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)

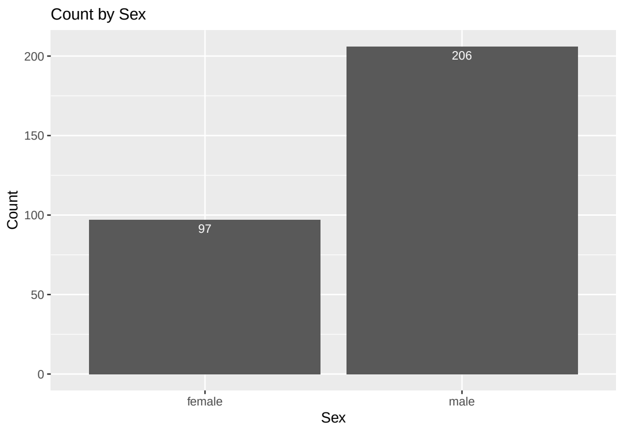
```r
processed.cleveland = read.csv("/Users/fionachang/Desktop/data607/hw4/processed.cleveland.data",
header=FALSE)
```

Replace value based on condition

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

Count Plot

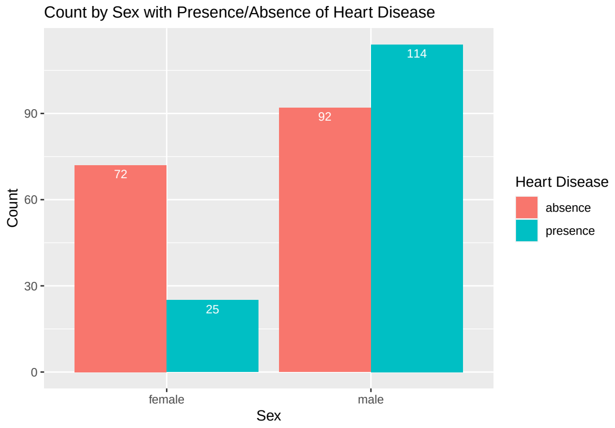
```r
library(ggplot2)
library(tidyverse)
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

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

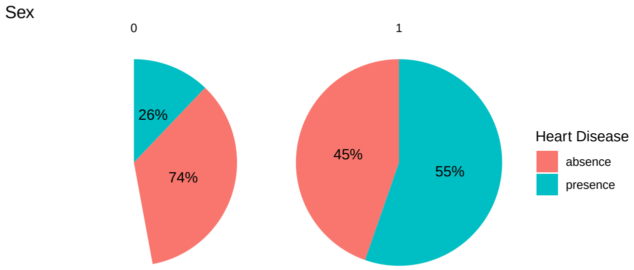
Group Count Plot

```
#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
```



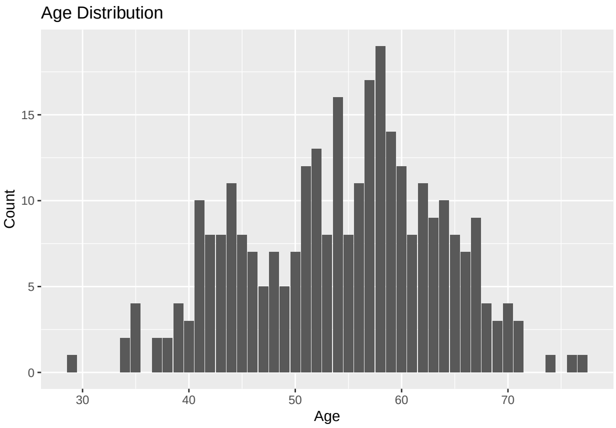
Pie Plot

```
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"))
```

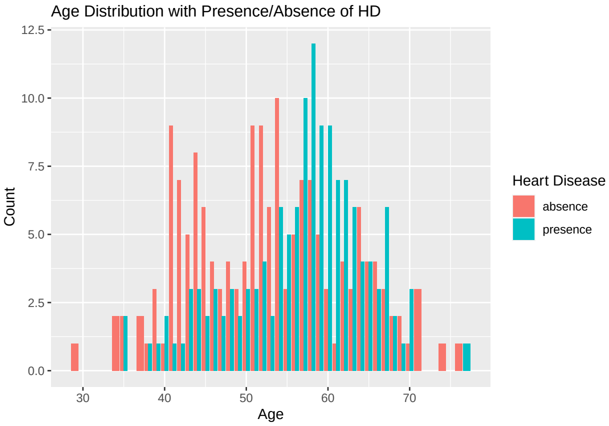


- The percentage of having heart disease in male is larger than female.

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



