Capstone 1 – Final Report

By Richard Wolff – 2018-04-18

Obesity is at epidemic levels in the United States and it doesn't 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 27,000 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 the .describe() function on a pandas dataframe of our data, we are able to see all the columns that may have less than the original

---

[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

28200 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

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

information.

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
7.) Calories consumed per capita by census tract

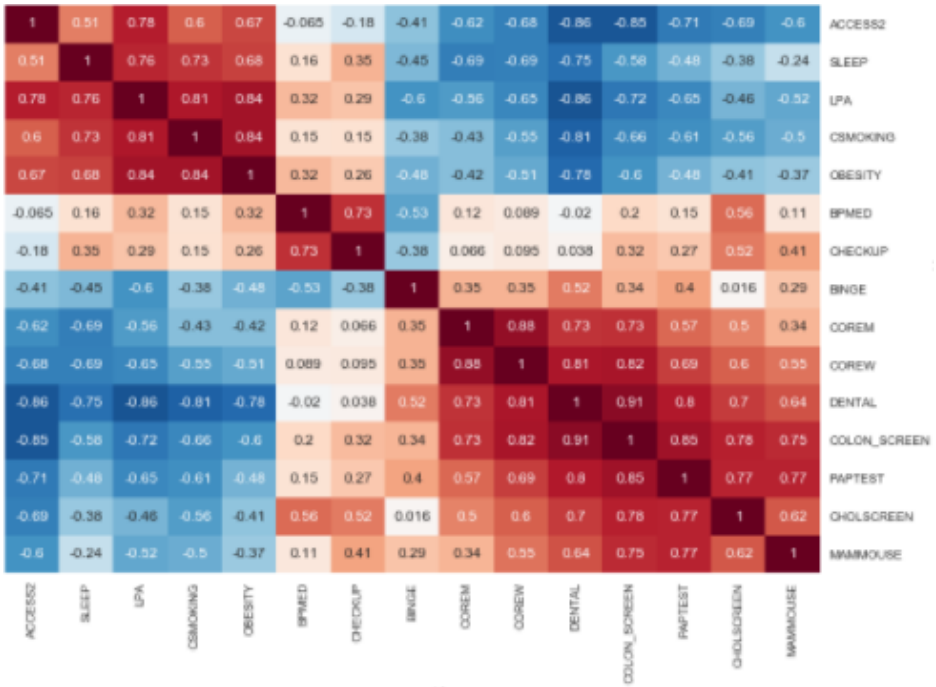
By including some of this information, we may be able to build a better prediction of obesity rate. For example, there is information that shows the obesity rate can range from 11.7% to 48.1% depending on race and 32.3% to 40.2% depending on age.[3]

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.

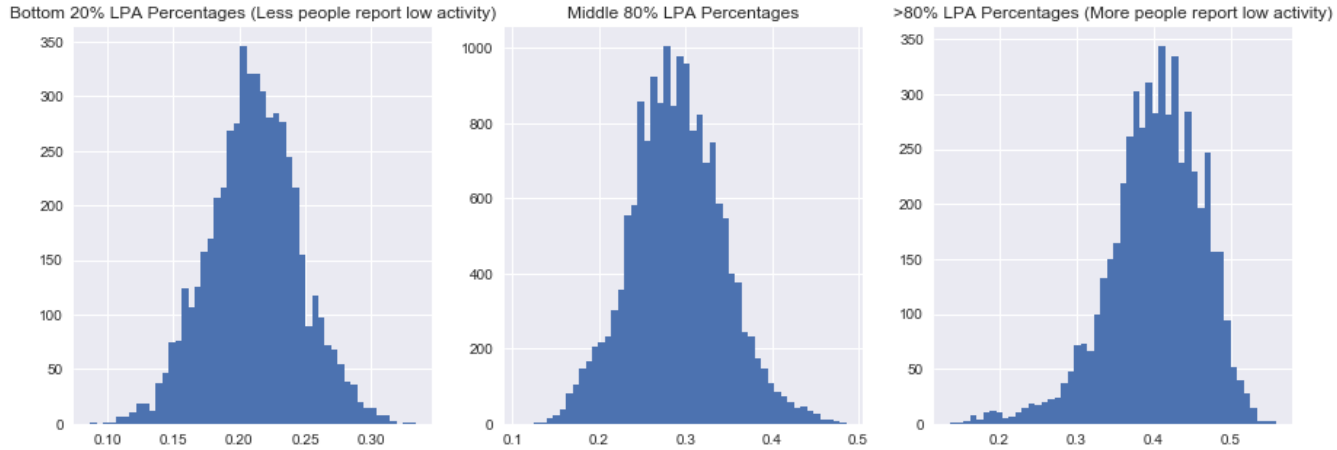
[3] CDC Obesity Info: https://www.cdc.gov/obesity/data/adult.html

