SI 370 Project Proposal
October 22, 2024
Allison Lee, Christina Ng, Annabel Zhuang, Aren Shah

Dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
Summary:
The data comes in a csv file. The size is 1190 rows, specifically 918 non-duplicate (unique) rows. The columns are as follows:

- Age: numerical
- Sex: text/categorical
- ChestPainType: text/categorical
- RestingBP: numerical
- Cholesterol: numerical
- FastingBS: numerical
- RestingECG: text/categorical
- MaxHR: numerical
- ExerciseAngina: text/categorical
- Oldpeak: numerical
- ST_Slope: text/categorical
- HeartDisease: numerical

Exploratory Questions:

1. **What variable is the strongest predictor of heart failure?**
    a. an exploratory data analysis method: linear regression
    b. a possible visualization method: Seaborn - scatterplots of each variable with best fit lines (heart failure on y axis)
2. **What proportion of people with heart failure have chest pain?**
    a. an exploratory data analysis method: chi-square test
    b. a possible visualization method: Matplotlib - show a bar graph and see if results are statistically significant with error bars
3. **How can cholesterol predict heart failure?**
    a. an exploratory data analysis method: Machine Learning
    b. a possible visualization method: Matplotlib - show a bar graph and see if results are statistically significant with error bars
4. **What is the optimal blood pressure based on the heart failure outcomes? Does this match with the recommendations available online?**
    a. an exploratory data analysis method: linear regression
    b. a possible visualization method: scatterplot with blood pressure vs heart failure outcomes