Capstone Project 2 - Milestone Report
Predicting Health using the National Health and Nutrition Survey

Problem Statement
There are many factors that impact the health and wellness of human society. It should be possible to reduce the occurrence of particular health concerns if those factors with high impact can be identified. The goal of this project is to use statistical inference and machine learning to explore if predictions in certain illnesses can be made related to habits, nutrition, BMI, and bloodwork.

Proposed Solution
Perform exploratory analysis and predictive modeling from the National Health and Nutrition Survey (NHANES) dataset that assesses the health and nutritional status of people in the United States.

- Accurate predictive models can help people in general to benefit if looking for ways to better their life.
- It can extend to professionals in the medical, nutritional, and fitness arenas to discuss the potential effects of life choices relating to health conditions.
- This could lead to people wanting to research particular findings in more depth in order to better their lives. Potentially adding longevity and making a positive difference for our population at large.

Datasets
A collection of datasets from the National Health and Nutrition Survey (NHANES) was obtained from Kaggle. https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey.
This information was used in the project to study variables that could potentially help predict health conditions and improve human lives.
More information on the NHANES survey can be found on the Center for Disease Control and Prevention's website https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm.

Wrangling Steps Performed:

Data, Demographics, and Labs

1. Three csv files were downloaded and read into a normalized pandas dataframe. They were merged together to create one dataframe.
2. Columns were explored to determine which would be useful on this project. Those identified for use were then renamed from their initial codes to identifiable word strings.
3. A feature generation for BMI at age 25 was created using the height and weight found in the data.

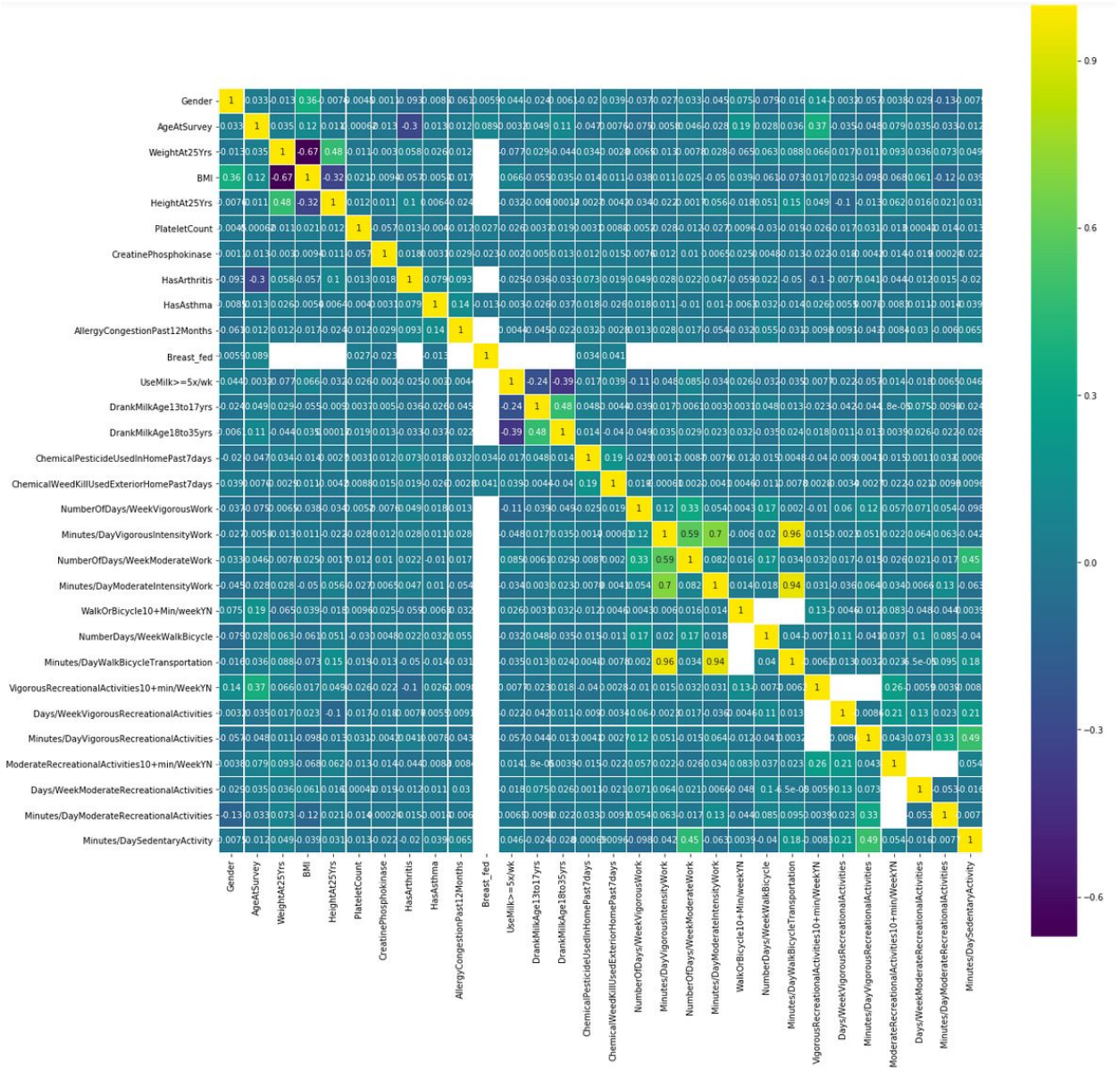
4. The data was checked and scanned for any null information present. This was later used when looking at categories for those having asthma and those having arthritis. Null values had their entire rows removed when those columns were used for explorations.

