IDMP_Assignment1

Mansi Pravin Thanki - 002128043

2023-01-23

Q1) Write a function of the following form: imputeNA(data, use.mean = FALSE) • data: A data.frame for which to impute the missing values • use.mean: Use the mean instead of the median for imputing continuous values The function should return a modified copy of data with missing values (NAs) imputed. Continuous variables (numeric types) should be imputed using the median or mean (according to use.mean) of the non-missing values. Categorical variables (character or factor types) should be imputed using the mode. (You may find it useful to first create a function for calculating the mode.)

```
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
```

```
# this function calculates the mode.
# reference: https://www.tutorialspoint.com/r/r_mean_median_mode.htm
Mode <- function(val, removeNA = FALSE) {
  if(removeNA){ # if removeNA = True, remove NA values
    val = val[!is.na(val)]
  }

  value <- unique(val)
  return(value[which.max(tabulate(match(val, value)))])
}
```

```
imputeNA <- function(data, use.mean = F){
  for(i in 1:ncol(data)){
    if(is.numeric(data[[i]])){
      if(use.mean){
        # if use.mean is True, uses mean to impute
        data[is.na(data[,i]), i] <- mean(data[,i], na.rm = TRUE)
      }
      else{
```

```
        # if use.mean is False, uses median to impute
        data[is.na(data[,i]), i] <- median(data[,i], na.rm = TRUE)
      }

    }
    else{
      # if data is categorical, uses mode to impute
      data[is.na(data[,i]), i] <- Mode(data[,i],T)
    }
  }
  print(data)
}
```

```
# testing function without use.mean=TRUE. Answer should impute NA values with median values.
# Categorical values will be imputed with Mode value
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
```

```
# testing function without use.mean=TRUE. Answer should impute NA values with mean values.
# Categorical values will be imputed with Mode value

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
```

Q2) Write a function of the following form: countNA(data, byrow = FALSE) • data: A data.frame for which to count the number of missing values • byrow: Should missing values be counted by row (TRUE) or by column (FALSE)? The function should return a named numeric vector giving the count of missing values (NAs) for each row or each column of data (depending on the value of byrow). The names of the result should be the rownames() or colnames() of data, whichever is appropriate.

```
countNA <- function(data, byrow = FALSE){
  margin_value = 2 # if margin value = 2, missing values are counted by column
  if(byrow){        # checks if byrow value is true, sets margin value to 1
    margin_value = 1 # if margin value = 1, missing values are counted by row
  }
  apply(X = is.na(data), margin_value, FUN = sum)
}
```

```
countNA(testdf)
```

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

```
countNA(testdf, TRUE)
```

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

Using the police_killings dataset, we would like to visualize the distribution of Americans killed by police by race and income. First, use the na.omit() function to remove missing data from the dataset. Then, visualize the count of Americans killed of each race/ethnicity, broken out by national quintile of household income (use the nat_bucket column). Do you notice any differences in the distribution of police killings based on income level?

Solution:

Do you notice any differences in the distribution of police killings based on income level?

- Yes, one can observe that there is a **linearly decreasing relationship** between police killings and income levels for all race and ethnicities. Higher the income, lesser is the count of Americans killed by police. In the below barplot,
- the x-axis represents the National Quantile of Household Income.
- y-axis represents the Count of Americans Killed by Police
- Except for White Americans, the lowest income level 1 shows highest police killings for all Race and Ethnicities.
- As we move forward to higher household income levels, the police killings count reduces. Hence, there is a **negative or inverse relationship** between police killings count and household income