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%.

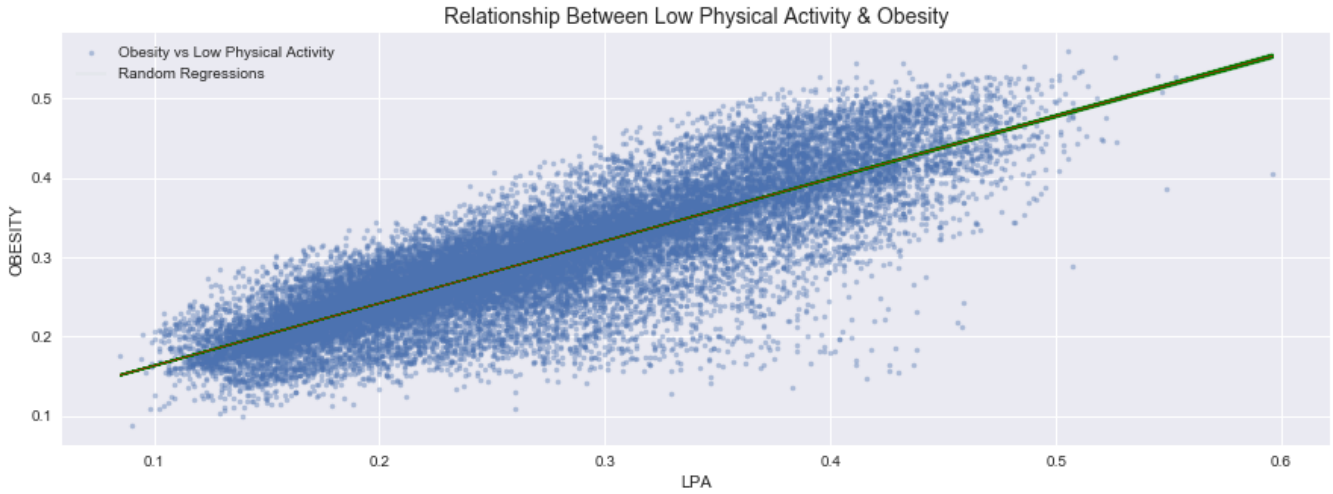
We can start looking at obesity rates in correlation to different metrics. I looked at low physical activity, low sleep, population and geography may affect the obesity rate.

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.

