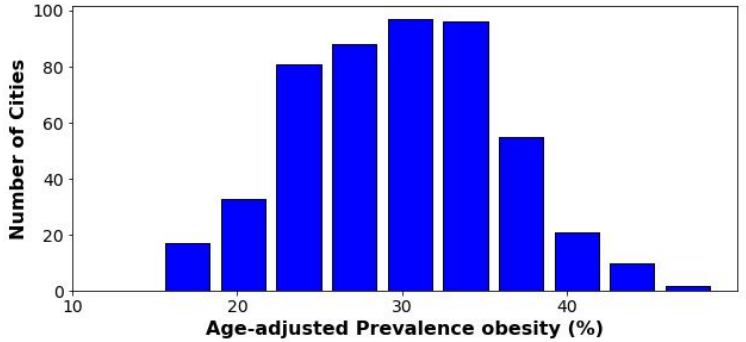
Supervised Learning: Support Vector Machine

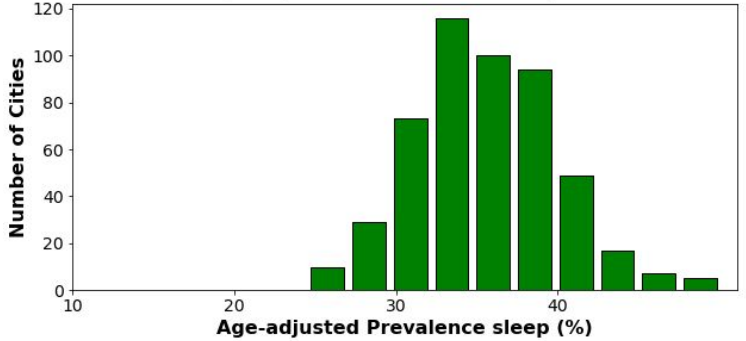
Folium, Matplotlib, and Seaborn for visualization

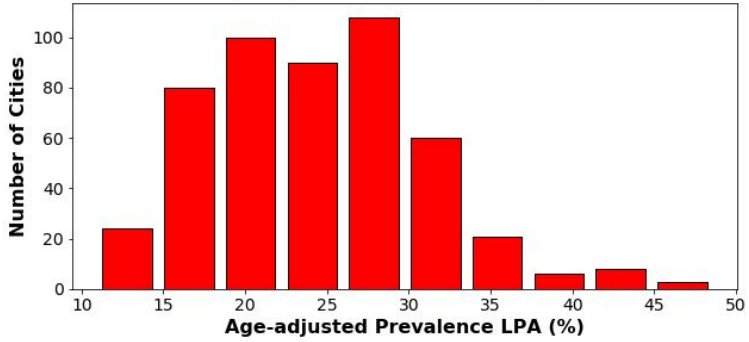
Histogram and Basic Statistics

	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

Lack of sleep very prevalent







Obesity analysis

	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

Location data with Foursquare provides venues!

```
obesity_venues.shape
```

```
(1380, 7)
```

