# Machine Learning in Vision and Eye Health

Jason Huang
Gianna Nguyen

## Introduction

This project uses multiple methods to predict the group with poor eye vision using data sets with evaluation using cross validation. In the end, we predict how likely you will have poor vision and/or eye health based on various factors such as gender, socioeconomic status, age etc. This will help public health officials decide how to distribute resources to those who need it the most. The inputs will be various minority group labels such as “Female”, “People with Diabetes” etc. and the output will be the probability they have or could develop poor vision and/or eye problems. The labels are independent, so they do not affect one another but instead we are focusing on predicting for one label at a time.

We used KNN, linear regression, and logistic regression. We will use all other machine learning methods in the future. After interpreting our results, we found that there is indeed a correlation between certain groups that are minorities or have pre-existing health conditions and having eye problems. We found that it was hard to obtain a specific probability of having eye problems, but the higher probability of certain groups compared to other more well off, healthy groups shows that our hypothesis was correct.

## Related Work

A similar problem was described in the AMA Journal of Ethics, titled “Can AI help reduce disparities in general medical and mental health care?” One method used in the study linked below is “logistic regression with L1 regularization (implemented by Python package scikit-learn[29] with a hyperparameter of C = 1) using an 80/20 split for training and testing data over 50 trials.” What they did was compare error rates for psychiatric readmission with error rates for ICU mortality based on race, gender, and insurance payer type.

The goal of the study to decide if there were disparities in health care among minority groups is similar to the goal of our project. However, their data included dependent labels while ours included independent labels (i.e., race, gender, and insurance payer type all have an effect on the end result while for ours there is only one label being looked at one time)We decided to try a similar method but with linear regression instead of logistic regression after trying both models but had a similar set up compared to their data and 80/20 split. We also used scikit-learn based on the efficiency of this library in other similar studies.

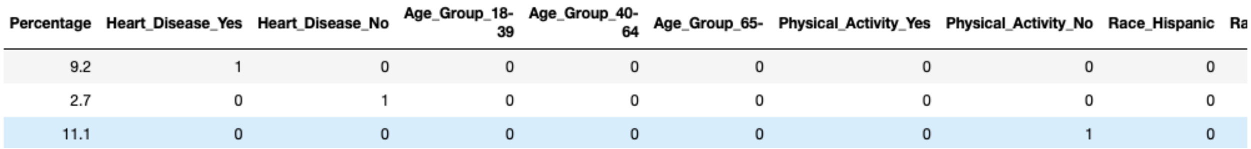
*Reference: https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02* *https://mlmedicine.org/2017/11/04/healthcare-ai-python-version-based-python-sk-learn-library/*

## Data Sets

The original data set contains 58097 eye health data and 11 classification labels. They are: Heart Disease (Yes/No/Total), Age Group (18-39/40-64/65-older/Total), Physical Activity (Yes/No/Total), Race/Ethnicity (Non-Hispanic Black/Non-Hispanic White/Hispanic/Other/Total), Smoking Status (Former/Current/Never/Total), Educational Level (High School and Below/More than High School/Total), Stroke (Yes/No/Total), Gender (Male/Female/Total), Diabetes (Yes/No/Total), Fair or Poor General Health (Yes/No/Total), and Hypertension (Yes/No/Total).

We modified the data set by getting rid of any with null values and then assigning 1 to 0 based on the classification label (Figure 1, 2) (Refer to A.2 at the end). Along with the classification label, there is the percentage of that sample size suffering from eye/vision problems.

Reference: https://catalog.data.gov/dataset/behavioral-risk-factors-vision-amp-eye-health-c8237



| Percentage | Heart_Disease_Yes | Heart_Disease_No | Age_Group_18-39 | Age_Group_40-64 | Age_Group_65- | Physical_Activity_Yes | Physical_Activity_No | Race_Hispanic | Ra |
|---|---|---|---|---|---|---|---|---|---|
| 9.2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 2.7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 11.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |

Figure 1: Example rows of what our data set looked like after being processed for regression problems

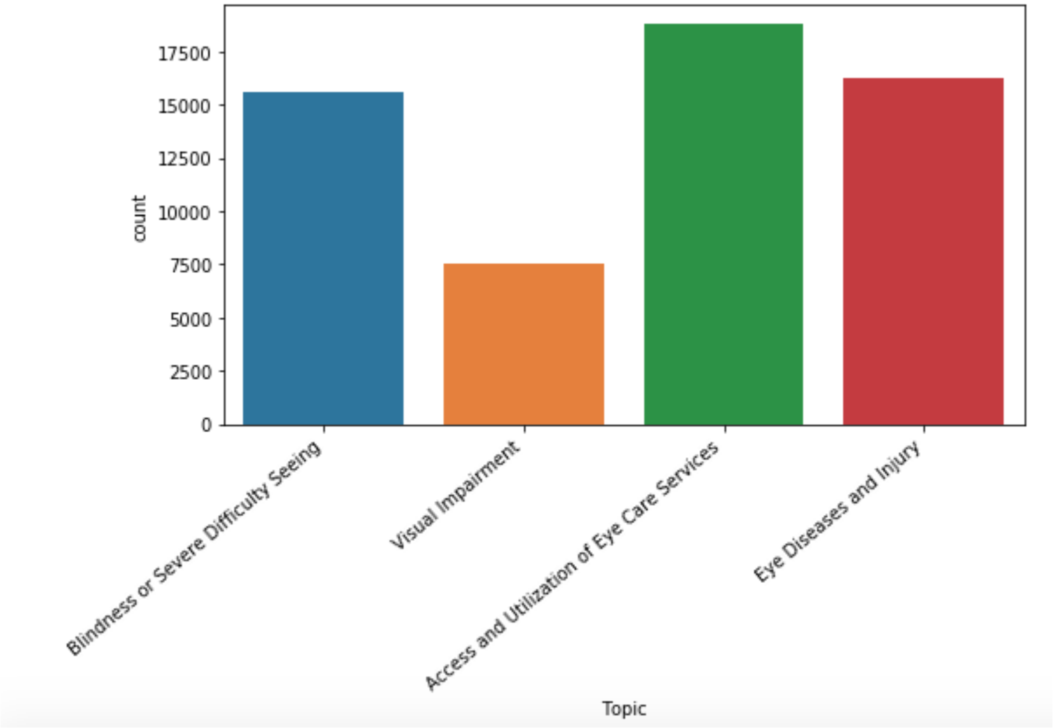
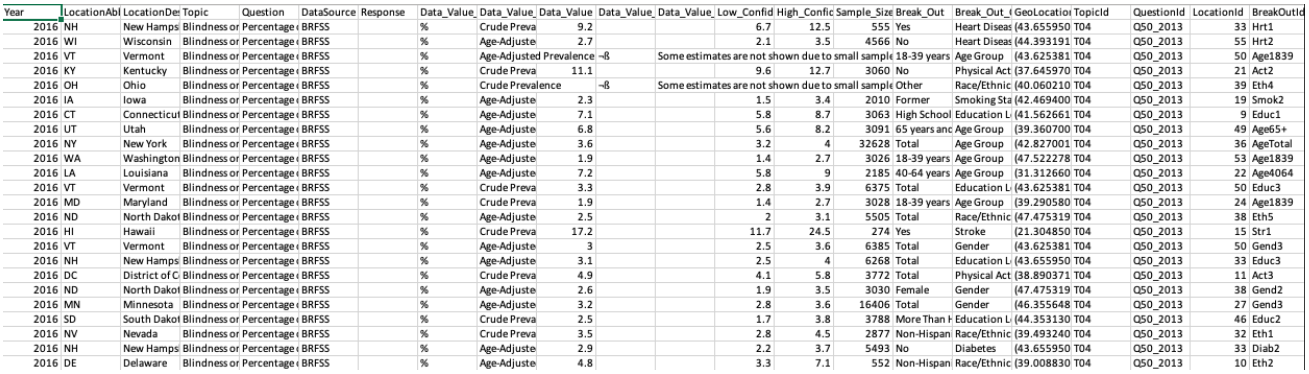


