Introduction to Data Science HW 4

```
# Enter your name here: Nora Lin
```


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

```
# 2. I did this homework with help from the book and the professor and these Internet sources:
#https://statisticsglobe.com/na-omit-r-example/
```

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).

```
#Loads in the library in which r objects are converted to/from JSON
library(jsonlite)
#Creating a new dataframe called "dataset" and retreiving this data from an extranal url
dataset <- url("https://intro-datascience.s3.us-east-2.amazonaws.com/role.json")
#Creating a new dataframe in which we convert the R objects from the JSON format of the URL data
readlines <- jsonlite::fromJSON(dataset)
#Creating a new dataframe called "df" from the "readlines" dataframe which has various smaller data set
df <- readlines$objects$person
```

```
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.g
##  $ middlename  : chr  "D." "E." "J." "A." ...
##  $ name        : chr  "Sen. Michael "Mike" Crapo [R-ID]" "Sen. Charles "Chuck" Grassley [R-IA]" "Sen
##  $ namemod     : chr  "" "" "" "" ...
##  $ nickname    : chr  "Mike" "Chuck" "" "" ...
##  $ osid        : chr  "N00006267" "N00001758" "N00009918" "N00026050" ...
##  $ pvsid       : chr  "26830" "53293" "53353" "15841" ...
##  $ sortname    : chr  "Crapo, Michael "Mike" (Sen.) [R-ID]" "Grassley, Charles "Chuck" (Sen.) [R-IA]
##  $ twitterid   : chr  "MikeCrapo" "ChuckGrassley" "SenatorLeahy" "LisaMurkowski" ...
##  $ youtubeid   : chr  "senatorcrapo" "senchuckgrassley" "SenatorPatrickLeahy" "senatormurkowski" ...
```

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 "Mike" Crapo [R-ID]             Mike N00006267 26830
## 2 Sen. Charles "Chuck" 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 "Chuck" Schumer [D-NY]            Chuck N00001093 26976
##                                  sortname     twitterid           youtubeid
## 1     Crapo, Michael "Mike" (Sen.) [R-ID]     MikeCrapo        senatorcrapo
## 2 Grassley, Charles "Chuck" (Sen.) [R-IA] ChuckGrassley    senchuckgrassley
## 3            Leahy, Patrick (Sen.) [D-VT]  SenatorLeahy SenatorPatrickLeahy
## 4           Murkowski, Lisa (Sen.) [R-AK] LisaMurkowski    senatormurkowski
## 5             Murray, Patty (Sen.) [D-WA]   PattyMurray  SenatorPattyMurray
## 6  Schumer, Charles "Chuck" (Sen.) [D-NY]    SenSchumer      SenatorSchumer
```

B. Explain the dataset o What is the dataset about?

```
# The dataset is about person information about uS governemtn congress members.
```

o How many rows are there and what does a row represent?

```
# There are 100 observations, meaning 100 rows. Each row represents 1 individual's information.
```

o How many columns and what does each column represent?

```
# There are 17 columns. Bioguideid is an identification number for each person, birthday is a 10 digit
```

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

```
vals <- substr(df$birthday,1,4)
# creates a new variable called "vals" which substrings a character vector which is the birthday column
```

D. Create a new attribute 'age' - how old the person is **Hint:** You may need to convert it to numeric first.

```r
vals <- as.numeric(vals)
age <- 2021 - vals
```

E. Create a function that reads in the role json dataset, and adds the age attribute to the dataframe, and returns that dataframe

```r
#this is what I want the function to do:
df$ages <- age

#function;
df_add_age <- function(dataset,new_attribute,content){
  dataset$new_attribute <- content
}

