HW 5

March 11, 2021

1 IST 387 HW 5

[2]:
```
# Enter your name here: Connor Hanan
```

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

[3]:
```
# 1. I did this homework by myself, with help from the book and the professor.
```

(Chapter 11 of Introduction to Data Science)

Reminders of things to practice from previous weeks: Descriptive statistics: mean( ) max( ) min( ) Coerce to numeric: as.numeric( )

1.1 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).

[4]:
```
library(RCurl) #load library RCurl
library(jsonlite) #load library jsonlite
dataset <- getURL("https://ist387.s3.us-east-2.amazonaws.com/data/role.json")␣
 ↪#calls API and stores reponse as variable
readlines <- jsonlite::fromJSON(dataset) #strips the JSON formatting of the API␣
 ↪response as a list of dataframes
df <- readlines$objects$person #subsets the list to select the dataframe we want
```

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

[6]:
```
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" …
```

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?

[7]:
```
#the dataset is about information about senators
#there are 100 rows, each one represents a single senator
#there are 17 columns, each one represents info about a senator
```

1.2 Part 2: Investigate the resulting dataframe

C. How many senators are women?

[8]:
```
library(tidyverse)
```

```
  Attaching packages                                    tidyverse
1.3.0

  ggplot2 3.3.2      purrr   0.3.4
  tibble  3.0.4      dplyr   1.0.2
  tidyr   1.1.2      stringr 1.4.0
  readr   1.4.0      forcats 0.5.0

  Conflicts
tidyverse_conflicts()
  tidyr::complete() masks
RCurl::complete()
```

```
  dplyr::filter()   masks
stats::filter()
  purrr::flatten()  masks
jsonlite::flatten()
  dplyr::lag()      masks stats::lag()
```

[17]:
```
df %>% filter(gender=='female') %>% nrow()
```

24

D. How many senators have a YouTube account?

[13]:
```
df %>% filter(!is.na(youtubeid)) %>% nrow()
```

73

E. How many women senators have a YouTube account?

[14]:
```
df %>% filter(gender=='female') %>% filter(!is.na(youtubeid)) %>% nrow()
```

16

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

[25]:
```
youtubeWomen <- df %>% filter(gender=='female') %>% filter(!is.na(youtubeid))
```

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

[28]:
```
youtubeWomen$year <- substr(youtubeWomen$birthday,1,4) #creates a new column of␣
↪the birthyear of each senator in youtubeWomen by taking a substring of␣
↪characters 1-4 in their full birthday
```

[29]:
```
youtubeWomen
```

A data.frame: 16 × 18

bioguideid <chr>	birthday <chr>	cspanid <int>	firstname <chr>	gender <chr>	gender_label <chr>	lastname <chr>	lin <c
M001153	1957-05-22	1004138	Lisa	female	Female	Murkowski	htt
M001111	1950-10-11	25277	Patty	female	Female	Murray	htt
D000622	1968-03-12	94484	Tammy	female	Female	Duckworth	htt
C000127	1958-10-13	26137	Maria	female	Female	Cantwell	htt
F000062	1933-06-22	13061	Dianne	female	Female	Feinstein	htt
S000770	1950-04-29	45451	Debbie	female	Female	Stabenow	htt
B001230	1962-02-11	57884	Tammy	female	Female	Baldwin	htt
B001243	1952-06-06	31226	Marsha	female	Female	Blackburn	htt
H001042	1947-11-03	91216	Mazie	female	Female	Hirono	htt
G000555	1966-12-09	1022862	Kirsten	female	Female	Gillibrand	htt
K000367	1960-05-25	83701	Amy	female	Female	Klobuchar	htt
S001191	1976-07-12	68489	Kyrsten	female	Female	Sinema	htt
W000817	1949-06-22	1023023	Elizabeth	female	Female	Warren	htt
F000463	1951-03-01	1034067	Deb	female	Female	Fischer	htt
C001035	1952-12-07	45738	Susan	female	Female	Collins	htt
S001181	1947-01-28	22850	Jeanne	female	Female	Shaheen	htt