Figure 2: All the topics to predict probability of eye-related problems in general.

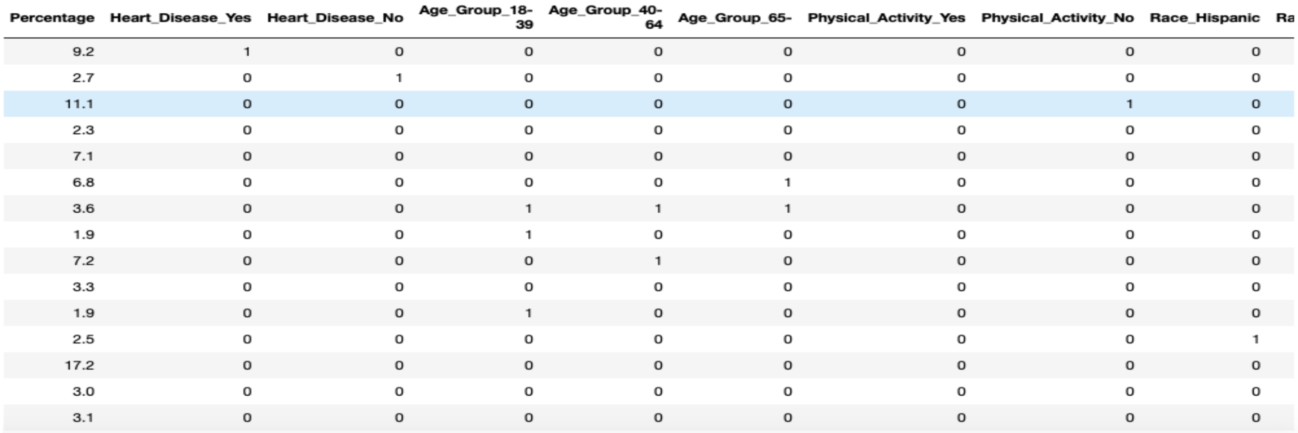
## Technical Approach

By giving a detailed look at our dataset and play with it using seaborn libraries (visualizing data and distributions), our y value from dataset appears to be a continuous float number indicating the percentage (or likelihood) of having eye health problems. But the other columns appear to be a mess since it contains multiple category names, empty data, and non-useful data in the data set. The first thing in our mind is to separate the 11-category name from 1 column and then separate them further in detail since each of the category has different values. In order to transform the dataset to ideal dataset that can be used by our machine learning algorithms, the first step is to clean, modify, and tune the data. This the core and one of the most important steps in our project, which is data preprocessing. We preprocessed the data from the CSV file and transformed them into our desired forms by doing data loading, data cleaning, and data encoding. In particular, using one-hot encoding to ensure X variables are binaries. We used seaborn's heat map to identify empty cells in the data frame, and then we used data cleaning scripts and wrote helper functions for one-hot data preprocess. In the end, we recreated a csv file for the data after being preprocessed. Feature scaling was also helpful because it helped normalize the data while speeding up algorithm's calculation. Before going to the next step, we also created graphs and charts to see how the distribution goes.



| Year | LocationAbl | LocationDe: | Topic | Question | DataSource | Response | Data_Value_ | Data_Value_ | Data_Value | Data_Value_ | Data_Value_ | Low_Confid | High_Confic | Sample_Size | Break_Out | Break_Out_( | GeoLocatio | TopicId | QuestionId | LocationId | BreakOutId |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2016 | NH | New Hamps | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 9.2 | | | 6.7 | 12.5 | 555 | Yes | Heart Diseas | (43.655950 | T04 | Q50_2013 | 33 | Hrt1 |
| 2016 | WI | Wisconsin | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 2.7 | | | 2.1 | 3.5 | 4566 | No | Heart Diseas | (44.393191 | T04 | Q50_2013 | 55 | Hrt2 |
| 2016 | VT | Vermont | Blindness or | Percentage ( | BRFSS | | % | Age-Adjusted Prevalence | | ¬ß | Some estimates are not shown due to small sample | | | | 18-39 years | Age Group | (43.625381 | T04 | Q50_2013 | 50 | Age1839 |
| 2016 | KY | Kentucky | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 11.1 | | | 9.6 | 12.7 | 3060 | No | Physical Act | (37.645970 | T04 | Q50_2013 | 21 | Act2 |
| 2016 | OH | Ohio | Blindness or | Percentage ( | BRFSS | | % | Crude Prevalence | | ¬ß | Some estimates are not shown due to small sample | | | | Other | Race/Ethnic | (40.060210 | T04 | Q50_2013 | 39 | Eth4 |
| 2016 | IA | Iowa | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 2.3 | | | 1.5 | 3.4 | 2010 | Former | Smoking Sta | (42.469400 | T04 | Q50_2013 | 19 | Smok2 |
| 2016 | CT | Connecticut | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 7.1 | | | 5.8 | 8.7 | 3063 | High School | Education L | (41.562661 | T04 | Q50_2013 | 9 | Educ1 |
| 2016 | UT | Utah | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 6.8 | | | 5.6 | 8.2 | 3091 | 65 years anc | Age Group | (39.360700 | T04 | Q50_2013 | 49 | Age65+ |
| 2016 | NY | New York | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 3.6 | | | 3.2 | 4 | 32628 | Total | Age Group | (42.827001 | T04 | Q50_2013 | 36 | AgeTotal |
| 2016 | WA | Washington | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 1.9 | | | 1.4 | 2.7 | 3026 | 18-39 years | Age Group | (47.522278 | T04 | Q50_2013 | 53 | Age1839 |
| 2016 | LA | Louisiana | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 7.2 | | | 5.8 | 9 | 2185 | 40-64 years | Age Group | (31.312660 | T04 | Q50_2013 | 22 | Age4064 |
| 2016 | VT | Vermont | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 3.3 | | | 2.8 | 3.9 | 6375 | Total | Education L | (43.625381 | T04 | Q50_2013 | 50 | Educ3 |
| 2016 | MD | Maryland | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 1.9 | | | 1.4 | 2.7 | 3028 | 18-39 years | Age Group | (39.290580 | T04 | Q50_2013 | 24 | Age1839 |
| 2016 | ND | North Dakot | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 2.5 | | | 2 | 3.1 | 5505 | Total | Race/Ethnic | (47.475319 | T04 | Q50_2013 | 38 | Eth5 |
| 2016 | HI | Hawaii | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 17.2 | | | 11.7 | 24.5 | 274 | Yes | Stroke | (21.304850 | T04 | Q50_2013 | 15 | Str1 |
| 2016 | VT | Vermont | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 3 | | | 2.5 | 3.6 | 6385 | Total | Gender | (43.625381 | T04 | Q50_2013 | 50 | Gend3 |
| 2016 | NH | New Hamps | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 3.1 | | | 2.5 | 4 | 6268 | Total | Education L | (43.655950 | T04 | Q50_2013 | 33 | Educ3 |
| 2016 | DC | District of C | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 4.9 | | | 4.1 | 5.8 | 3772 | Total | Physical Act | (38.890371 | T04 | Q50_2013 | 11 | Act3 |
| 2016 | ND | North Dakot | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 2.6 | | | 1.9 | 3.5 | 3030 | Female | Gender | (47.475319 | T04 | Q50_2013 | 38 | Gend2 |
| 2016 | MN | Minnesota | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 3.2 | | | 2.8 | 3.6 | 16406 | Total | Gender | (46.355648 | T04 | Q50_2013 | 27 | Gend3 |
| 2016 | SD | South Dakot | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 2.5 | | | 1.7 | 3.8 | 3788 | More Than H | Education L | (44.353130 | T04 | Q50_2013 | 46 | Educ2 |
| 2016 | NV | Nevada | Blindness or | Percentage ( | BRFSS | | % | Crude Preva | 3.5 | | | 2.8 | 4.5 | 2877 | Non-Hispani | Race/Ethnic | (39.493240 | T04 | Q50_2013 | 32 | Eth1 |
| 2016 | NH | New Hamps | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 2.9 | | | 2.2 | 3.7 | 5493 | No | Diabetes | (43.655950 | T04 | Q50_2013 | 33 | Diab2 |
| 2016 | DE | Delaware | Blindness or | Percentage ( | BRFSS | | % | Age-Adjuste | 4.8 | | | 3.3 | 7.1 | 552 | Non-Hispani | Race/Ethnic | (39.008830 | T04 | Q50_2013 | 10 | Eth2 |

