Intro to Data Science - HW 5

```
# Enter your name here: Bhavya Shah
```

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

```
# 3. I did this homework with help from Shrey Sheth but did not cut and paste any code.
```

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).

```
#install.packages('RCurl') # This command is used to install Rcurl package
#install.packages('jsonlite') # This command is used to install jsonlite package
library(RCurl) # This command is used to load the functions of the Rcurl package
library(jsonlite) # This command is used to load the functions of the jsonlite package
dataset <- getURL("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json") # This line is gets the data from the website we need to use
readlines <- jsonlite::fromJSON(dataset) # This command helps us to get the data in the structured format which we had received originally in json format
df <- readlines$objects$person # This command gets the dataframe which is stored in person column thereby giving us the actual data in the dataframe
```

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
## 1     Sen. Michael â€œMikeâ€\u009d Crapo [R-ID]             Mike N00006267
## 2 Sen. Charles â€œChuckâ€\u009d Grassley [R-IA]            Chuck N00001758
## 3                    Sen. Patrick Leahy [D-VT]                  N00009918
## 4                   Sen. Lisa Murkowski [R-AK]                  N00026050
## 5                     Sen. Patty Murray [D-WA]                  N00007876
## 6  Sen. Charles â€œChuckâ€\u009d Schumer [D-NY]            Chuck N00001093
##   pvsid                                        sortname     twitterid
## 1 26830     Crapo, Michael â€œMikeâ€\u009d (Sen.) [R-ID]     MikeCrapo
## 2 53293 Grassley, Charles â€œChuckâ€\u009d (Sen.) [R-IA] ChuckGrassley
## 3 53353                    Leahy, Patrick (Sen.) [D-VT]  SenatorLeahy
## 4 15841                   Murkowski, Lisa (Sen.) [R-AK] LisaMurkowski
## 5 53358                     Murray, Patty (Sen.) [D-WA]   PattyMurray
## 6 26976  Schumer, Charles â€œChuckâ€\u009d (Sen.) [D-NY]    SenSchumer
##             youtubeid
## 1        senatorcrapo
## 2    senchuckgrassley
## 3 SenatorPatrickLeahy
## 4    senatormurkowski
## 5  SenatorPattyMurray
## 6      SenatorSchumer
```

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 senator details including their youtube and twitter id with the other basic details
# There are 100 rows and each row represents details of senators like name, DOB, gender, social media ids.
# There are 17 columns where each one represents a category of information and when collectively used gives all the categories that are required by the dataset
```

Part 2: Investigate the resulting dataframe

A. Describe what you see when you run the **table()** function on the **gender** variable.

```
table(df$gender)
```

```
##
## female   male
##     24     76
```

```
# The count of number of rows whose gender are male and female is visible
```

A1. Generate the count of number of females and number of males, using the tidyverse **group_by()**, **summarise()** and **n()** functions.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   1.0.1 
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.3.0      ✔ stringr 1.5.0 
## ✔ readr   2.1.3      ✔ forcats 1.0.0 
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ tidyr::complete() masks RCurl::complete()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ purrr::flatten()  masks jsonlite::flatten()
## ✖ dplyr::lag()      masks stats::lag()
```

```
df %>%
  group_by(gender) %>%
  summarise(n=n())
```

```
## # A tibble: 2 × 2
##   gender     n
##   <chr>  <int>
## 1 female    24
## 2 male      76
```

B. How many senators are women?

```
# According to the above dataset 24 senators are female.
```

C. How many senators don’t have a YouTube account?
Hint: You can use the **is.na** function to locate the rows for which the YouTube account is missing and then wrap it in the **nrow()** or **sum** function to count the number of missing instances.

```
noytaccount <- is.na(df$youtubeid)
sum(noytaccount)
```

```
## [1] 27
```

```
# 27 senators do not have a youtube account
```

D. Using the approach in C, i.e.using the **is.na()** function, show how many senators ***do*** have a YouTube account. **Hint:** You can reverse the **is.na()** function by placing a **!** in front of it **- !is.na( )**.

```
ytaccount<-!is.na(df$youtubeid)
sum(ytaccount)
```

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

E. How many women senators have a YouTube account?

