PREVENTATIVE PULSE
PREDICTIVE ANALYTICS FOR HEART HEALTH

Presented by

Tyler Gabriele, John Apel,

Brian Lee, and Richard Dickson

Links to data:
[CDC Heart Disease Indicators]
[CDC BRFSS]
[CDC Questionnaires]

DATA DESCRIPTION

The dataset originally is from the CDC and stems from a Behavioral Risk Factor Surveillance System (BRFSS) questionnaire, who conducts annual telephone surveys to collect data on health of US Residents. We will be using two sets of data, a training test set and a testing set, to run supervised learning to make predictions on new unseen data. For both sets of data, the most relevant variables were chosen. For the training set we will be using the 2020 test data set. This dataset has 18 columns and 319,796 rows and has a mix of numerical and categorical data. The testing set that we will be using, is the 2022 test data with NaN's. This dataset has 40 columns and 445,133 rows with a mix of numerical and categorical data.

Regarding the reliability of the data, these links to the original data source provide an understanding as to where our original findings were generated from. Data organization and cleaning was performed on the original datasets from the CDC website using pandas library in python to select the most relevant variables with an impact on heart disease.

TOPICS FROM QUESTIONNAIRE USED FOR STAKEHOLDERS

The Stakeholders for this project include all medical professionals throughout the US. Since Heart Disease is the number one cause of death for most ethnicities in the US, there's an obvious importance for doctors and nurses to detect early signs within their patience. The questionnaire asks a variety of topics like:

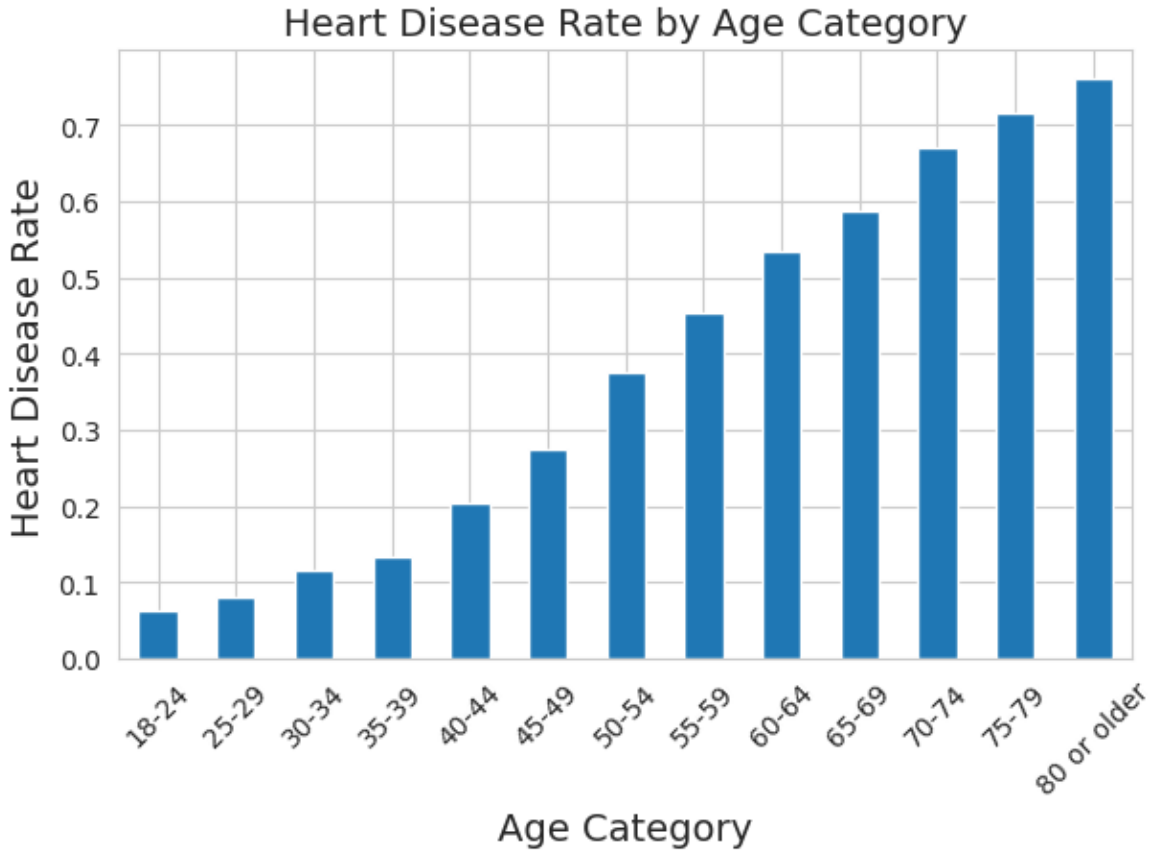
- Patient Medical History
  - Diabetes
  - Kidney Disease
  - Skin Cancer
  - Stroke
  - BMI, Asthma