Figure 3: The original dataset that has a lot of useless data and empty cells.



| Percentage | Heart_Disease_Yes | Heart_Disease_No | Age_Group_18-39 | Age_Group_40-64 | Age_Group_65- | Physical_Activity_Yes | Physical_Activity_No | Race_Hispanic | Ra |
|---|---|---|---|---|---|---|---|---|---|
| 9.2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 2.7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 11.1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| 2.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 7.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6.8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 3.6 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | |
| 1.9 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 7.2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 3.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 1.9 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 2.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| 17.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

Figure 4: Dataset after one-hot encoding.

Initially, we believed it was supervised learning problem and regression problem since it had labeled data and the output of the model is a number. Hence, our first intuition was using one of the regression algorithms - Linear Regression. Or, in particular, multi-linear regression. Linear Regression summarizes the relationship between Y and multiple X values. In our case, the relationship between the likelihood of having eye health problems and different groups (for example, people who have heart disease) of people. On the first try, our result looks quite alright with some minor difference between actual value and predict value.

We also thought about using classification method because the dataset had a lot of categories and groups. Since one of our goals was to predict the likelihood of getting eye health problems, we can also do some modification to make it a classification problem. In particular, we set a boundary line (for values lower than the predetermined y value, we set it to 0 - less likely to get eye health problems; for values higher than the predetermined y value, we set it to 1 - more likely to get eye health problems). We decided to use K nearest neighbor since it was one of the simple classification algorithms. First, we read and split the data into Xtr, Ytr, Xva, Yva. Then we trained with various K: 1, 2, 5, 10, 50, 100, 200. After plotting the training vs. validation mean squared error, we realized that this algorithm would not be the best fit for our data since the error was very high.

So, we moved on to another classification algorithm - logistic regression. By reprocessing the y value of the data frame, we are able to get a binary outcome. In the end, we believed logistic regression might be a better approach since it gave better results and high accuracy rate (accuracy =0.82).

## Software

*(a) code we wrote*

| **Software Name** | **Files** | **Purpose** | **How we used the library/code we wrote** |
|---|---|---|---|
| one-hot encoding function | project.ipynb | to preprocess the data and put it into a new csv file | we wrote a helper function that encode the data to 1 and 0 automatically |
| Total numbers of category | project.ipynb | to generate all the categories and its possible sub-categories | we wrote a portion of code to get all the categories from the dataset for further use. |

*(b) code from other people that we used*

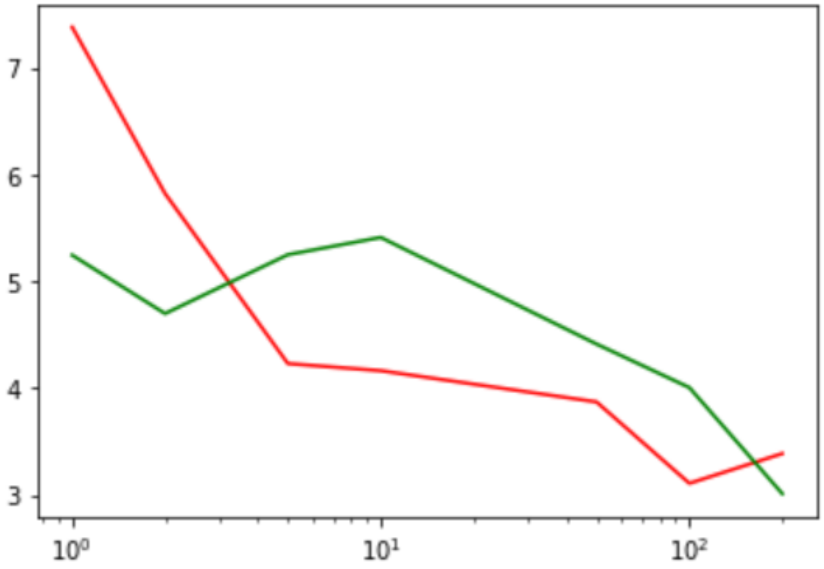
| **Software Name** | **Files** | **Purpose** | **How we used the library/code we wrote** |
|---|---|---|---|
| pandas | 184aDataSet.csv,<br>184a_KNN.ipynb,<br>project.ipynb | reading the data from the csv file and generating data frames | changing the data format and splitting into training and validation data. |
| matplotlib.pyplot | 184aDataSet.csv,<br>184a_KNN.ipynb,<br>project.ipynb | plotting charts | computing the training vs. validation error to chart |
| seaborn | project.ipynb | visualizing data | making charts (heat map, countplot...) to see if there were any null values in the data and then removing them |
| sklearn | 184a_KNN.ipynb<br>project.ipynb | creating different models, generate reports, and giving results | trained and predicted values using linear regression, logistic regression, and KNN |