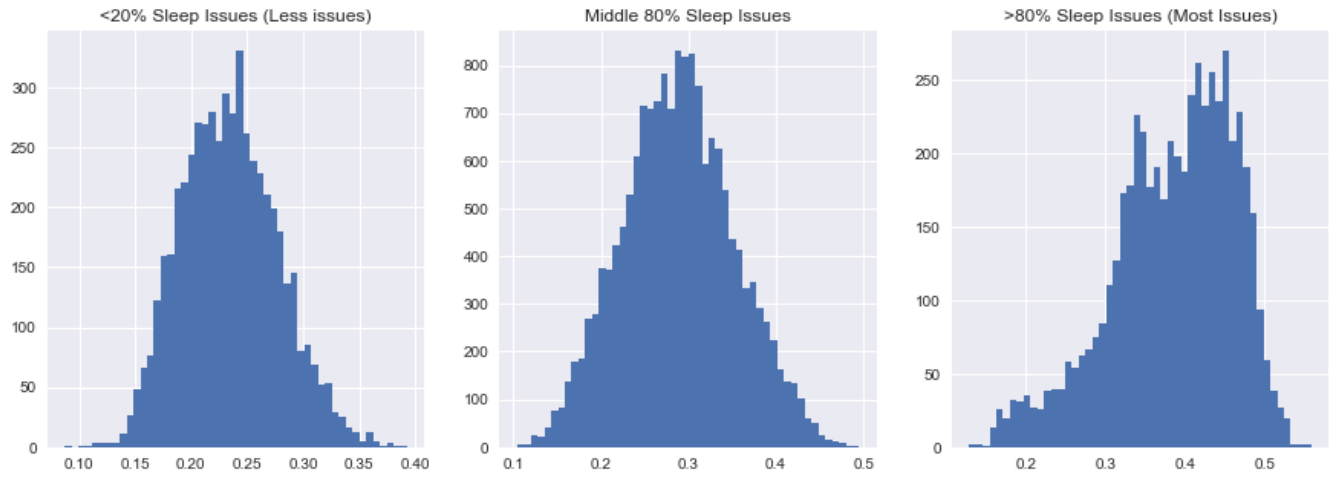



```
Pearson Correlation: 0.841
P-Value: 0.00000
```

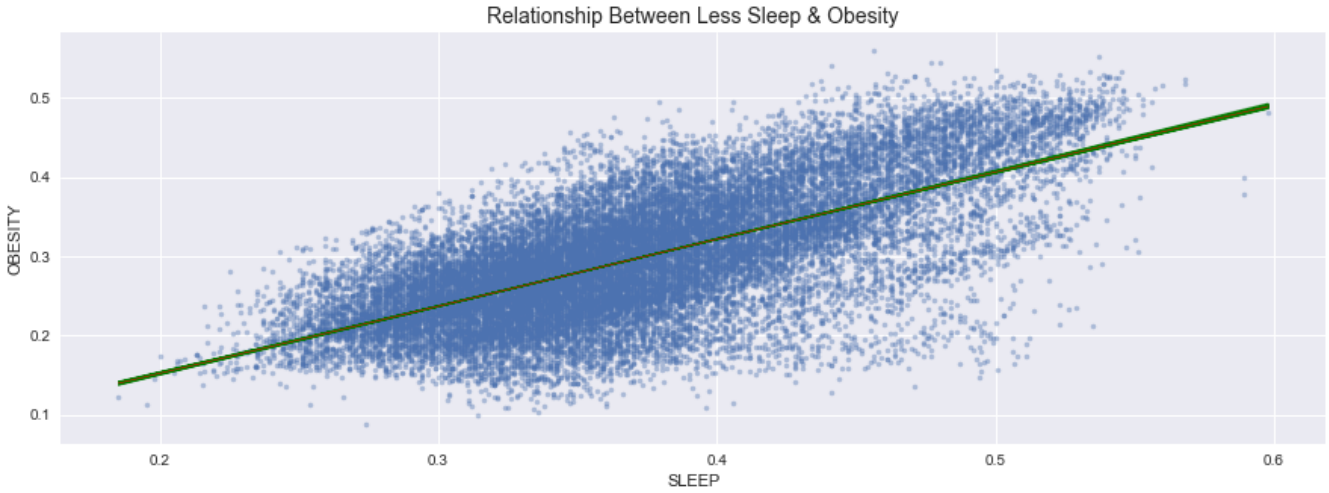
Lack of Sleep and Its Relationship with Obesity

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.