- Demographics
  - Gender
  - Race
  - Age Group

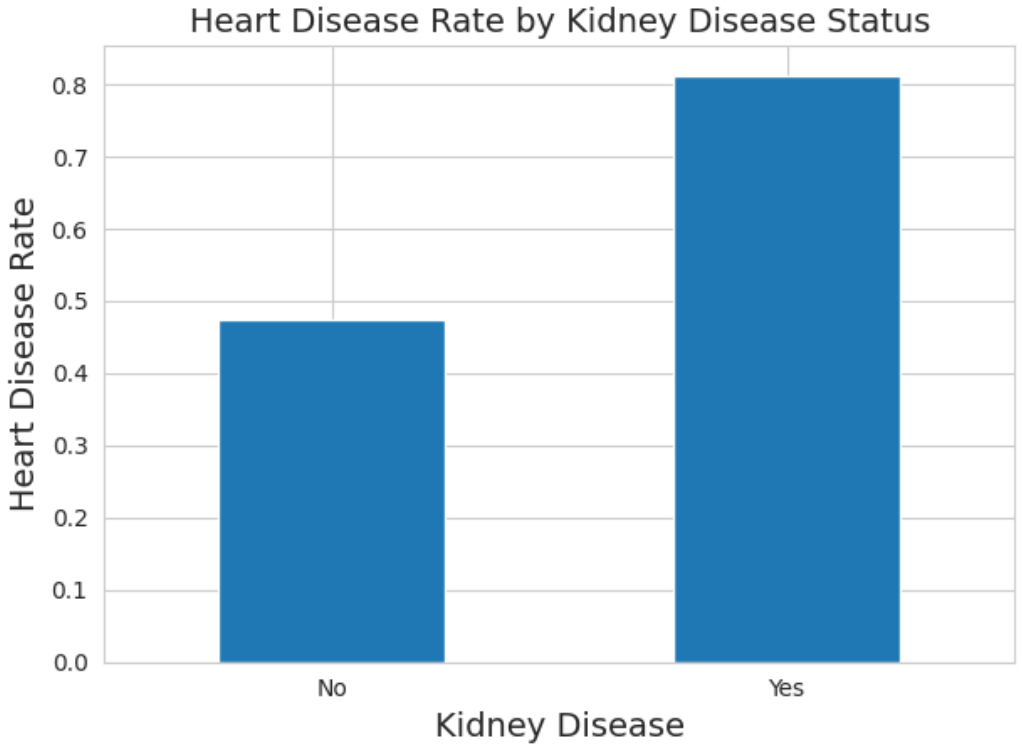
- Patient Lifestyle Habits
  - Smoking
  - Exercise
  - Sleep
  - Alcohol Use

