HW1-Spandan-Maaheshwari

Spandan Maaheshwari

2022-09-21

PROBLEM 1 ->

```
testdf <- data.frame(
  row.names=c("Jack", "Rosa", "Dawn", "Vicki", "Blake", "Guillermo"),
  age=c(24, 23, NA, 25, 32, 19),
  city=c("Harlem", NA, "Queens", "Brooklyn", "Brooklyn", NA),
  gpa=c(3.5, 3.6, 4.0, NA, 3.8, NA))

testdf
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23     <NA> 3.6
## Dawn       NA   Queens 4.0
## Vicki      25 Brooklyn  NA
## Blake      32 Brooklyn 3.8
## Guillermo  19     <NA>  NA
```

```
countNA <- function (data,byrow=FALSE){
  if (byrow){
    rowSums(is.na(data))
    }else {
      colSums(is.na(data))
    }
}

countNA(testdf)
```

```
##  age city  gpa
##    1    2    2
```

```
countNA(testdf, byrow=TRUE)
```

```
##      Jack      Rosa      Dawn     Vicki     Blake Guillermo
##         0         1         1         1         0         2
```

PROBLEM 2 ->

Mode function referenced from Stack Overflow

```
calc.mode <- function(vect){
  names(table(vect)[table(vect) == max(table(vect))])
}

inputNA <- function(data, use.mean = FALSE)
{
   col.class = lapply(data, class)

  for(i in 1:ncol(data))
  {
    if (col.class[i] != 'numeric')
    {
      data[i][is.na(data[i])] <- calc.mode(data[,i])
    }else
    {
      if(use.mean)
      {

        data[i][is.na(data[i])] <- mean(data[,i], na.rm = TRUE)
      }else
      {

      data[i][is.na(data[i])] <- median(data[,i], na.rm = TRUE)

      }

    }

  }
  print(data)
}

inputNA(testdf)
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23 Brooklyn 3.6
## Dawn       24   Queens 4.0
## Vicki      25 Brooklyn 3.7
## Blake      32 Brooklyn 3.8
## Guillermo  19 Brooklyn 3.7
```

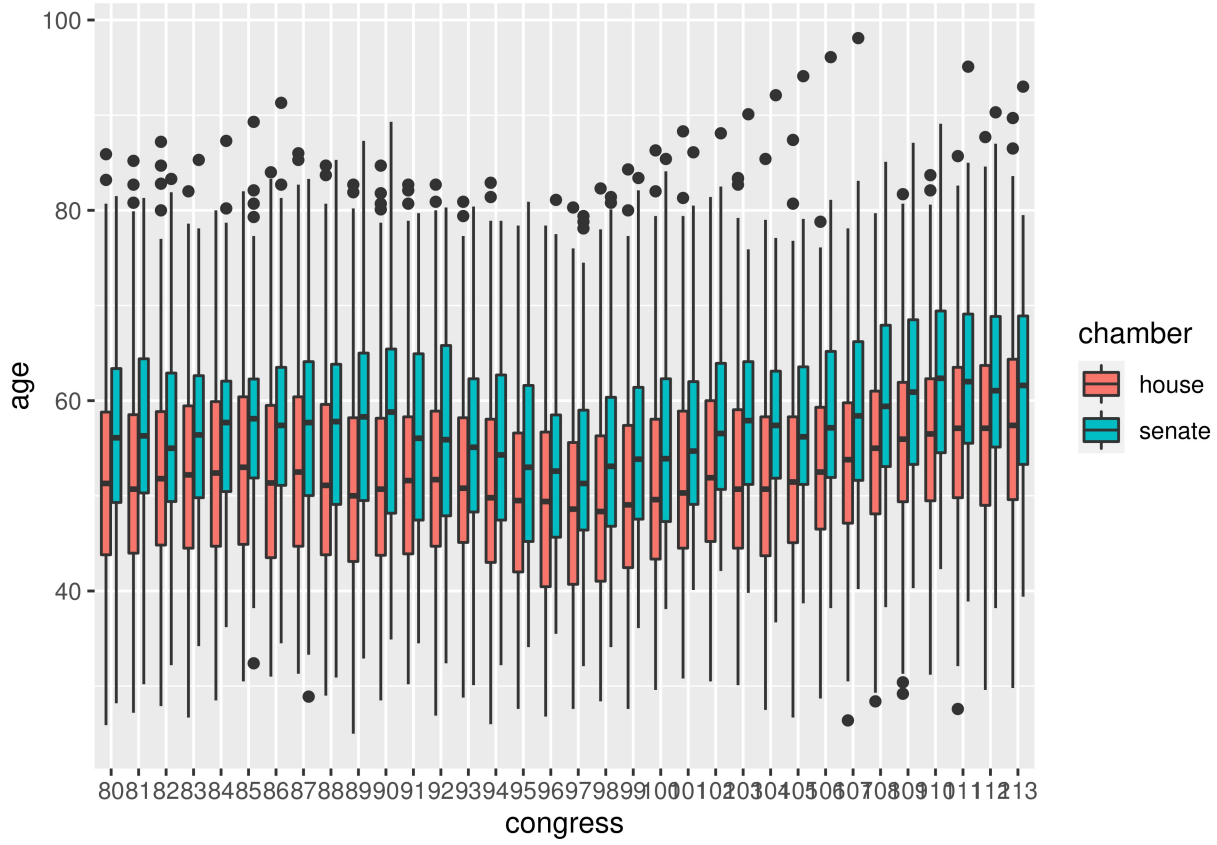
```
inputNA(testdf, use.mean = TRUE)
```

```
##            age     city   gpa
## Jack      24.0   Harlem 3.500
## Rosa      23.0 Brooklyn 3.600
## Dawn      24.6   Queens 4.000
## Vicki     25.0 Brooklyn 3.725
## Blake     32.0 Brooklyn 3.800
## Guillermo 19.0 Brooklyn 3.725
```

