HW1-Solution

Aditya Singh

2022-09-22

Part A

Problem 1

For Problem 1 part A, I have created a `data.frame` with `row.names` as student names and columns as scores in some test.

```
df <- data.frame(row.names=c('Tom','Brad','Mat','Billy'),
                 Physics=c('A',NA,'A+','C'),
                 Chemistry=c(NA,NA,'D','F'),
                 Maths= c('B','C',NA,'F'),
                 Biology=c("A",NA,"B",NA)
                 )
print(df)
```

```
##       Physics Chemistry Maths Biology
## Tom         A      <NA>     B       A
## Brad     <NA>      <NA>     C    <NA>
## Mat        A+         D  <NA>       B
## Billy       C         F     F    <NA>
```

Below is the `countNA` function which takes arguments `data.frame` and `byrow` to return a numeric vector of missing values(NAs) for each row or each column of data(depending on the *byrow* argument).

```
countNA <- function(df,byrow=FALSE){
  if(byrow){
    return (rowSums(is.na(df)))
  }
  else{
    return (sapply(df,function(df) sum(is.na(df))))
  }
}
```

The Following is the output when *byrow* is `FALSE`.

```
print(df)
```

```
##       Physics Chemistry Maths Biology
## Tom         A      <NA>     B       A
```

```
## Brad     <NA>      <NA>     C    <NA>
## Mat        A+         D  <NA>       B
## Billy       C         F     F    <NA>
```

```
countNA(df)
```

```
##   Physics Chemistry     Maths   Biology
##         1         2         1         2
```

The Following is the output when *byrow* is `TRUE`.

```
print(df)
```

```
##       Physics Chemistry Maths Biology
## Tom         A      <NA>     B       A
## Brad     <NA>      <NA>     C    <NA>
## Mat        A+         D  <NA>       B
## Billy       C         F     F    <NA>
```

```
countNA(df,TRUE)
```

```
##   Tom  Brad   Mat Billy
##     1     3     1     1
```

Problem 2

Lets take a sample `data.frame` to solve this problem.

```
df = data.frame(row.names=c(1:8),
                city=c('Mumbai','Pune',NA,'Mumbai',NA,'Chennai','Chennai','Banglore'),
                Income=c(1000000,765097,888690,NA,676538,968271,634158,4271821),
                Age = c(85,77,82,NA,75,80,73,69)
                )
print(df)
```

```
##       city  Income Age
## 1   Mumbai 1000000  85
## 2     Pune  765097  77
## 3     <NA>  888690  82
## 4   Mumbai      NA  NA
## 5     <NA>  676538  75
## 6  Chennai  968271  80
## 7  Chennai  634158  73
## 8 Banglore 4271821  69
```

Now first, lets write a function `getmode` which will return the mode of categorical column provided in argument.

```
getmode <- function(x){
  val <- unique(x[!is.na(x)])
  return (val[which.max(tabulate(match(x,val)))])
}
```

Below is a imputation function

```
imputeNA <- function(df, use.mean= FALSE){
  for(i in c(1:length(names(df)))){
      if(class(df[,i])=="character"){
        df[ , i][is.na(df[ , i])]<-getmode(df[,i])
      }
      else{
        if(use.mean){
          df[ , i][is.na(df[ , i])]<-mean(df[,i],na.rm=TRUE)
        }
        else{
          df[ , i][is.na(df[ , i])]<-median(df[,i],na.rm=TRUE)
        }
      }

    }
  return (df)
}
```

First imputation with `median`

```
print(df)
```

```
##       city  Income Age
## 1   Mumbai 1000000  85
## 2     Pune  765097  77
## 3     <NA>  888690  82
## 4   Mumbai      NA  NA
## 5     <NA>  676538  75
## 6  Chennai  968271  80
## 7  Chennai  634158  73
## 8 Banglore 4271821  69
```

```
imputeNA(df)
```

```
##       city  Income Age
## 1   Mumbai 1000000  85
## 2     Pune  765097  77
## 3   Mumbai  888690  82
## 4   Mumbai  888690  77
## 5   Mumbai  676538  75
## 6  Chennai  968271  80
## 7  Chennai  634158  73
## 8 Banglore 4271821  69
```

