Introduction to Data Science HW 4

```
# Enter your name here: Victoria Haley
```

Copyright Jeffrey Stanton, Jeffrey Saltz, and Jasmina Tacheva

Attribution statement: (choose only one and delete the rest)

```
# 2. I did this homework with help from the book and the professor and these Internet sources: statisti
```

Reminders of things to practice from previous weeks: Descriptive statistics: mean( ) max( ) min( ) Coerce to numeric: as.numeric( )

Part 1: Use the Starter Code

Below, I have provided a starter file to help you.

Each of these lines of code **must be commented** (the comment must that explains what is going on, so that I know you understand the code and results).

```
library(jsonlite)
dataset <- url("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json")
readlines <- jsonlite::fromJSON(dataset)
df <- readlines$objects$person
```

A. Explore the **df** dataframe (e.g., using head() or whatever you think is best).

```
summary(df)
```

```
##   bioguideid          birthday            cspanid          firstname        
##  Length:100         Length:100         Min.   :    260   Length:100        
##  Class :character   Class :character   1st Qu.:  25277   Class :character  
##  Mode  :character   Mode  :character   Median :  68489   Mode  :character  
##                                        Mean   : 584001                     
##                                        3rd Qu.:1004138                     
##                                        Max.   :9269028                     
##                                        NA's   :11                          
##     gender          gender_label         lastname             link          
##  Length:100         Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##   middlename            name              namemod            nickname        
##  Length:100         Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##      osid              pvsid             sortname          twitterid        
##  Length:100         Length:100         Length:100         Length:100        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   youtubeid
##  Length:100
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# Using summary(), I'm able to view the overview of non-numeric variables as well as numerical summarie
```

B. Explain the dataset o What is the dataset about? o How many rows are there and what does a row represent? o How many columns and what does each column represent?

```
# The dataset is about the information of senators. There are 100 rows (observations), each with the in
```

C. What does running this line of code do? Explain in a comment:

```
vals <- substr(df$birthday,1,4)
# Running this line of code stores the first 4 elements in the birthday column of the df dataset as a v
```

D. Create a new attribute 'age' - how old the person is **Hint:** You may need to convert it to numeric first.

```
age <- 2022 - as.numeric(vals)
```

E. Create a function that reads in the role json dataset, and adds the age attribute to the dataframe, and returns that dataframe

```
agefunc <- function(inputDF) {
  df <- data.frame(df, readlines$objects$role_type, age)
  return(df)
}
```

F. Use (call, invoke) the function, and store the results in df

```
df <- agefunc(df)
```

Part 2: Investigate the resulting dataframe 'df'

A. How many senators are women?

```
sum(df$gender == "female")
```

```
## [1] 24
```

```
#24 senators are women
```

B. How many senators have a YouTube account?

```
sum(is.na(df$youtubeid) == "FALSE")
```

```
## [1] 73
```

```
#73 senators have a YouTube account
```

C. How many women senators have a YouTube account?

```
tapply(df$gender == "female", (is.na(df$youtubeid) == "FALSE"), sum)
```

```
## FALSE  TRUE
##     8    16
```

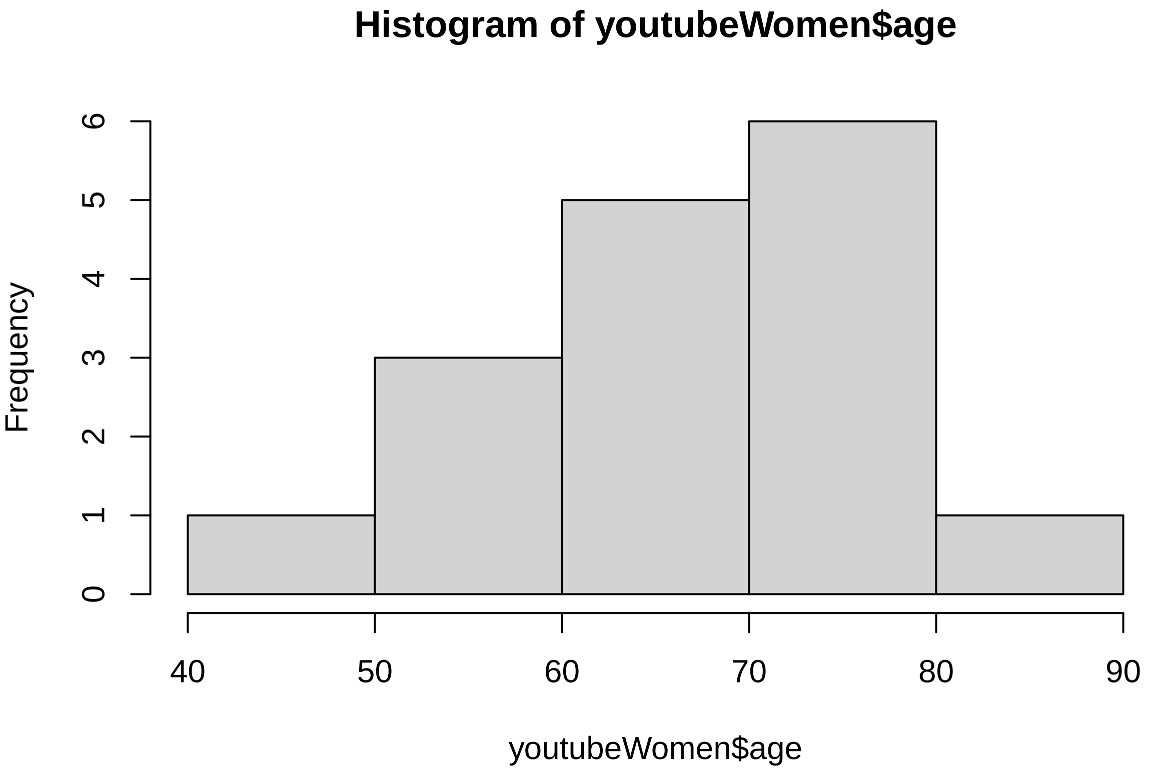
```
#16 women senators have a YouTube account
```

