Data Visualization and logistic regression with R

Data set:

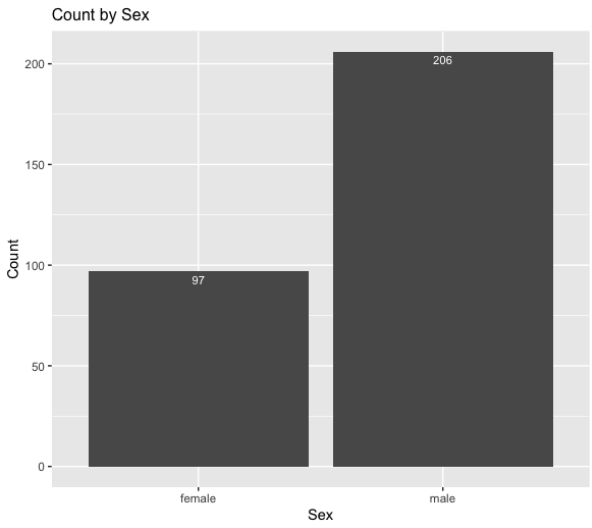
https://archive.ics.uci.edu/ml/datasets/Heart+Disease (processed.cleveland.data)

Replace value based on condition

```
processed.cleveland$V14[processed.cleveland$V14 != 0] <- 1
```

Count Plot

```
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

#Count plot
ggplot(count_df, aes(x = V2, y = n)) +
  geom_col(position = "dodge") + #group bar plot
  theme(plot.title = element_text(size=12)) + #white background, title size
  labs(title="Count by Sex", y="Count",x="Sex") + #change x, y axis title
  scale_x_discrete(labels=c('female', 'male')) + #change x axis labels
  geom_text(aes(label = n),
            colour = "white", size = 3,
            vjust = 1.5, position = position_dodge(.9)) #bar labels
```

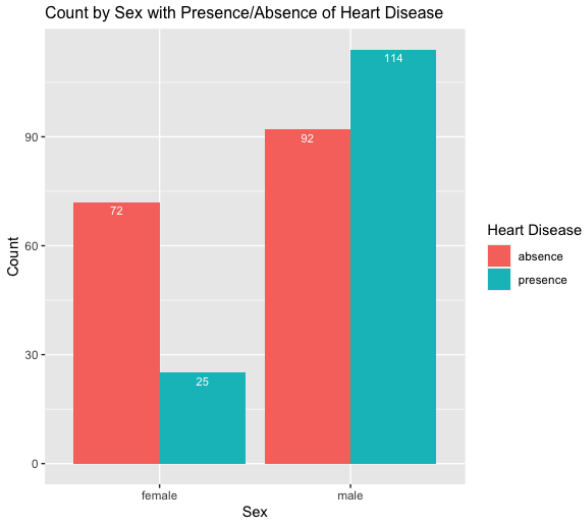


- Total number of male: 206, Total number of female: 97
- The total number of male is twice larger the total number of female

Group count plot

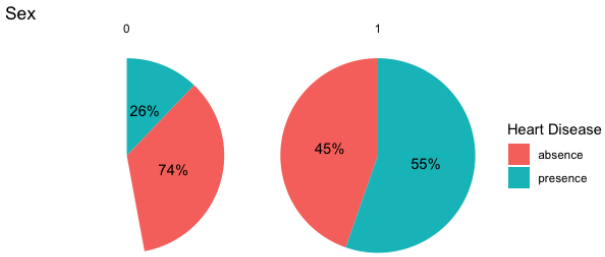
```
#count by group
count_df <- processed.cleveland %>% group_by(V2) %>% count(V14)
```

```
#continuous value → categorical value
count_df$V2 <- as.factor(count_df$V2)
count_df$V14 <- as.factor(count_df$V14)

#group count plot
ggplot(count_df, aes(x = V2, y = n, fill = V14)) +
  geom_col(position = "dodge") + #group bar plot
  theme(plot.title = element_text(size=12)) + #white background, title size
  labs(title="Count by Sex with Presence/Absence of Heart Disease", y="Count",x="Sex") + #change x, y axis title
  guides(fill=guide_legend(title="Heart Disease")) + #change legend title
  scale_x_discrete(labels=c('female', 'male')) + #change x axis labels
  scale_fill_discrete(labels=c('absence', 'presence')) + #change legend labels
  geom_text(aes(label = n),
            colour = "white", size = 3,
            vjust = 1.5, position = position_dodge(.9)) #bar labels
```