PROBLEM 3 ->

For the boxplots shown below for distribution of ages for each congress number, median age turns out to be a wavy structure where it first increases, a dip is seen between in the middle with lowest at congress number #97 and attains the maximum value at the last number #113 of Senate chamber. Another observation is that the median age of Senate chamber is larger than the House chamber throughout the congress number.

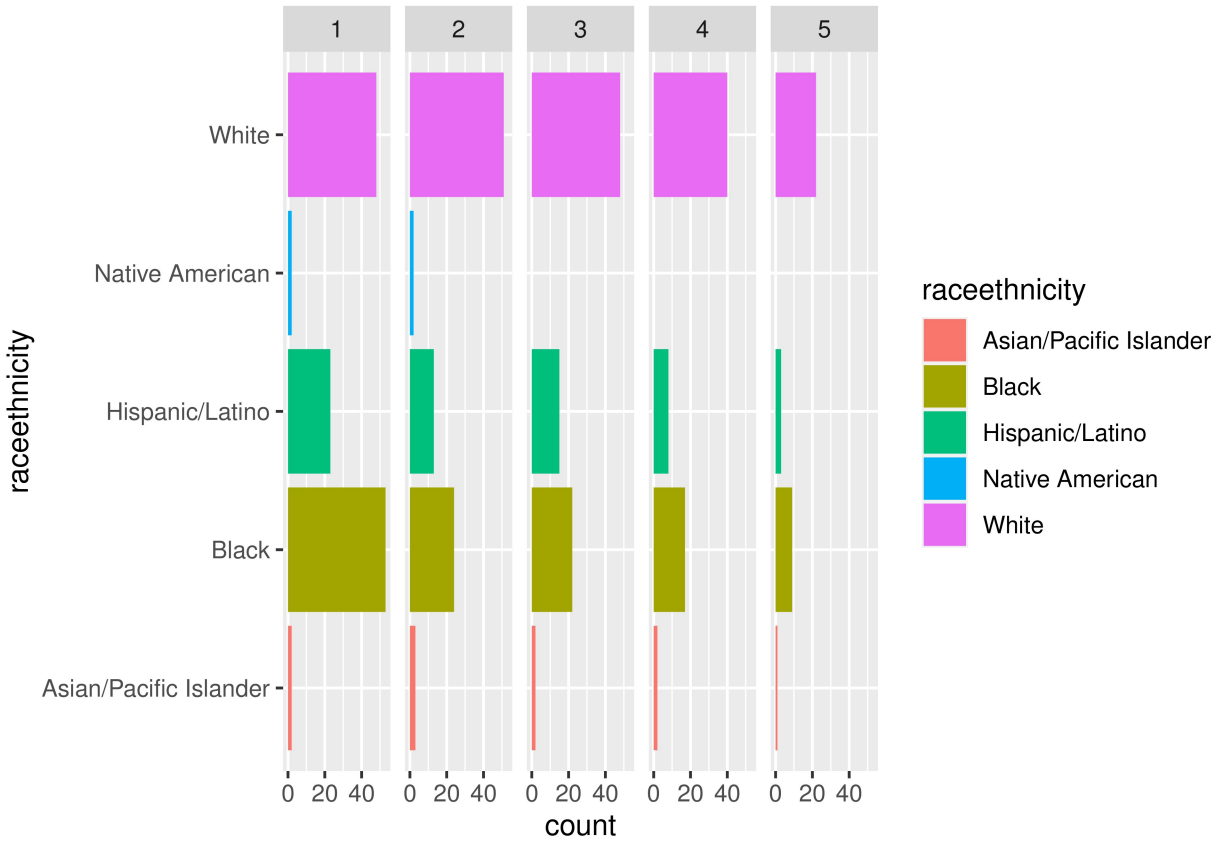
```
library(fivethirtyeight)
library(ggplot2)
congress_age$congress = as.factor(congress_age$congress)
ggplot(data=congress_age,mapping=aes(x=congress, y=age, fill=chamber)) +
  geom_boxplot()
```



PROBLEM 4 ->

For count of Americans killed of each race/ethnicity, we can see that as the national quintile of household income increases, the total count of killing decreases and in the case of Native American it becomes null.