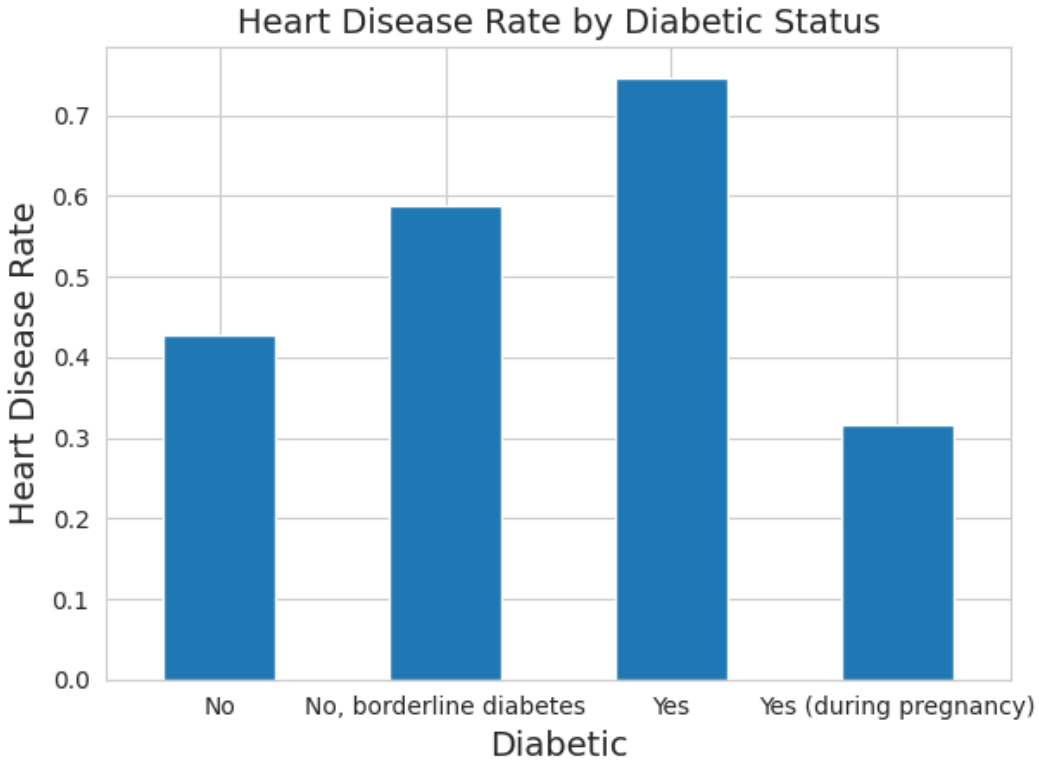
EXPLORATORY DATA ANALYSIS

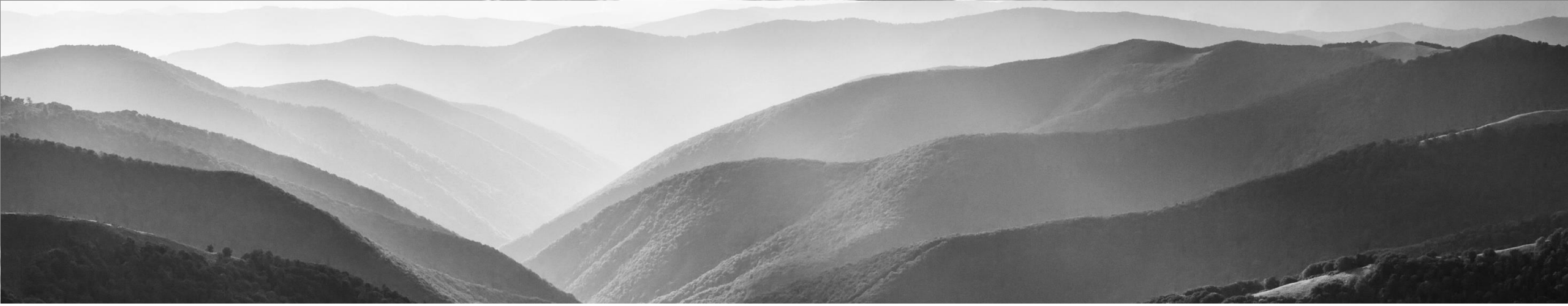
WHEN PLOTTING A FEW OF OUR VARIABLES, COMPARED TO OUR TARGET VARIABLE, IT'S APPARENT HOW MUCH AGE IMPACTS PEOPLE WHO ARE DIAGNOSED WITH A HEART DISEASE CONDITION.

