Name: Elias Fedai

Date: 07/28/19

Title: Analysis of Heart Disease

Section 2

- **How to import and clean my data**.
  In my research project it was very crucial to ensure that the datasets I imported be cleaned. I used the read.csv to import my sets and then combined all three data sets to form one data set. I skipped the header because I wanted to label them myself. Once the import was done, I needed to clean the data. The purpose of this was to remove any unwanted space, data, outlier or anything that could possible mislead or obscure my results and then to modify predictor variables.

  Note: upon further analysis and evaluation I have decided to replace my third data set https://healthdata.gov/dataset/national-health-interview-survey-nhis-nationalcardiovascular-disease-surveillance-data to the Swiss heart disease dataset https://archive.ics.uci.edu/ml/datasets/heart+Disease. The data from the swiss data set will give me more valuable data results in regards to my research questions and harmonizes much better with the other two data sets.

- **What does the final data set look like?**

RStudio Source Editor

combinedclean

Filter

	Age	Gender	CP	Trestbps	Chol	FBS	RestECG	Thalach	Exang	Oldpeak	Slope	CA	Thal	target	Disease
1	63	Male	typical angina	145	233	true	V.hypertrophy	150	no	2.3	downsloping	0	fixed defect	0	No Disease
2	67	Male	asymptomatic	160	286	false	V.hypertrophy	108	yes	1.5	flat	3	normal	1	Heart Disease
3	67	Male	asymptomatic	120	229	false	V.hypertrophy	129	yes	2.6	flat	2	reversable defect	1	Heart Disease
4	37	Male	non-anginal pain	130	250	false	normal	187	no	3.5	downsloping	0	normal	0	No Disease
5	41	Female	atypical angina	130	204	false	V.hypertrophy	172	no	1.4	upsloping	0	normal	0	No Disease
6	56	Male	atypical angina	120	236	false	normal	178	no	0.8	upsloping	0	normal	0	No Disease
7	62	Female	asymptomatic	140	268	false	V.hypertrophy	160	no	3.6	downsloping	2	normal	1	Heart Disease
8	57	Female	asymptomatic	120	354	false	normal	163	yes	0.6	upsloping	0	normal	0	No Disease
9	63	Male	asymptomatic	130	254	false	V.hypertrophy	147	no	1.4	flat	1	reversable defect	1	Heart Disease
10	53	Male	asymptomatic	140	203	true	V.hypertrophy	155	yes	3.1	downsloping	0	reversable defect	1	Heart Disease
11	57	Male	asymptomatic	140	192	false	normal	148	no	0.4	flat	0	fixed defect	0	No Disease
12	56	Female	atypical angina	140	294	false	V.hypertrophy	153	no	1.3	flat	0	normal	0	No Disease

Showing 1 to 14 of 298 entries, 15 total columns

Note: displayed in the above image is only a partial/condensed image of the total data set, showing all 15 columns. Total clean data set contains 298 observations.

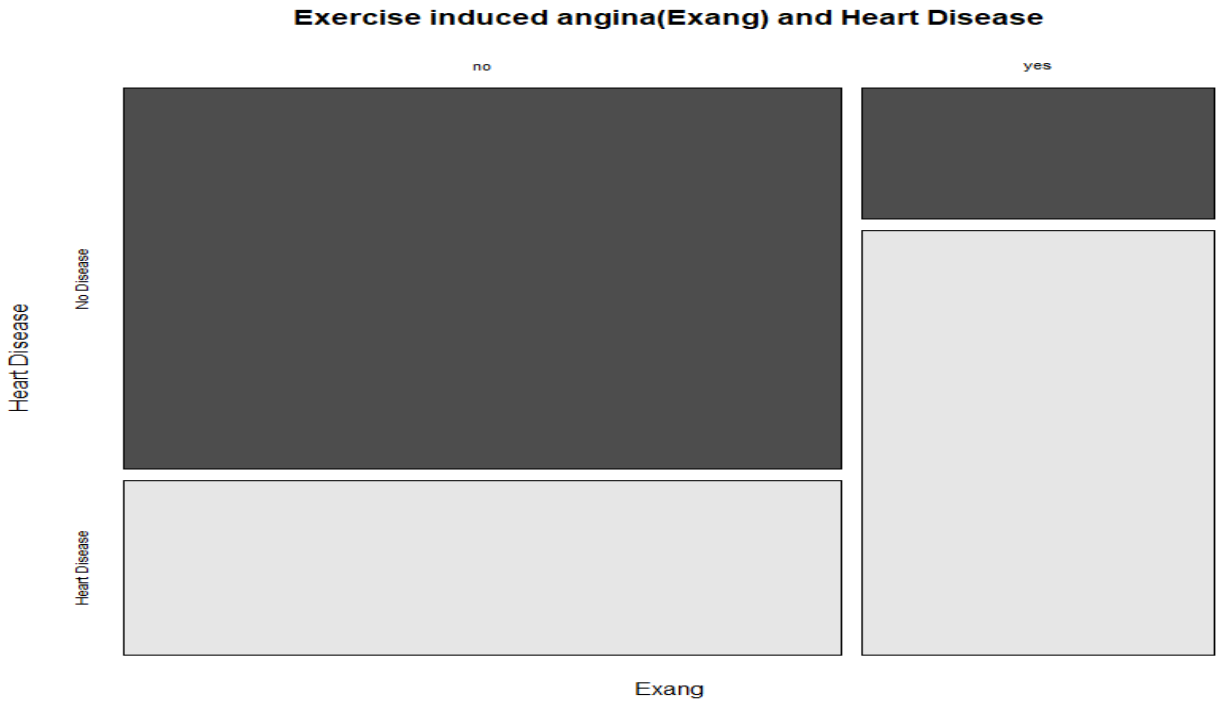
- **What information is not self-evident?**
  From the information thus far, I can get a general idea of the distributions, and possible skewness of the variables, but none of my research questions are evident. Further analysis is needed.

- **What are different ways you could look at this data?**
  There are several ways to look at this data. As an entirety or by specific variables. In my case I'm going to first address the data set as whole, and then I will begin addressing those variables in my research questions by reviewing each variable individually(univariate), then proceeding to combined(multivariate), and then proceed into a regression analysis.

- **How do you plan to slice and dice the data?**
  Some of the data in the visualization portion will be split by class of the target features needing to be evaluated per research questions.

- **How could you summarize your data to answer key questions?**
  One of the first and most frequent used feature for me is the summary stat to gain insight on how to handle my data in regards to my goal or research questions (descriptive stats.) I then will proceed to summarize the data by the different types of tables and plots that are in relation to my specific research questions (visual data.)

  What types of plots and tables will help you to illustrate the findings to your questions?
  Various plots (i.e. Bar, box, histograms, scatter, etc.), tables. I will incorporate univariate and multivariate into plots to attain better visual data. Below is an example plot.



- **Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**
  Yes, during my initial observation using my plots, summaries, and tables I established a baseline of which variables may display some likelihood of prediction of heart disease. I will be using regression analysis to complete my research questions.

- **Questions for future steps.**
  1) Which type of regression analysis do I use?
  2) Which variables would I incorporate into my analysis?