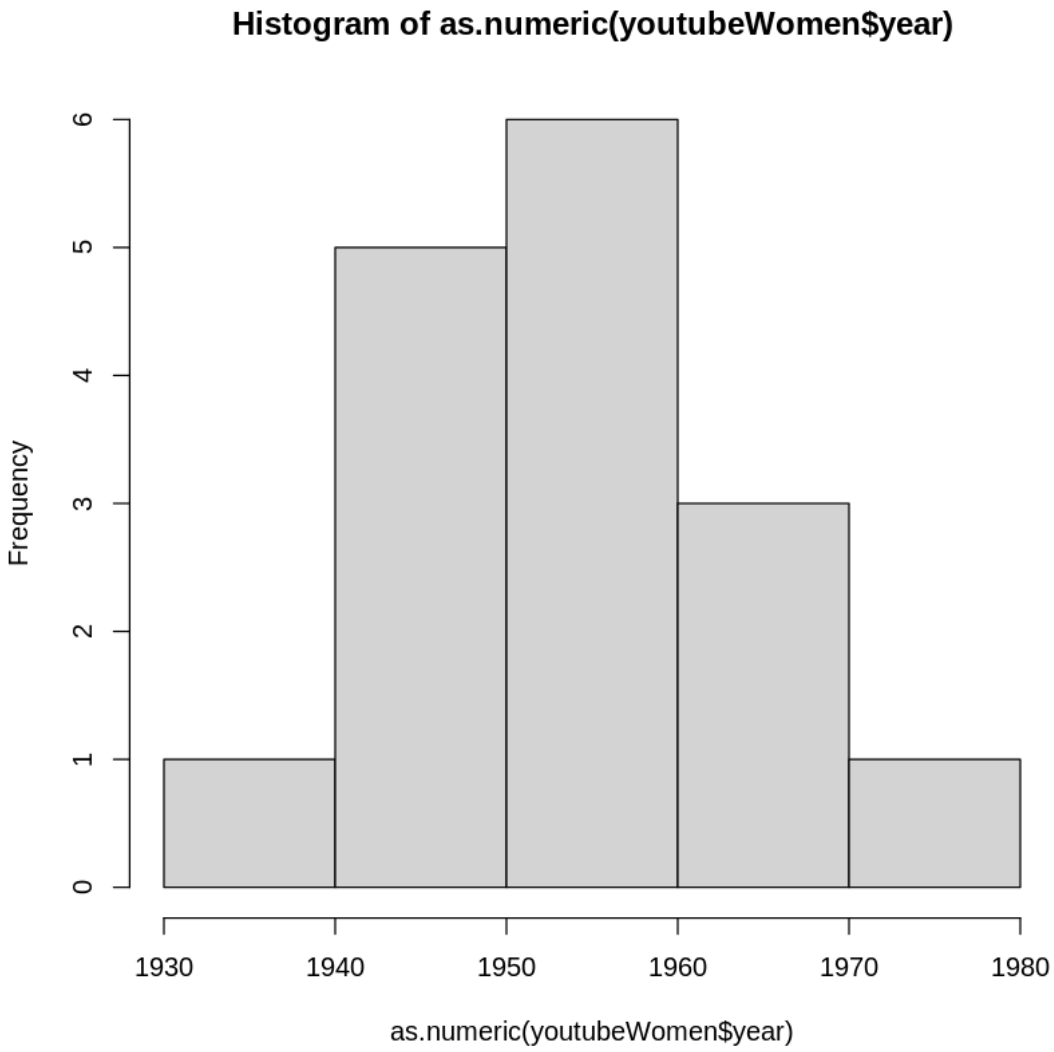
H. Use this new variable to calculate the mean **birthyear** in **youtubeWomen**. **Hint:** You may need to convert it to numeric first.

[31]:
```
mean(as.numeric(youtubeWomen$year))
```

1954.875

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

[48]:
```
hist(as.numeric(youtubeWomen$year)) #seems to be a fairly normal distribution,␣
↪with a slight bias to the older ages
```


Histogram of as.numeric(youtubeWomen$year)
Frequency
0
1
2
3
4
5
6
1930
1940
1950
1960
1970
1980
as.numeric(youtubeWomen$year)