Capstone 1 – Milestone Report

By Richard Wolff – 2018-02-21

Obesity is at epidemic levels in the United States and it doesn't not show signs of slowing. Over 1/3rd of adults are obese while another third is overweight.[1]This is a big problem. Obesity is heavily associated with a host of other problems including heart disease, stroke type 2 diabetes, and some types of cancer. Some of these are the leading cause of preventable death.[2]

In the ongoing quest to better understand this epidemic, my research is aimed to inform health professionals and those who are just curious about what could be at the root cause of the increasing obesity rate.

The Data

In my search for information surrounding obesity, I came across the 500 cities project which aggregated health data on the city and census tract level. This was conducted by the CDC Foundation and the Robert Wood Johnson Foundation with the goal to provide small area estimates for chronic disease risk factors, health outcomes and clinical preventative service for the largest 500 cities in the US.

My use for the data will be to compare the percentages of different health outcomes, chronic disease risk factors, and preventative actions across all 27000 census tracts to predict the Obesity rate.

Data Definitions

The set of 500 city health data can be broken out into dimensions and metrics. Under dimensions, there are two types of information.

1.) Location information such as city, state, census fips info, etc.
2.) Metric definitions: Category of metric (health outcome or preventative activity), definition of metric, and other items to better understand what's being analyzed.

What's in the data?

When looking at just the metrics, we see there are 28200 entries, most are census tracts but some are also aggregated cities, as well as a line item for the US as a whole.

```
Int64Index: 28200 entries, 0 to 28999
Data columns (total 31 columns):
```

```
Census Tract    27198
City             1000
US                  2
Name: geographiclevel, dtype: int64
```

We also need to contend with there are two types measured values, one for age adjusted prevalence and crude prevalence. This age adjusted value only happens at the city and US level of data. By working with the census level data completely, we do not have to filter one of these out.

Cleaning the dataset

While cleaning the dataset, I wanted to check if there were any missing values. By using .describe() function on a pandas dataframe of our data, we are able to see all the columns that may have less than the original 28200

---

[1] Taken from National Instiute of Diabetes and Digestive and Kidney Diseases:
https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity

[2] According to the CDC:
https://www.cdc.gov/obesity/data/adult.html

rows of data at the. By chaining .sort_values('count',axis1) to the end of the .describe() statement, we bring the lowest count values to the front of the view.

Sure enough, there were 7 out of 29 measure columns with varying amounts of missing data.

	COREW	COREM	TEETHLOST	MAMMOUSE	COLON_SCREEN	PAPTEST	ACCESS2
count	28091.000000	28097.000000	28140.000000	28163.000000	28181.000000	28193.000000	28199.000000

To decipher where this missing data was located, I filtered by null values and printed the value counts by geographic level (city, census, or us). All missing data were indeed coming from census tract level information.

```
Census Tract    336
Name: geographiclevel, dtype: int64
```

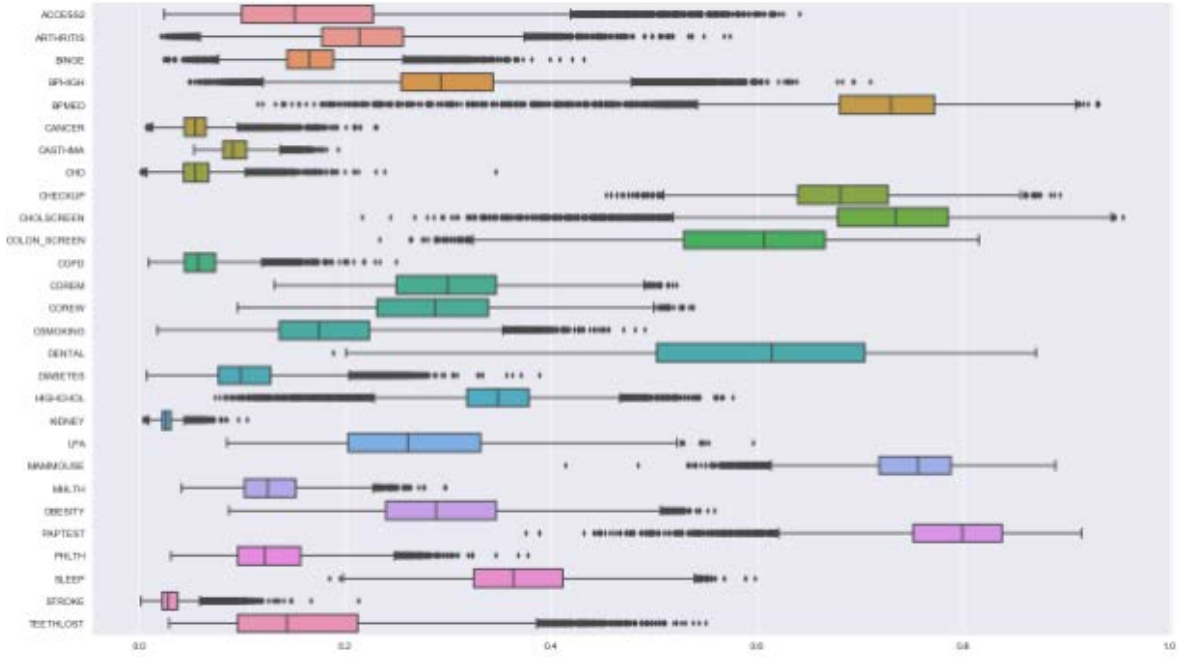
Next, I wanted to see if there was any census tract that was missing enough data to disregard that geography all together. It seems the most data that was missing for any one census tract was 21% of its data. With that said, I decided just to fill all the missing data vs disregard the entire tract.