Now Imputation with `mean`

```
print(df)
```

```
##       city  Income Age
## 1   Mumbai 1000000  85
## 2     Pune  765097  77
## 3     <NA>  888690  82
## 4   Mumbai      NA  NA
## 5     <NA>  676538  75
## 6  Chennai  968271  80
## 7  Chennai  634158  73
## 8 Banglore 4271821  69
```

```
imputeNA(df,TRUE)
```

```
##       city  Income      Age
## 1   Mumbai 1000000 85.00000
## 2     Pune  765097 77.00000
## 3   Mumbai  888690 82.00000
## 4   Mumbai 1314939 77.28571
## 5   Mumbai  676538 75.00000
## 6  Chennai  968271 80.00000
## 7  Chennai  634158 73.00000
## 8 Banglore 4271821 69.00000
```

Part B

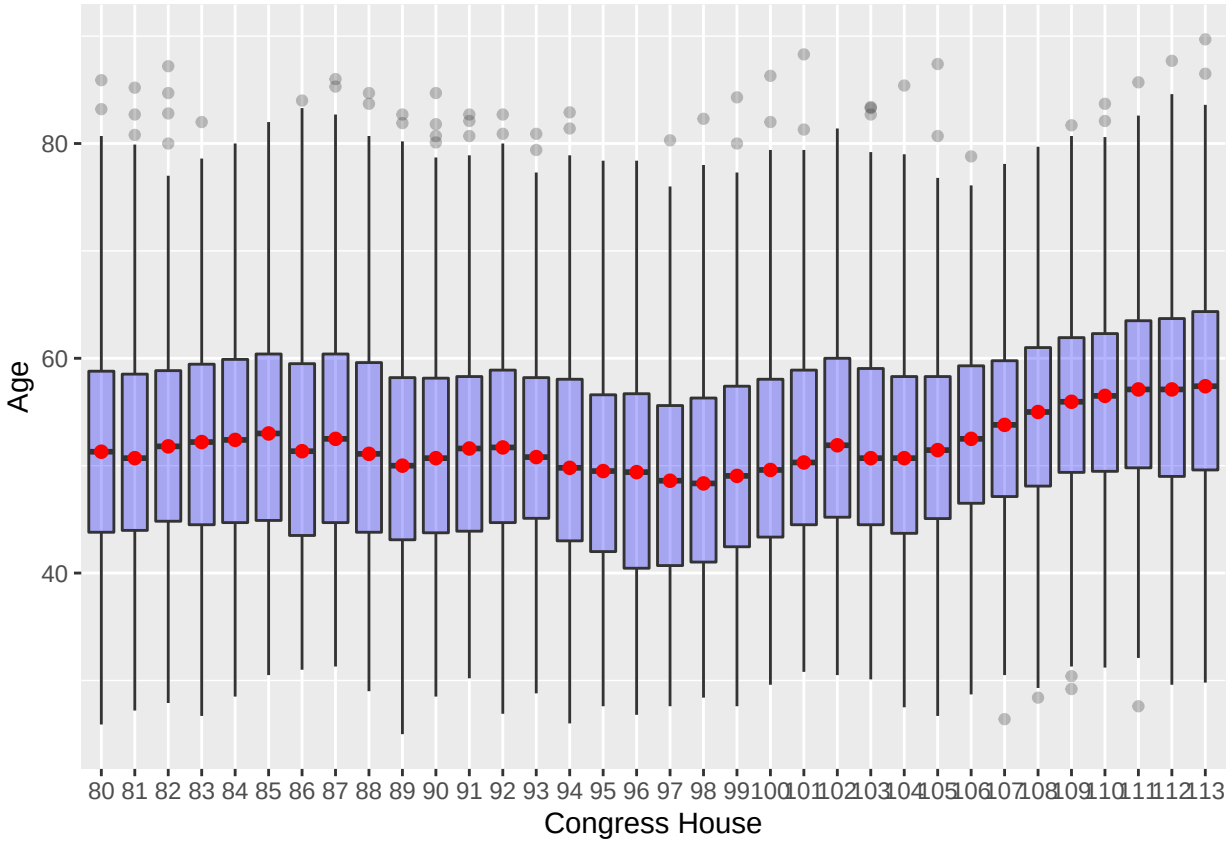
Problem 3

```
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

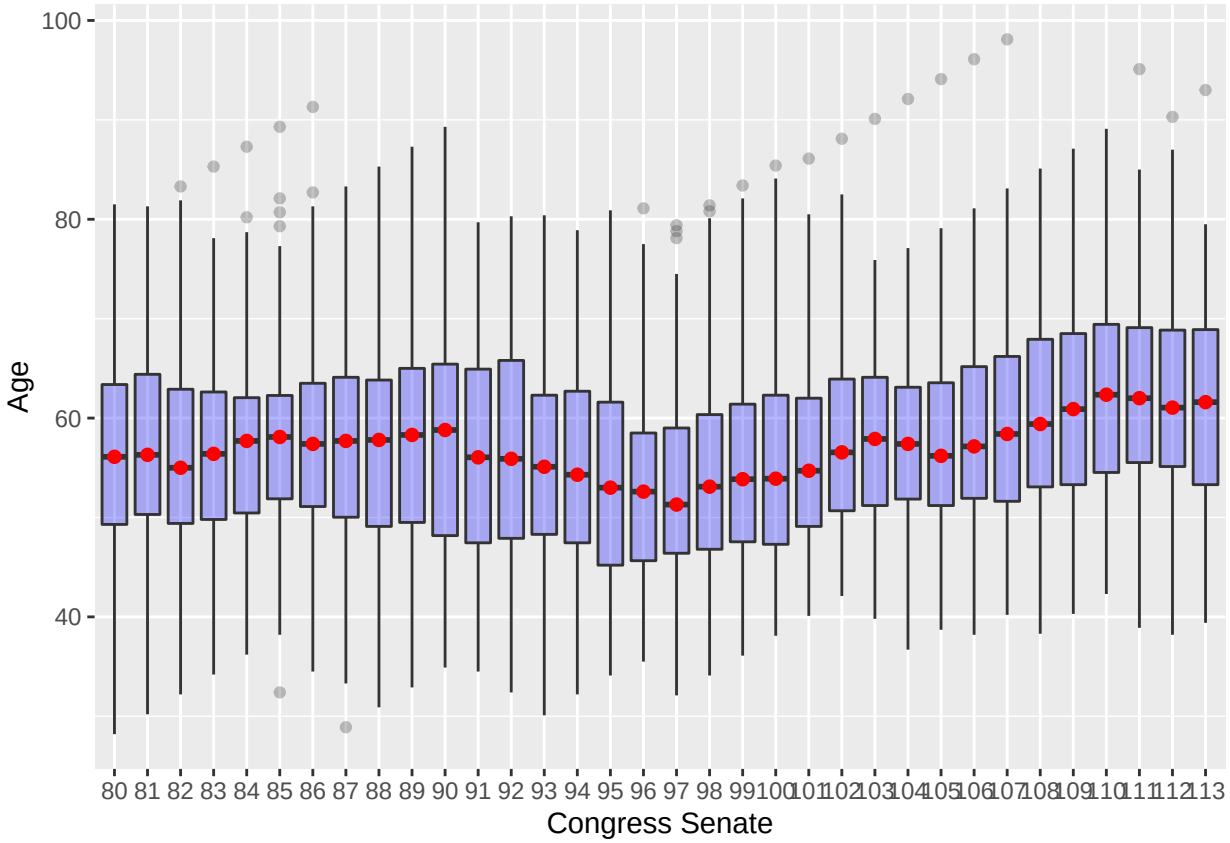
```
df <- get(data('congress_age'))
library(ggplot2)

x <-ggplot(data= subset(df,df$chamber=='house'), mapping = aes(x=as.factor(congress),y=age),colour= char
x + geom_boxplot(alpha = 0.3,fill='blue')+labs(x="Congress House",y='Age')+
stat_summary(fun=median, geom="point", shape=20, size=3, color="red", fill="red")
```



From the above `boxplot` we can make see that the median age increases from *80* to *86* then decreases from *92* to *98* then increases again from *99* to *113*