Dealing with Outliers

The dataset had categories that had significantly less data to use. Categories with more data were used in order to provide a substantial sampling to match the illnesses chosen to look at of asthma and arthritis.

Exploratory Data Analysis

Correlation Plot



High correlations:

- Minutes per day of vigorous intensity work and minutes per day walking or bicycling for transportation at 0.96.
- Minutes per day of moderate intensity work and minutes per day walking or bicycling for transportation at 0.94.

The above correlation plot suggests that people who have vigorous and moderate intensity type of active work, often walk or bicycle for transportation.

Low correlations:

- Drank Milk regularly during ages 18 - 25 and Using Milk 5 or more times per week now at -0.39.

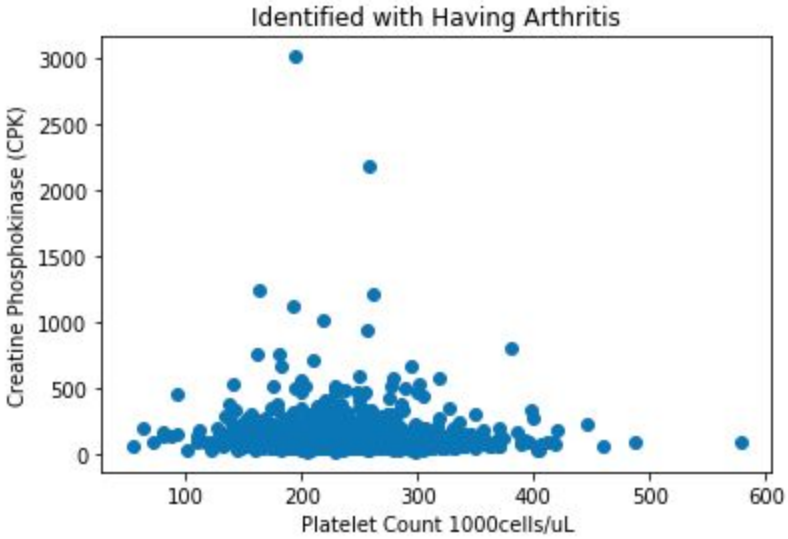
- Drank Milk regularly during ages 13-17 and Using Milk 5 or more times per week now at -0.24.

The above correlation plot suggests that people who drank milk during their younger years has little connection to whether they drink milk now.

Scatterplot

Arthritis and Blood Counts

One area this study further explored the correlation with those who reported having been told they had arthritis and their blood counts.



The scatterplot shows most participants identified with arthritis are in the lower left quadrant, which indicates lower platelet and CPK counts. Based on this visual alone, a statement such that a participant with low levels of CPK and low platelet counts have a good chance of being told by a health professional that they have arthritis. Further, there are no participants in the upper right quadrant, which suggests that higher counts for both CPK and platelets relates to those not being diagnosed with arthritis. There are definitely a few outliers in the upper left quadrant that are not significant and would not be included when making generalizations. There are more outliers in the lower right quadrant, perhaps significant enough to look further into whether the higher platelet counts play a role in diagnosing possible arthritis.