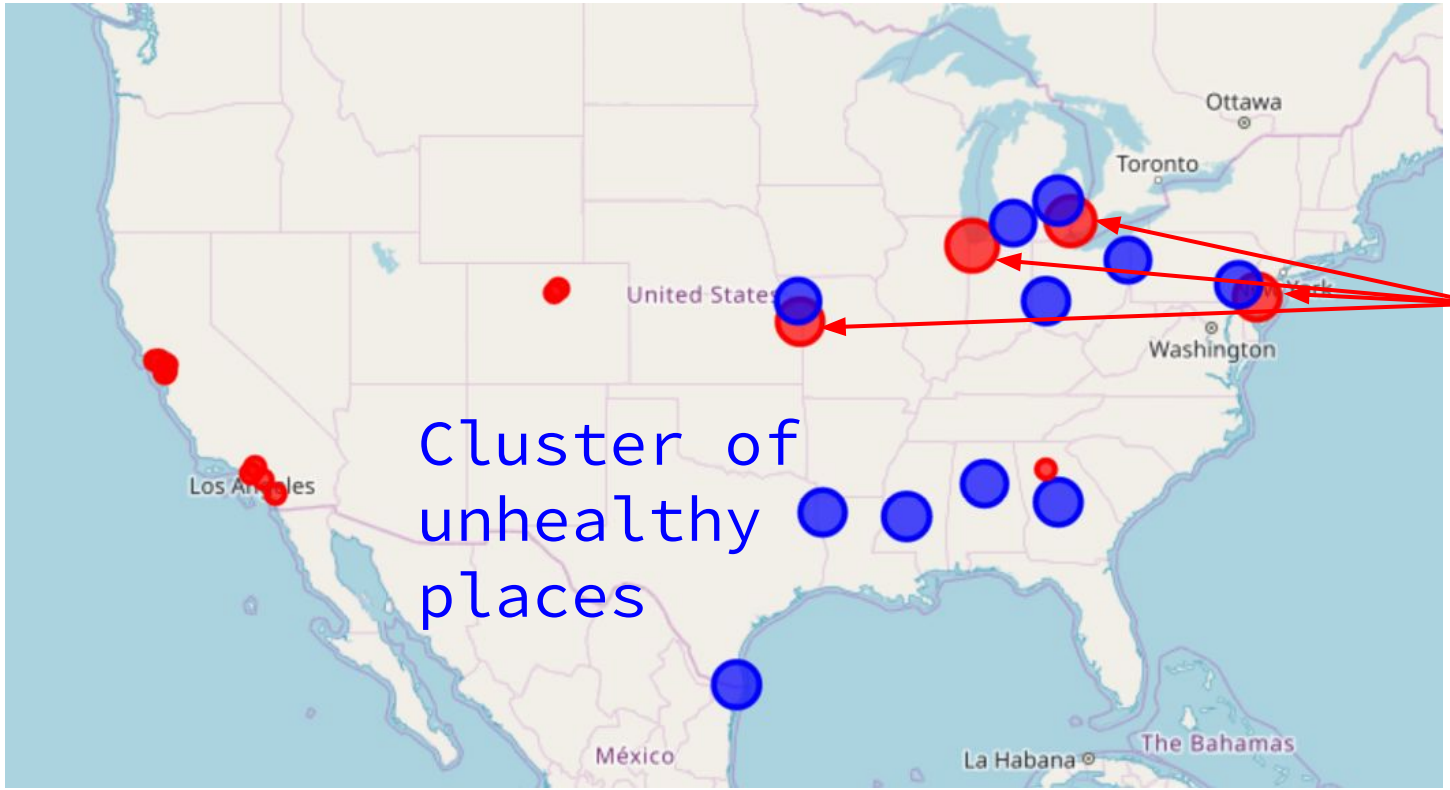
```
obesity_venues.head()
```

	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

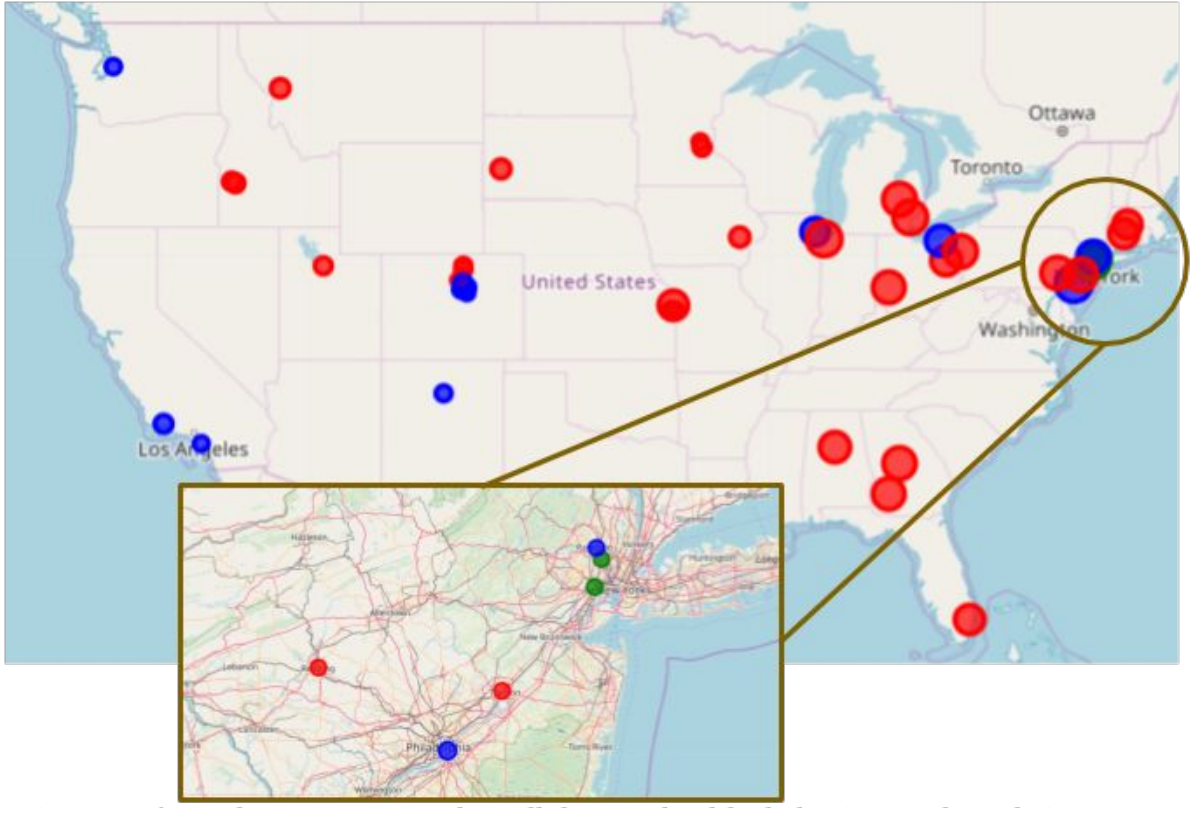
```
print('There are {} uniques categories.'.format(len(obesity_venues['Venue Category'].unique())))
```