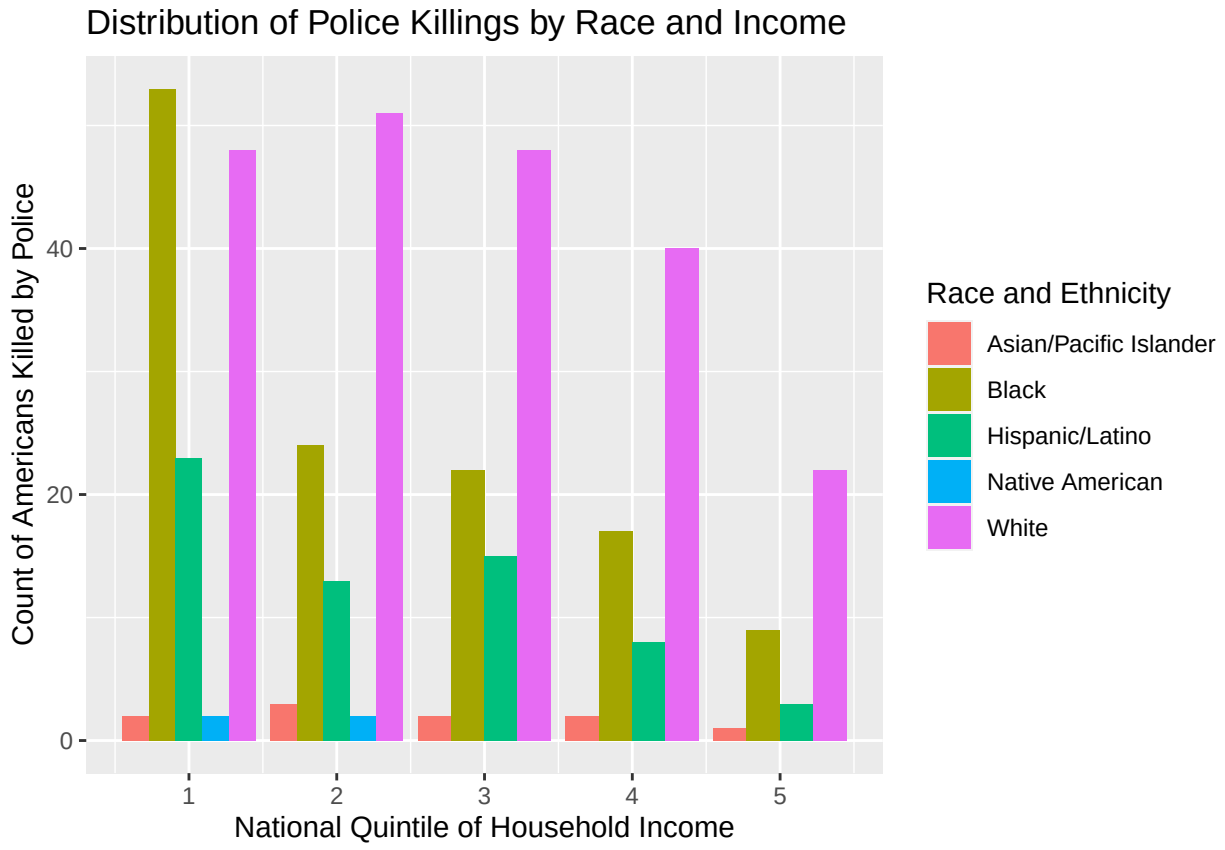
```
# Install and load the fivethirtyeight package
install.packages("fivethirtyeight" , repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/2cgj8871421bvklfhk05xfzc0000gn/T//RtmpVdWwtT/downloaded_packages
```

```
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')
```

```
data(police_killings) # Loading the dataset
police_killings_withoutNA <- na.omit(police_killings) # omitting NA values

# Plotting the count of Americans killed of each race/ethnicity,
# broken out by national quintile of household income
library(ggplot2)
ggplot(police_killings_withoutNA, aes(x = nat_bucket, fill = raceethnicity)) +
  geom_bar(position = "dodge") +
  labs(title="Distribution of Police Killings by Race and Income",
       x="National Quintile of Household Income",
       y="Count of Americans Killed by Police",
       fill="Race and Ethnicity")
```



Q4. Using the congress_age dataset, we would like to visualize the distribution of ages in US Congress. Use box-and-whiskey plots to visualize the distribution of ages for each congress number (#80 through #113), broken out by the congress chamber (House and Senate). How does the median age of congress members change over time? Do you notice any differences between the two chambers? Hint: Use as.factor(congress) to treat it as a categorical variable.

```
install.packages("fivethirtyeight" , repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/2cgj8871421bvklfhk05xfzc0000gn/T//RtmpVdWwtT/downloaded_packages
```

```
library(fivethirtyeight)
data(congress_age)
```

Solution:

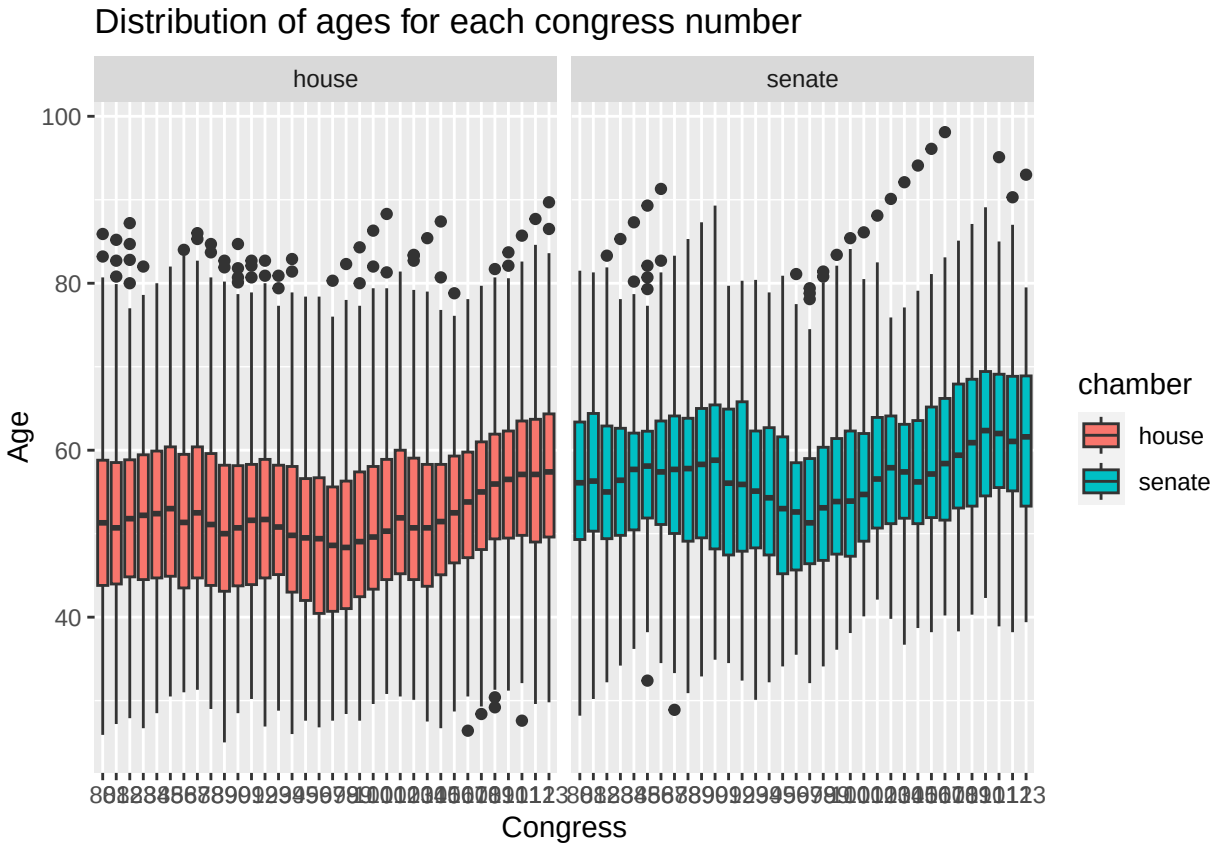
How does the median age of congress members change over time?

- For both the chambers, the changes in pattern of the median age is pretty similar.
- After the 103 congress number, the median age pretty much as increased in linear fashion for both the chambers.
- For both the chambers, 97 congress had the lowest median of all

Do you notice any differences between the two chambers?

- There are not many differences.
- The nature of outliers is very distant in senate in comparison with house.
- The median age of senate for senate is slightly higher than median age of house
- House contains members who are younger in age than members of senate.

```
library(ggplot2)
ggplot(congress_age, aes(x=as.factor(congress), y=age, fill=chamber)) +
  geom_boxplot() +
  labs(title="Distribution of ages for each congress number",
        x="Congress",
        y="Age") +
  facet_wrap(~ chamber)
```




```
install.packages("fivethirtyeight" , repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/r9/2cgj8871421bvklfhk05xfzc0000gn/T//RtmpVdWwtT/downloaded_packages
```

```
library(fivethirtyeight)
data(bechdel)
```

Using the bechdel dataset, we would like to investigate if there is a relationship between passing the Bechdel test and the amount of money spent and made from a movie. The Bechdel test is a basic set of criteria designed to reveal trends of gender bias in the movies. The test asks: does a movie (1) have at least two female characters (2) who talk to each other (3) about something other than a man? Plot the worldwide gross (in 2013 dollars) as the dependent variable against the movie budget (in 2013 dollars) as the independent variable, using color to indicate whether the movie passes the Bechdel test or not. Describe the relationship between movie budget and movie gross, and whether passing the Bechdel test seems to have an affect on this relationship.