In general, high CPK levels in the muscle suggest the presence of inflammatory muscle disease, but they can also be caused by trauma, injection into the muscle, or muscle disease due to hypothyroidism. Conversely, low levels of CPK can be indicative of rheumatoid arthritis. https://www.arthritis-health.com/glossary/creatine-phosphokinase

Machine Learning Classification and K-Means Clustering focusing on Asthma Diagnosis

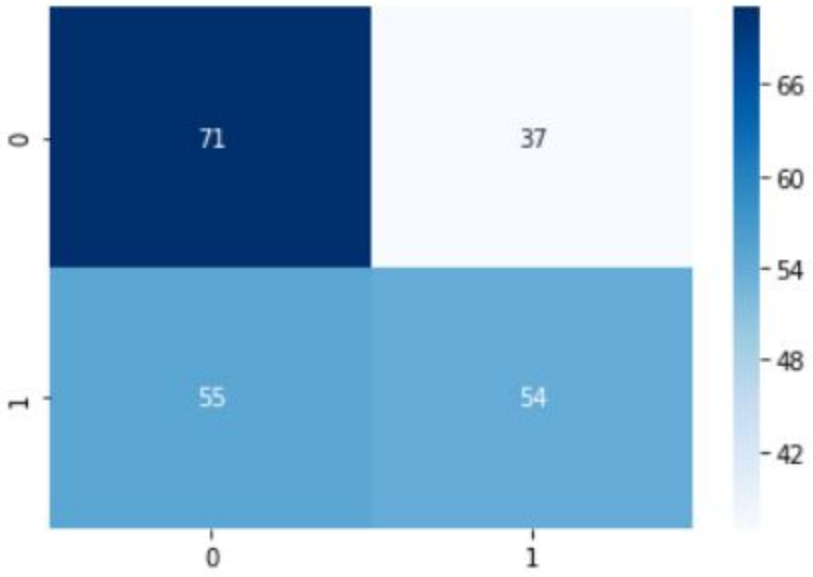
Confusion Matrix Plots

Visualizing the matrix as a heat map. The classification rate is around 57%, which is a good Accuracy in that it is more than 50%. The precision, or how often the model is correct, predicts if people with certain habits or bloodwork will or will not get asthma 56% of the time. If there are people that will or will not get asthma, the Logistic Regression model can identify it 66% of the time.

Accuracy: 0.576036866359447
Precision: 0.5634920634920635
Recall: 0.6574074074074074

A Confusion Matrix is a performance measurement tool used in machine learning classification. It is useful for visualizing details of how well a classifier performs for one with any number of classes greater than 2. There are two classes here; the person will get asthma or the person will not get asthma. The classifier made a total of 217 predictions.

The prediction showed “yes” 91 times, and “no” 126 times. The true positive (TP), lower right corner, is when the prediction was to have asthma and the person was diagnosed with asthma. The true negative (TN), upper left corner, is when the prediction was to not have asthma, and the person was not diagnosed with asthma. The false positive(FP), is a Type I error, where the prediction was to have asthma, yet that person did not. The false negative(FN), is a Type II error, where the prediction was for no asthma, and that person did have the asthma diagnosis.



