DS5110 Homework 1

Parth Shah

2023-01-28

Part A

Problem 1

```
imputeNA <- function(data, use.mean = FALSE) {
  mode <- function(x) {
    unique_values <- unique(na.omit(x))
    return (unique_values[which.max(tabulate(match(x, unique_values)))])
  }
  for (col in colnames(data)) {
    if (is.factor(data[[col]]) || is.character(data[[col]])) {
      data[[col]][is.na(data[[col]])] <- mode(data[[col]])
    } else if (is.numeric(data[[col]])) {
      if (use.mean) {
        data[[col]][is.na(data[[col]])] <- mean(data[[col]], na.rm = TRUE)
      } else {
        data[[col]][is.na(data[[col]])] <- median(data[[col]], na.rm = TRUE)
      }
    }
  }
  return(data)
}
```

```
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
```

```
imputeNA(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
```

```
imputeNA(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 2

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

```
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(testdf)
```

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

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

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

Part B

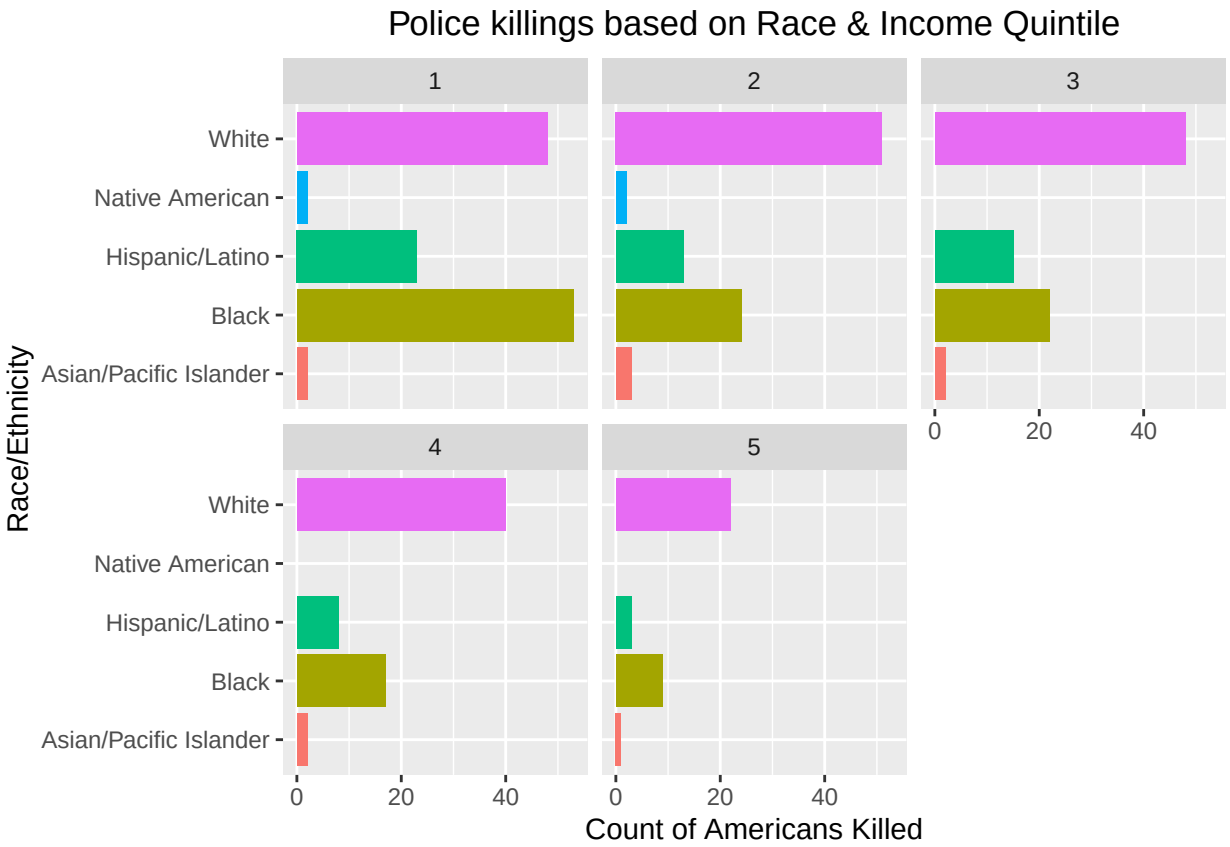
Problem 3

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fivethirtyeight)

police_killings_copy <- na.omit(police_killings)
ggplot(police_killings_copy) +
  aes(raceethnicity, fill = raceethnicity) +
  facet_wrap( ~ nat_bucket) +
  geom_bar() +
  ggtitle('Police killings based on Race & Income Quintile') +
  xlab('Race/Ethnicity') +
  ylab('Count of Americans Killed') +
  scale_fill_discrete(name = "Race/Ethnicity") +
  coord_flip() +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none")
```