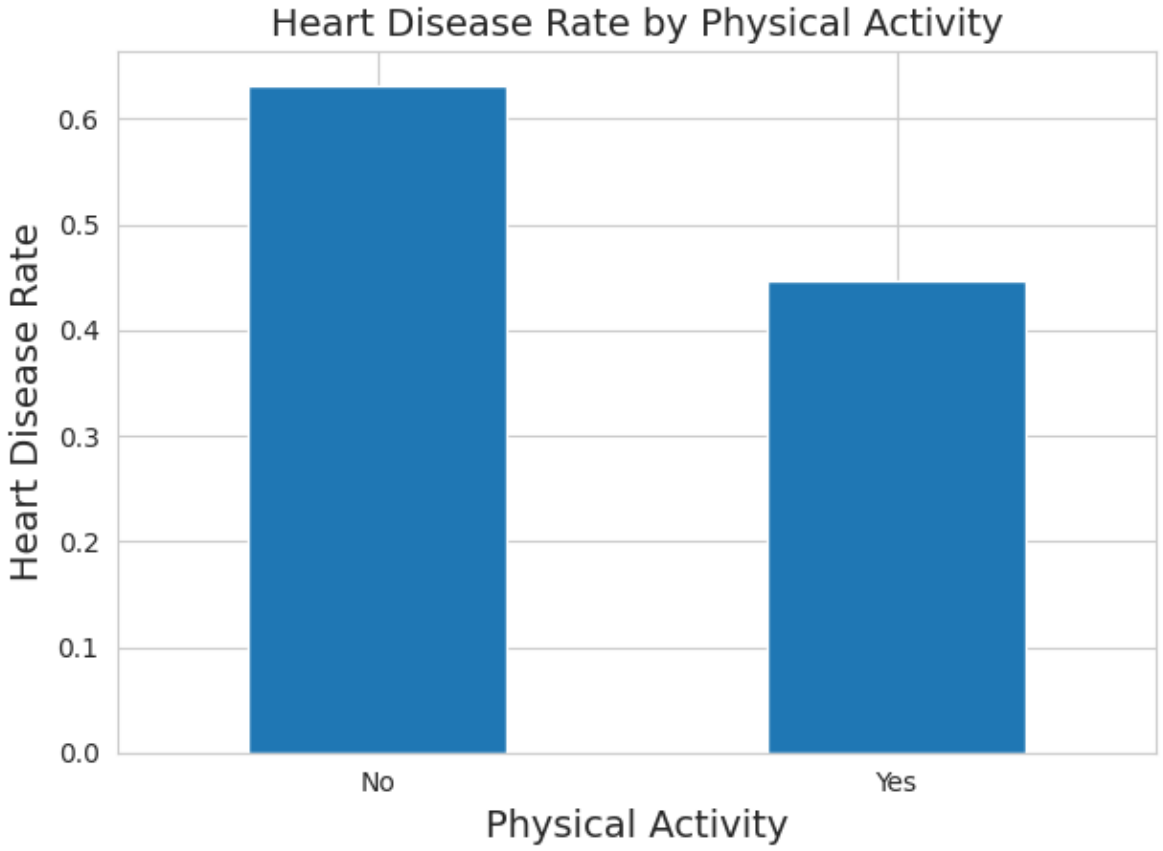


OTHER IMPORTANT FACTORS INCLUDED KIDNEY DISEASE AND DIABETES

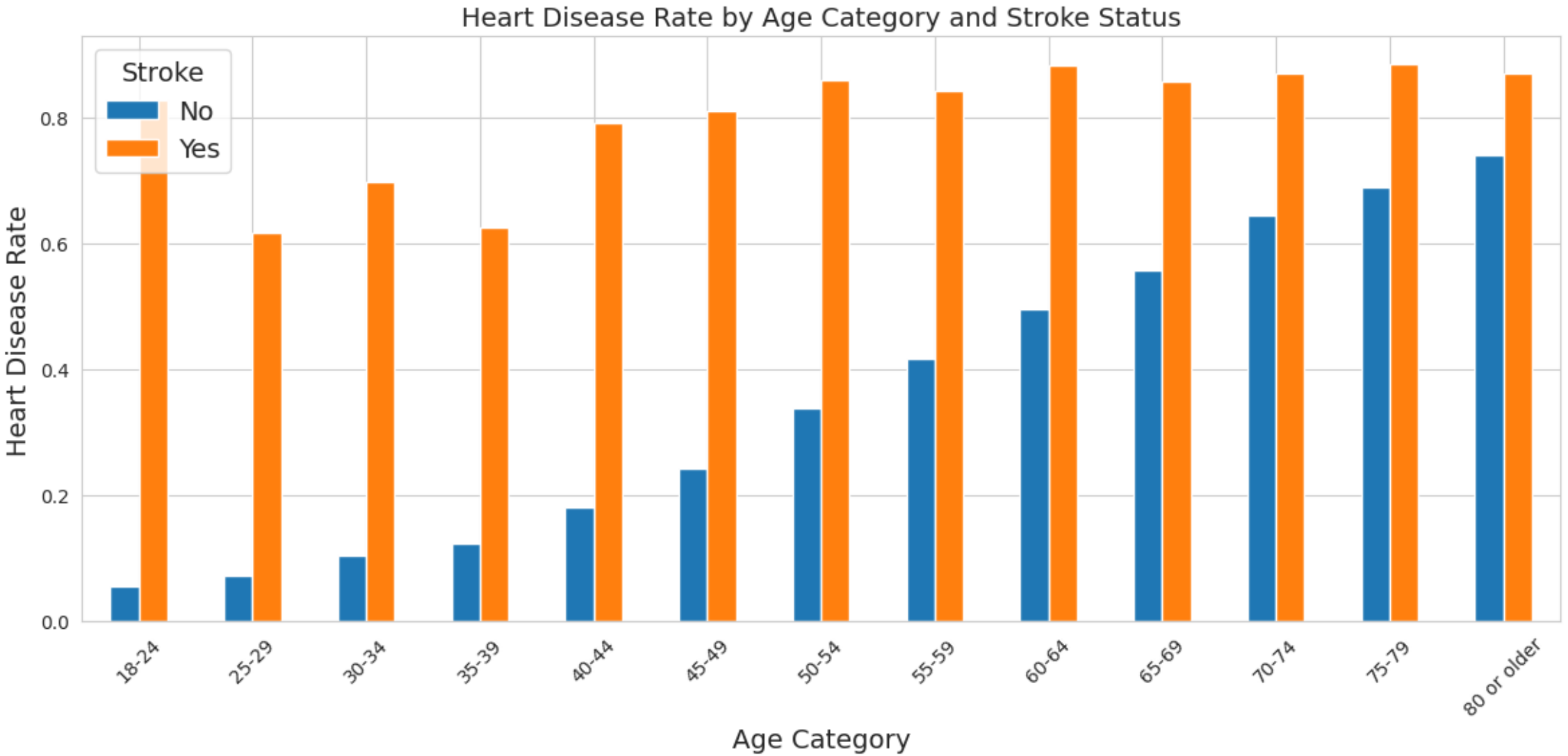








A FEW OF THE VARIABLES THAT WE'RE MEASURING COULD BE INCONCLUSIVE BASED ON THE HONESTY OF THE PATIENTS ANSWERING THE QUESTIONNAIRE



The Age x Stroke interaction with Heart Disease risk increases with Age for both groups. This breakdown reveals that strokes act as an independent risk factor which compounds with age.

METHODS OF ANALYSIS

METHODS & ROADBLOCKS

Going into this project, we wanted to use multiple classifiers in building our model, along with other techniques that include Tuning, Imputation, and Encoding.

After hitting some snags with our F-1 score and seeing a high duration time, we added Gridsearch for Hypertuning, PCA for Dimensionality Reduction along with SMOTE implementation. Also, we quickly observed that our data was heavily imbalanced with no's over yes' within our target variable for predicting heart disease