```
x <-ggplot(data= subset(df,df$chamber=='senate'), mapping = aes(x=as.factor(congress),y=age),colour= cha
x + geom_boxplot(alpha = 0.3,fill='blue')+labs(x="Congress Senate",y='Age')+
stat_summary(fun=median, geom="point", shape=20, size=3, color="red", fill="red")
```



From the above `boxplot` we can make see that the median age increases from *80* to *89* then decreases from *91* to *96* then increases again from *105* to *113*

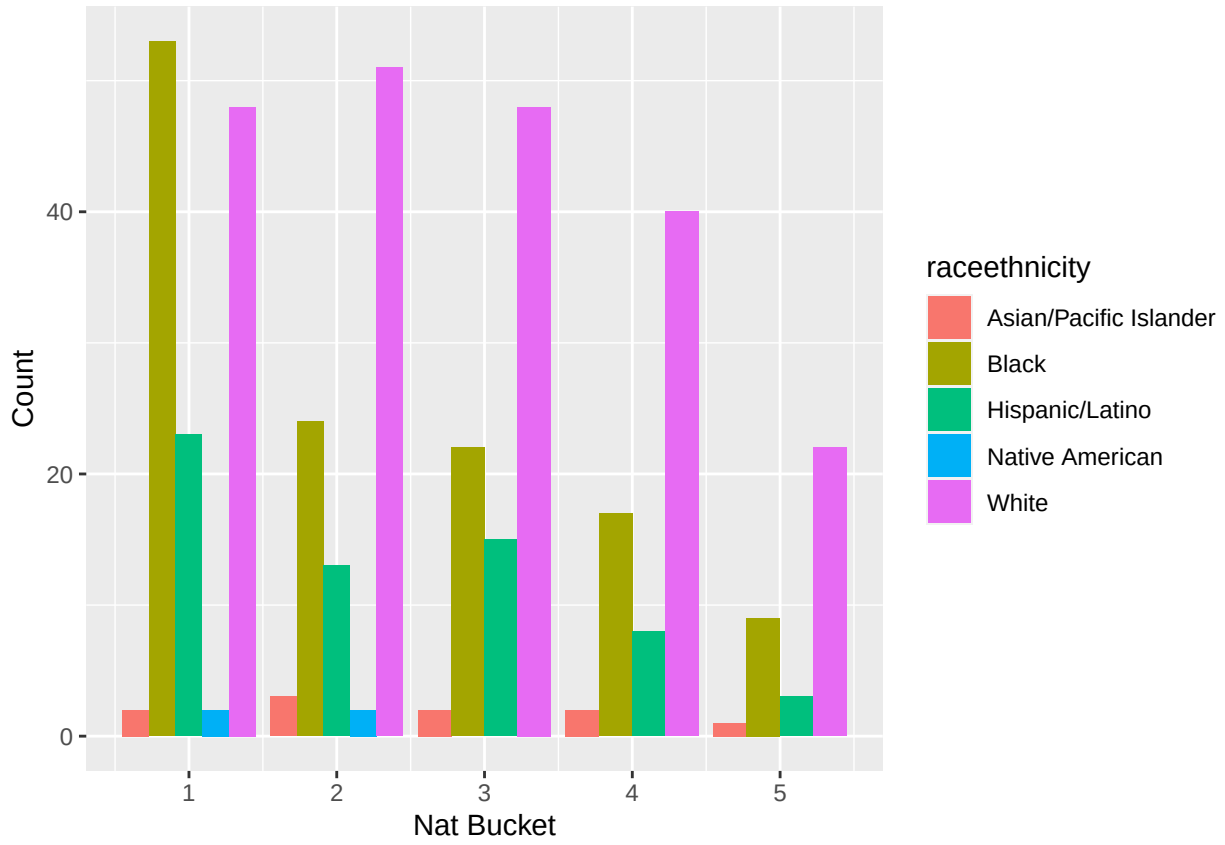
In general the `median` age for Senate is more than that of house

Problem 4