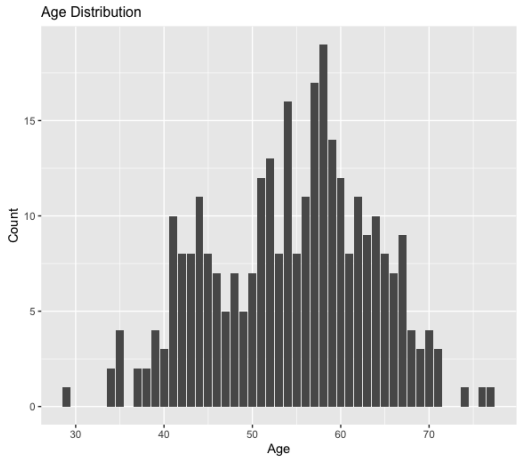
Pie Plot

```
count_df <- processed.cleveland %>% group_by(V2) %>% count(V14)
                                %>% ungroup(V14) %>% mutate(perc = `n` / sum(`n`))
                                %>% arrange(perc) %>% mutate(labels = scales::percent(perc))

ggplot(data=count_df, aes(x=" ", y=n, group=V14,  fill=V14)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(aes(label = labels),position =position_stack(vjust = 0.5)) +
  coord_polar("y", start=0) +
  facet_grid(.~ V2) +
  theme_void() +
  ggtitle("Sex") +
  guides(fill=guide_legend(title="Heart Disease")) +
  scale_fill_discrete(labels = c("absence", "presence"))
```

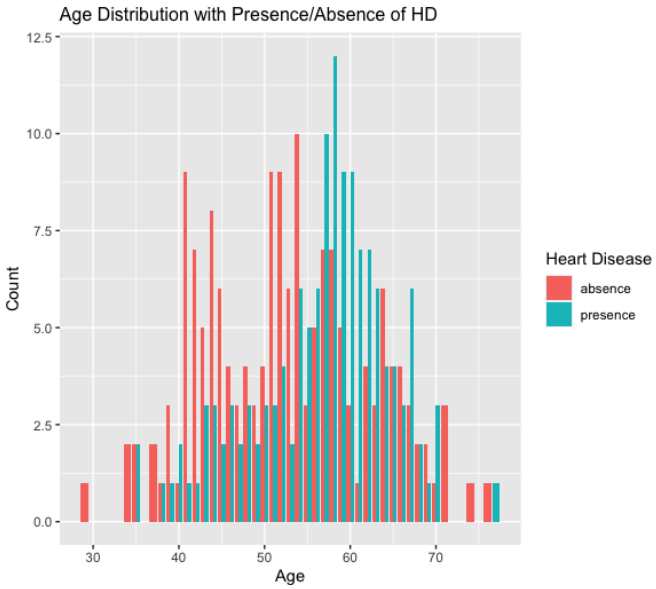


- The percentage of having heart disease in male is larger than female.

```
ggplot(processed.cleveland, aes(x = V1)) +
  geom_bar() +
  labs(title="Age Distribution", y="Count",x="Age")
```




```
ggplot(count_df, aes(x = V1, y = n, fill = V14)) +
  geom_col(position = "dodge") + #group bar plot
  theme(plot.title = element_text(size=12)) +
  labs(title="Age Distribution with Presence/Absence of HD ", y="Count",x="Age") + #change x, y axis title
  guides(fill=guide_legend(title="Heart Disease")) + #change legend title
  scale_fill_discrete(labels=c('absence', 'presence')) #change legend labels
```

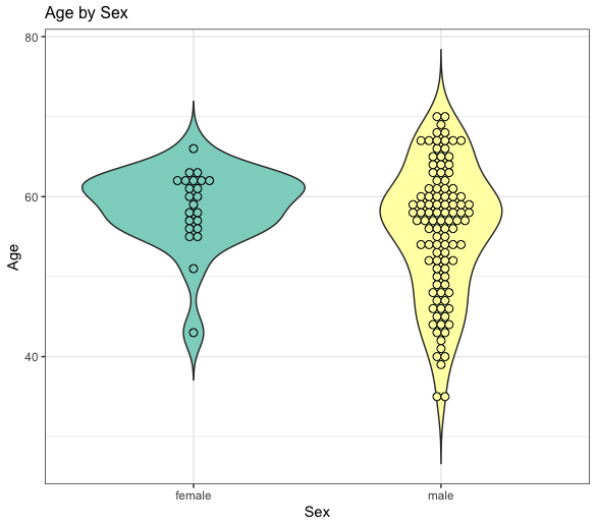


- In the age group of 30-55, The salmon bar is higher than the blue bar, which shows that there are less cases in heart disease
- From age 55 to age 65, there are more presence of heart disease than the absence of heart disease.
- And after age 65, it seems that the presence and absence of heart disease is equal
- So, by the graph we assume that age around 55 to 65 are more likely to have heart disease.