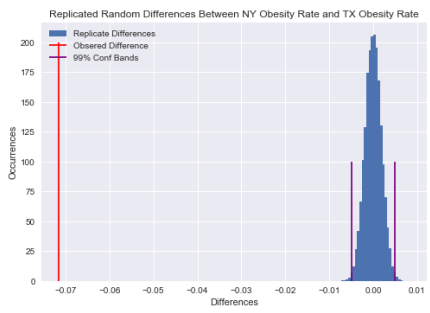

```
Pearson Correlation: -0.139
P-Value: 0.00000
```

Geography and Its Relationship with Obesity

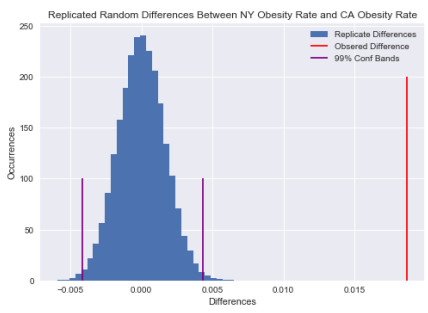
Finally, I wanted to see if geography could indicate a higher obesity rate. There are many anecdotes that populations from the south are more obese than populations from the north. If this is true then we should see differences in the mean obesity rates by state.

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.

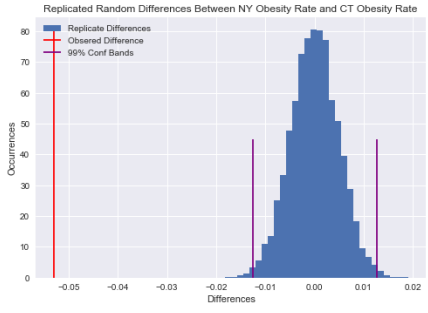
In all tested cases we can reject the null hypothesis that there is no difference in obesity rates between states. There is a very observable difference between NY and TX with TX having 710bps higher mean obesity than NY. In the case of NY and CA, NY mean obesity rate is 190BP higher than CA. Finally,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

Applying Data Science Models

Next, I applied different data science algorithms that are commonly used today. Specifically, I'll use Linear Regression and its related Lasso/Ridge methods. To add in more of a black box model, I'll apply a Random Forests model as well to see if there we can increase our models accuracy.

Before I applied any model, I needed to make sure the data was ready to be fed into the different algorithms that exists in the sklearn and scipy packages.

Features That Do Not Cause Obesity

First, we know that there are a few features that have are known, according to the CDC[24], to be effects of Obesity and not causes. They include

- CHD: Population reported having Coronary heart disease
- BPHIGH: Population reported having High blood pressure
- BPMED: Population reporting being on blood medicaiton
- ARTHRITIS: Population reported having arthritis
- DIABETES: Population reported having diabetes
- HIGHCHOL: Population reporting having high cholesterol
- PHLTH: Population reporting having poor health
- KIDNEY: Population reporting having Kidney disease
- STROKE: Population reporting having a stroke

