# Introduction to Data Science HW 4

```
# Enter your name here: Cy Seeley
```

**Copyright Jeffrey Stanton, Jeffrey Saltz, and Jasmina Tacheva**

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

```
# 1. I did this homework by myself, with help from the book and the professor.
```

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).

```
#This code is simply calling the jsonlite package in R
library(jsonlite)
#This is creating an object called dataset which uses the url function and follows a given url and brin
dataset <- url("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json")
#This line of code is converting our URL dataset to a dataframe which is usable in R
readlines <- jsonlite::fromJSON(dataset)
#This code just trims the original data set to something a bit smaller and easier to work with
df <- readlines$objects$person
```

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

head(df)

B. Explain the dataset o What is the dataset about? #This is a dataset about different congressman and contains variables such as their name and social media id's o How many rows are there and what does a row represent? #There is 100 rows and each of them represents a different congressman o How many columns and what does each column represent? #There is 17 different columns and each one is a different varaible such as their name, social media id's, and genders

C. What does running this line of code do? Explain in a comment: #it creates a list of the years in which each congressman was born

```
vals <- substr(df$birthday,1,4)
```

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

valsnum <- as.numeric(vals) thisyear <- format(Sys.Date(), "%Y") thisyearnum <- as.numeric(thisyear) age <- thisyearnum - valsnum

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

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

  vals <- substr(df$birthday, 1, 4)
  valsnum <- as.numeric(vals)
  thisyear <- format(Sys.Date(), "%Y")
  thisyearnum <- as.numeric(thisyear)
  age <- thisyearnum - valsnum

  df$age <- age

  return(df)
}
```

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

```r
df <- addnew()
```

## Part 2: Investigate the resulting dataframe ‘df’

A. How many senators are women?

sum(df$gender == “female”) #There are 24 female senators

B. How many senators have a YouTube account?

#73 senators have a youtube account

```r
sum(!is.na(df$youtubeid))
```

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

C. How many women senators have a YouTube account? #There is 16 female senators with a youtube

```r
womenyoutube <- subset(df, gender=='female' & !is.na(youtubeid))
```

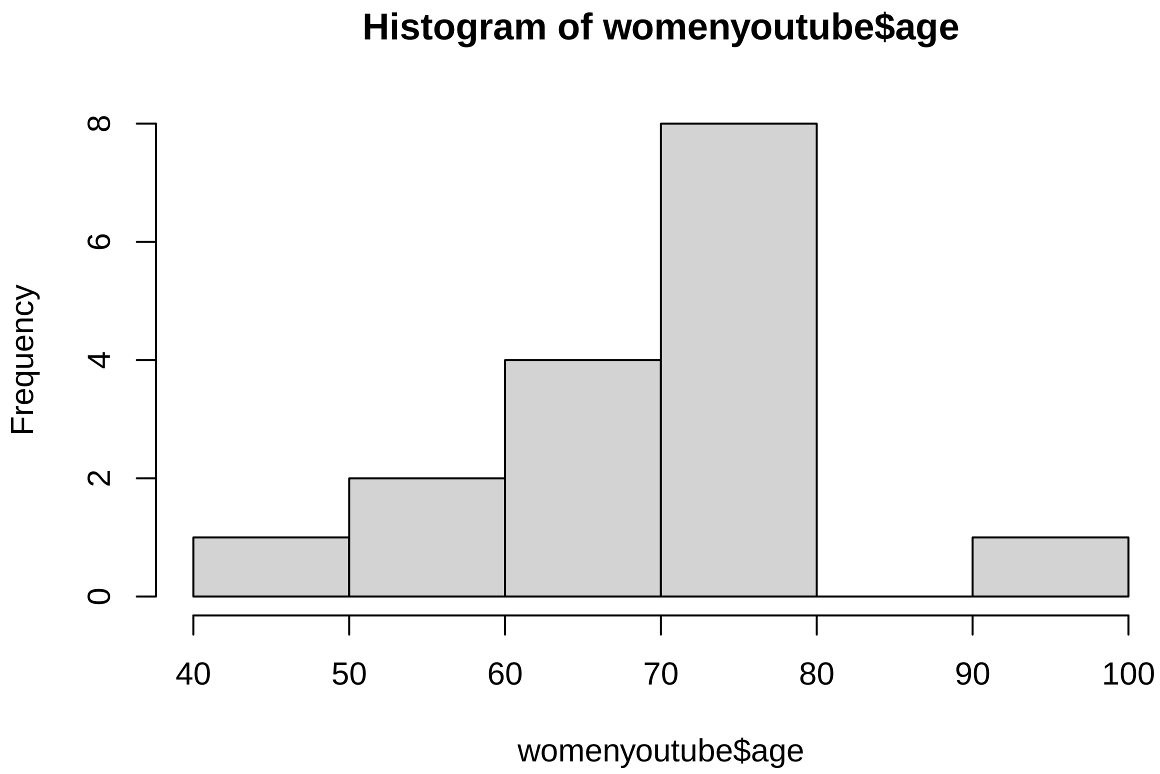
D. Create a new dataframe called **youtubeWomen** that only includes women senators who have a YouTube account.

