Intro to Data Science - HW 5

```
# Enter your name here: Chaithra Kopparam Cheluvaiah
```


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).

```
# loading the RCurl package used for accessing internet data
library(RCurl)

# loading the jsonlite package required for decoding json
library(jsonlite)

# getting the JSON data from a URL
dataset<-getURL("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json")

# parsing JSON data to R object using a function in jsonlite package
readlines <- jsonlite::fromJSON(dataset)

# getting person data frame which is inside 'objects' list
df <- readlines$objects$person
```

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

```
# head(df) # shows first few rows of the dataframe
# tail(df) # shows last few rows of the dataframe

# shows numerical summaries of numeric variable and overview of
# non-numeric variables
# summary(df)

# library(tidyverse)
# glimpse(df) # shows structure (rows and columns) of data frame
```

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?

```
# 1) demography of senior senators
# 2) 100 rows; each row represents senior senator from different states
# 3) 17 columns; each column represents variables related to demography
# of senators
```

Part 2: Investigate the resulting dataframe

C. How many senators are women?

```
femaleSenators <- df[df$gender=="female",] # subsetting dataframe with logical
# condition to filter females from gender column

nrow(femaleSenators) # number of rows
```

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

```
length(femaleSenators$gender) # number of items in gender column
```

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

D. How many senators have a YouTube account?

```
youtubeDf <- df[!is.na(df$youtubeid),] # sub-setting data frame to retrieve all
# the senators having youtube with the help of is.na function

nrow(youtubeDf) # number of rows
```

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

E. How many women senators have a YouTube account?

```
# logical condition to check senators that have youtube id and gender is female
areWomenYoutubers <- !is.na(df$youtubeid) & df$gender=="female"

# creating new data frame with above criteria
femaleYoutubers <- df[areWomenYoutubers,]

# number of rows
nrow(femaleYoutubers)
```

```
## [1] 16
```

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

```
# filter all the rows in youtubeDS having gender as females
youtubeWomen <- youtubeDf[youtubeDf$gender=="female",]
```

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

```
# extracting first 4 characters from birthdays which represents year of birth
# and storing it in a new column called year
youtubeWomen$year <- substr(youtubeWomen$birthday,1,4)
youtubeWomen$year
```

```
##  [1] "1957" "1950" "1968" "1958" "1933" "1950" "1962" "1952" "1947" "1966"
## [11] "1960" "1976" "1949" "1951" "1952" "1947"
```

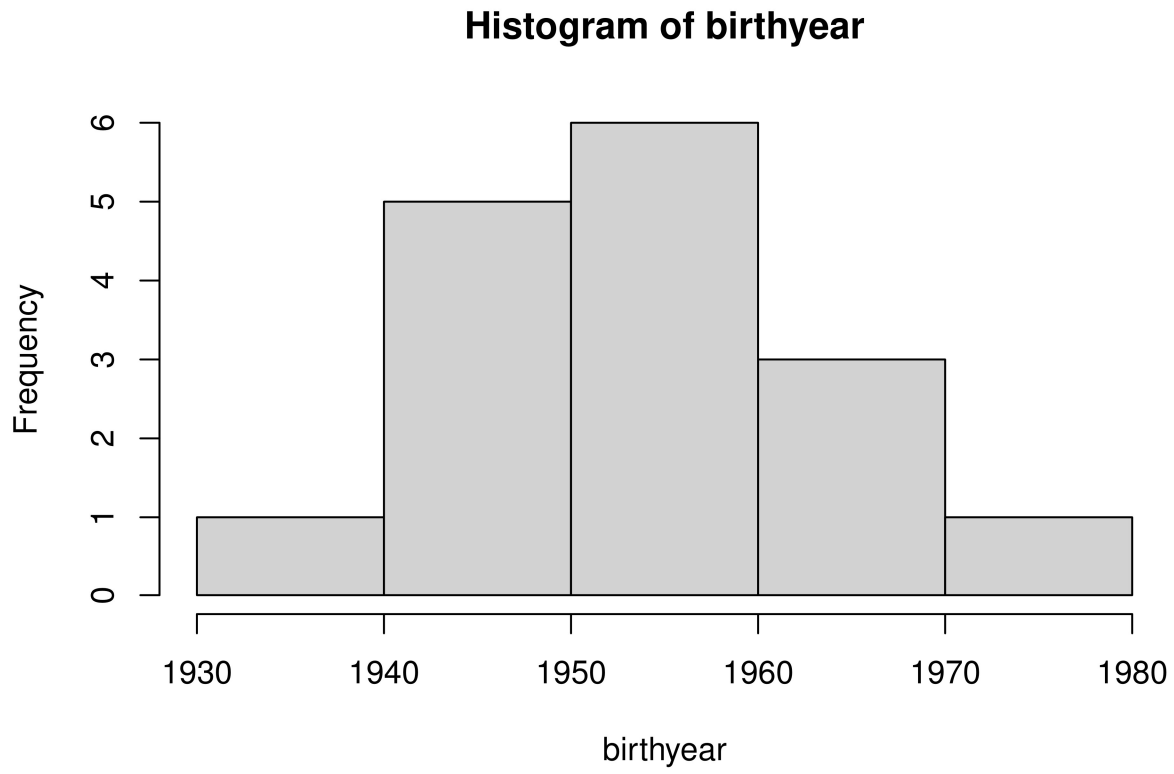
H. Use this new variable to calculate the mean **birthyear** in **youtubeWomen**. **Hint:** You may need to convert it to numeric first.

```
# converting year variable from string to numeric
birthyear <- as.numeric(youtubeWomen$year)
mean(birthyear) # average of birth year
```

```
## [1] 1954.875
```

I. Make a histogram of the **birthyears** of senators in **youtubeWomen**. Add a comment describing the shape of the distribution.

```
hist(birthyear) # plotting histogram with birth years
```



```
# shape of the distribution is not perfectly Symmetric. hence, we can consider
# it as very close to normal distribution.
```