uniqueid	stateabbr	cityname	geographiclevel	tractfips	cityfips	statedesc	latitude	longitude	numMissing	% Missing
0150000-01097003605	AL	Mobile	Census Tract	1.097004e+09	150000.0	Alabama	-88.182347	30.695163	5	0.172414
0658072-06037402404	CA	Pomona	Census Tract	6.037402e+09	658072.0	California	-117.809722	34.059742	5	0.172414
0669196-06083002013	CA	Santa Maria	Census Tract	6.083002e+09	669196.0	California	-120.449959	34.881549	6	0.206897

When filling in the missing data, I grouped by the city level and took the simple average of all the census tracts in that city. I looped through each column and performed this transformation.

```
cleanedData.groupby('cityname')[col].transform(lambda x: x.fillna(np.mean(x)))
```

My final step in cleaning this data set was to see if there were any outliers. I plotted boxplots on all the data.



There are many data points for each metric that would could be labeled as an outlier but because there are so many, this probably means that the data is correct and it's just a function of the population in individual census tracts. We should include all of those outlier data points in our analysis.

Other Potential Data Sets

There are other potential datasets that could provide more meaningful analysis of the 500 city dataset. I'd be interested in pulling the following type of information in:

1.) Income by census tract
2.) Demographics by census tract
3.) Employment rate by census tract
4.) Rate of population who attends a workout facility
5.) Rate of population who visits with a nutritionist
6.) Walkability ratings by census tract

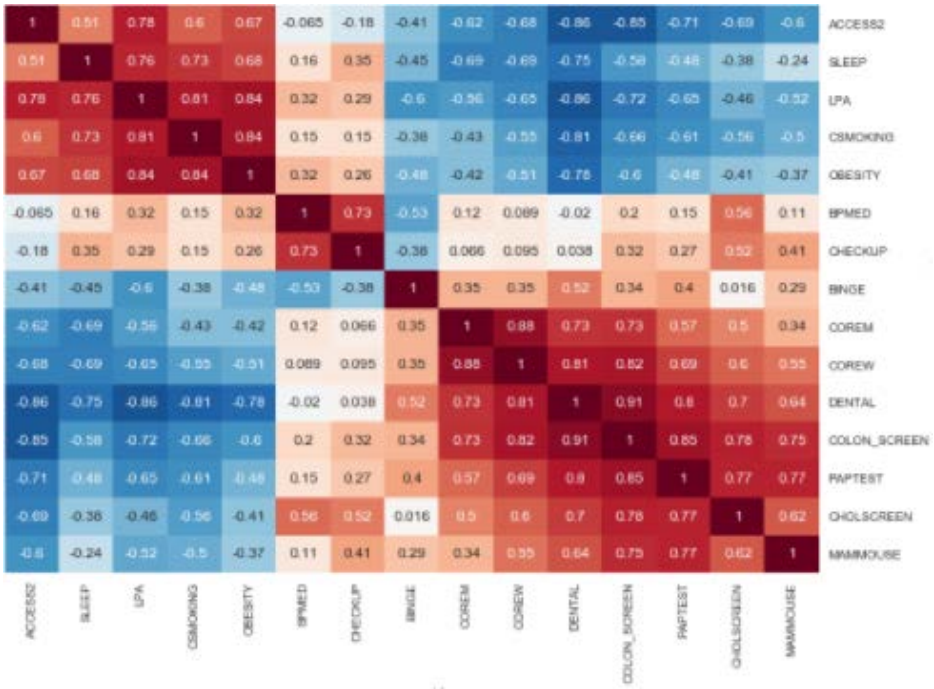
By including some of this information, we may be able to build a better prediction of obesity rate. For example, there is information that shows the obesity rate can range from 11.7% to 48.1% depending on race and 32.3% to 40.2% depending on age.[3]