TN = 71 FP= 37
FN = 55 TP= 54

K-Means Clustering

The elbow method showed optimization at 4 clusters.

```
[0 2 2 0 2 0 2 0 2 2 0 0 0 0 1 0 0 2 2 0 0 2 2 2 0 2 0 2 2 0 2 0 0 2 0 0
 2 0 2 0 2 2 0 0 0 2 2 2 0 1 2 1 2 2 2 2 0 0 0 2 2 0 2 1 0 0 2 0 1 0 0 1 2
 2 2 2 2 2 3 2 2 2 0 2 2 2 0 2 2 2 2 0 2 0 2 0 0 0 2 0 0 0 2 2 0 0 0 2 2 0
 2 2 1 0 1 0 2 0 2 2 1 2 2 2 1 2 2 2 2 0 1 2 2 0 2 0 2 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 3 2 0 0 0 3 2 0 2 2 2 2 0 2 1 2 2 0 0 0 2 2 2 0 0 0 0 2 0 2 0 2 2
 2 0 2 2 0 0 2 0 0 0 2 2 0 2 2 2 2 2 2 2 0 0 2 0 0 2 0 0 0 2 2 0 2 2 1 3 2 2
 2 0 3 2 0 2 0 0 0 1 2 2 2 2 1 0 2 2 0 0 0 2 2 2 0 2 2 0 0 2 2 2 2 2 2 2 0 1
 0 0 2 2 2 2 0 0 0 0 2 0 0 0 0 2 0 0 2 0 0 2 2 2 0 0 2 2 0 3 2 0 0 0 2 2 0
 0 0 0 2 1 2 0 0 2 2 2 0 2 2 0 0 0 2 0 2 2 0 0 2 0 2 0 2 2 0 1 2 2 0 0 0 2
 2 2 0 2 2 0 2 2 0 2 0 2 2 0 2 3 2 0 2 0 2 2 2 2 0 2 2 2 2 0 1 2 2 2 2 0 0
 2 0 0 0 2 2 2 2 2 0 2 2 2 0 2 2 2 2 2 0 0 2 2 2 2 2 0 0 2 0 0 0 0 2 0 0 2 0 2
 0 0 0 1 2 2 2 2 2 0 2 2 0 0 0 0 0 0 0 0 1 0 0 0 0 2 2 2 0 2 2 2 2 0 0 2 2 0
 3 0 0 0 0 2 0 0 0 0 0 0 0 2 2 0 0 2 2 0 2 3 2 2 0 0 0 0 0 2 2 1 2 2 0 0 2 2 0
 2 1 2 2 2 0 2 2 0 0 0 2 0 2 0 2 0 1 0 2 2 2 2 2 0 0 3 0 2 2 0 2 0 2 2 2 2 2
 2 0 2 2 0 2 2 0 0 0 2 2 1 0 2 2 2 0 0 0 2 0 3 2 1 0 0 0 0 2 0 0 0 0 0 2 0
 0 3 0 1 2 2 0 0 2 0 0 2 2 0 2 0 2 2 0 0 2 0 0 3 2 0 2 0 2 0 0 0 2 0 0 0 0 0
 2 0 0 2 2 0 0 2 0 2 0 0 0 2 2 2 2 2 2 0 2 0 2 0 3 2 0 0 0 2 2 2 0 0 0 0 3
 2 2 0 3 0 2 2 0 0 0 0 2 0 0 0 2 2 0 0 0 0 0 2 0 0 2 0 0 0 2 2 2 2 2 2 0 0 3
 2 0 2 0 0 2 2 0 0 2 0 0 0 2 2 2 0 0 0 2 3 2 2 0 0 0 0 2 0 2 0 0 0 0 3 2 2
 2 0 0 0 0 0 2 0 2 0 0 2 2 2 2 2 0 1 2 2 0 0 2 0 2 0 0 0 0 2 0 0 2 0 2 0 0
 0 2 2 2 0 2 2 2 0 0 0 0 2 2 2 2 2 0 2 0 2 2 0 0 0 0 2 2 0 2 0 2 2 2 2 2 2 0
 2 0 2 2 2 0 0 2 2 2 2 0 2 2 2 2 0 0 0 0 0 0 3 0 0 0 2 0 2 0 2 2 0 0 0 2 2
 2 0 0 2 2 0 0 2 0 0 2 0 0 0 0 3 0 0 2 2 1 0 0 0 0 2 2 2 2 2 2 0 2 2 2 2 0 2
 0 2 0 2 0 0 2 2 0 2 2 2 0 2 0 0 2]
```