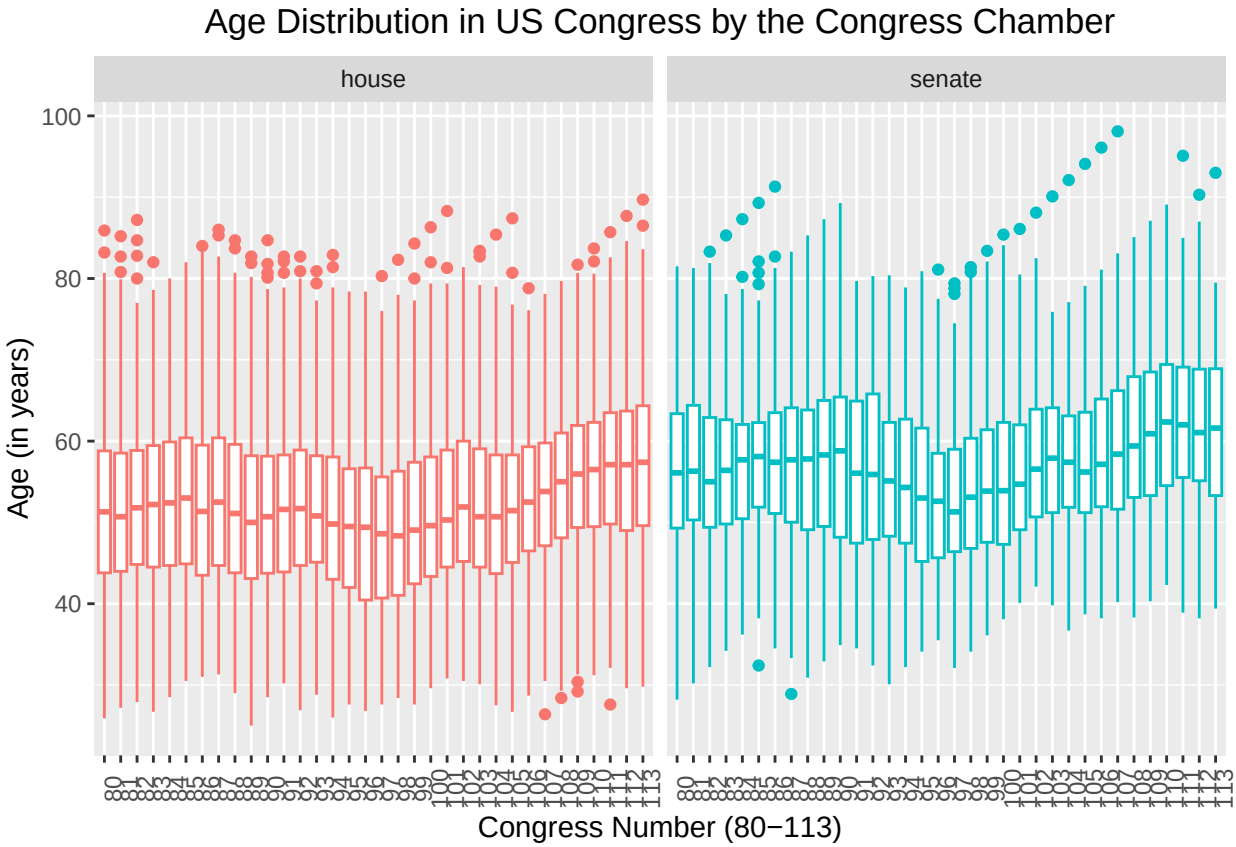


Overall in all the income quintiles, the Police killings of White and Black Americans are significantly higher than other races. Moreover, as we go from lower to higher income quintiles, the number of police killings reduces considerably. In income quintile 1 (lowest income households), most Black Americans are killed by the police, and as we go from quintile 2 to 5, the number of White Americans killed is high than other races. These observations are based solely on the national household income quintiles and the race of the people included in the dataset.