Another example would be to include income and employment rate data. Does having a higher income lead to better or worse obesity outcomes? What about the employment rate?

These data sets should be kept in mind for any future analysis.

Initial Findings

Using a seaborn clustermap of correlations, I'm able to quickly find what may be negatively or positively related to obesity. We start seeing some high correlations. Being a current smoker, low physical activity, and lack of sleep correlate with a higher rate of obesity while getting dental exams and colon screenings are inversely correlated with high obesity.



There are other correlations not pictured here. Interesting, highly correlated metrics include having diabetes, COPD, missing teeth, high blood pressure, and generally reported poor health. It's important to remember that while the math works for these correlations, contextually we know some of them may be the result of being obese and not a cause of it.

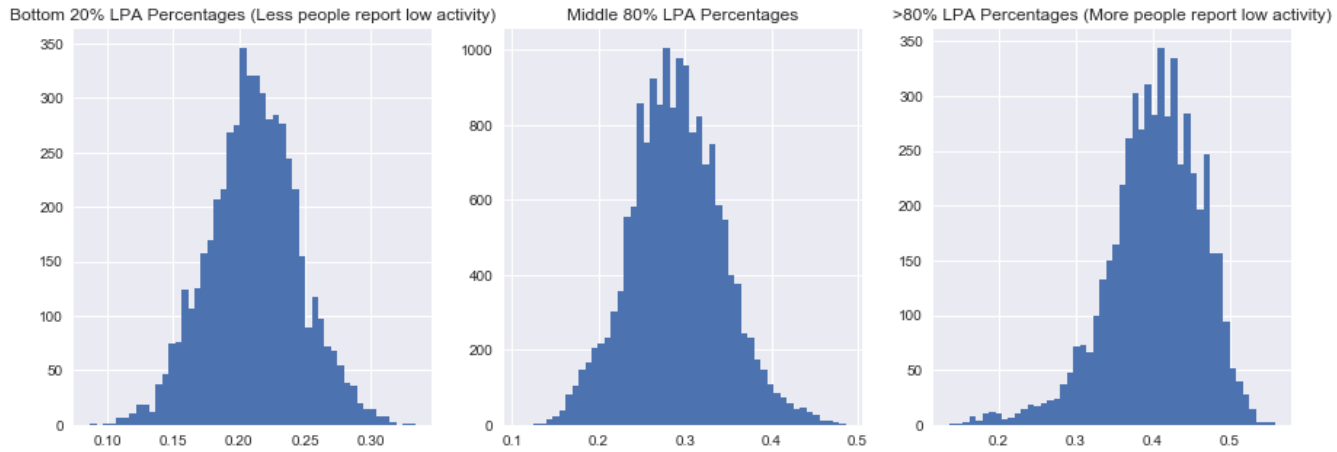
[3] CDC Obesity Info: https://www.cdc.gov/obesity/data/adult.html

Looking at the histogram of obesity rates, we see an almost normal distribution but with a slight skew to the right. But we see a simple mean obesity rate of about 28%.