## Evaluation

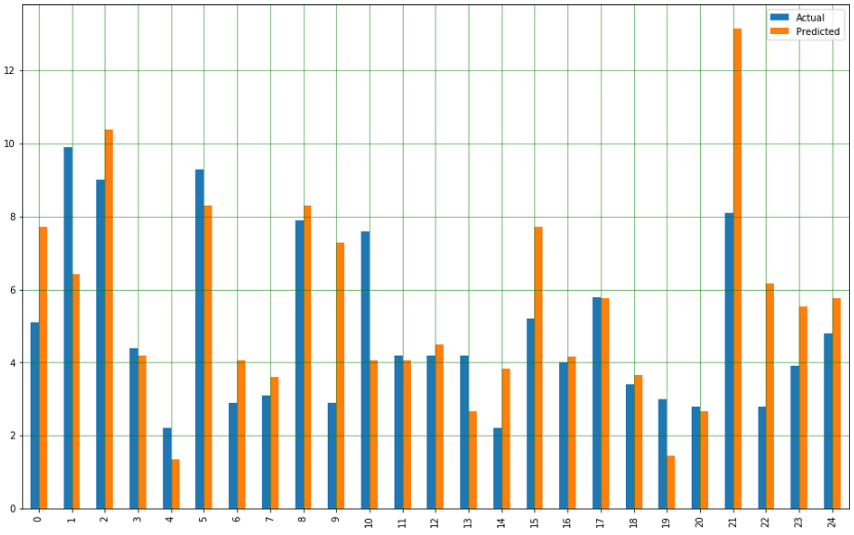
(a) How we set up our experiments

We used cross-validation for most of our project. We split the training data and test data into many portions and randomly shuffled them using cross-validation to get the final result for our models and algorithms. We used cross validation to separate the data into groups and compare the scores of each group. In order to evaluate the classification model, we used Classification Accuracy to calculate the percentage of correct predictions, and visualized the performance of the classification model, null accuracy, percentage of zeros/ones, false positive rate and etc. We finalized our evaluation with a success method plot for showing the percentage of successful predictions for each method we used.

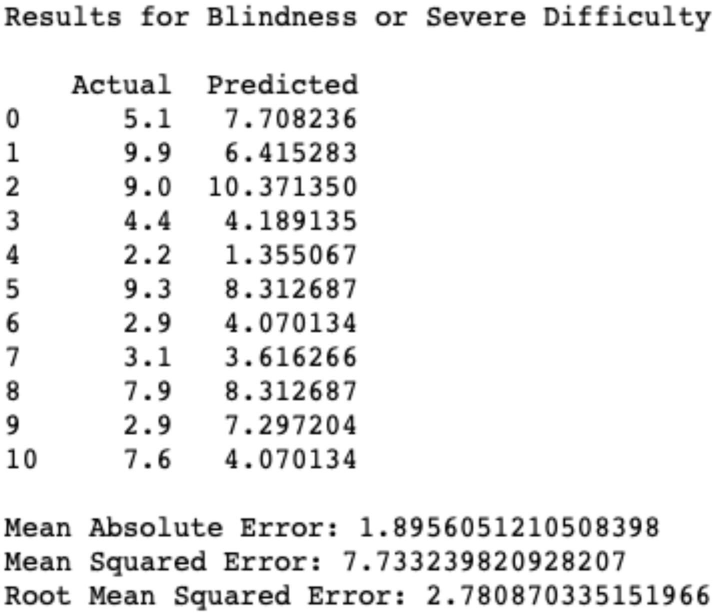
(b) Results we obtained



For KNN, performance highly depended on K. As you can see from this chart, we tried different K but, in the end, the mean squared error was abnormally high (>1). This is why we decided not to use KNN in the end.



This chart shows our actual vs. predicted results for linear regression.

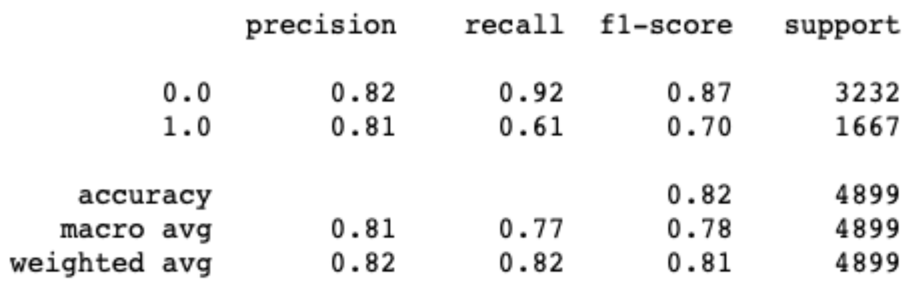


```
Results for Blindness or Severe Difficulty Seeing:

    Actual  Predicted
0      5.1   7.708236
1      9.9   6.415283
2      9.0  10.371350
3      4.4   4.189135
4      2.2   1.355067
5      9.3   8.312687
6      2.9   4.070134
7      3.1   3.616266
8      7.9   8.312687
9      2.9   7.297204
10     7.6   4.070134

Mean Absolute Error: 1.8956051210508398
Mean Squared Error: 7.733239820928207
Root Mean Squared Error: 2.780870335151966
```

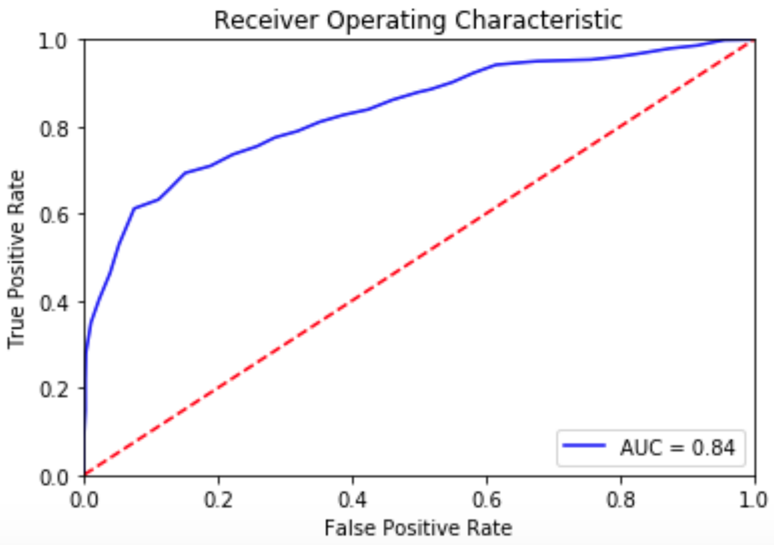
Above is the result of using linear regression



