Project Milestone 2

Data Cleaning

Data cleaning

1. 1.1 Handling missing data
   1. The steps used
      * Selected the dataset and used the “missing data handling” feature under the “transform data” tab. We then selected all the variables and chose to “delete” any records with missing data.

1.2 Random sampling (if you get a large dataset, because the limitation of XLMiner is 10,000 records.)

* + 1. Describe the steps you used
       - Our dataset contained less than 1000 records, so we did not have to use random sampling.
       - Files have been Uploaded

1.3. Identify numerical variables vs. categorical variables’

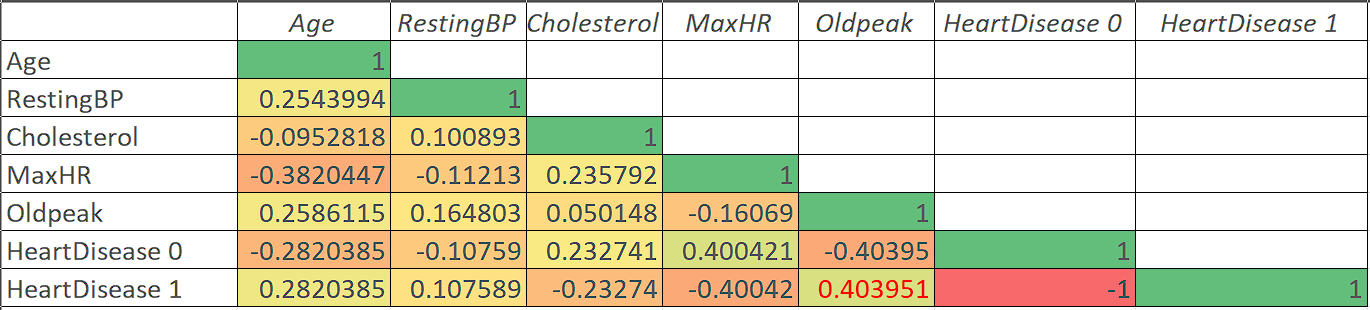
* + - We went through the values under each variable and categorized all the discrete, ordinal, and nominal as categorical variables so we transformed each of them into dummies to get effective results after running a regression analysis and categorized all the variables with continuous ranges as numerical variables.
    - So, our **numerical variables** are **Age**, **RestingBP**, **Cholesterol**, **MaxHR** and **Oldpeak** and our **categorical variables** are **Sex**, **ChestPainType**, **FastingBS**, **RestingECG**, **ExerciseAngina** and **ST\_Slope.**

1.4 Based on your questions, which variable is the dependent variable? Which variables could be used as independent variables? (If any variables in your original dataset is not used in your analysis, please list your reasons)

* + Dependent Variable: **HeartDisease**
  + Independent Variables: **Age, RestingBP, Cholesterol, MaxHR, Oldpeak, Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST\_Slope**