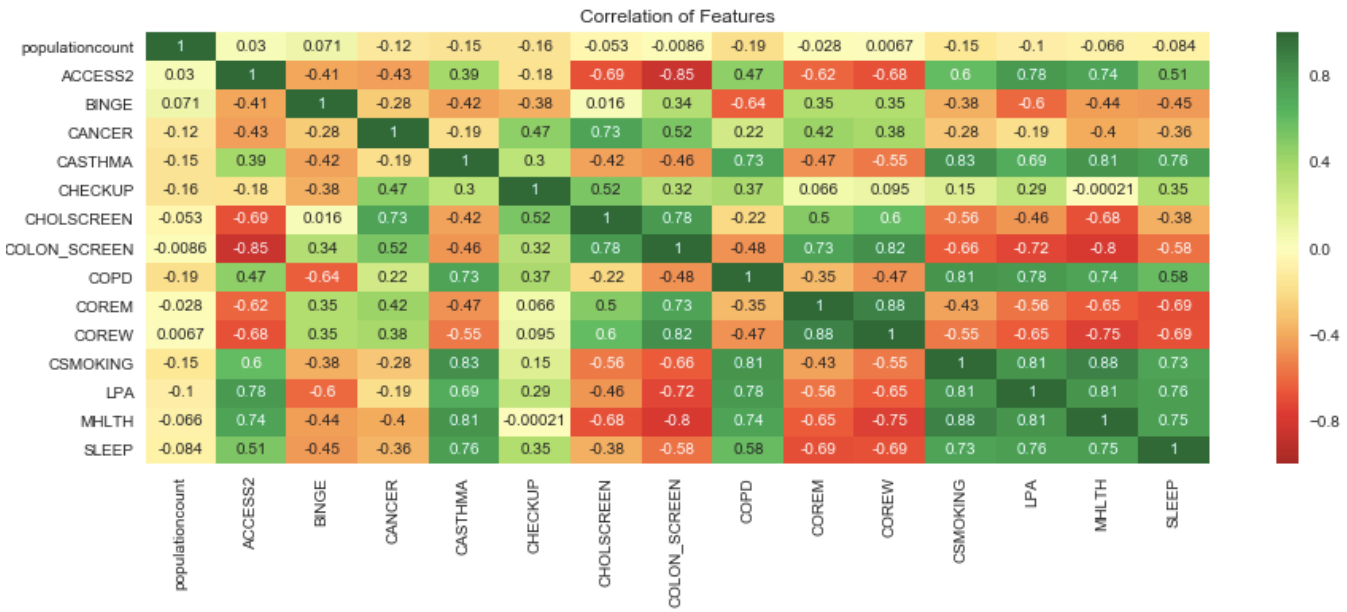
There are a few features that have should have no effects when it comes to causing a population to have higher incidences of Obesity. They include:

- MAMMOUSE: Mammography use of women aged 50 - 74
- PAPTEST: Papanicolaou smear use among adult women aged 21 - 65 Years
- BPMED: Taking bp med, I dropped this here because it's highly correlated with BPHIGH.
- TEETHLOST: All teeth lost among adults aged >=65 Years
- DENTAL: Visits to dentist or dental clinic among adults aged >=18 Years

All of these features were dropped from our analysis.

Correlation of Features

Next, I wanted to reduce the features that have high correlations with one another. Using a heat map with a correlation matrix, we're able to identify the features that have high correlation.



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

There are 4 highly positive correlated features. They Include:

- Mental health & current smoker (.88)
- Mental health & asthma (.81)
- COREM & COREW (.88)
- Colon_Screen & ACCESS2 (-.85)
- COPD & SMOKING (.81)

From this, I've dropped MHLTH (mental health), COREM (% of men 65+ who received core preventative services), Colon Screening, and COPD from the feature set.

The features we're left with include Access2 (% of population with current lack of health insurance,) BINGE (% of adults who reported binge drinking), CANCER (% of population reported having cancer), CASTHMA (% of population having asthma), CHECKUP (Visits to doctor for routine checkup within the past Year), CHOLSCREEN (Cholesterol screening among adults aged >=18 Years), COREW (% of older women up to date on preventative services), CSMOKING (% of population who currently smokes), LPA (% of population reporting low physical activity), and SLEEP (% of population reporting <7 hours of sleep daily).

Regression Modeling

After pruning the feature set to a handful of items that may be linked to a higher obesity rate in a given population, I performed my initial linear regression using the statsmodels OLS feature. I chose this first as it gives a great summary of the model and may helped guide my thinking through this process.

First, I set my features to a variable X and set my target to a variable y. I then split these variables into a 75% training set and a 25% test set to perform final model testing to be sure our model has a good general fit.