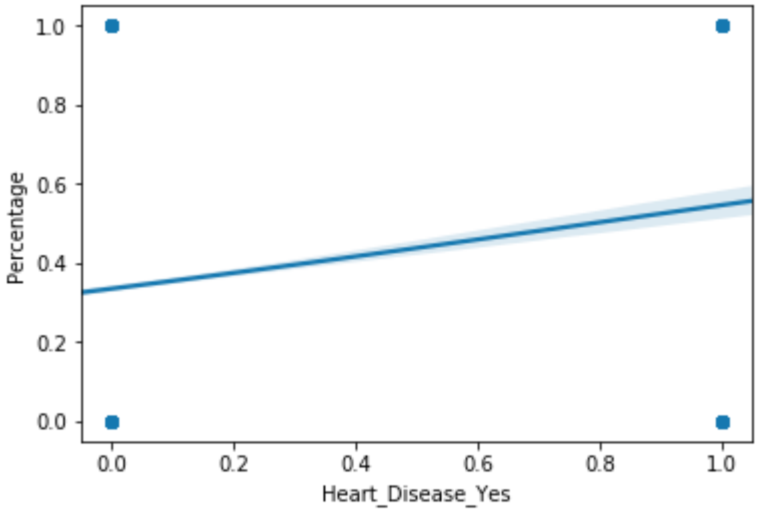
```
              precision    recall  f1-score   support

         0.0       0.82      0.92      0.87      3232
         1.0       0.81      0.61      0.70      1667

    accuracy                           0.82      4899
   macro avg       0.81      0.77      0.78      4899
weighted avg       0.82      0.82      0.81      4899
```



Here is the result we obtained using logistic regression. Compared to our other machine learning algorithms, logistic regression had the best performance so far since it had pretty good accuracy rate and AUC score.



We tried using different scaling of the features to remove things such as outliers, but it ended up performing worse after we did so. We hypothesize that this is because the data is so widely distributed that trying to scale it affects the results by too much and thus ends up being less accurate. (StandardScaler, RobustScaler, MinMaxScaler from sklearn).

## Conclusion

Overall, we would say that our hypothesis was met with certain minority groups and those with pre-existing health conditions having higher probabilities of eye problems. However, the specific results of our algorithms were not what we expected. We learned that KNN was not the best algorithm for this type of problem. We learned that logistic regression and linear regression would be more suitable for our project due to lower variance and that is why its results were overall better for our specific dataset.

The major limitations of current approaches in solving the problem of socioeconomic disparities in health care is based on the type, amount, and accuracy of data obtained. From our research when looking at methods used on similar problems, there are larger amounts of healthcare data form certain groups such as Caucasians or how data is highly limited from groups like illegal immigrants that may be hesitant to seek out health care. Similarly, our data is limited due to how there is only one label when the result of the algorithm would be much more accurate if it was a combination such as “Hispanic Male with Heart Disease” compared to “Female with High Physical Activity.”

If we were in charge of a research lab, we would further improve this project by trying to cluster the data. This way we could see whether certain labels have similar probabilities of having an eye problem and then we could research whether the labels are truly independent or are they actually related to one another. Also, we did not have time to do so but we would have explored other methods such as decision trees to see if they performed better and have it weighted so that data points from a smaller sample size will have less effect.

## Contributions

**Jason Huang**: Cleaned, preprocessed the data set, got rid of null values and made a new data set with the proper rows and values actually needed. Generated charts showing the distribution of the data. Wrote machine learning models using linear regression and logistic regression which we ultimately decided to use in our final project because it performed the best. Worked on testing the code and interpreting results using graphs, charts, and statistics.

**<u>Gianna Nguyen</u>**: found the data set and did the primary charts of the data so we could decide what method to use. Used KNN model but both the training and validation error was very high, so we decided to use linear regression instead. Wrote the README and organized the files to submit.

# Appendix

Figure A.1: KNN Training (red) and Validation MSE (green) for different K's

As one can see from the graph, the error (y-axis) is over one and quite high so the performance was very low. After such a performance, we decided to choose a different model.