```
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
```

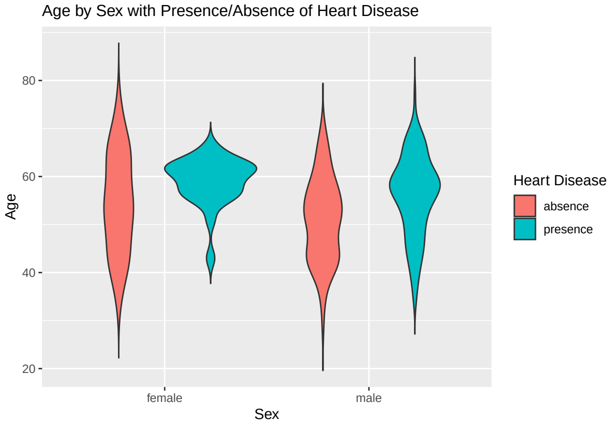


- In the age group of 30-55, The salmon bar is higher than the blue bar, which shows that there are less cases in heart disease
- From age 55 to age 65, there are more presence of heart disease than the absence of heart disease.
- And after age 65, it seems that the presence and absence of heart disease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

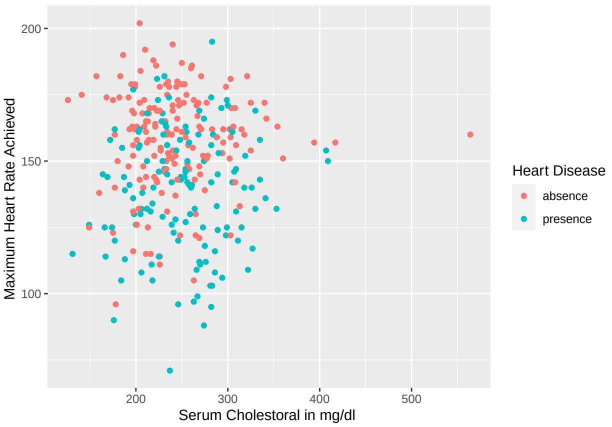
```
#continuous value → categorical value
processed.cleveland$V2 <- as.factor(processed.cleveland$V2)

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

```
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) -2.095626e+01 1.455402e+03 -0.014398947 0.9885116996
## V1          -2.240863e-03 2.817895e-02 -0.079522579 0.9366169731
## V21          6.772925e-01 6.029205e-01  1.123353016 0.2612875956
## V3           9.039905e-01 2.490691e-01  3.629476256 0.0002839969
## V4           3.127568e-02 1.358799e-02  2.301715516 0.0213512206
## V5          -2.402152e-05 5.418975e-03 -0.004432854 0.9964631058
## V6          -8.699676e-01 6.666560e-01 -1.304972332 0.1919022691
## V7           2.776360e-01 2.255504e-01  1.230926679 0.2183502883
## V8          -2.989100e-03 1.279640e-02 -0.233589094 0.8153039953
## V9           1.312440e+00 4.758622e-01  2.758024021 0.0058151924
## V10          4.093545e-01 2.562990e-01  1.597175694 0.1102265490
## V11          6.321283e-01 4.370363e-01  1.446397705 0.1480656851
## V120.0       1.232633e+01 1.455398e+03  0.008469391 0.9932424848
## V121.0       1.424906e+01 1.455398e+03  0.009790490 0.9921884442
## V122.0       1.548646e+01 1.455398e+03  0.010640709 0.9915101024
## V123.0       1.488419e+01 1.455398e+03  0.010226885 0.9918402683
## V133.0      -2.320205e+00 1.775950e+00 -1.306458154 0.1913968005
## V136.0      -1.332575e+00 1.926078e+00 -0.691859260 0.4890257173
## V137.0      -3.895787e-01 1.789677e+00 -0.217680985 0.8276776800
```

```
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
## -2.6183  -0.4270  -0.1224   0.3551   2.5357
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.096e+01  1.455e+03  -0.014 0.988512
## V1          -2.241e-03  2.818e-02  -0.080 0.936617
## V21          6.773e-01  6.029e-01   1.123 0.261288
## V3           9.040e-01  2.491e-01   3.629 0.000284 ***
## V4           3.128e-02  1.359e-02   2.302 0.021351 *
## V5          -2.402e-05  5.419e-03  -0.004 0.996463
## V6          -8.700e-01  6.667e-01  -1.305 0.191902
## V7           2.776e-01  2.256e-01   1.231 0.218350
## V8          -2.989e-03  1.280e-02  -0.234 0.815304
## V9           1.312e+00  4.759e-01   2.758 0.005815 **
## V10          4.094e-01  2.563e-01   1.597 0.110227
## V11          6.321e-01  4.370e-01   1.446 0.148066
## V120.0       1.233e+01  1.455e+03   0.008 0.993242
## V121.0       1.425e+01  1.455e+03   0.010 0.992188
## V122.0       1.549e+01  1.455e+03   0.011 0.991510
## V123.0       1.488e+01  1.455e+03   0.010 0.991840
## V133.0      -2.320e+00  1.776e+00  -1.306 0.191397
## V136.0      -1.333e+00  1.926e+00  -0.692 0.489026
## V137.0      -3.896e-01  1.790e+00  -0.218 0.827678
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 333.83  on 241  degrees of freedom
## Residual deviance: 147.74  on 223  degrees of freedom
## AIC: 185.74
##
## Number of Fisher Scoring iterations: 14
```

```
#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
## 33 28
```

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

```
## [1] 0.7704918
```

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

```
## [1] 0.2295082
```