METHODS & ROADBLOCKS

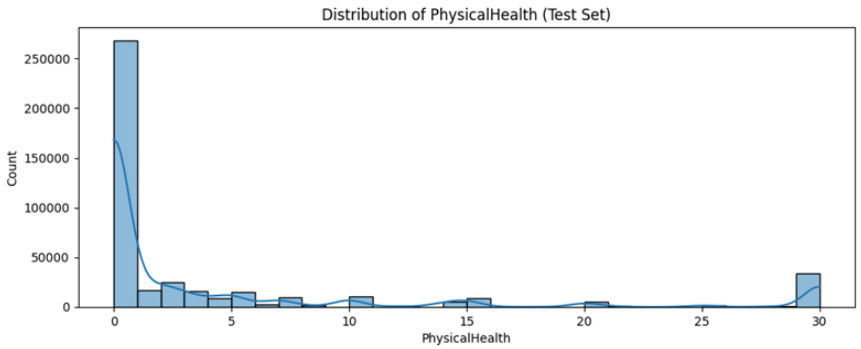
So, our model initially performed well at predicting No's, but not so well at predicting yes's

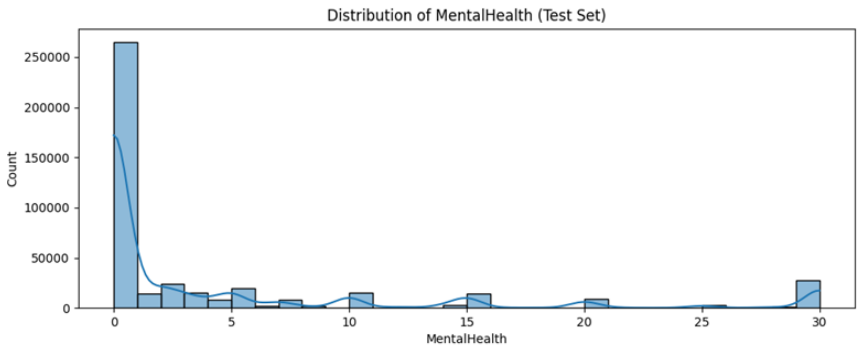
- As shown by our Random Forest Classifier Performance

```
Random Forest Classifier Performance:

Accuracy: 0.896914738378153

Classification Report:
              precision    recall  f1-score   support

          No       0.93      0.96      0.94      8773
         Yes       0.36      0.27      0.31       821

    accuracy                           0.90      9594
   macro avg       0.65      0.61      0.63      9594
weighted avg       0.88      0.90      0.89      9594


Confusion Matrix:
 [[8386  387]
 [ 602  219]]
```

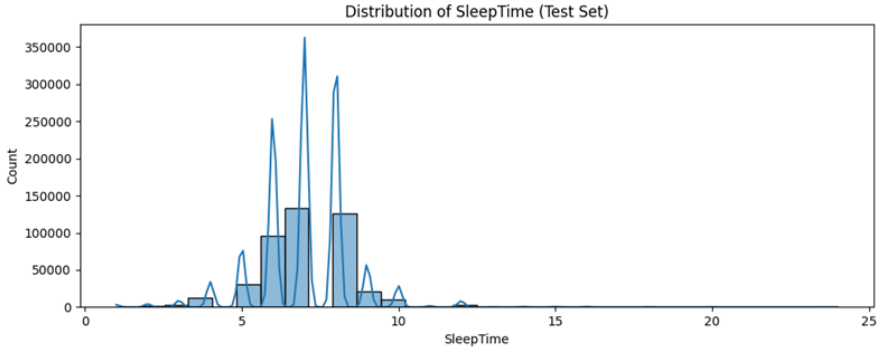
NUMERIC COLUMNS TEST SET

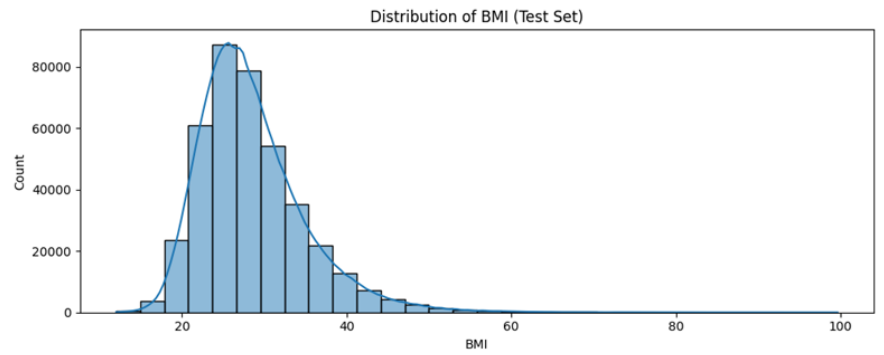
Physical Health Days and Mental Health Days distribution is highly right skewed and zero-inflated (large spike of values at 0)