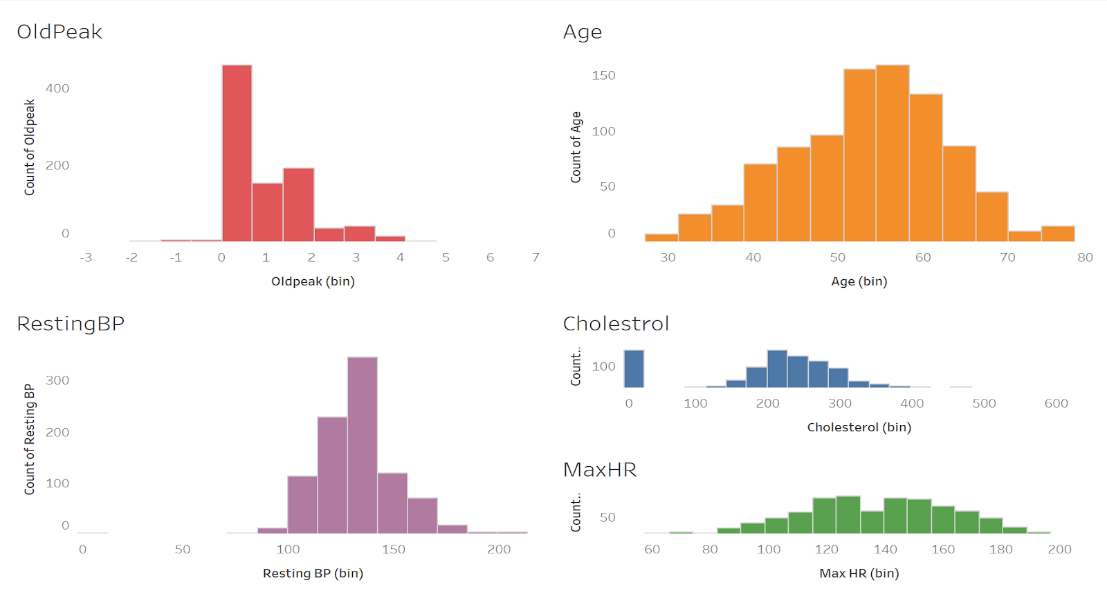
1.3. Data visualization

1.3.1 Correlation table of numerical variables,



* + - Based on the correlation matrix, the occurrence of a heart disease is highly correlated by the patients’ **OldPeak** value followed by their **Age.** "Oldpeak" measures how much the heart's electrical activity changes during exercise due to lack of blood supply, providing valuable information for diagnosing heart disease.

1.3.2 Histogram of numerical variables (choose at least three variables you are most interested),

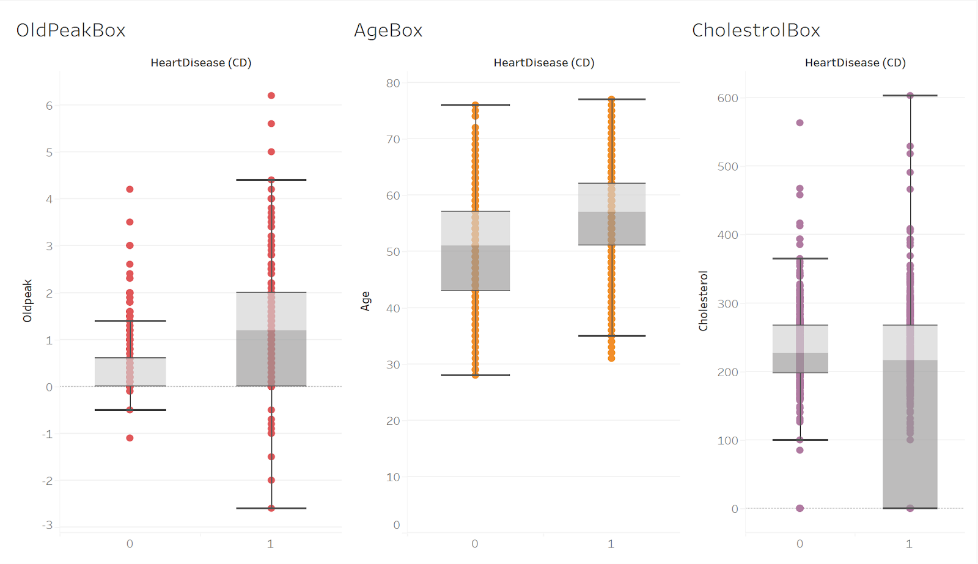
****

* + - All the distributions look like normal distributions so there does not seem to be a requirement for a log transformation.

1.3.3 Scatterplot of the Dependent variable (if your DV is a numerical variable) vs independent (choose at laeast 3 variables you are most interested in, comment on relations)

* + - No possibility for a scatter plot.
    - Our dependent variable is a categorical variable.

1.3.4 Boxplot of the Dependent variable (if your DV is a categorical variable) vs independent variables (choose at least three variables you are most interested)

****

* + - For the Old Peak Box, the median is much higher in the 1 group than the 0 group.
    - For age, those who have heart disease have a higher median of people, compared to those who do not have CD. Also, the variability is relatively the same.
    - For Cholesterol, the medians are about the same but those who have heart disease, have a much more variability in cholesterol.
    - For Old Peak those with CD have a higher variability as well. Overall, the boxplots of those with a CD are higher in variability and range overall, besides age.

1.3.5 From the previous plots, detect outliers and find out whether they are errors or extreme values

* + The outliers lie outside of the whiskers on the boxplot.
  + Both Old Peak and Cholesterol contain outliers. The outliers for both are not errors but some are just higher values than the average person. Those cholesterol levels are possible and can occur in individuals, as well as the Old Peak outliers.

2. In general, are there any potential issues with the dataset? Like if you could get data from the company directly, how would you extend current dataset?

* 1. (We have not experienced any issues with the data so far because the dataset we found was clean to start with and the number of variables taken into consideration seem to be sufficient to build a prediction model by performing logistic regression. We might come across possibilities as we work on the dataset further which we can document along the way.)