Training our model

Initial statsmodels OLS Regression

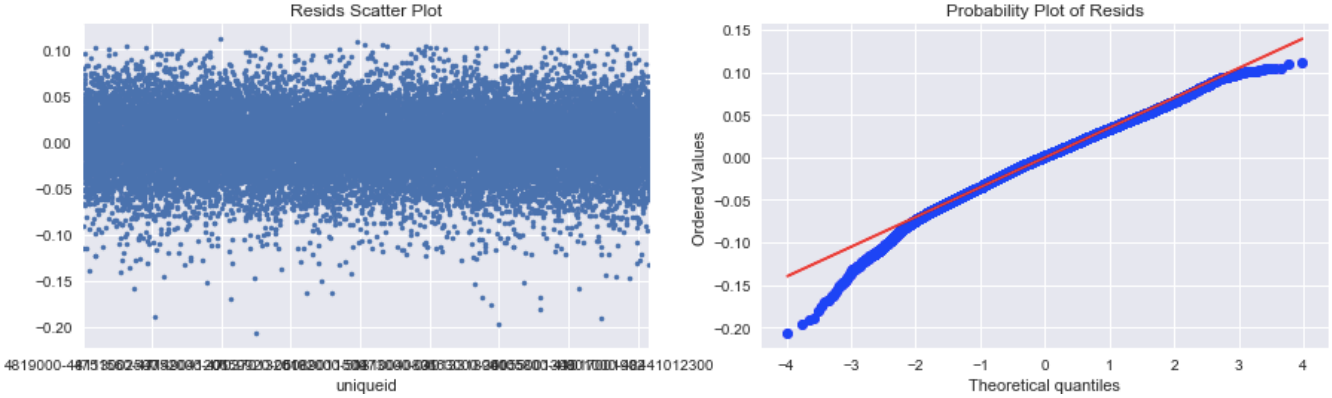
OLS Regression Results

Dep. Variable:	OBESITY	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	1.401e+05
Date:	Tue, 17 Apr 2018	Prob (F-statistic):	0.00
Time:	14:58:06	Log-Likelihood:	39366.
No. Observations:	20398	AIC:	-7.871e+04
Df Residuals:	20387	BIC:	-7.862e+04
Df Model:	11		
Covariance Type:	nonrobust		

On my first attempt, the regression returned a r^2 of .987 which signifies a pretty good fit for this model. I thought it may be over fit at this point, but I wanted to keep going to see where these results may lead.

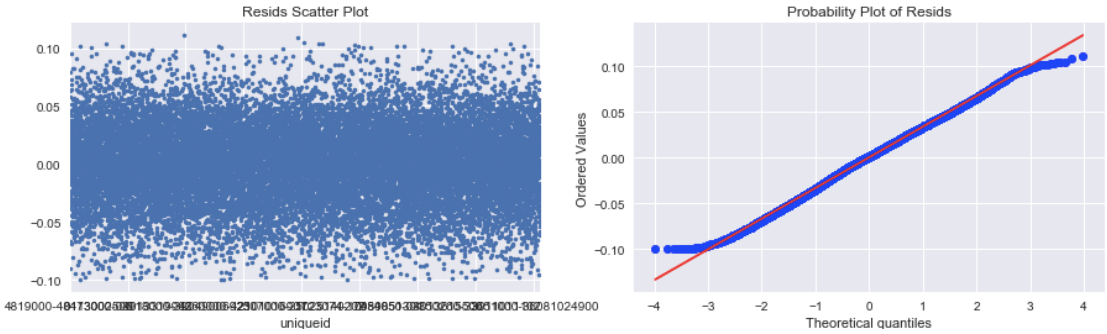
Next, I created a scatter and probability plot of the residuals. It seems show that there may be some outliers on the lower end of the residuals, at about -.10.

Residual Scatter and Probability plot after the first Linear Regression Model



Indeed, after I discarded the data with residuals of < -.10, the plots seemed to read as much more normal residual plots.

Residual Scatter and Probability plot after discarding outlier residuals.

