Intro to Data Science - HW 5

```
# Enter your name here: Hrishikesh Telang
```

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).

```
library(RCurl) #We are calling the Rcurl package
library(jsonlite) #We are calling the jsonlite package
dataset <- getURL("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json") #We are calling the
readlines <- jsonlite::fromJSON(dataset) #This function converts JSON data to an R object
df <- readlines$objects$person #This line stores all the JSON data into df
```

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

```
head(df)
```

```
##   bioguideid   birthday cspanid firstname gender gender_label  lastname
## 1    C000880 1951-05-20   26440   Michael   male         Male     Crapo
## 2    G000386 1933-09-17    1167   Charles   male         Male  Grassley
## 3    L000174 1940-03-31    1552   Patrick   male         Male     Leahy
## 4    M001153 1957-05-22 1004138      Lisa female       Female Murkowski
## 5    M001111 1950-10-11   25277     Patty female       Female    Murray
## 6    S000148 1950-11-23    5929   Charles   male         Male   Schumer
##                                                                  link middlename
## 1    https://www.govtrack.us/congress/members/michael_crapo/300030         D.
## 2 https://www.govtrack.us/congress/members/charles_grassley/300048         E.
## 3    https://www.govtrack.us/congress/members/patrick_leahy/300065         J.
## 4   https://www.govtrack.us/congress/members/lisa_murkowski/300075         A.
## 5     https://www.govtrack.us/congress/members/patty_murray/300076
## 6  https://www.govtrack.us/congress/members/charles_schumer/300087         E.
##                                            name namemod nickname      osid pvsid
## 1     Sen. Michael â€œMikeâ€<9d> Crapo [R-ID]             Mike N00006267 26830
## 2 Sen. Charles â€œChuckâ€<9d> Grassley [R-IA]            Chuck N00001758 53293
## 3                   Sen. Patrick Leahy [D-VT]                  N00009918 53353
## 4                  Sen. Lisa Murkowski [R-AK]                  N00026050 15841
## 5                    Sen. Patty Murray [D-WA]                  N00007876 53358
## 6  Sen. Charles â€œChuckâ€<9d> Schumer [D-NY]            Chuck N00001093 26976
##                                          sortname     twitterid
## 1     Crapo, Michael â€œMikeâ€<9d> (Sen.) [R-ID]     MikeCrapo
## 2 Grassley, Charles â€œChuckâ€<9d> (Sen.) [R-IA] ChuckGrassley
## 3                   Leahy, Patrick (Sen.) [D-VT]  SenatorLeahy
## 4                  Murkowski, Lisa (Sen.) [R-AK] LisaMurkowski
## 5                    Murray, Patty (Sen.) [D-WA]   PattyMurray
## 6  Schumer, Charles â€œChuckâ€<9d> (Sen.) [D-NY]    SenSchumer
##             youtubeid
## 1        senatorcrapo
## 2    senchuckgrassley
## 3 SenatorPatrickLeahy
## 4    senatormurkowski
## 5  SenatorPattyMurray
## 6      SenatorSchumer
```

```
tail(df)
```

```
##     bioguideid   birthday cspanid firstname gender gender_label     lastname
## 95     T000278 1954-09-18      NA     Tommy   male         Male   Tuberville
## 96     H000273 1952-02-07      NA      John   male         Male Hickenlooper
## 97     H000601 1959-08-14      NA      Bill   male         Male      Hagerty
## 98     P000145 1973-03-22      NA Alejandro   male         Male      Padilla
## 99     O000174 1987-02-16      NA       Jon   male         Male       Ossoff
## 100    W000790 1969-07-23      NA   Raphael   male         Male      Warnock
##                                                                     link
## 95   https://www.govtrack.us/congress/members/tommy_tuberville/456796
## 96  https://www.govtrack.us/congress/members/john_hickenlooper/456797
## 97       https://www.govtrack.us/congress/members/bill_hagerty/456798
## 98  https://www.govtrack.us/congress/members/alejandro_padilla/456856
## 99         https://www.govtrack.us/congress/members/jon_ossoff/456857
## 100   https://www.govtrack.us/congress/members/raphael_warnock/456858
##     middlename                                        name namemod nickname
## 95      Hawley                Sen. Tommy Tuberville [R-AL]
## 96      Wright               Sen. John Hickenlooper [D-CO]
## 97     Francis                    Sen. Bill Hagerty [R-TN]
## 98             Sen. Alejandro â€œAlexâ€<9d> Padilla [D-CA]             Alex
## 99                                  Sen. Jon Ossoff [D-GA]
## 100   Gamaliel                 Sen. Raphael Warnock [D-GA]
##     osid  pvsid                                       sortname      twitterid
## 95  <NA> 188306                Tuberville, Tommy (Sen.) [R-AL]  SenTuberville
## 96  <NA>   <NA>               Hickenlooper, John (Sen.) [D-CO]           <NA>
## 97  <NA> 128466                    Hagerty, Bill (Sen.) [R-TN] SenatorHagerty
## 98  <NA>   <NA> Padilla, Alejandro â€œAlexâ€<9d> (Sen.) [D-CA] SenAlexPadilla
## 99  <NA>   <NA>                      Ossoff, Jon (Sen.) [D-GA]           <NA>
## 100 <NA>   <NA>                 Warnock, Raphael (Sen.) [D-GA] SenatorWarnock
##     youtubeid
## 95       <NA>
## 96       <NA>
## 97       <NA>
## 98       <NA>
## 99       <NA>
## 100      <NA>
```

```
View(df)
```

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. The dataset contains the information of all the US Senators
#2. There are 100 rows and every row represents a senator
#3. There are 17 columns and every column represents the unique ID, attributes (or personal information
```

Part 2: Investigate the resulting dataframe

C. How many senators are women?

