Heart Disease Analysis Report

ISDS 7070 - Group 1

Mitchell Montanio

Erin Vasut

Justin Joseph

Frame the Problem

- **Overview**
  - Nation's leading cause of death (1-in-4)
  - Age, gender, health history, etc.
- **Objective**
  - Create a model that accurately predicts whether any given person in this dataset has heart disease.

The 3 Q's

1. Are age and gender good predictors of TenYearCHD?
2. What model can create the best predictions?
3. Does health history play a significant role in causing heart disease? If so, what specific aspects of health history are most significant?

Explore the Data

Initial Dataset Observations

- 16 columns
- 4,238 rows
- 582 rows containing nulls
- TenYearCHD = Coronary Heart Disease

Missing Details in Dataset

- Date/Time
- Race/Ethnicity
- Null values

Wrangle the Data

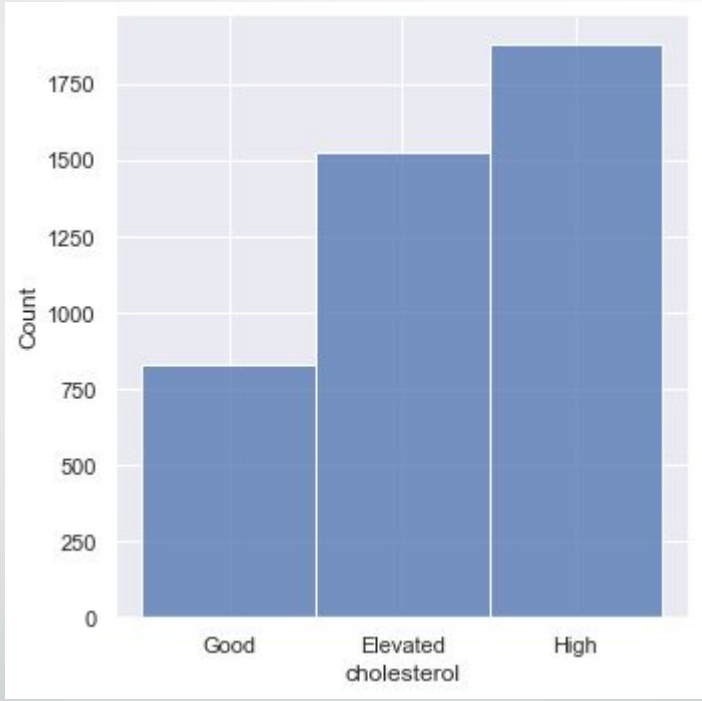
Four data frames were created to explore different methods of of data wrangling:

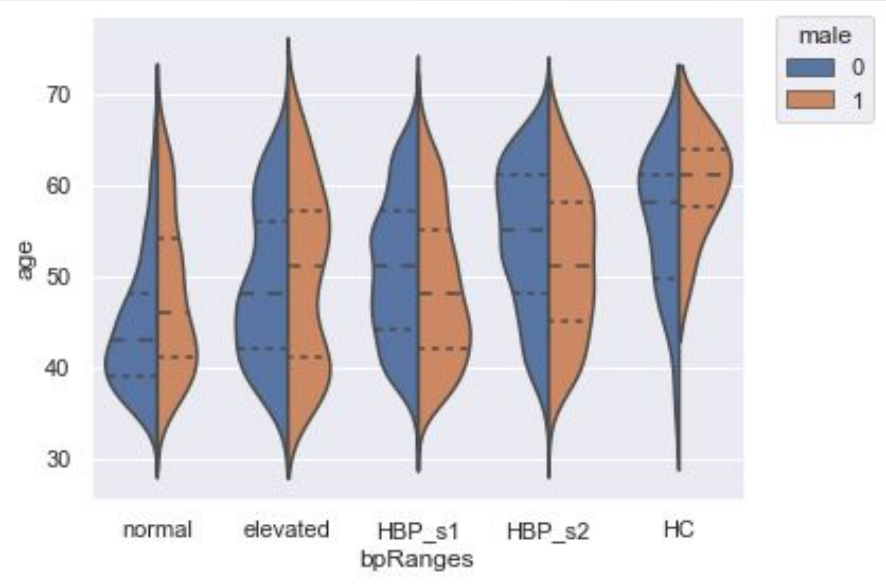
- **df**: Nulls were replaced with either mean or mode of the column
- **df2**: Rows containing nulls were dropped
- **df3**: Copy of df2, but columns containing strings were binned, then bins were converted to numeric values
- **df4**: Copy of df, but all columns containing strings were dropped

