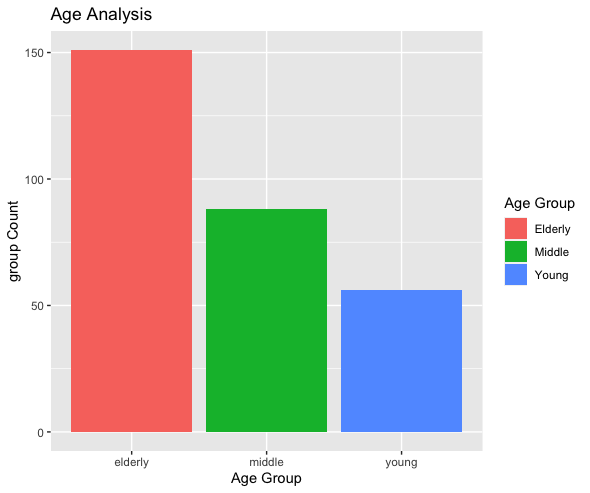


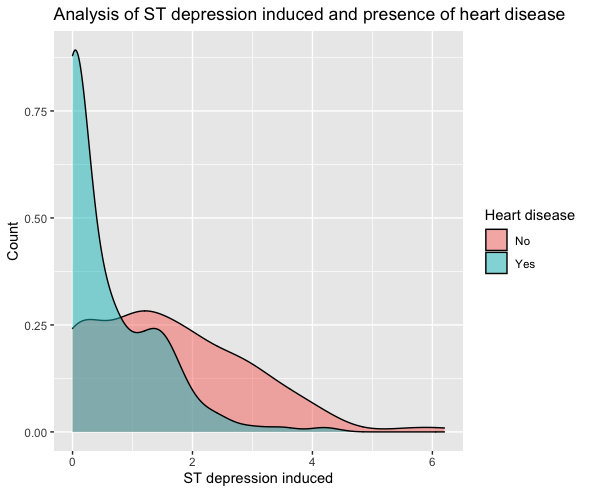
| # Displaying the correlation matrix corr <- cor(heart\_df) # Visualize the correlation matrix corrplot(corr) |
| --- |

The correlation map helps when measuring the extent of relationship between multiple variables in one given figure. This helps when trying to decide which features to remove when trying when going about feature reduction for better statistical analysis and modeling. We can see from the given correlation plot that our dependent features are least related with fbs, chol, trestbps, restecg because they are between -0.2, and 0.2.



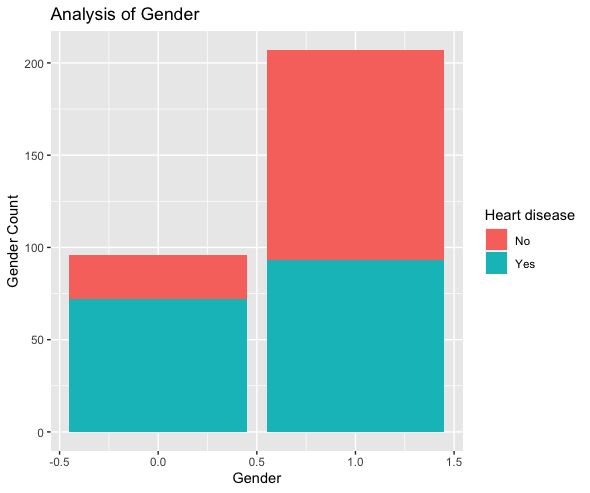
| # Group the different ages in three groups (young, middle, old) young <- heart\_df[which((data$age<45)), ] middle <- heart\_df[which((data$age>=45)&(data$age<55)), ] elderly <- heart\_df[which(data$age>55), ] groups <- data.frame(age\_group = c("young","middle","elderly"), group\_count = c(NROW(young$age), NROW(middle$age), NROW(elderly$age)))  # Plotting different age groups ggplot(groups, aes(x=groups$age\_group, y=groups$group\_count, fill=groups$age\_group)) +   ggtitle("Age Analysis") +  xlab("Age Group") +  ylab("group Count") +  geom\_bar(stat="identity") +  scale\_fill\_discrete(name = "Age Group", labels = c("Elderly", "Middle", "Young")) |
| --- |

I converted feature age of continuous data into categorical data of young, middle-aged, and elderly. Doing so also gave it a more clean look and simplified look of the given age data. There seems to be more elderly patients than both middle aged and young grouping. This gives us a general look at what kind of patients we are looking at in terms of age.



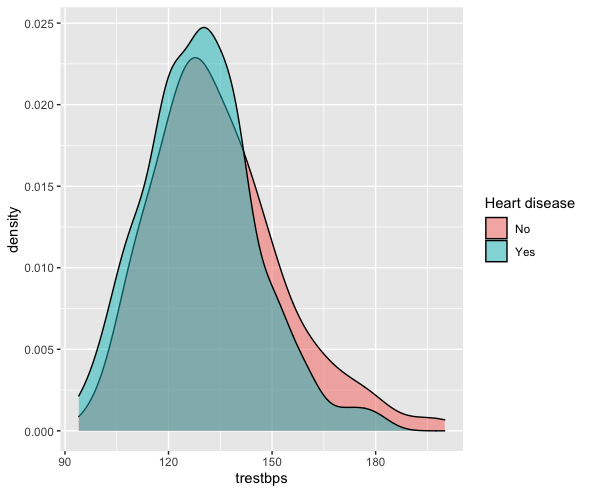
| ggplot(heart\_df, aes(x = oldpeak, fill = target)) +  geom\_density(alpha=0.5) +  xlab("ST depression induced") +  ylab("Count") +  ggtitle("Analysis of ST depression induced and presence of heart disease") +  scale\_fill\_discrete(name = "Heart disease", labels = c("No", "Yes")) |
| --- |

From the above graph, we can see that as the value of the old peak increases the chances of heart disease decreases. Thus highlighting the importance of this given feature when creating a statistical model or running any basic statistical data. I am surprised to see how this feature seems to impact the target feature of heart disease.



| # Analysis of gender ggplot(heart\_df, aes(x= heart\_df$sex, fill=heart\_df$target)) +   geom\_bar() +  xlab("Gender") +  ylab("Gender Count") +  ggtitle("Analysis of Gender") +  scale\_fill\_discrete(name = "Heart disease", labels = c("No", "Yes")) |
| --- |

When thinking about gender I would expect more males to have greater heart disease due to the fact they are generally physically bigger individuals and therefore their heart strain must be greater and thus increasing the chances of disease. But upon looking at the data for gender when compared against heart disease, we see that our dataset has more females then males, nearly a 1 to 2 ratio, making our data imbalanced for this feature and as a result we can not use this is untrustworthy for decision making when related to heart disease.



Looking at the graph above we can see very little change between people with heart disease and those without in terms of high and low resting blood pressure. Thus the data shows that looking at a patient’s individual resting blood pressure is not an effective way of determining heart disease.

| ggplot(heart\_df, aes(x = trestbps, fill = target)) +  geom\_density(alpha=0.5) +  scale\_fill\_discrete(name = "Heart disease", labels = c("No", "Yes")) |
| --- |