```
#Using nrow
nrow(df[df$gender == 'female',])
```

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

```
#Using sum
sum(df$gender == 'female')
```

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

D. How many senators have a YouTube account?

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

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

```
#Using nrow
nrow(df[is.na(df$youtubeid) == FALSE,])
```

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

E. How many women senators have a YouTube account?

```
nrow(df[df$gender == 'female' & is.na(df$youtubeid) == FALSE,])
```

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

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

```
youtubeWomen <- df[df$gender == 'female' & is.na(df$youtubeid) == FALSE,]
View(youtubeWomen)
```

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

```
youtubeWomen$year <- substr(youtubeWomen$birthday,1,4)
#The code of line essentially subsets the year from the 'birthday' column and adds it in a new 'year' c
View(youtubeWomen)
```

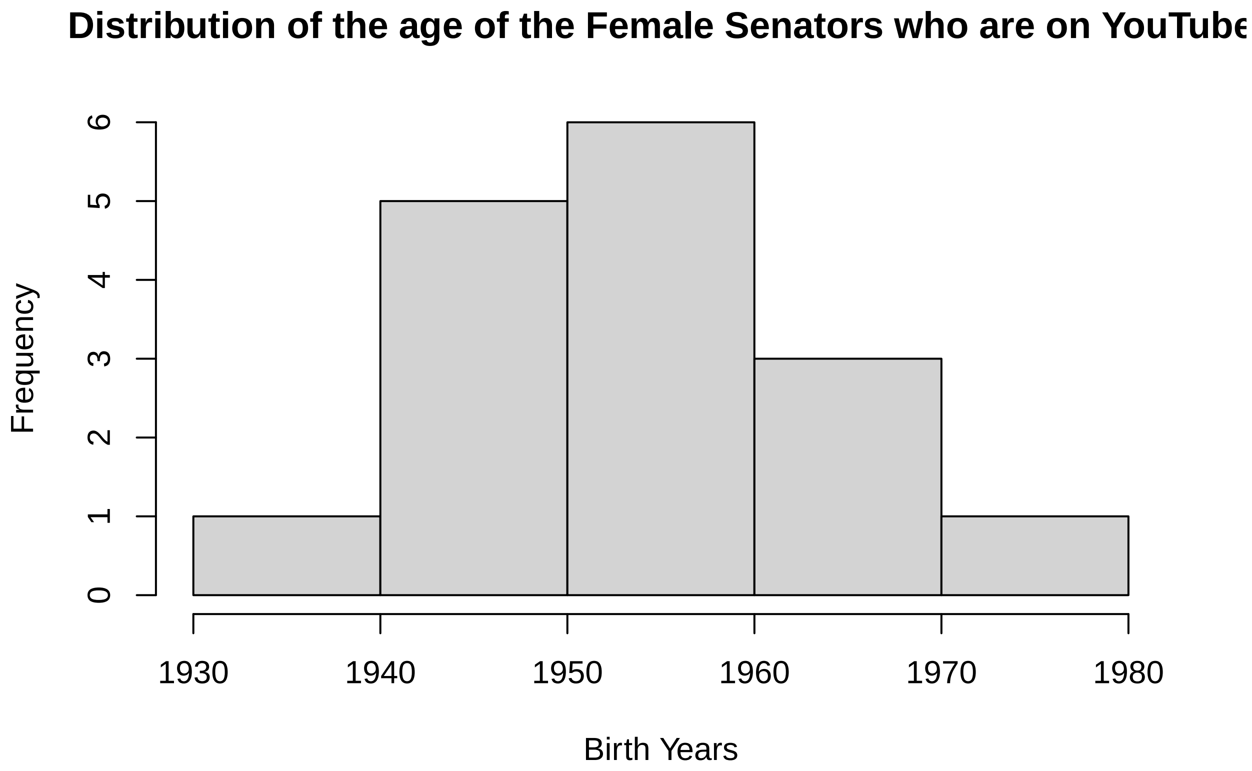
H. Use this new variable to calculate the mean **birthyear** in **youtubeWomen**. **Hint:** You may need to convert it to numeric first.

```
youtubeWomen$year <- as.numeric(as.character(youtubeWomen$year))
mean(youtubeWomen$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(youtubeWomen$year,
     main='Distribution of the age of the Female Senators who are on YouTube',
     xlab='Birth Years')
```



```
#The shape of the bell curve is slightly shifted towards the left, which means that there is a greater
```