#df
```

F. Use (call, invoke) the function, and store the results in df

```r
df_add_age(df,age,age)
#check:
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 "Mike" Crapo [R-ID]             Mike N00006267 26830
## 2 Sen. Charles "Chuck" 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 "Chuck" Schumer [D-NY]            Chuck N00001093 26976
##                                  sortname     twitterid           youtubeid
## 1     Crapo, Michael "Mike" (Sen.) [R-ID]     MikeCrapo        senatorcrapo
## 2 Grassley, Charles "Chuck" (Sen.) [R-IA] ChuckGrassley    senchuckgrassley
## 3            Leahy, Patrick (Sen.) [D-VT]  SenatorLeahy SenatorPatrickLeahy
## 4           Murkowski, Lisa (Sen.) [R-AK] LisaMurkowski    senatormurkowski
## 5             Murray, Patty (Sen.) [D-WA]   PattyMurray  SenatorPattyMurray
## 6  Schumer, Charles "Chuck" (Sen.) [D-NY]    SenSchumer      SenatorSchumer
##   ages
## 1   70
```

```
## 2   88
## 3   81
## 4   64
## 5   71
## 6   71
```

Part 2: Investigate the resulting dataframe 'df'

A. How many senators are women?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_women <- df%>%
  filter(gender == "female")

 # filter(df,gender_label =='Female')
head(df_women)
```

```
##   bioguideid   birthday cspanid firstname gender gender_label     lastname
## 1    M001153 1957-05-22 1004138      Lisa female       Female    Murkowski
## 2    M001111 1950-10-11   25277     Patty female       Female       Murray
## 3    D000622 1968-03-12   94484     Tammy female       Female    Duckworth
## 4    H001076 1958-02-27   67481  Margaret female       Female       Hassan
## 5    C001113 1964-03-29  105698 Catherine female       Female Cortez Masto
## 6    C000127 1958-10-13   26137     Maria female       Female     Cantwell
##                                                                      link
## 1         https://www.govtrack.us/congress/members/lisa_murkowski/300075
## 2           https://www.govtrack.us/congress/members/patty_murray/300076
## 3        https://www.govtrack.us/congress/members/tammy_duckworth/412533
## 4        https://www.govtrack.us/congress/members/margaret_hassan/412680
## 5 https://www.govtrack.us/congress/members/catherine_cortez_masto/412681
## 6         https://www.govtrack.us/congress/members/maria_cantwell/300018
##   middlename                                  name namemod nickname      osid
## 1         A.           Sen. Lisa Murkowski [R-AK]                   N00026050
## 2                        Sen. Patty Murray [D-WA]                   N00007876
## 3                    Sen. Tammy Duckworth [D-IL]                   N00027860
## 4       Wood Sen. Margaret "Maggie" Hassan [D-NH]           Maggie N00038397
## 5              Sen. Catherine Cortez Masto [D-NV]                   N00037161
## 6                      Sen. Maria Cantwell [D-WA]                   N00007836
##   pvsid                                sortname       twitterid
```

```
## 1 15841           Murkowski, Lisa (Sen.) [R-AK]   LisaMurkowski
## 2 53358             Murray, Patty (Sen.) [D-WA]     PattyMurray
## 3 57442          Duckworth, Tammy (Sen.) [D-IL]    SenDuckworth
## 4 42552 Hassan, Margaret "Maggie" (Sen.) [D-NH]   Senatorhassan
## 5 69579   Cortez Masto, Catherine (Sen.) [D-NV]  sencortezmasto
## 6 27122           Cantwell, Maria (Sen.) [D-WA] SenatorCantwell
##            youtubeid ages
## 1   senatormurkowski   64
## 2 SenatorPattyMurray   71
## 3       repduckworth   53
## 4               <NA>   63
## 5               <NA>   57
## 6    SenatorCantwell   63
```

```
#There are 24 senators that are women
```

B. How many senators have a YouTube account?

```
nrow(df) -sum(is.na(df$youtubeid))
```

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

```
#There are 73 Senators with Youtube account
```

C. How many women senators have a YouTube account?

```
nrow(df_women) - sum(is.na(df_women$youtubeid))
```

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

```
#There are 16 women senators with Youtube accounts.
```