D. Create a new dataframe called **youtubeWomen** that only includes women senators who have a YouTube account.

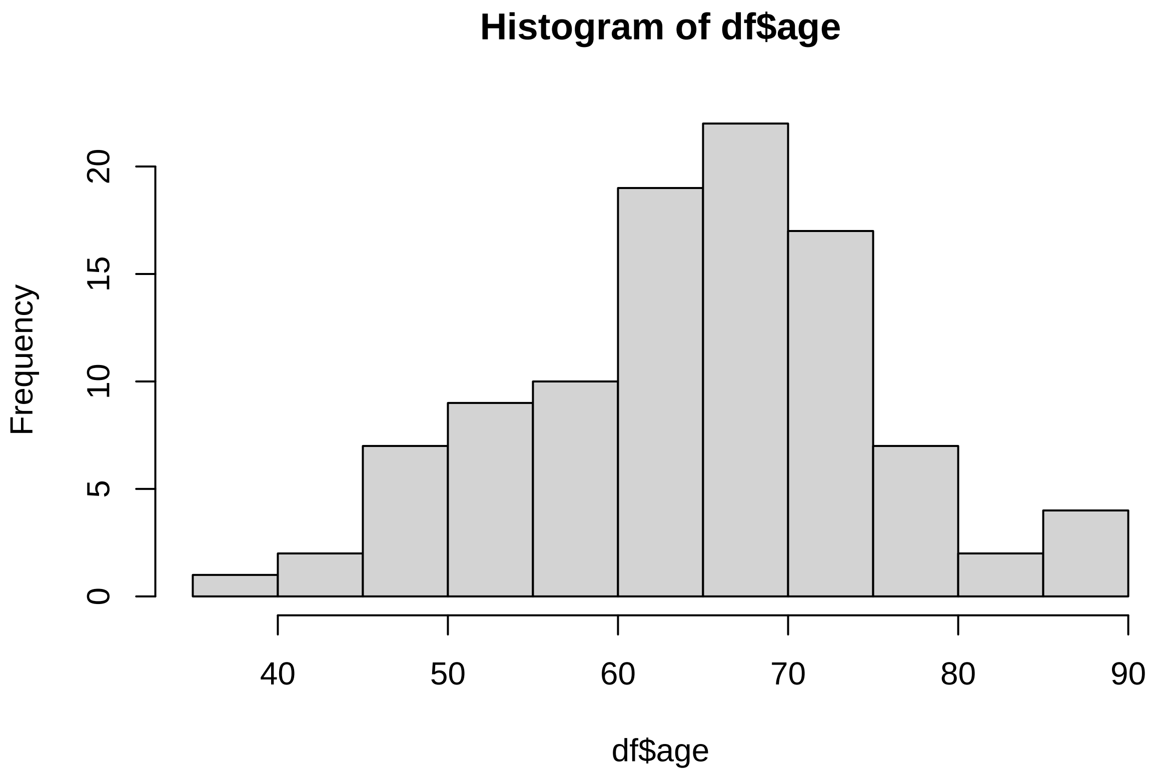
```
womenOnly <- subset(df, df$gender !="male")
youtubeWomen <- womenOnly[complete.cases(womenOnly$youtubeid), ]
```

E. Make a histogram of the **age** of senators in **youtubeWomen**, and then another for the senetors in **df**. Add a comment describing the shape of the distributions.

```
hist(youtubeWomen$age)
```



```
hist(df$age)
```



```
#Both histograms have a "bell" like shape, however the histogram for df$age is more normally distribute
```