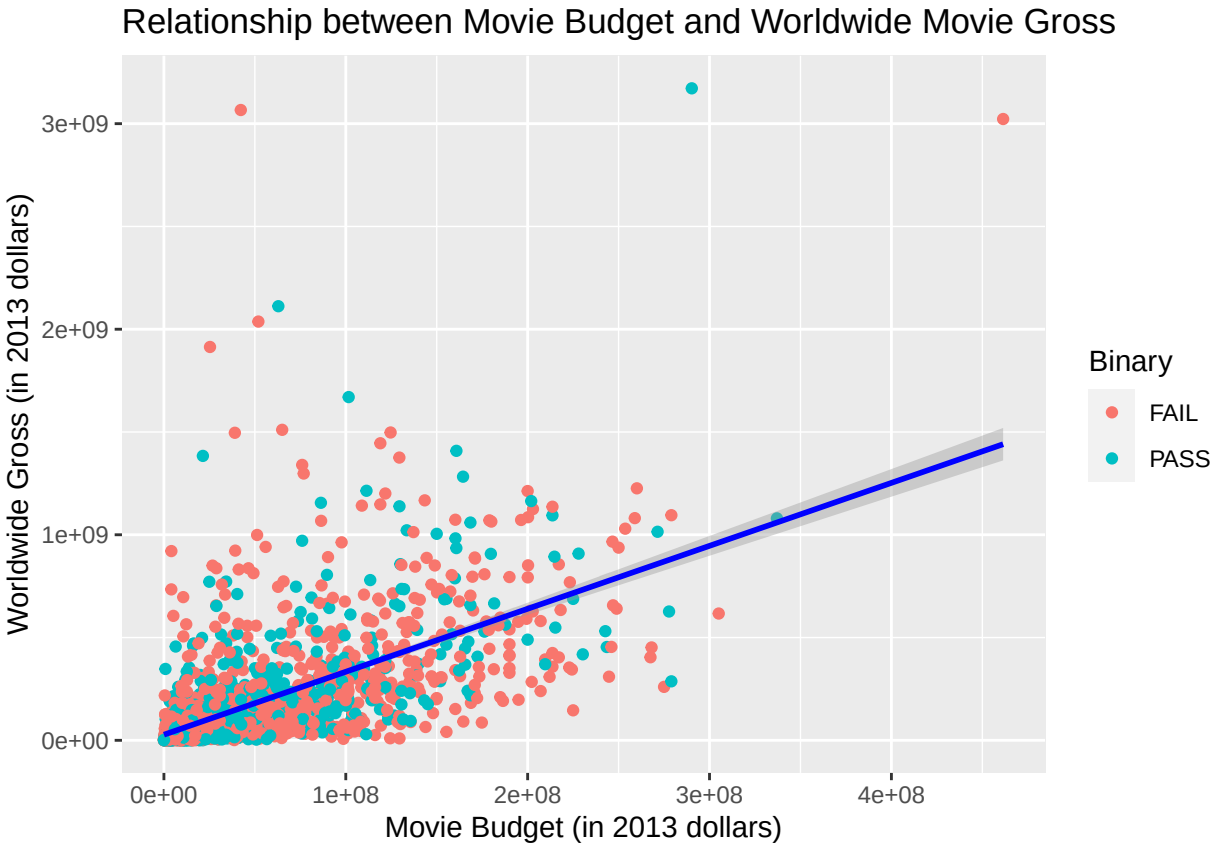
Solution:

- Passing the Bechdel test seems to have **no effect** on the relationship between movie budget and movie gross.
- In the scatter plot below, the green points indicate the movies that passed the Bechdel test.
- The red points denote the movies that failed the Bechdel test.

- As you can see, it is very difficult to infer any relationship whether passing the test affects the movie budget and movie gross relationship. Both the passed and failed points are clustered together.
- I also tried using log() on x and y values to scale down and get a better view of the relationship. However, even after scaling down the failed and passed test points were too clustered together to make out any inference.
- So, the answer is **NO EFFECT** on the relationship.

```
library(ggplot2)

ggplot(bechdel, aes(x=budget_2013, y=intgross_2013)) +
  geom_point(aes(color = factor(binary))) +
  stat_smooth(method = "lm",
        col = "blue",
        se = TRUE,
        size = 1) +
  labs(title="Relationship between Movie Budget and Worldwide Movie Gross",
        x="Movie Budget (in 2013 dollars)",
        y="Worldwide Gross (in 2013 dollars)",
        color="Binary")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.

## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: Removed 11 rows containing non-finite values (‘stat_smooth()‘).

## Warning: Removed 11 rows containing missing values (‘geom_point()‘).
```


Relationship between Movie Budget and Worldwide Movie Gross
Worldwide Gross (in 2013 dollars)
3e+09
2e+09
1e+09
0e+00
0e+00
1e+08
2e+08
3e+08
4e+08
Movie Budget (in 2013 dollars)
Binary
FAIL
PASS