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

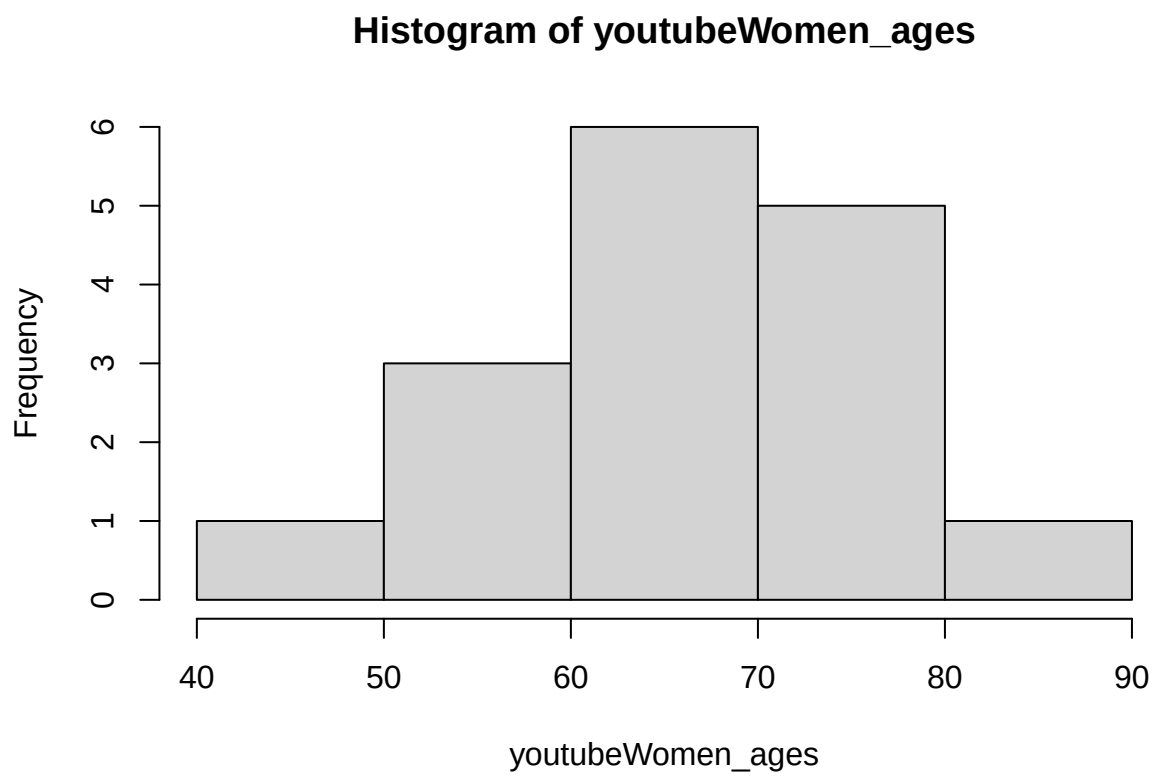
```
youtubeWomen <- df_women[!is.na(df_women$youtubeid),]
head(youtubeWomen)
```

```
##   bioguideid   birthday cspanid firstname gender gender_label  lastname
## 1    M001153 1957-05-22 1004138      Lisa female       Female Murkowski
## 2    M001111 1950-10-11   25277     Patty female       Female    Murray
## 3    D000622 1968-03-12   94484     Tammy female       Female Duckworth
## 6    C000127 1958-10-13   26137     Maria female       Female  Cantwell
## 7    F000062 1933-06-22   13061    Dianne female       Female Feinstein
## 8    S000770 1950-04-29   45451    Debbie female       Female  Stabenow
##                                                              link middlename
## 1   https://www.govtrack.us/congress/members/lisa_murkowski/300075         A.
## 2     https://www.govtrack.us/congress/members/patty_murray/300076
## 3  https://www.govtrack.us/congress/members/tammy_duckworth/412533
## 6   https://www.govtrack.us/congress/members/maria_cantwell/300018
```