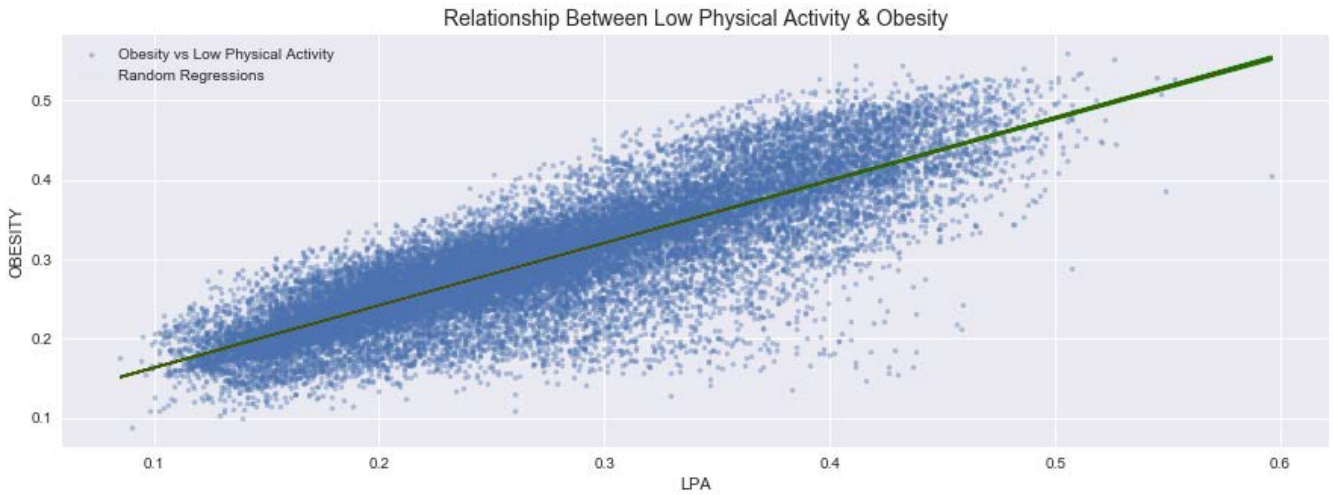
We can start looking at obesity rates in correlation to different metrics. I looked at low physical activity, low sleep, population and geography may affect the obesity rate.

Low Physical Activity

When separating low physical activity rates by the bottom 20%, top 20%, and middle 60% we start seeing some very clear trends appearing. For those reporting the lowest rates of low physical activity, the mean obesity rate is centered on 20%. The middle 60% sees their obesity rate centered around 30% while the top 20% reporting low physical activity obesity rate centers around 40% (with a slight left skew).



Next, we take a look at the correlation between the two metrics. We find a strong correlation of 0.841 and a P-value of close to 0.

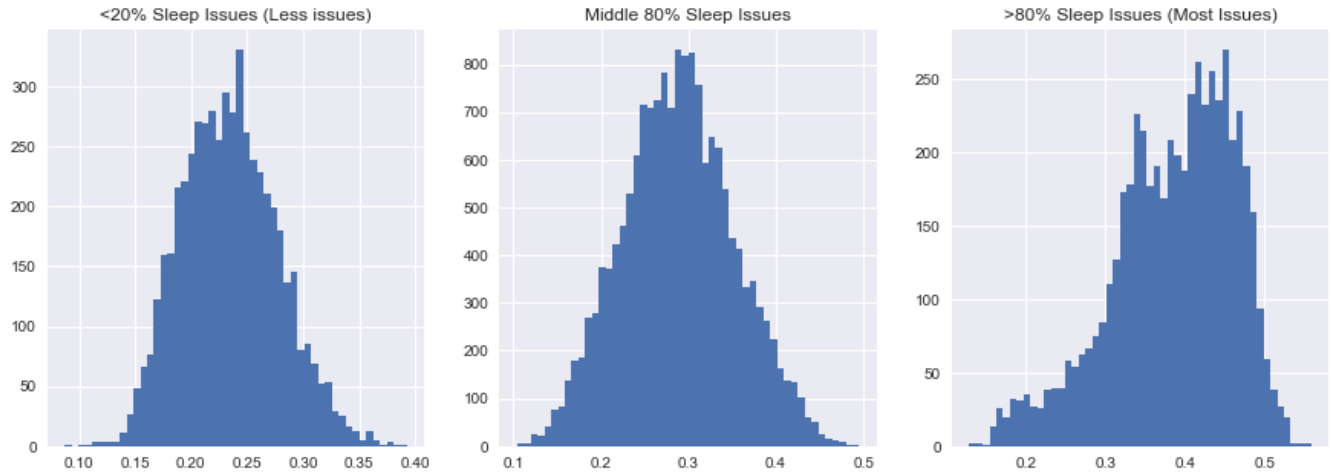



```
Pearson Correlation: 0.841
P-Value: 0.00000
```