Problem 4

```
library(tidyverse)
library(fivethirtyeight)

ggplot(congress_age) +
  aes(factor(congress), age, color = chamber) +
  facet_wrap( ~ chamber) +
  geom_boxplot() +
  ggtitle('Age Distribution in US Congress by the Congress Chamber') +
  xlab('Congress Number (80-113)') +
  ylab('Age (in years)') +
  scale_fill_discrete(name = "Chamber") +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5),
        legend.position = "none")
```

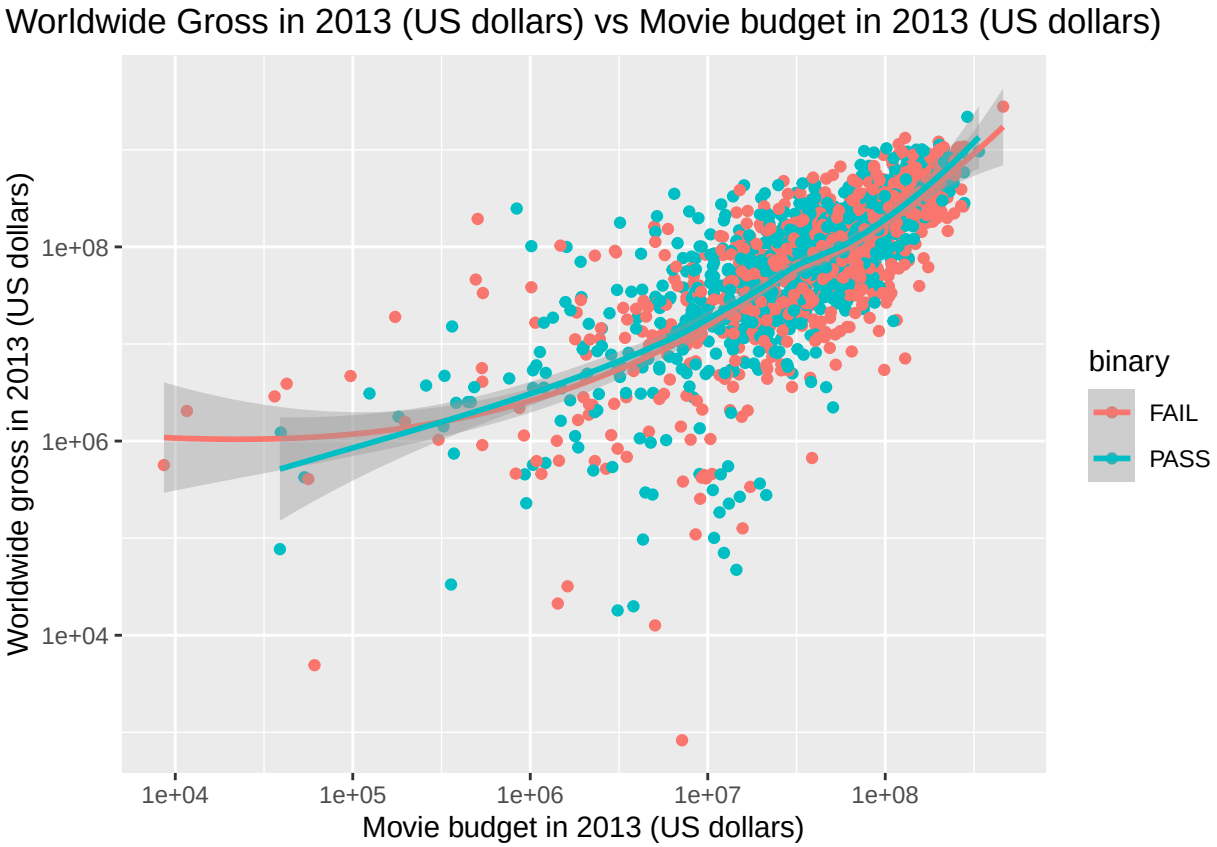


Overall, the Senate members seem slightly older than the House members. The median age of both houses remains more or less constant, with the ages going slightly lower around Congress number 95 to 100 in both chambers, and after that, they gradually increase. Most of the Congress numbers have some outliers in both chambers (more in House than Senate), with a majority being more aged than the median age, with just a handful being younger than the median. These observations are solely based on the age (in years) of the Congress number 80 to 113 from the House and Senate, included in the dataset.

Problem 5

```
library(tidyverse)
library(fivethirtyeight)

bechdel_copy <- na.omit(bechdel)
ggplot(bechdel_copy) +
  aes(budget_2013, intgross, color = binary) +
  geom_point() +
  geom_smooth() +
  ggtitle('Worldwide Gross in 2013 (US dollars) vs Movie budget in 2013 (US dollars)') +
  xlab("Movie budget in 2013 (US dollars)") +
  ylab("Worldwide gross in 2013 (US dollars)") +
  scale_x_continuous(trans = 'log10') +
  scale_y_continuous(trans = 'log10') +
  theme(plot.title = element_text(hjust = 0.5))
```



There is a positive correlation between the budget of films in US dollars and their international gross in US dollars in 2013. However, passing the Bechdel test does not appear to impact this relationship. It is evident from the data points for films that pass the test (represented in green) and those that fail (in red) are closely clustered at the top right corner of the plot, and the rest are more or less evenly distributed across the plot. It is important to note that these observations are based solely on the movie budgets and their worldwide gross in 2013 in US dollars, and Bechdel test results are included in the dataset.