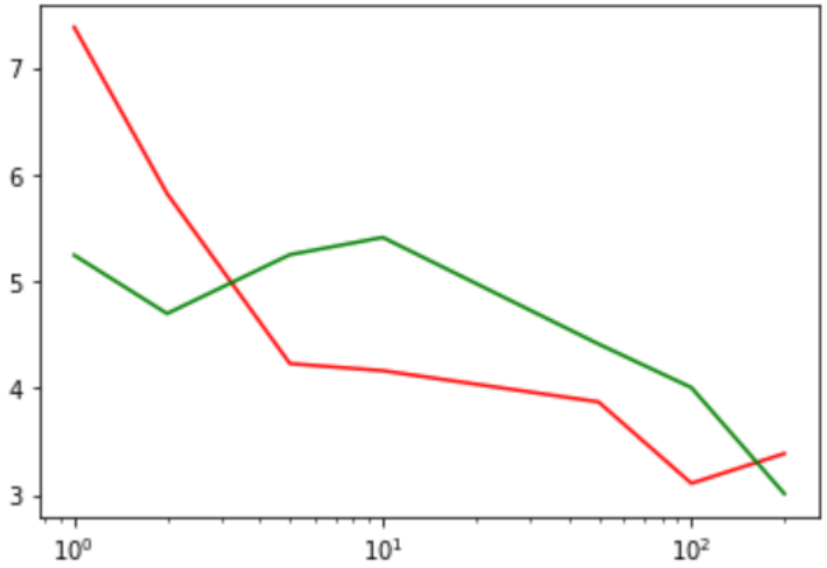
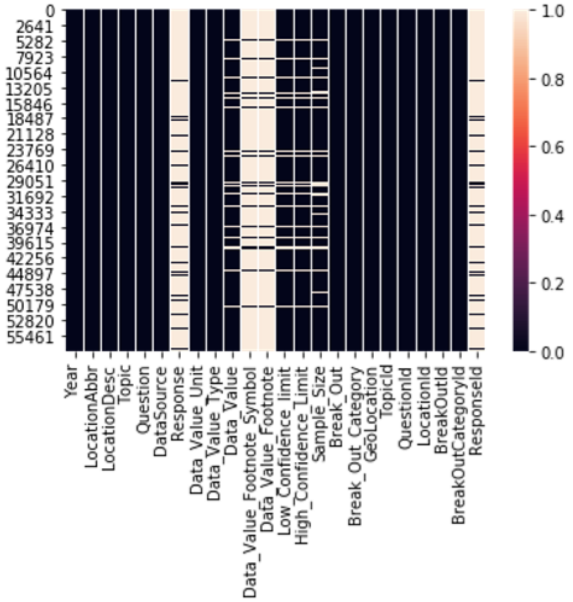
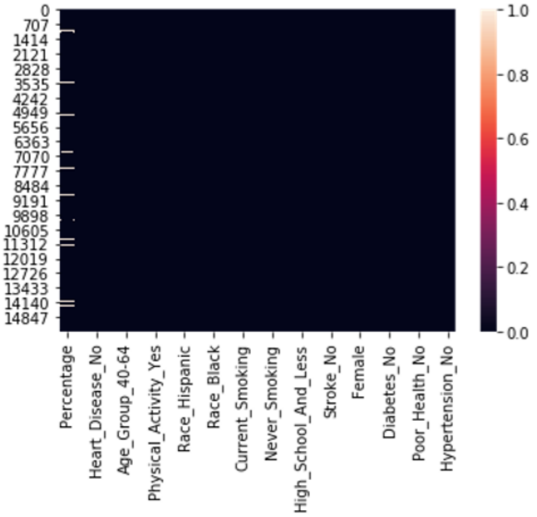
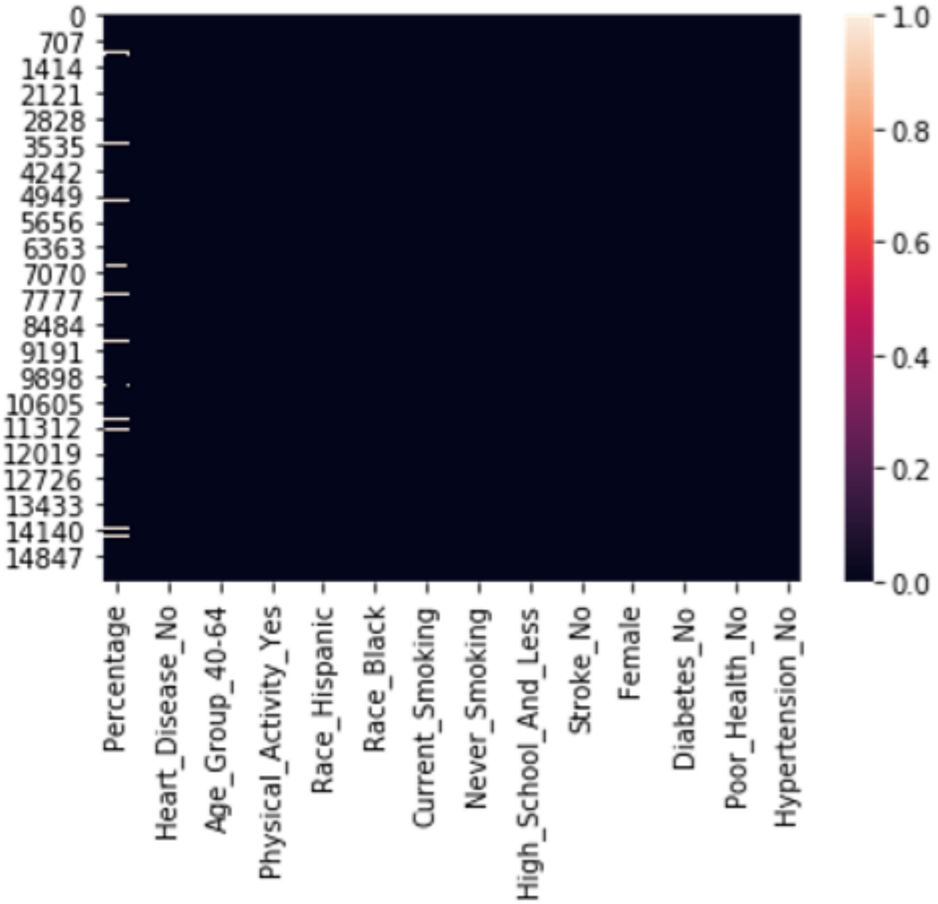
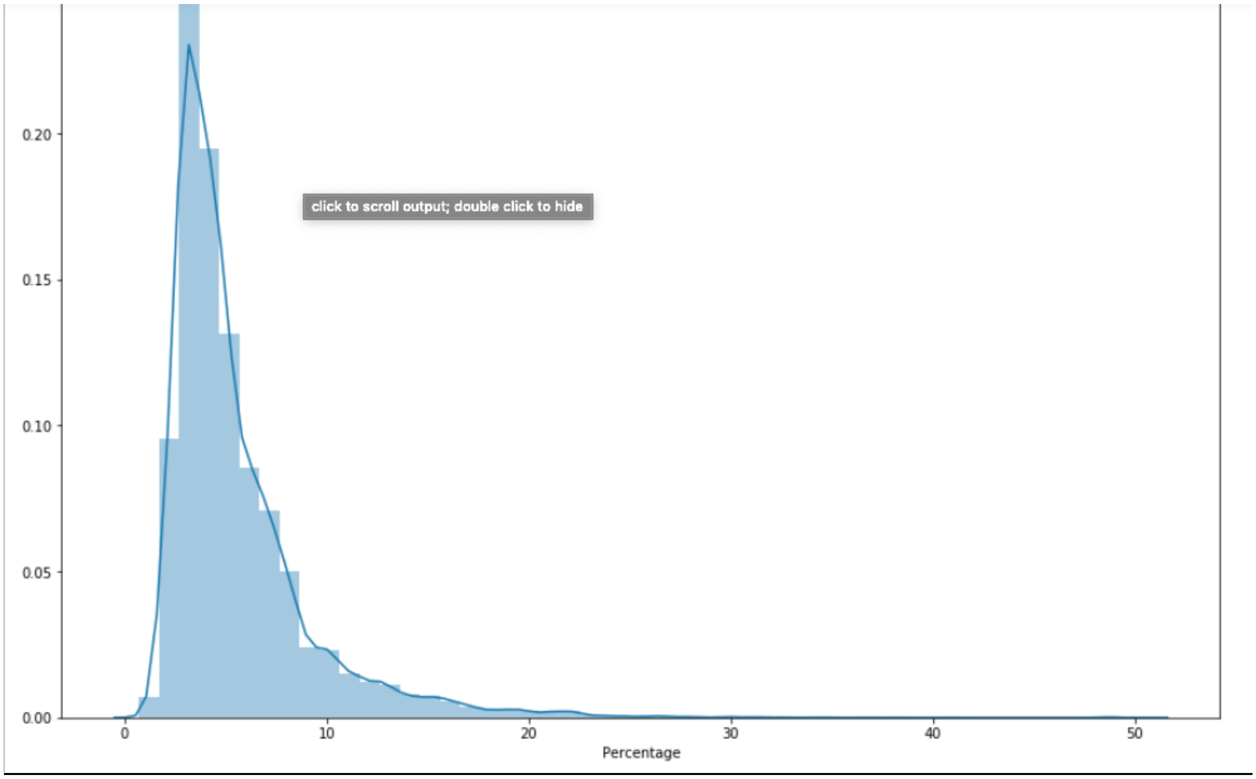


