# Intro to Data Science - HW 5

##### Copyright 2021, Jeffrey Stanton, Jeffrey Saltz, and Jasmina Tacheva

# Rutwik Ghag

### 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) # install the RCurl library  
library(jsonlite) # install the JSON library  
dataset <- getURL("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json") # copy the contents of the JSON file from the mentioned URL into the variable  
readlines <- jsonlite::fromJSON(dataset) # convert the JSON data into an R object   
df <- readlines$objects$person # store the data in readlines in a data frame

1. 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

str(df)

## 'data.frame': 100 obs. of 17 variables:  
## $ bioguideid : chr "C000880" "G000386" "L000174" "M001153" ...  
## $ birthday : chr "1951-05-20" "1933-09-17" "1940-03-31" "1957-05-22" ...  
## $ cspanid : int 26440 1167 1552 1004138 25277 5929 1859 1962 45465 92069 ...  
## $ firstname : chr "Michael" "Charles" "Patrick" "Lisa" ...  
## $ gender : chr "male" "male" "male" "female" ...  
## $ gender\_label: chr "Male" "Male" "Male" "Female" ...  
## $ lastname : chr "Crapo" "Grassley" "Leahy" "Murkowski" ...  
## $ link : chr "https://www.govtrack.us/congress/members/michael\_crapo/300030" "https://www.govtrack.us/congress/members/charles\_grassley/300048" "https://www.govtrack.us/congress/members/patrick\_leahy/300065" "https://www.govtrack.us/congress/members/lisa\_murkowski/300075" ...  
## $ middlename : chr "D." "E." "J." "A." ...  
## $ name : chr "Sen. Michael â€œMikeâ€<9d> Crapo [R-ID]" "Sen. Charles â€œChuckâ€<9d> Grassley [R-IA]" "Sen. Patrick Leahy [D-VT]" "Sen. Lisa Murkowski [R-AK]" ...  
## $ namemod : chr "" "" "" "" ...  
## $ nickname : chr "Mike" "Chuck" "" "" ...  
## $ osid : chr "N00006267" "N00001758" "N00009918" "N00026050" ...  
## $ pvsid : chr "26830" "53293" "53353" "15841" ...  
## $ sortname : chr "Crapo, Michael â€œMikeâ€<9d> (Sen.) [R-ID]" "Grassley, Charles â€œChuckâ€<9d> (Sen.) [R-IA]" "Leahy, Patrick (Sen.) [D-VT]" "Murkowski, Lisa (Sen.) [R-AK]" ...  
## $ twitterid : chr "MikeCrapo" "ChuckGrassley" "SenatorLeahy" "LisaMurkowski" ...  
## $ youtubeid : chr "senatorcrapo" "senchuckgrassley" "SenatorPatrickLeahy" "senatormurkowski" ...

#View(df)

1. 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 senators and their information like name, gender, twitter id

## Part 2: Investigate the resulting dataframe

1. How many senators are women?

sum(df$gender=="female")

## [1] 24

1. How many senators have a YouTube account?

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

## [1] 73

1. How many women senators have a YouTube account?

sum(df$gender=="female" & !is.na(df$youtubeid))

## [1] 16

1. 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),]

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

youtubeWomen$year <- substr(youtubeWomen$birthday,1,4)  
# adds a new column, year, in the df youtubeWomen to store the birth year

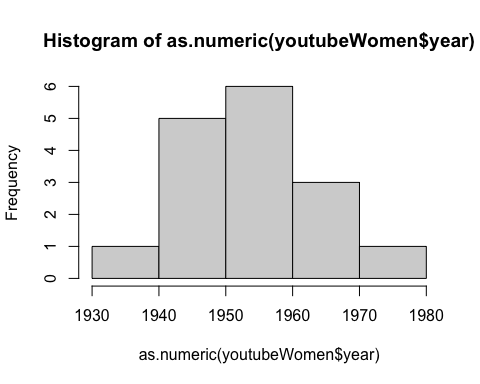
1. Use this new variable to calculate the mean **birthyear** in **youtubeWomen**. **Hint:** You may need to convert it to numeric first.

mean(as.numeric(youtubeWomen$year))

## [1] 1954.875

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

hist(as.numeric(youtubeWomen$year))



#the histogram is like an evenly distributed bell curve