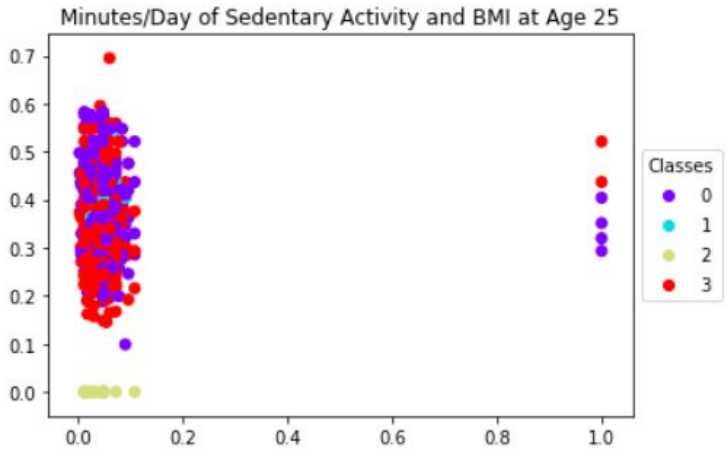
Here the four clusters are separated into their individual identities. This then allows matching up to the features represented.

```
array([[0.49290954, 0.31531325, 0.04206305, 0.11369193, 0.11919315,
        0.08679707, 0.09688264, 0.07243276, 0.04623347, 0.12438875,
        0.35969412, 0.0107878 , 0.10360636],
       [0.77715517, 0.31944241, 0.02809126, 0.08189655, 0.12068966,
        0.11637931, 0.0862069 , 0.09051724, 0.04117653, 0.06465517,
        0.        , 1.        , 0.08189655],
       [0.82732274, 0.31747039, 0.03213677, 0.11033007, 0.11430318,
        0.10971883, 0.10085575, 0.07365526, 0.04598846, 0.11308068,
        0.39395728, 0.00912912, 0.05409535],
       [0.68154762, 0.3003715 , 0.04304954, 0.14285714, 1.        ,
        0.10714286, 0.11309524, 0.08333333, 0.05457689, 0.125     ,
        0.40756962, 0.0083228 , 0.0952381 ]])
```

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Highest to Lowest Cluster Ranking			
age at survey	0.4865239290	0.77715517	0.82732274	0.68154762	2	1	3	0
platelet count	0.3089125890	0.31944241	0.31747039	0.3003715	1	2	0	3
creatine phosphokinate	0.0368235249	0.02809126	0.03213677	0.04304954	**3**	0	2	1
chem pesticide use in home past week	0.1152392950	0.08189655	0.11033007	0.14285714	**3**	0	2	1
chem weed kill use outside past week	0.1209068010	0.12068966	0.11430318	1	**3**	0	1	2
vig rec activity 10+ min in a week Y/N	0.0903652393	0.11637931	0.10971883	0.10714286	1	2	3	0
walk bike 10+min in a week Y/N	0.0947732997	0.0862069	0.10085575	0.11309524	**3**	2	0	1
moderate rec act 10+minin a week Y/N	0.0746221662	0.09051724	0.07365526	0.08333333	1	3	0	2
seden activity typical day	0.0452259332	0.04117653	0.04598846	0.05457689	**3**	2	0	1
use milk 5+times / week	0.1218513850	0.06465517	0.11308068	0.125	**3**	0	2	1
BMI at age 25	0.3511277320	0	0.39395728	0.40756962	**3**	2	0	1
weight at age 25	0.0112681394	1	0.00912912	0.0083228	1	0	2	3
has been diagnosed with arthritis	0.1086272040	0.08189655	0.05409535	0.0952381	0	3	1	2

Cluster 3 has the highest values for particular categorical features. This finding means those with asthma appear the most in these feature categories.

KMeans Plot



In the above Kmeans plot, one cluster is uniquely identified at the bottom and it then prompts investigation to analyze cluster centers in the table. Minutes per day of Sedentary Activity and a participant's BMI at age 25.

Conclusions:

Arthritis

The participants with low levels of creatine phosphokinase and low platelet counts have a good chance of being told by a health professional that they have arthritis. Further, it suggests that

higher counts for both creatine phosphokinase and platelets relates to those not being diagnosed with arthritis.

Correlations

The study suggests that people who have vigorous and moderate intensity type of active work, often walk or bicycle for transportation. It further suggests that people who drank milk during their younger years have little connection to whether they drink milk now.

Asthma

The participants with asthma ranked higher in particular categories from this study. It is seen in the k-means clustering findings. This indicates that indeed it cannot be ruled out that particular categories influence having asthma such as levels of creatine phosphokinase, chemical pesticide and weed kill use, sedentary activity in a day, using milk regularly, and BMI at age 25.