Data Visualization on Heart Disease Data Set using R

Chih-Hua Chang

2022-09-30

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

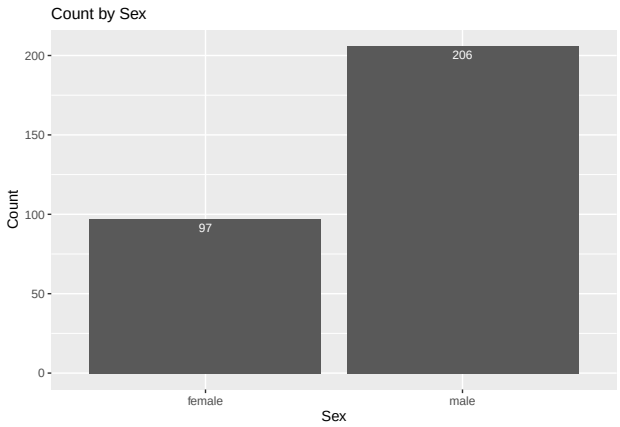
```
processed.cleveland = read.csv("/Users/fionachang/Desktop/data607/hw4/processed.cleveland.data",
header=FALSE)
```

Replace value based on condition

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

Count Plot

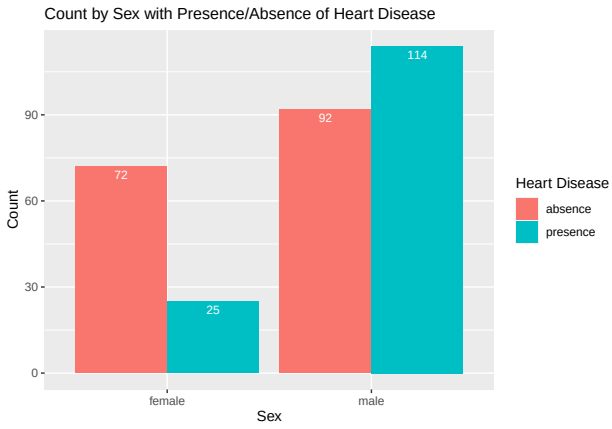
```
library(ggplot2)
library(tidyverse)
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

count_df <- processed.cleveland %>% group_by(V2) %>% count(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

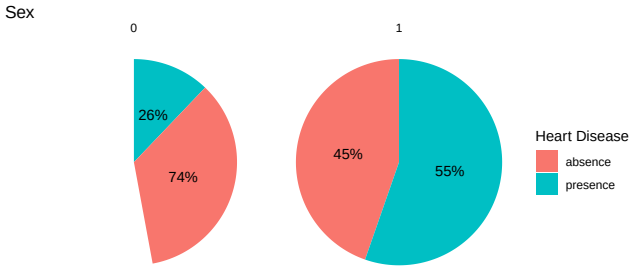
Group Count Plot

```
#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
```



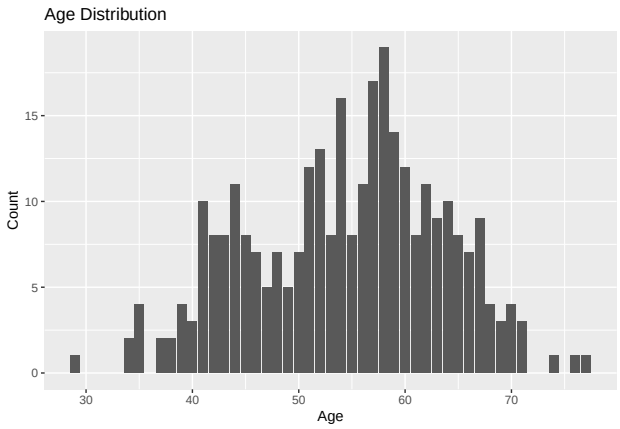
Pie Plot

```
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)
processed.cleveland$V14 <- as.factor(processed.cleveland$V14)

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"))
```

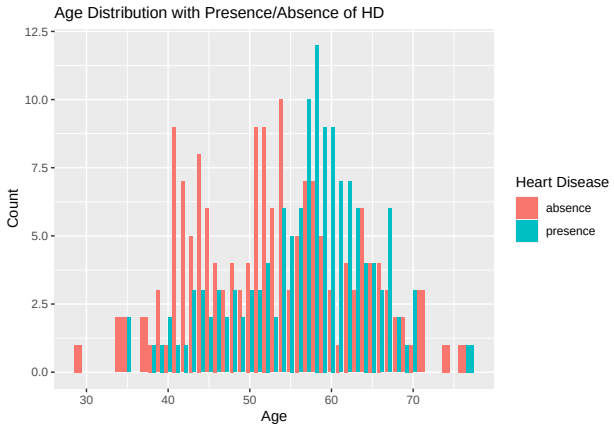


- The percentage of having heart disease in male is larger than female.

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