Sleep time distribution centered around 6-hours , slight right skew, and outliers with a few responses as high as 20 hours which could be unrealistic or input errors. BMI distribution is right skewed with majority of values between 20-35 and peak around 25-28 which is in overweight category





EVALUATION RESULTS OF ML MODELS

Defined 6 models
Model Evaluation Results:

Evaluating Logistic Regression...
F1-Macro: 0.7644 (+/- 0.0036)
Accuracy: 0.7644 (+/- 0.0036)
ROC-AUC: 0.8407 (+/- 0.0025)

Evaluating Random Forest...
F1-Macro: 0.7371 (+/- 0.0039)
Accuracy: 0.7373 (+/- 0.0039)
ROC-AUC: 0.8056 (+/- 0.0042)

Evaluating Gradient Boosting...
F1-Macro: 0.7628 (+/- 0.0038)
Accuracy: 0.7628 (+/- 0.0038)
ROC-AUC: 0.8384 (+/- 0.0023)

Evaluating K-Nearest Neighbors...
F1-Macro: 0.7226 (+/- 0.0031)
Accuracy: 0.7226 (+/- 0.0031)
ROC-AUC: 0.7811 (+/- 0.0038)

Evaluating Decision Tree...
F1-Macro: 0.6673 (+/- 0.0042)
Accuracy: 0.6673 (+/- 0.0042)
ROC-AUC: 0.6678 (+/- 0.0042)

Evaluating XGBoost...
F1-Macro: 0.7571 (+/- 0.0041)
Accuracy: 0.7573 (+/- 0.0041)
ROC-AUC: 0.8315 (+/- 0.0022)
Best model: Logistic Regression
Best F1-Macro score: 0.7644

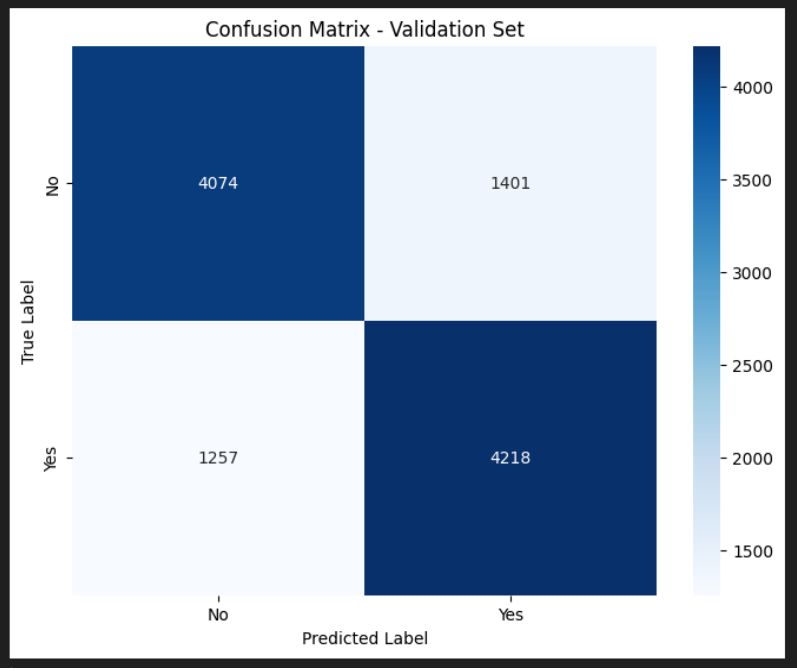
Key Metrics

- F1-Macro: Averages F1-scores across all classes, treating them equally .
- Accuracy: Overall percentage of correctly predicted samples.
- ROC-AUC: a performance metric for binary classification models that measures the model's ability to distinguish between positive and negative classes across all possible classification thresholds.

```
Best parameters: {'classifier__solver': 'liblinear', 'classifier__penalty': 'l2', 'classifier__C': 0.1}
```