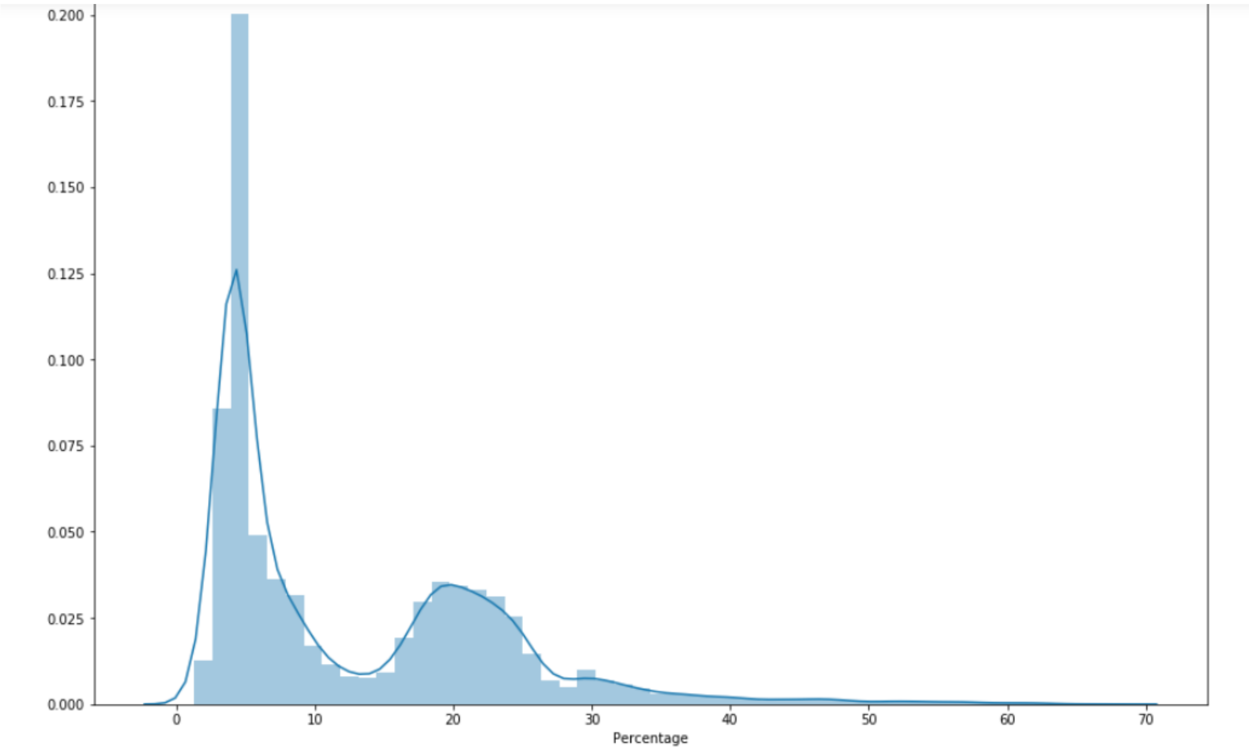
Figure A.2: Before and After of Processing Data for Desired labels and getting rid of null values

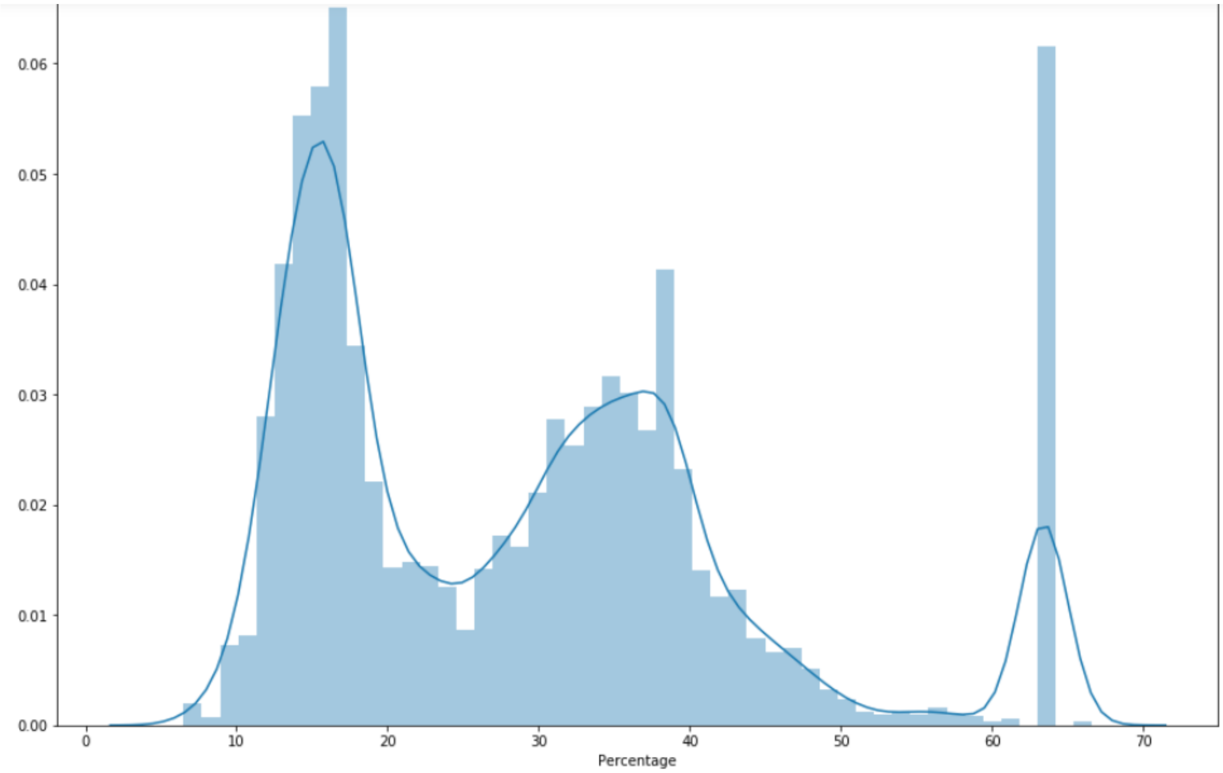

0
2641
5282
7923
10564
13205
15846
18487
21128
23769
26410
29051
31692
34333
36974
39615
42256
44897
47538
50179
52820
55461
Year
LocationAbbr
LocationDesc
Topic
Question
DataSource
Response
Data_Value_Unit
Data_Value_Type
Data_Value
Data_Value_Footnote_Symbol
Data_Value_Footnote
Low_Confidence_limit
High_Confidence_Limit
Sample_Size
Break_Out
Break_Out_Category
GeoLocation
TopicId
QuestionId
LocatioId
BreakOutId
BreakOutCategoryId
ResponseId
1.0
0.8
0.6
0.4
0.2
0.0