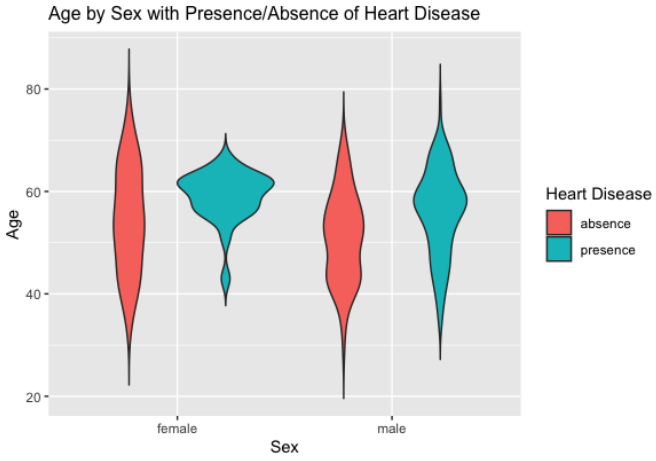
Violin Plot

http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization

```
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

#violin plot
ggplot(df_presence, aes(x=V2, y=V1, fill=V2)) +
  geom_violin(trim=FALSE) +
  geom_dotplot(binaxis='y', stackdir='center', dotsize=1, binwidth = 1) + #dot in the violin plot
  theme_bw() + #white background
  scale_fill_brewer(palette="Set3") + #change filling color
  theme(legend.position="none",plot.title = element_text(size=12)) +
  labs(title="Age by Sex", y="Age",x="Sex") + #set plot title, x axis, and y axis
  scale_x_discrete(labels=c('female', 'male')) #change x axis labels
```



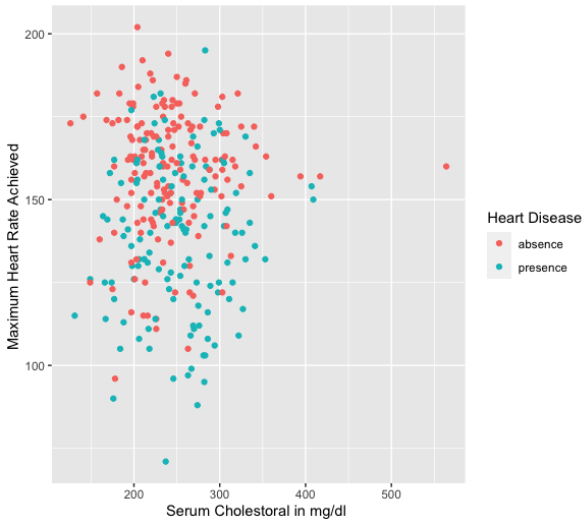

```
ggplot(processed.cleveland, aes(x=V2, y=V1, fill=V14)) +
  geom_violin(trim=FALSE) +
  theme(plot.title = element_text(size=12)) +
  labs(title="Age by Sex with Presence/Absence of Heart Disease", y="Age",x="Sex") +
  scale_x_discrete(labels=c('female', 'male')) + guides(fill=guide_legend(title="Heart Disease")) +
  scale_fill_discrete(labels=c('absence', 'presence'))
```



- The shape of female and male in presence of heart disease is very different.
- The shape of female is more extreme, the age range of having heart disease are around 55-65, and there is much less case out of the age range
- However, the age range of having heart disease in male is more prevalent, though there are still a relatively larger area in age 60's
- The relationship of age and the presence of heart disease is more sensitive in female than in male.

Scatter Plot

```
ggplot(processed.cleveland, aes(V5,V8,color = V14)) +
  geom_point()+
  labs(y="Maximum Heart Rate Achieved ",x="Serum Cholestoral in mg/dl ") +
  guides(color = guide_legend(title = "Heart Disease")) +  #legend title
  scale_colour_discrete(labels=c('absence', 'presence'))  #lengend labels
```



Logistic Regression

```
#sample training data
train <- sample(303, 242)

#train logistic regression
logRegDef<-glm(V14 ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11
                     + V12 + V13,
  family=binomial, data=processed.cleveland, subset=train)

#training result summary
summary(logRegDef)$coefficients
summary(logRegDef)

#test logistic regression
logRegDef.predict<-predict(logRegDef,
  newdata=processed.cleveland[-train,], type="response")

#y predict by sigmoid
ypred<-ifelse(logRegDef.predict<1/2, 0, 1)

#y values
table(processed.cleveland$V14[-train])

#accuracy
mean(ypred == processed.cleveland[-train,]$V14)
sum(ypred!=processed.cleveland$V14[-train])/(303-242)
```