```
womenytaccount<-df[df$gender=="female",]
woytaccount<-!is.na(womenytaccount$youtubeid)
sum(woytaccount)
```

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

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

```
youtubeWomen<-data.frame(womenytaccount %>% drop_na(youtubeid))
```

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

```
youtubeWomen$year <- substr(youtubeWomen$birthday,1,4)
# this command is uded to get first four characters of a string in a particular column and row
```

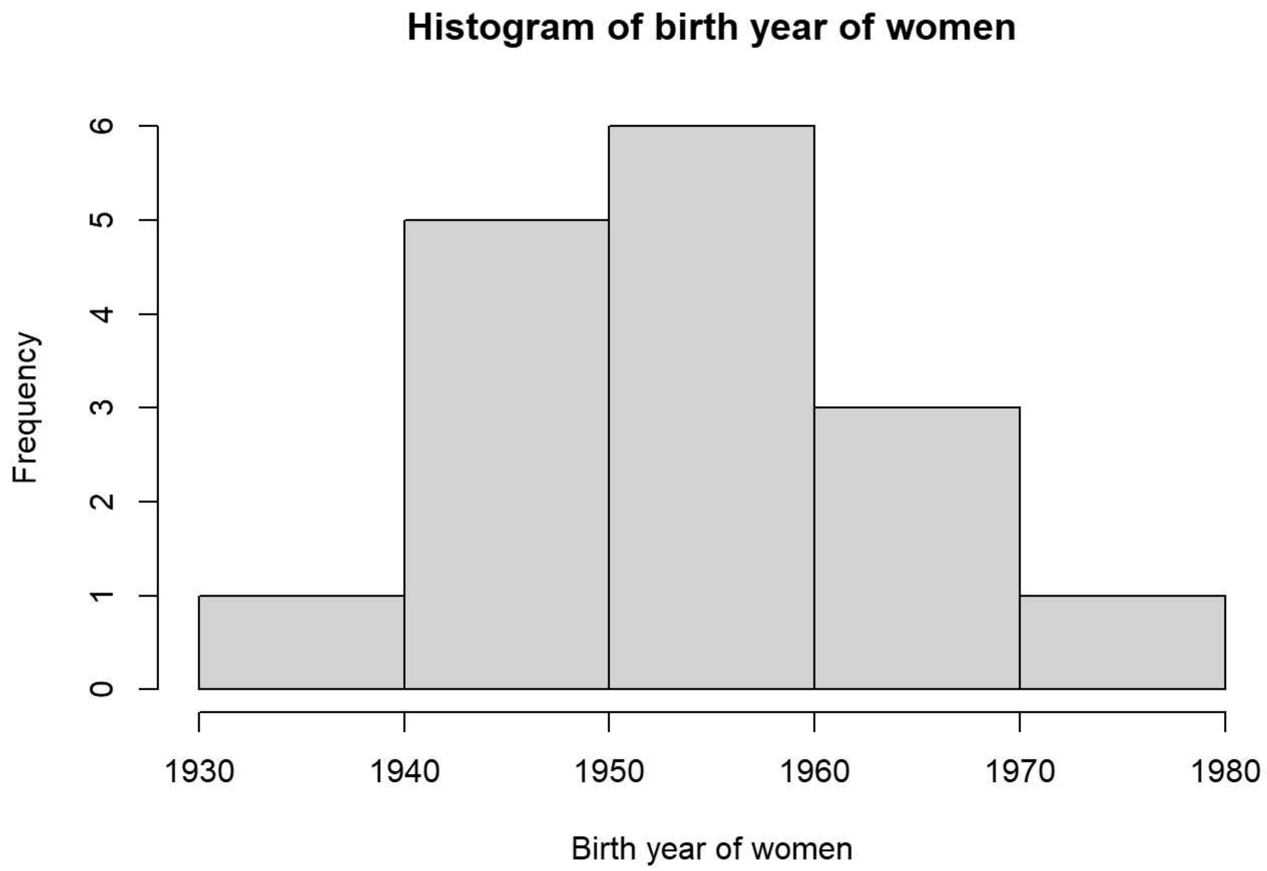
H. Use this new variable to calculate the mean **birthyear** in **youtubeWomen**.
Hint: You may need to convert it to numeric first using the **as.numeric()** function.

```
mean(as.numeric(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.