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

K-Means on obesity data with two clusters



K-Means on all unhealthy behaviors with three clusters



Fast Food was typically observed in places that were unhealthy

Support Vector Machine on venue data

Problem Level = No. of. AAP values > 30.0

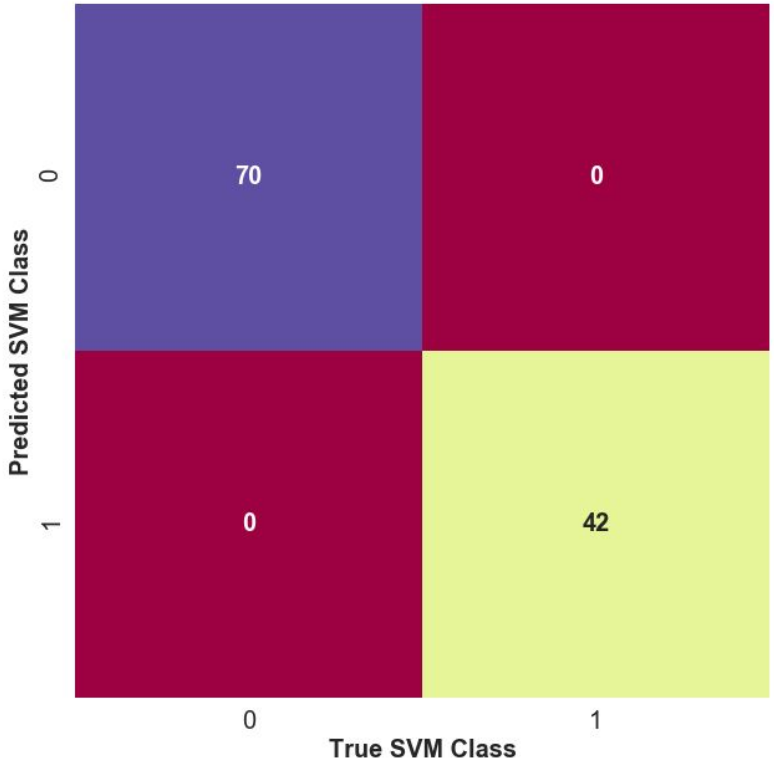
SVM Class = 0 if Problem Level >= 2

= 1 if Problem Level < 2

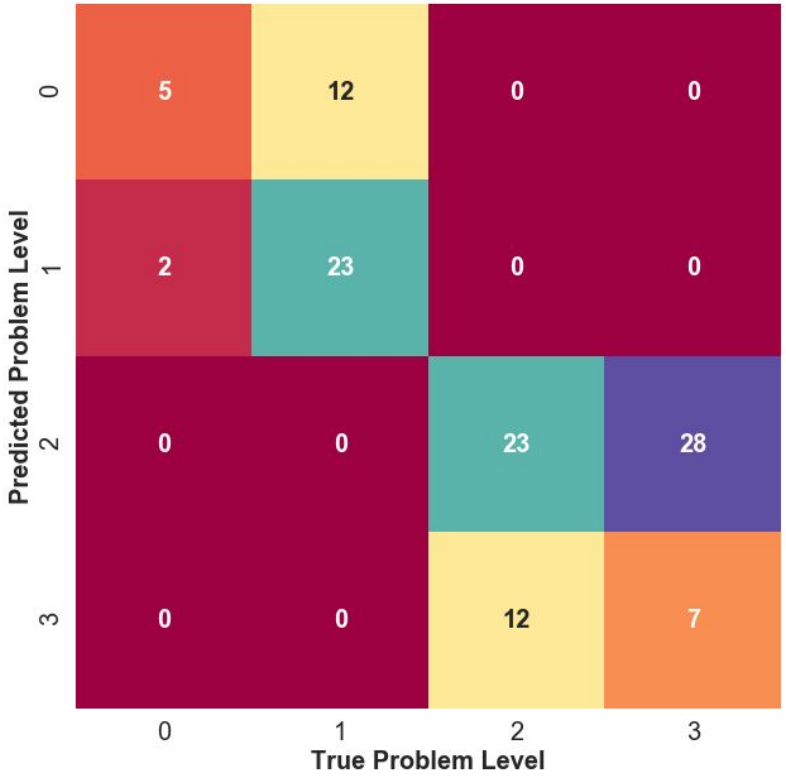
Kernel = Radial Basis Function

C = 1, Gamma = ‘auto’

SVM on ‘SVM Class’: Very high precision and recall



SVM on ‘Problem Level’



Misclassification between 0 and 1 & between 2 and 3

Inferences

Locations clustered using K-Means provided reasonable clustering.

SVM classification on ‘SVM Class’ very accurate.

SVM classification on ‘Problem Level’ less accurate, but still very useful.

Inferences

New location not in the 500 cities can be classified using SVM model and venues data.

For Unhealthy locations,

- Reduction in Fast Food and unhealthy food
- Gym, Park, sports, and fitness centers

Conclusions

Example locations that can improve with Gyms and Food quality

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

Opportunity to setup new businesses here!!!