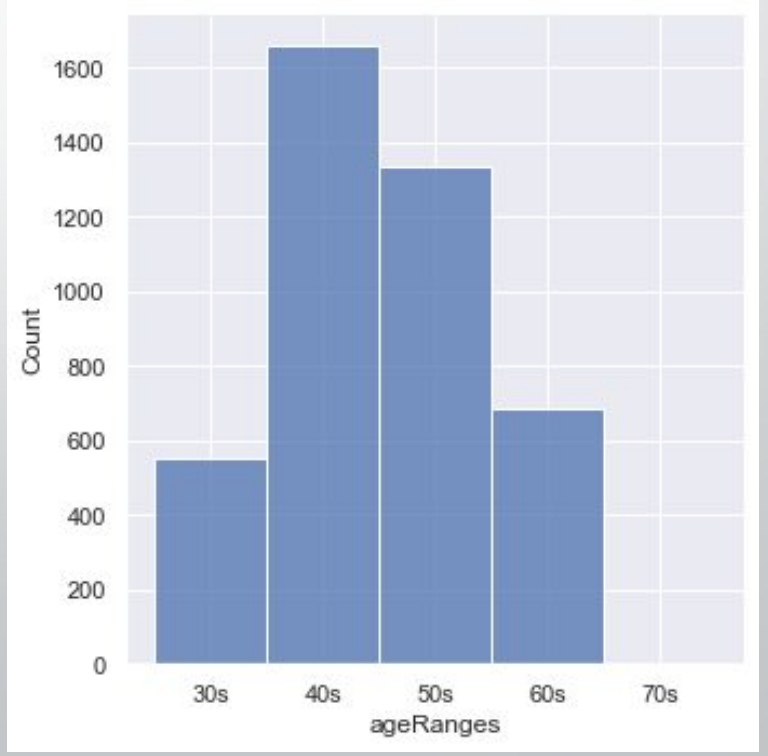
Wrangle the Data *cont.*

New columns:

- ageRanges
- cholesterol
- bpRanges







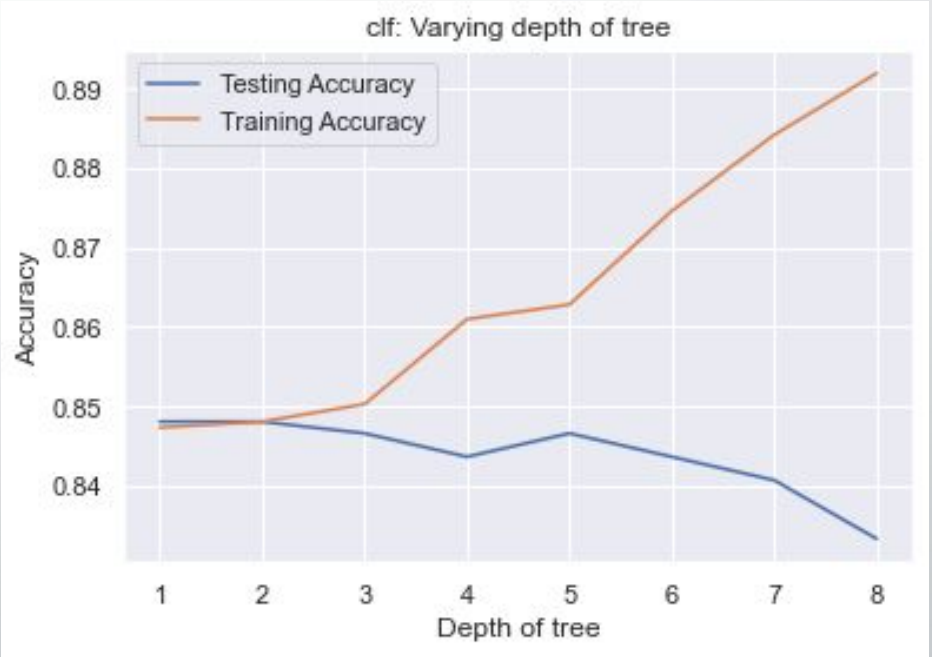
Linear Regression Model

- Not effective at predicting TenYearCHD
- Predictor variable (x): diaBP
- Response variable (y): sysBP
- Mean Squared Error (MSE): 163.781
- Coefficient of Determination (COD): 0.6232

*All models used a default train:test ratio of 80:20

Decision Tree Model

- Best data frame: df4
- Response variable: TenYearCHD
- Depth: 2
- Hit rate: 85.2%



*All models used a default train:test ratio of 80:20

Random Forest Model

Default Parameters:

- Estimators: 100
- Depth: 25
- Predictors: 12
- Best data frame: df/df4

Results:

- COD: 65.5%
- MSE: 0.044

Top Five Predictors:

1. sysBP
2. age
3. diaBP
4. prevalentHyp
5. glucose

*All models used a default train:test ratio of 80:20

Conclusion

- The 3 Q's
- Best model: decision tree
- Best predictors: sysBP, age, diaBP, prevalentHyp, glucose
- Best data frame and wrangling method: df/df4, fill nulls
- Least effective: linear regression, dummy variables