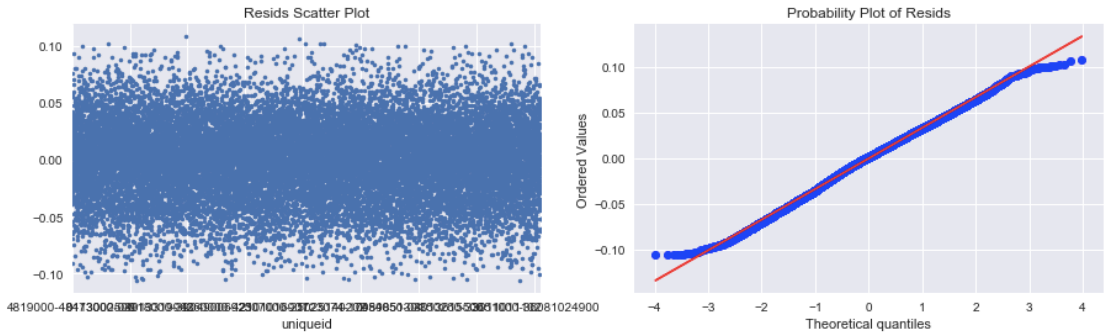


Statsmodels results after removing outliers

```
AIC Change: -1336.6065
                            OLS Regression Results
==============================================================================
Dep. Variable:                 OBESITY   R-squared:                       0.988
Model:                             OLS   Adj. R-squared:                  0.988
Method:                  Least Squares   F-statistic:                 1.539e+05
Date:                 Tue, 17 Apr 2018   Prob (F-statistic):               0.00
Time:                         14:58:09   Log-Likelihood:                 40035.
No. Observations:               20242   AIC:                         -8.005e+04
Df Residuals:                   20231   BIC:                         -7.996e+04
Df Model:                          11
Covariance Type:            nonrobust
```

Rerunning the statsmodels OLS, we saw an AIC change of -1337, a r^2 increase from .987 to .988 and a more normal residuals plot.

Residual Scatter and Probability plot after updating the model.



Coefficients of our model

	low	high
CASTHMA	0.641193275	0.751992613
CSMOKING	0.504944768	0.540993787
LPA	0.285464308	0.320737523
ACCESS2	0.160539792	0.179349859
CHOLSCREEN	0.101238641	0.127024479
COREW	0.034027081	0.056435883
populationcount	-0.000001622	-0.000001139
CHECKUP	-0.007213327	0.023056344
SLEEP	-0.124738157	-0.088470267
BINGE	-0.142903353	-0.113291311
CANCER	-0.247184246	-0.142176431

Deciphering the Results of The Model

Our model's coefficients include two strong predictors of a high obesity rate in a given census. They include populations with a high rate of asthma and a higher rate of smokers. My theory is that if a population has conditions where it may be harder to perform aerobic activities they may see a higher obesity rate.

It's not a stretch to entertain that those who smoke and those who have asthma may have a harder time breathing while performing physical activity. At the same time, the third highest predictor of a population with a high obesity rate is those with higher rates of low physical activity. Again, this theme of populations that may exercise less may have higher incidences of obesity.

It's important to say that before we say for sure that populations with high rates of asthma and smoking definitely exercise less, we'd want to model the prediction ability of those as features of low physical activity.

Some features that may decrease obesity rates include populations that have higher incidences of cancer, binge drinking, and less than 7 hours of sleep daily. The reason that populations with higher incidences of cancer may reduce the obesity rate of that population is that once someone has cancer, they'll receive treatments such as chemo therapy. It's known that when going through treatments such as cancer, patients tend to lose weight.[5]

The other two features may be a bit harder to connect. First, one may expect that higher rates of binge drinking might lead to higher rates of obesity. One reason it may not is that populations where the % of population that binge drinks is higher (and maybe more acceptable) have a "work hard, play hard mentality". In other words, they'll drink more during the weekends but they'll "earn it" during the week by going to the gym routinely and eating healthily.

Finally, while as the % of a population reporting less than 7 hours of sleep daily increases, we should see the obesity rate of a population fall. This may be attributed to a belief that there is not enough time in a day to get everything done. The only way to get your physical activity in is to wake up early or stay up later to get everything completed in a day. It may also be related to the fact that while we are awake, we are burning more calories than while we're asleep, so those populations with less sleep burn more calories and have a lower % of obesity.

These are all theory at this point until there is further data to prove these out. At this point I'll accept that they are predictors of obesity and theorize on why later. Also, I will not be making any recommendations for populations to get less sleep, binge drink, or induce cancer as a key to reducing the obesity rate.