```r
count_df <- processed.cleveland %>% group_by(V1) %>% count(V14)
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
    guides(fill=guide_legend(title="Heart Disease")) + #change legend title
    scale_fill_discrete(labels=c('absence', 'presence')) #change legend labels
```

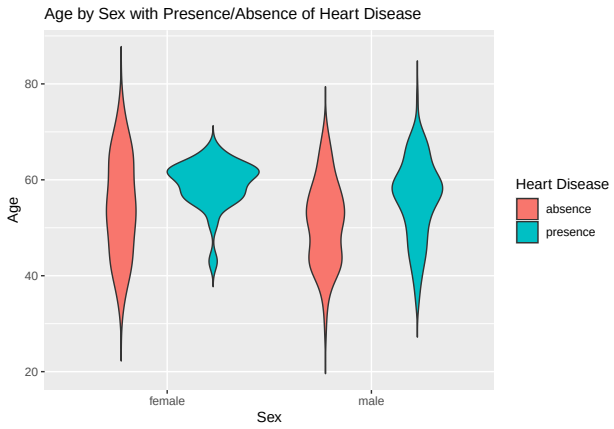


- In the age group of 30-55, The salmon bar is higher than the blue bar, which shows that there are less cases in heart disease
- From age 55 to age 65, there are more presence of heart disease than the absence of heart disease.
- And after age 65, it seems that the presence and absence of heart disease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

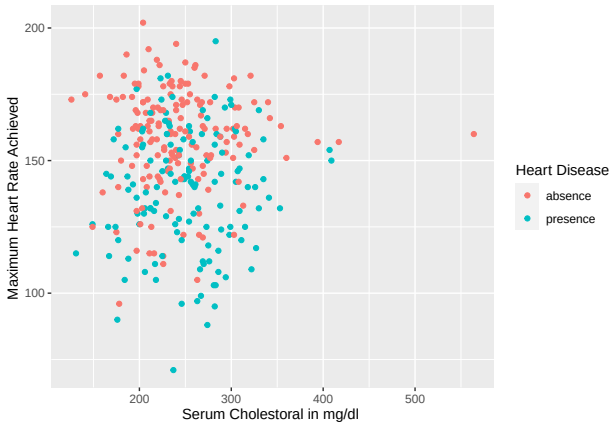
```r
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

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

```r
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
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -7.23849290 3.921641311 -1.8457815 0.0649239408
## V1          -0.01220039 0.028146488 -0.4334606 0.6646802123
## V21          1.90707238 0.591868646  3.2221210 0.0012724540
## V3           0.84266585 0.222985537  3.7790157 0.0001574495
## V4           0.02274179 0.012212290  1.8622054 0.0625741375
## V5           0.00622998 0.004082084  1.5261764 0.1269659476
## V6          -1.00534458 0.610733115 -1.6461275 0.0997375118
## V7           0.27808132 0.213811305  1.3005922 0.1933980647
## V8          -0.01470865 0.012174700 -1.2081326 0.2269962809
## V9           1.12420324 0.487779304  2.3047375 0.0211812806
## V10          0.39575896 0.243072952  1.6281489 0.1034933197
## V11          0.66535464 0.407131017  1.6342519 0.1022059661
## V120.0       0.21650169 1.705810029  0.1269202 0.8990035886
## V121.0       1.43711042 1.742302500  0.8248340 0.4094658140
## V122.0       3.41667591 1.869637673  1.8274535 0.0676316116
## V123.0       1.90753237 1.861851496  1.0245352 0.3055825604
## V133.0      -1.86339680 1.829769214 -1.0183780 0.3084983303
## V136.0      -2.75872472 2.013810665 -1.3699027 0.1707172684
## V137.0      -0.79595242 1.852227629 -0.4297271 0.6673941636
```

```
summary(logRegDef)
```

```
##
## Call:
## glm(formula = V14 ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 +
##     V10 + V11 + V12 + V13, family = binomial, data = processed.cleveland,
##     subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6345  -0.5074  -0.1238   0.3512   2.3213
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.238493   3.921641  -1.846 0.064924 .
## V1          -0.012200   0.028146  -0.433 0.664680
## V21          1.907072   0.591869   3.222 0.001272 **
## V3           0.842666   0.222986   3.779 0.000157 ***
## V4           0.022742   0.012212   1.862 0.062574 .
## V5           0.006230   0.004082   1.526 0.126966
## V6          -1.005345   0.610733  -1.646 0.099738 .
## V7           0.278081   0.213811   1.301 0.193398
## V8          -0.014709   0.012175  -1.208 0.226996
## V9           1.124203   0.487779   2.305 0.021181 *
## V10          0.395759   0.243073   1.628 0.103493
## V11          0.665355   0.407131   1.634 0.102206
## V120.0       0.216502   1.705810   0.127 0.899004
## V121.0       1.437110   1.742302   0.825 0.409466
## V122.0       3.416676   1.869638   1.827 0.067632 .
## V123.0       1.907532   1.861851   1.025 0.305583
## V133.0      -1.863397   1.829769  -1.018 0.308498
## V136.0      -2.758725   2.013811  -1.370 0.170717
## V137.0      -0.795952   1.852228  -0.430 0.667394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 334.67  on 241  degrees of freedom
## Residual deviance: 158.12  on 223  degrees of freedom
## AIC: 196.12
##
## Number of Fisher Scoring iterations: 6
```

```
#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])
```

```
##
##  0  1
```

```
## 36 25
```

```
#accuracy
mean(ypred == processed.cleveland[-train,]$V14)
```

```
## [1] 0.8032787
```

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

```
## [1] 0.1967213
```