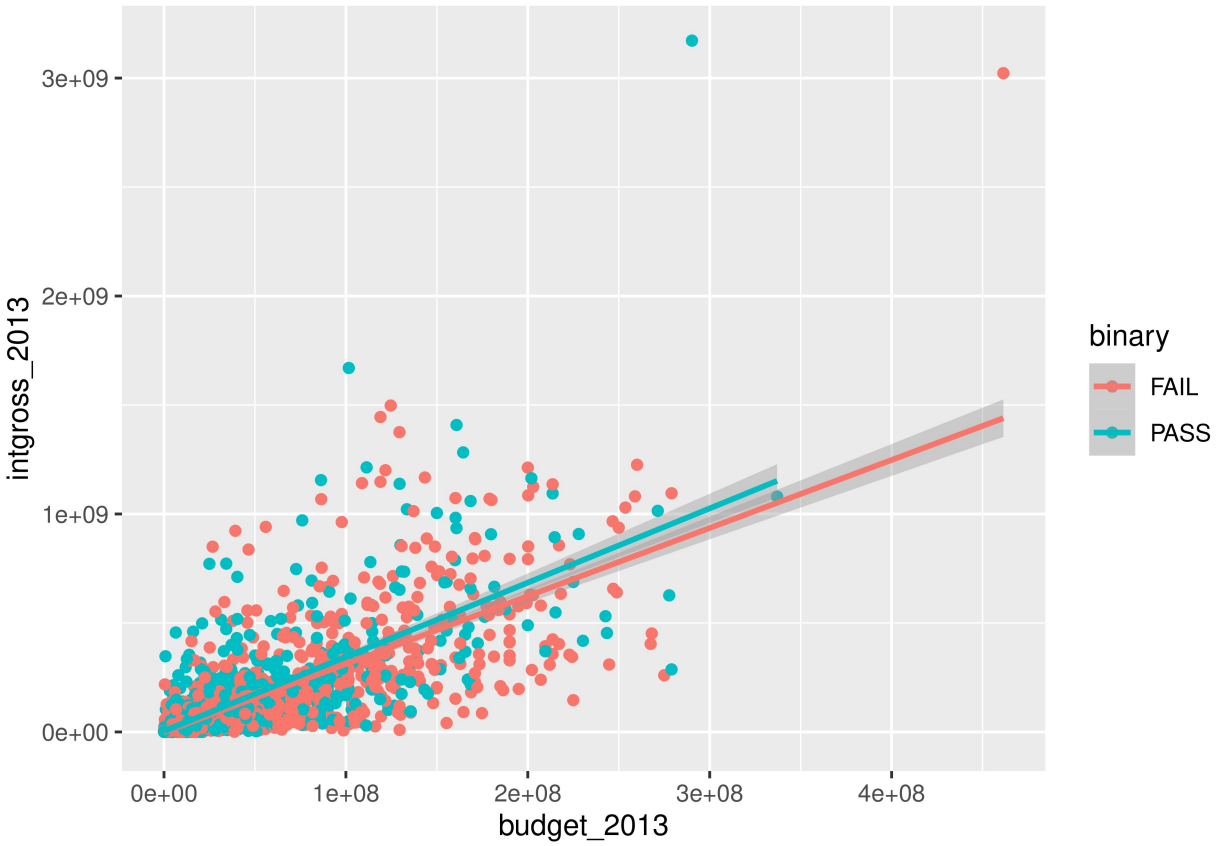
```
library(fivethirtyeight)
library(ggplot2)
ggplot(data=na.omit(police_killings),mapping=aes(y=raceethnicity, fill=raceethnicity)) +
  geom_bar() + facet_grid(~nat_bucket)
```



PROBLEM 5 ->

Movie gross & Movie budget has linear relationship among them, as budget of movie increases it's international gross increases. Overlayed smooth lines indicate linear relationship. Also passing the Bechdel Test has less impact on this relationship. We can observe that the budget for films that passed the Bechdel test is less than for films that failed. This suggests that the movie industry invested more in the movie with less women representation.

```
library(fivethirtyeight)
library(ggplot2)
ggplot(data = na.omit(bechdel),
       mapping = aes(x=budget_2013,
                     y=intgross_2013,
                     color=binary)) +
  geom_point() + geom_smooth(method = lm, formula = y ~ x)
```


3e+09
2e+09
1e+09
0e+00
intgross_2013
0e+00
1e+08
2e+08
3e+08
4e+08
budget_2013
binary
FAIL
PASS