Does the Model Generalize Well?

Now that we have a working model and coefficients, I'll port the data to sklearn to take advantage of the different features in this library. First, I wanted to fit the data the same data to a regression model using sklearn as we did in statsmodels.

---

[5]Coping with Cancer-Related Weight Changes and Muscle Loss:
https://www.cancercare.org/publications/140-coping_with_cancer-related_weight_changes_and_muscle_loss

First thing I noticed is that the r^2 dropped using sklearn. There was a decrease from .988 using statsmodels to .834 using the sklearn Linear Regression. Next, using cross fold validation to predict the r^2 across five different folds, I found there was a good, generalized fit using the model we trained.

```
Model Accuracy
Linear Regression R2: 0.8344
Cross Fold Validation
Linear Regression Cross Val Score: [ 0.82855239  0.83478279  0.83394659  0.8405201   0.83588151]
```

Looking at the coefficients from the sklearn model, we see some slight variations. Populations with higher rates of asthma and smoking still lead the predictors in high obesity rates but we saw those with lack of access to insurance jump ahead of low physical activity. The coefficient of low physical activity pretty did drop a bit moving from .30 to .23.

	Coefs
CASTHMA	1.191421800
CSMOKING	0.486911806
ACCESS2	0.323054755
LPA	0.236084636
CHOLSCREEN	0.227689364
COREW	0.128596521
BINGE	0.089076092
CHECKUP	0.004142167
populationcount	-0.000000731
SLEEP	-0.006522867
CANCER	-0.020872148

Interestingly, the coefficient for binge drinking moved from about -.128 to .089. This change signifies that perhaps the feature is not that important to the model. Less than 7 hours of sleep also shifted to a -.007 which could also mean it's not that important to the model.

As a final step inside the regression modeling, I used a Lasso and Ridge regression model to see if they would provide better scores. As a part of this, I used GridSearchCV to optimize for the alpha coefficient for both models, the ideal alpha was 0 which implies the regression does not benefit from the Ridge or the Lasso models.

Random Forest Model

Finally, to see if I could come up with a better score, I used a random forest decision model. My training score landed at a .986 with five fold cross validation ranging between .9789 and .9804 and a score of .9193 on the hold out data. There seems to be a little over fitting going on but overall the model is generalized fairly well.

	Feature Importances
LPA	0.611399313
CSMOKING	0.198529303
SLEEP	0.041874381
ACCESS2	0.032887805
CASTHMA	0.023670268
BINGE	0.020559657
CHECKUP	0.019965509
COREW	0.018021239
CHOLSCREEN	0.014246385
CANCER	0.011048485
populationcount	0.007797655

Although the random forest cannot tell us whether or not these features are correlated with a rise or decline in the obesity rate of a population, it can tell us which features are most important. Not surprisingly, low physical activity is most important to predicting the obesity rate of a population followed by the percentage of smokers in that population, while the rest of the features tail off at less than 4% importance per feature.

With the random forest model, I achieve better model accuracy, but can explain less of what drives the obesity rate up or down.

Using what is commonly held with activity levels and obesity, along with what we've seen in the regression model and random forests model, we can say that low physical activity has a strong relationship with the obesity rate. I'm not saying there is causation between the two variables because there is other data we don't have that may confound the true causes (such as caloric intakes). Until that data is collected, we won't be able to say for sure.

Summary

Throughout this capstone, I've gone through a few steps that helped me predict the obesity rate of a given population. They include:

- Defining 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
- Performed a regression & random forests model

Our models point to populations with higher low physical activity rates as a good predictor of the obesity rate along with higher rates of smoking, and higher rates of the uninsured. We cannot say for certain that these relationships are causal, but they are a good place to start investigating if we wanted to reduce the obesity rate.

For the next round of analysis, I'd suggest bringing in calorie consumption rates by population along with some of the other datasets mentioned above. We may also want to think about bringing other research performed around low physical activity and the obesity rate.