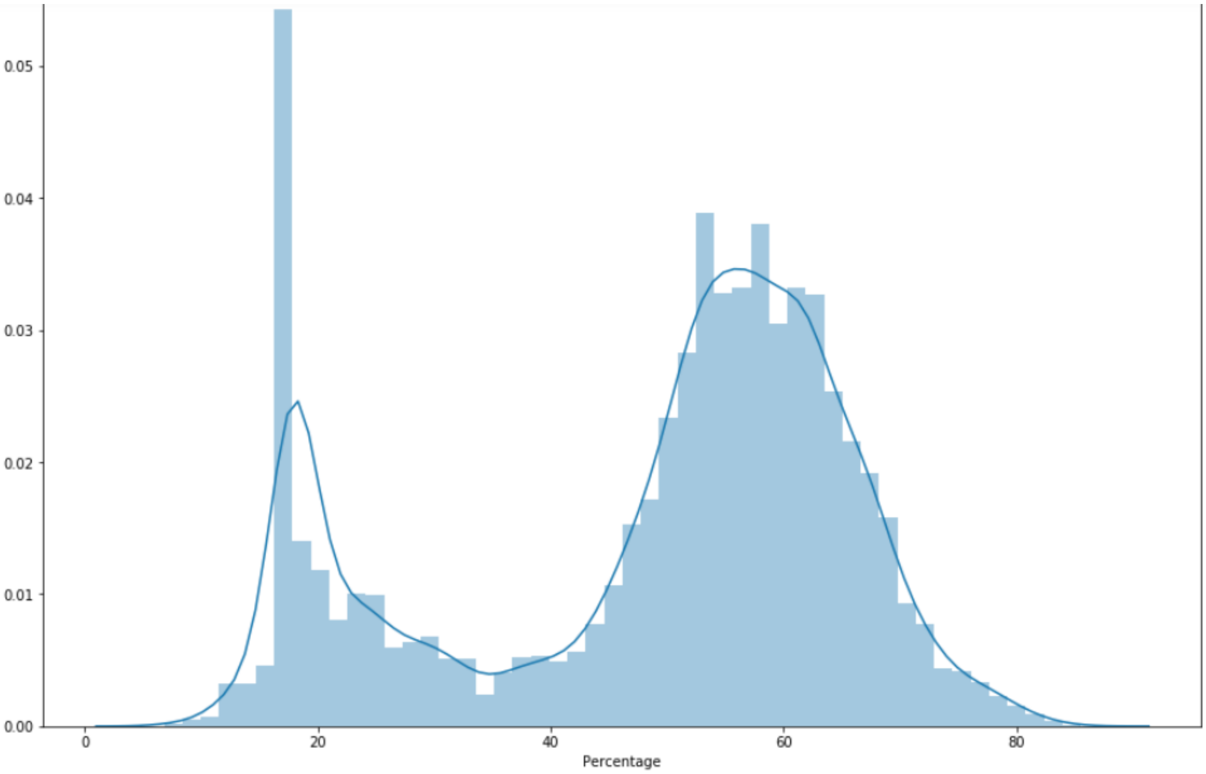

0
707
1414
2121
2828
3535
4242
4949
5656
6363
7070
7777
8484
9191
9898
10605
11312
12019
12726
13433
14140
14847
Percentage
Heart_Disease_No
Age_Group_40-64
Physical_Activity_Yes
Race_Hispanic
Race_Black
Current_Smoking
Never_Smoking
High_School_And_Less
Stroke_No
Female
Diabetes_No
Poor_Health_No
Hypertension_No
1.0
0.8
0.6
0.4
0.2
0.0

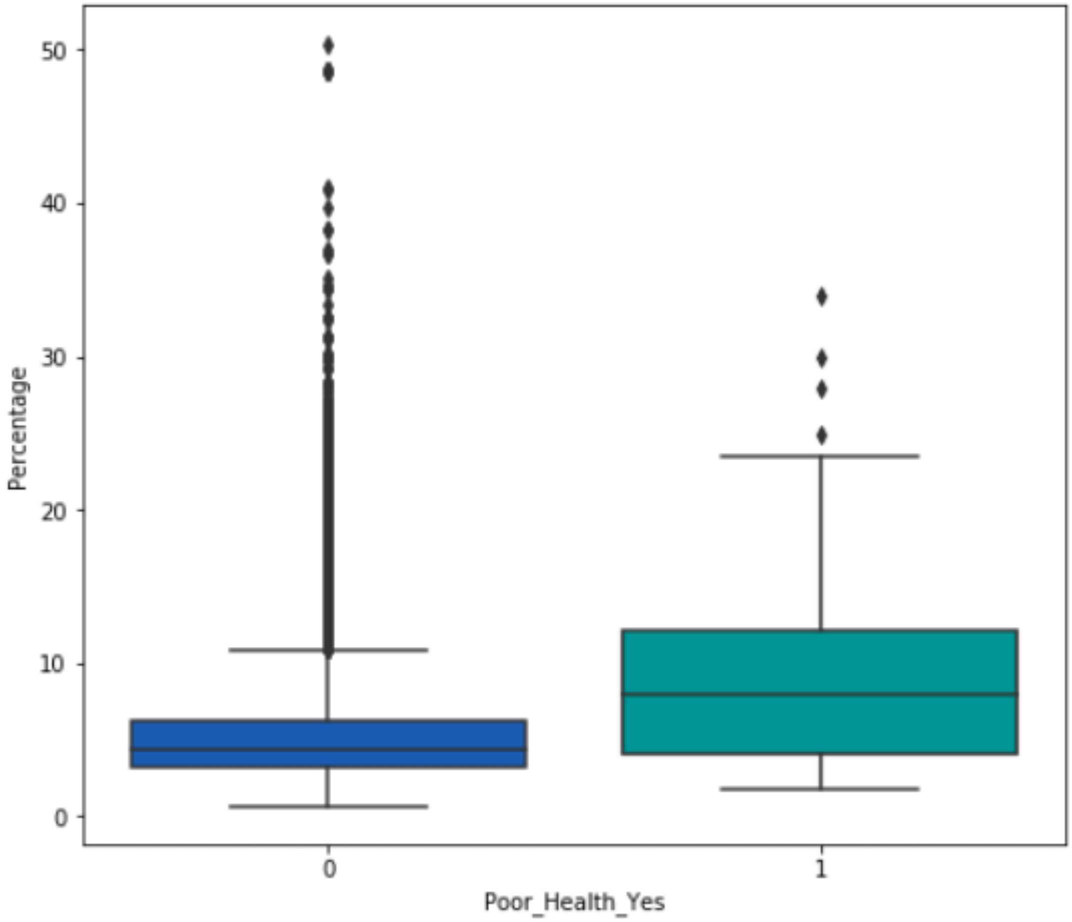

0
707
1414
2121
2828
3535
4242
4949
5656
6363
7070
7777
8484
9191
9898
10605
11312
12019
12726
13433
14140
14847
Percentage
Heart_Disease_No
Age_Group_40-64
Physical_Activity_Yes
Race_Hispanic
Race_Black
Current_Smoking
Never_Smoking
High_School_And_Less
Stroke_No
Female
Diabetes_No
Poor_Health_No
Hypertension_No
1.0
0.8
0.6
0.4
0.2
0.0


0.20
0.15
0.10
0.05
0.00
click to scroll output; double click to hide
0
10
20
30
40
50
Percentage


0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
50
60
70
Percentage


0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
Percentage


0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
Percentage


50
40
30
20
10
0
Percentage
0
1
Poor_Health_Yes