Validation Results and Prediction Summary

```
Final Validation Results:
Accuracy: 0.7573
F1-Macro: 0.7572
ROC-AUC: 0.8350

Classification Report:
              precision    recall  f1-score   support

          No       0.76      0.74      0.75      5475
         Yes       0.75      0.77      0.76      5475

    accuracy                           0.76     10950
   macro avg       0.76      0.76      0.76     10950
weighted avg       0.76      0.76      0.76     10950


Confusion Matrix:
[[4074 1401]
 [1257 4218]]
```



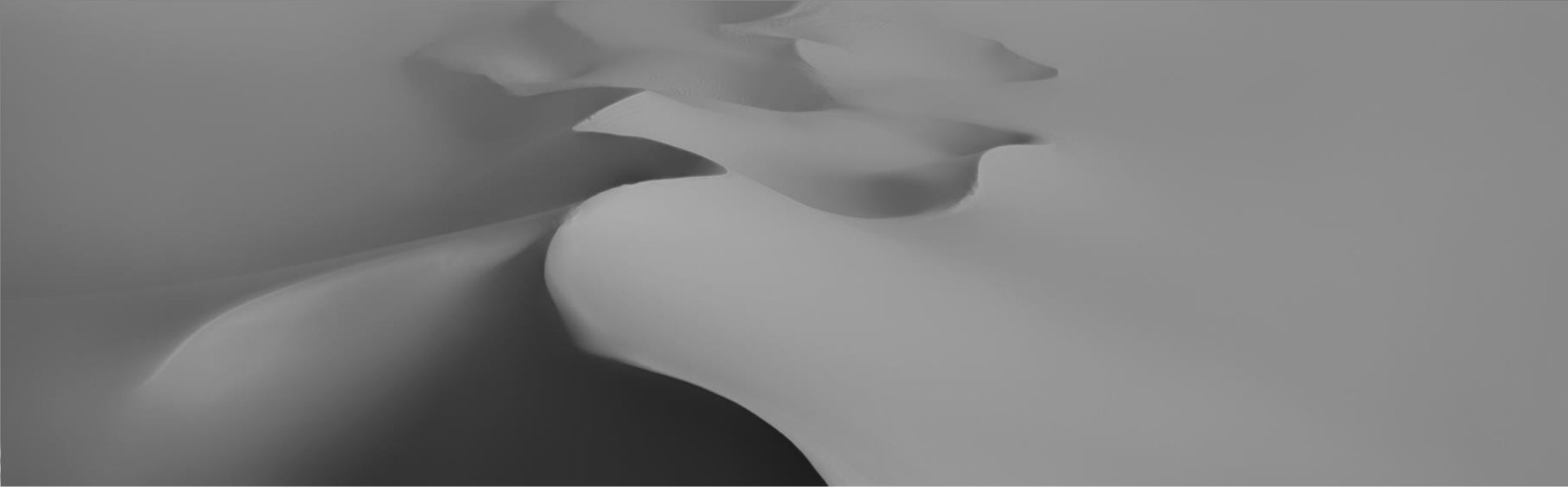

```
Prediction Summary:
Total predictions: 445132
Predicted positive cases: 31089
Predicted negative cases: 414043
Positive rate: 0.070

Predictions saved to 'heart_2022_with_predictions.csv'
```

CONCLUSION/RESULTS

Overall, as a group we set out to develop an accurate and true machine learning model to predict if people were to have heart disease. As presented in the presentation we were able to develop a model with an accuracy of above 75% and a F1 score above 75%. Using this we were able to distinguish on our test data out of 445,132 users that 31,089 did have heart disease. The 2022 dataset contained many more variables used to determine heart disease compared to the 2020 dataset which will lead to eventually a more thorough model when more data is eventually provided by the CDC.

TEAM ROLES

Name	Description	Notes
John Apel	Data Visualization/ Reporting	Data Visualization, Reporting
Tyler Gabriele	Data Collection/Program Editor	Working on Training dataset/ finding the data
Richard Dickson	Data Visualization/ Reporting	Data Visualization, Reporting
Brian Lee	Program Editor	Working on Test dataset/ develop model



THANK YOU

Presented by Team Preventative Pulse