```r
womenyoutube <- subset(df, gender=='female' & !is.na(youtubeid))
```

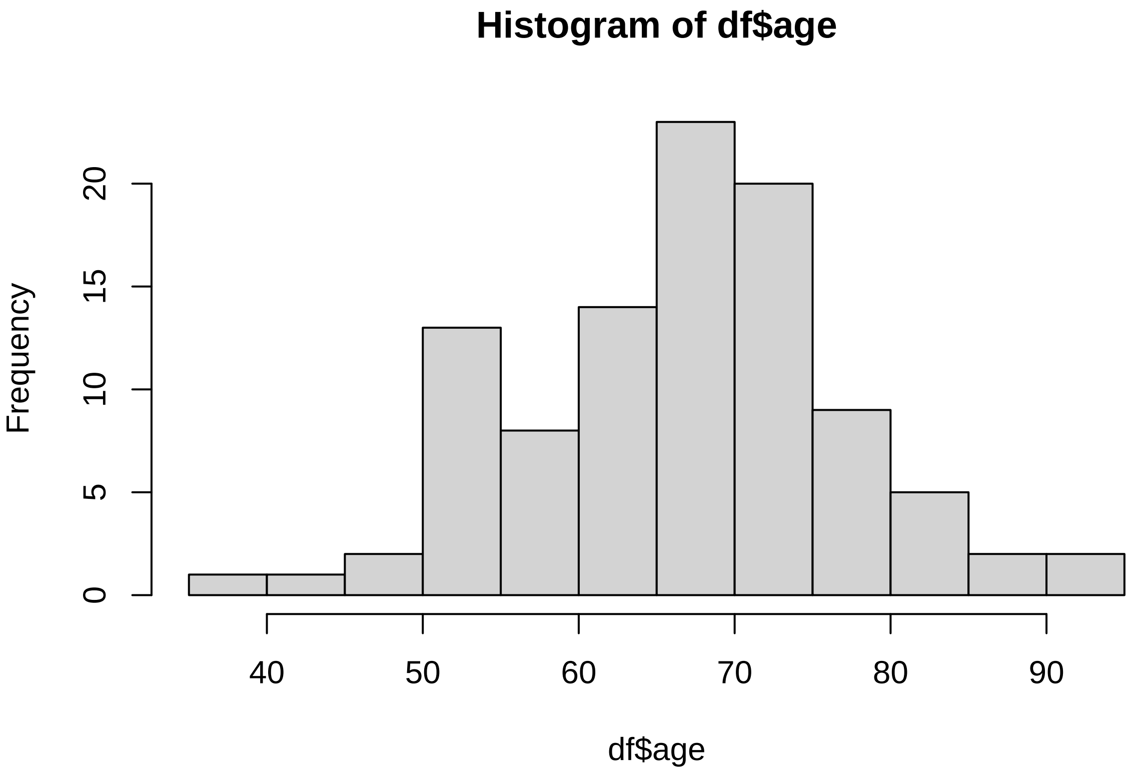
E. Make a histogram of the **age** of senators in **youtubeWomen**, and then another for the senetors in **df**. Add a comment describing the shape of the distributions.

#Both distributions are just a bit skewed to the left but they both seem to be approavhing normal as the number of senators included goes up

```
hist(womenyoutube$age)
```



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


Histogram of df$age
Frequency
0
5
10
15
20
40
50
60
70
80
90
df$age