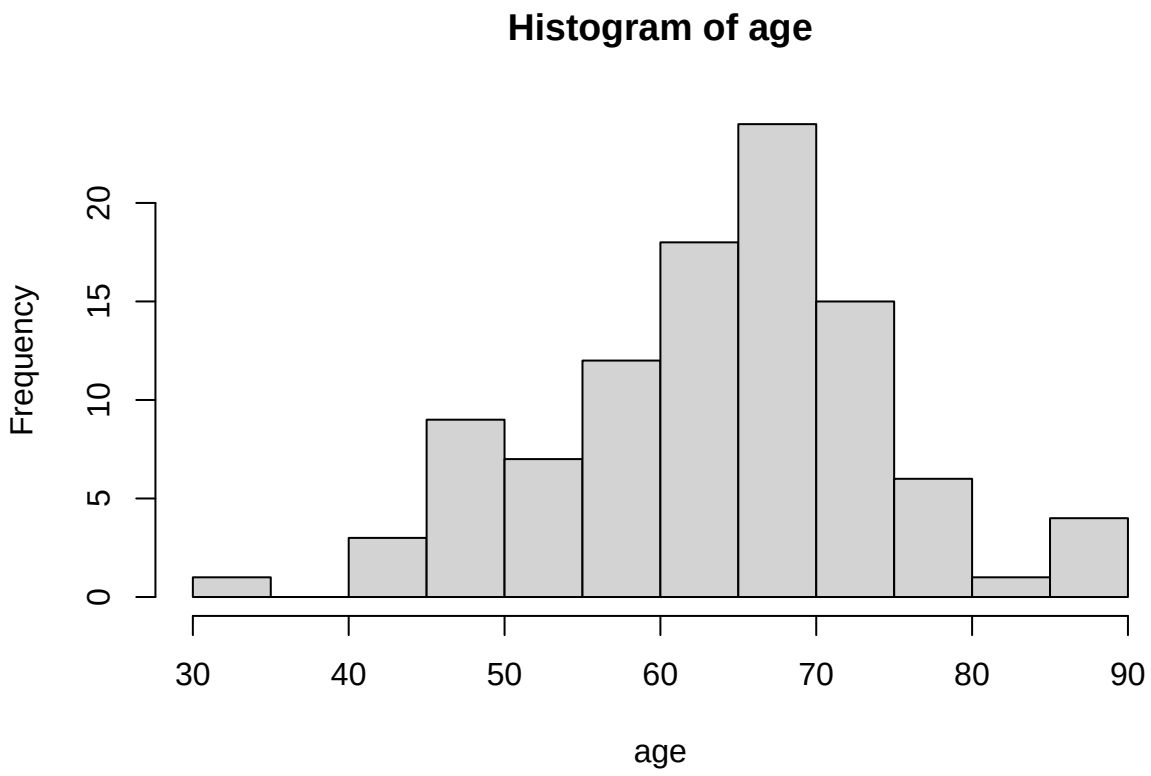
```
## 7 https://www.govtrack.us/congress/members/dianne_feinstein/300043
## 8  https://www.govtrack.us/congress/members/debbie_stabenow/300093        Ann
##                           name namemod nickname      osid pvsid
## 1   Sen. Lisa Murkowski [R-AK]                  N00026050 15841
## 2     Sen. Patty Murray [D-WA]                  N00007876 53358
## 3  Sen. Tammy Duckworth [D-IL]                  N00027860 57442
## 6   Sen. Maria Cantwell [D-WA]                  N00007836 27122
## 7 Sen. Dianne Feinstein [D-CA]                  N00007364 53273
## 8  Sen. Debbie Stabenow [D-MI]                  N00004118   515
##                          sortname       twitterid          youtubeid ages
## 1   Murkowski, Lisa (Sen.) [R-AK]   LisaMurkowski   senatormurkowski   64
## 2     Murray, Patty (Sen.) [D-WA]     PattyMurray SenatorPattyMurray   71
## 3  Duckworth, Tammy (Sen.) [D-IL]    SenDuckworth       repduckworth   53
## 6   Cantwell, Maria (Sen.) [D-WA] SenatorCantwell    SenatorCantwell   63
## 7 Feinstein, Dianne (Sen.) [D-CA]    SenFeinstein   SenatorFeinstein   88
## 8  Stabenow, Debbie (Sen.) [D-MI]     SenStabenow    senatorstabenow   71
```

E. Make a histogram of the **age** of senators in **youtubeWomen**, and then another for the senetors in **df**. Add a comment describing the shape of the distributions.

```
youtubeWomen_ages <- youtubeWomen$ages
hist(youtubeWomen_ages)
```



```
hist(age)
```




```
#the historgram of all senator ages has a wider range, as it includes the male senators as well. The hi
```