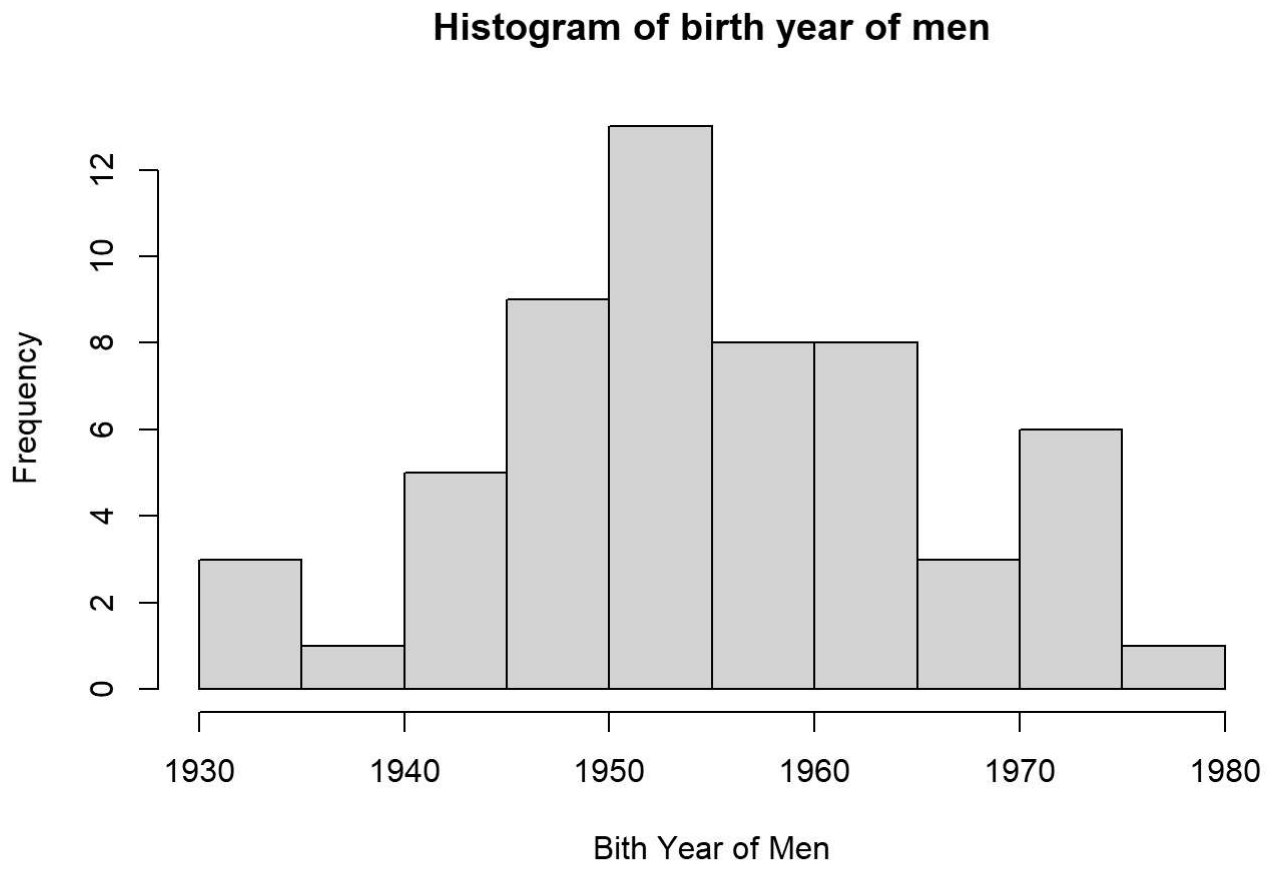
```
hist(as.numeric(youtubeWomen$year), main= "Histogram of birth year of women", xlab="Birth year o
f women")
```



```
# the shape of distribution is somewhat bimodal means it has two distinct peaks and others value
s are much lesser than the peak values
```

J. Create a dataframe called **youtubeMen** which only includes male senators with a youTube account. Repeat steps G & H for this dataframe and create a histogram of the birthyears in it. Compare the shape and properties of this histogram to the one in H.

```
men <- df[df$gender=="male",]
youtubeMen<-data.frame(men %>% drop_na(youtubeid))
youtubeMen$year <- substr(youtubeMen$birthday,1,4)
hist(as.numeric(youtubeMen$year), main= "Histogram of birth year of men", xlab = "Bith Year of M
en")
```



```
# the histogram of men birth year is random where the median of values lie between 1950 to 1955
# As compared to the women birth year most of the them lie between 1950 to 1960 and in women his
togram we can see there are 5 bar that represents the sample because women are only 16, But men
are 57 hence there are 10 bars which are distributed over sample frequency.
```

K. Take a look at this article (https://www.theguardian.com/us-news/ng-interactive/2018/nov/15/new-congress-us-house-of-representatives-senate) - explore its interactive features and focus specifically on the section on **gender**. Relating what you learned from the article back to our Senate data, who might feel left out and/or unrepresented based on the current gender composition of the Senate? Explain in a brief comment.

```
# According to the article there is also a section in gender that is trans+non-binary where ther
e were no person in that category. Other than the current senate data the website also displayed
categories such as religion, et5hnicity and orientation.
```