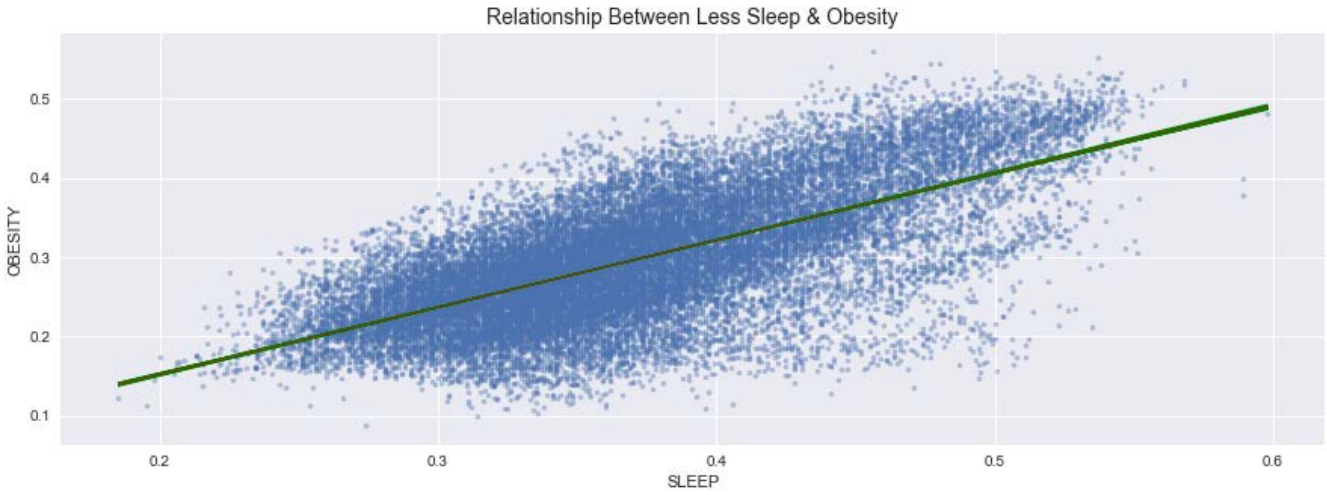
Lack of Sleep and Its Relationship with Obesity

We ran the same type of analysis on Lack of sleep and Obesity as we did with Low Physical Activity and found strikingly similar outcomes. Populations with the higher percent of less sleep tend to have a higher levels of obesity while those populations with and low percent of sleep issues have a lower obesity rate.

It's important to note that lack of sleep may not be causing obesity but in fact causing a low physical activity rate due to its population feeling more tired.



Looking at the correlation between lack of sleep and obesity, we see a moderate correlation of .682 with a P-value of close to 0.

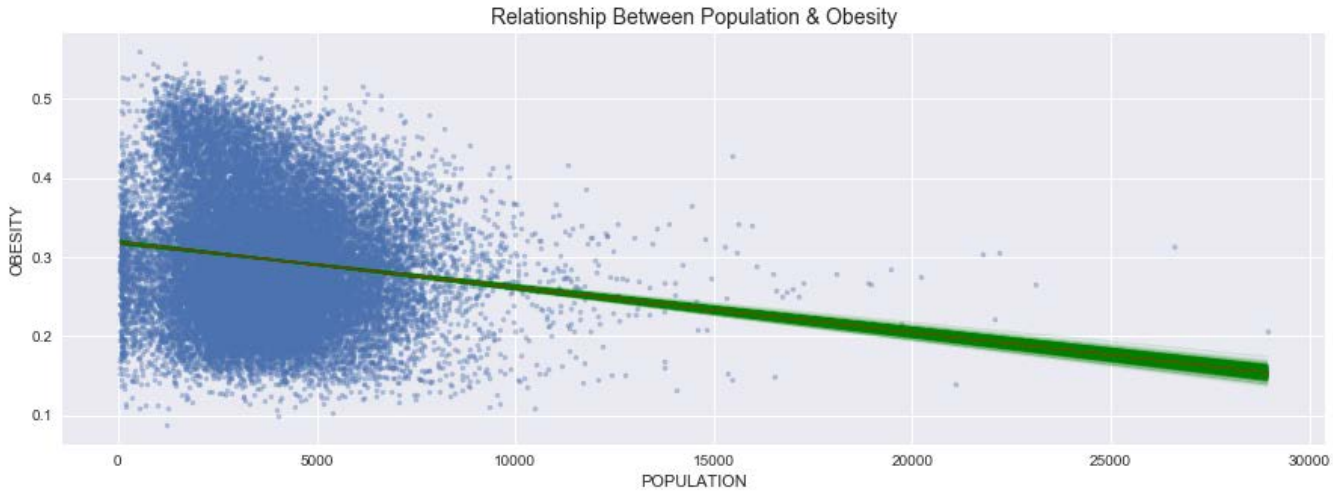



```
Pearson Correlation: 0.682
P-Value: 0.00000
```