```
library(fivethirtyeight)
df <- get(data('police_killings'))
```

```
data_omit <- na.omit(df)

df_base <- ggplot(data = data_omit, aes(x=nat_bucket))
df_base + geom_bar(stat="count",aes(fill=raceethnicity),position = position_dodge())+labs(x='Nat Bucket
```

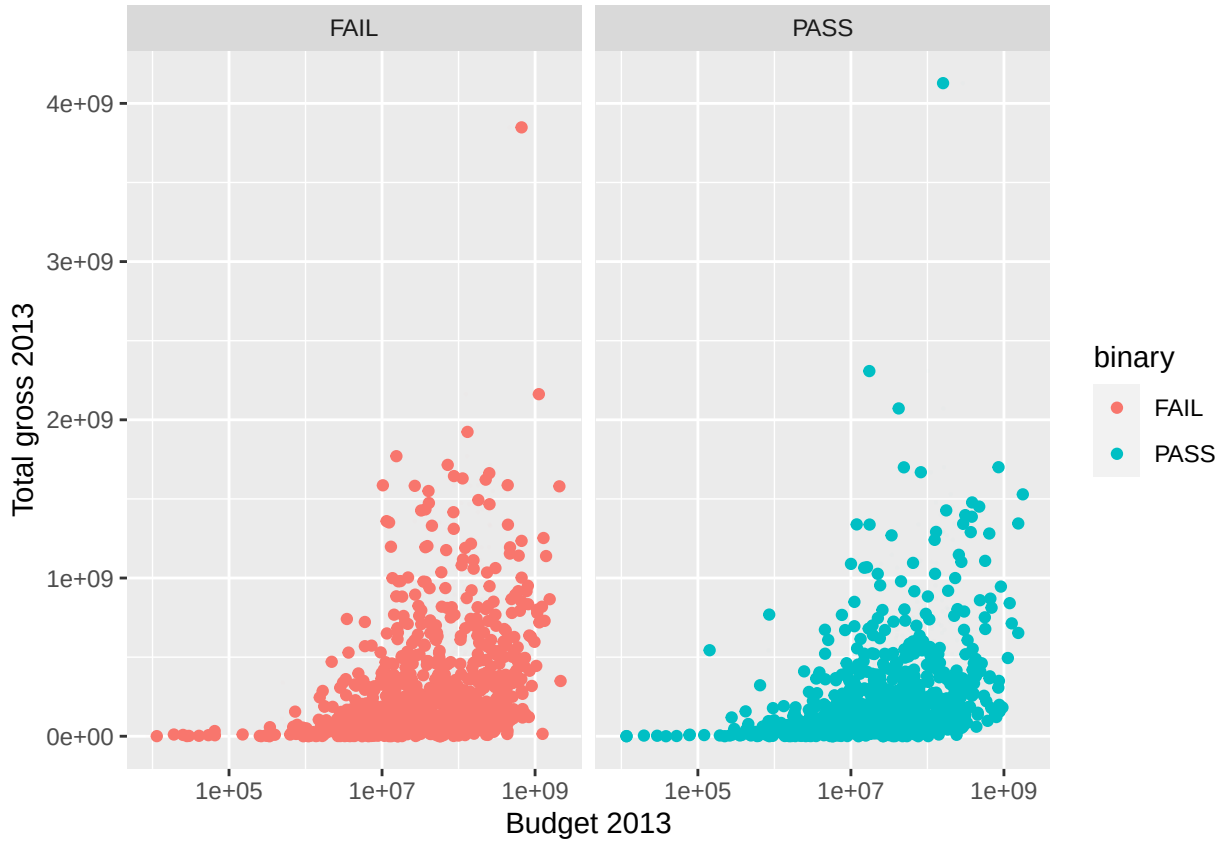


From the above bar plot we can infer that the killing increases for people belonging to lower `nat_bucket`, *specially in case of people belonging to Black ethnicity* the increase in killing is pretty drastic and evident.

Problem 5

```
library(fivethirtyeight)
df <- get(data("bechdel"))
df <- na.omit(df)
df$totalgross_2013 <- df$domgross_2013 + df$intgross_2013

df_base <- ggplot(data = df, aes(x=budget_2013,y=totalgross_2013,color=binary))
df_base + geom_point(size=0.02,alpha=0.01) + scale_x_log10() + facet_wrap(facets = vars(binary))+labs(x=
```



In general as the Budget of movie increases we can increase in Total gross of the movie too. However, there is no strong relation between success of movie and Bechdel test though some high budget movies that have failed the test have performed slightly better than movies with similar budget that have passed the test. And in case of low budget movies its opposite i.e movies which have passed the test have performed slightly better then the movies which have failed the test.