Population and Its Relationship with Obesity

I wanted to look at whether or not a county's population could have an effect on Obesity. The thought is that those with lower populations would have lesser access to fresh foods, preventative healthcare, and other deterrents to obesity that you would have from a larger population (think walkability of a population that is over populated).

What we found is there is generally no correlation at a coefficient of -0.139 and a P-value of 0.



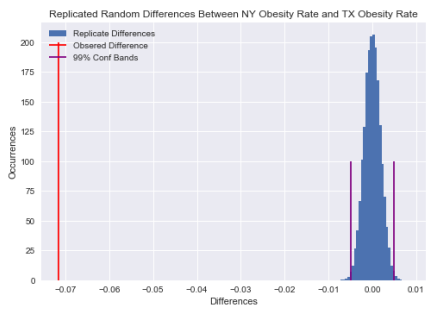

```
Pearson Correlation: -0.139
P-Value: 0.00000
```

Geography and Its Relationship with Obesity

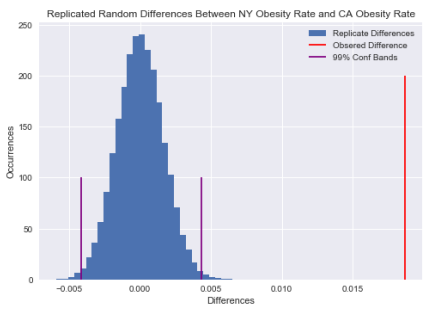
Finally, I wanted to see if geography could indicate a higher obesity rate. There are many anecdotes that populations from the south are more obese than populations from the north. If this is true then we should see differences in the mean obesity rates by state.

For each geography test, I stated the null hypothesis that there is no difference between states, shifted the obesity rates to the mean obesity rate for both states combined, and took random samples of the data to test the differences of means. I tested NY state vs CT, TX, and CA.

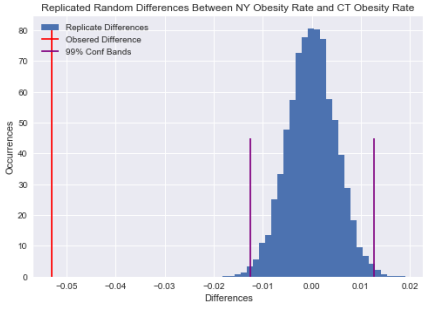
In all tested cases we can reject the null hypothesis that there is no difference in obesity rates between states. There is a very observable difference between NY and TX with TX having 710bps higher mean obesity than NY. In the case of NY and CA, NY mean obesity rate is 190BP higher than CA. Finally, In the case of NY vs CT, CT's mean obesity rate is 530BP higher than that of NY'S.




```
Observed Difference: -0.071
99% Conf Intervals: (-0.005, 0.005)
P Value: 0.000
```




```
Observed Difference: 0.019
99% Conf Intervals: (-0.004, 0.004)
P Value: 0.000
```




```
Observed Difference: -0.053
99% Conf Intervals: (-0.012, 0.013)
P Value: 0.000
```

Summary

So far in capstone 1 I've gone through a few steps that leads me into the machine learning of the course and building a model to predict the obesity rate. Steps we've taken include,

- Defining what the data
- Cleaning the data
- Identifying other possible data sources to help predict obesity
- Looked at initial correlations and hypothesis between obesity and other metrics

All of this has made pointed to Geography, Low Physical Activity, lack of sleep as potential indicators of obesity. As a final two steps, I'll build a prediction model and finally identify which items may indicate what can be causing obesity.