Introduction to Data Science HW 4


```
# Enter your name here: Benjamin Tisinger
```

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)
library(jsonlite)
dataset <- url("https://intro-datascience.s3.us-east-
2.amazonaws.com/role.json")
readlines <- jsonlite::fromJSON(dataset)
df <- readlines$objects$person
```

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

```
#Were finding the details of Senators in the US
#We can use nrow(df)to find how many rows and what they represent - 100
Senators
#We can also use ncol(df)to explore our columns and what is in there -17 Col
# I also included rownames() & colnnames() to show the details of each

summary(df)

##   bioguideid          birthday            cspanid         firstname
##  Length:100         Length:100         Min.   :    260   Length:100
##  Class :character   Class :character   1st Qu.:  25277   Class :character
##  Mode  :character   Mode  :character   Median :  68489   Mode  :character
##                                        Mean   : 584001
##                                        3rd Qu.:1004138
##                                        Max.   :9269028
##                                        NA's   :11
##     gender          gender_label         lastname             link
##  Length:100         Length:100         Length:100         Length:100
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
##   middlename            name              namemod            nickname
##  Length:100         Length:100         Length:100         Length:100
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      osid              pvsid             sortname          twitterid
##  Length:100         Length:100         Length:100         Length:100
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   youtubeid
##  Length:100
##  Class :character
##  Mode  :character
##
##
##
##

rownames(df)

##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"
"12"
##  [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"
"24"
##  [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"
"36"
##  [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"
"48"
##  [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"
"60"
##  [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"
"72"
##  [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"
"84"
##  [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"
"96"
##  [97] "97"  "98"  "99"  "100"

nrow(df)

## [1] 100

colnames(df)
```

```
##  [1] "bioguideid"   "birthday"     "cspanid"      "firstname"    "gender"
##  [6] "gender_label" "lastname"     "link"         "middlename"   "name"
## [11] "namemod"      "nickname"     "osid"         "pvsid"
"sortname"
## [16] "twitterid"    "youtubeid"

ncol(df)

## [1] 17
```

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

```
vals <- substr(df$birthday,1,4)
#were finding the substring for the data in Birthday and the year is #4
element value and we pass that into the value vals
```

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

```
number <- as.numeric(vals)
age <- (2022-number)
show(age)

##   [1] 71 89 82 65 72 72 88 73 72 72 67 68 67 61 63 61 58 50 54 57 76 51 59
65 51
##  [26] 67 50 54 71 64 58 64 75 89 72 60 70 70 79 68 81 71 49 75 56 62 66 62
67 70
##  [51] 51 75 46 73 78 61 71 52 64 65 70 68 43 75 58 70 70 78 67 88 80 66 73
79 75
##  [76] 68 59 69 65 50 76 64 45 60 53 58 52 62 68 50 62 64 63 68 68 70 63 49
35 53
```

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

```
df <- data.frame(df,age)
```

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

```
#df <- call(data.frame(df,age))
```

Part 2: Investigate the resulting dataframe 'df'

A. How many senators are women?

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

## [1] 24
```

B. How many senators have a YouTube account?

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

## [1] 73
```

C. How many women senators have a YouTube account?

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

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

D. 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,]
show(youtubewomen)
```

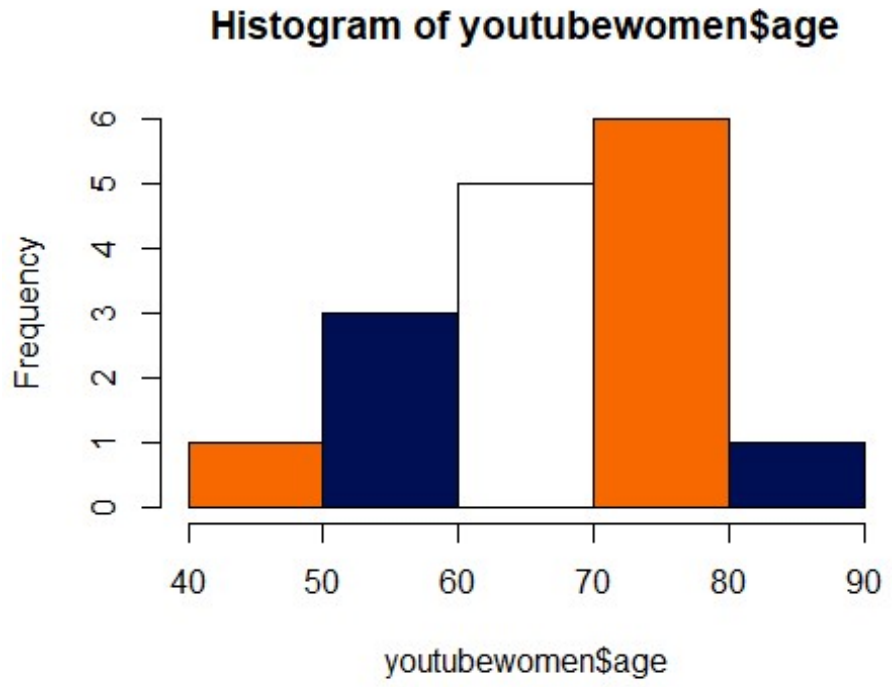
```
##    bioguideid   birthday cspanid firstname gender gender_label   lastname
## 4     M001153 1957-05-22 1004138      Lisa female       Female  Murkowski
## 5     M001111 1950-10-11   25277     Patty female       Female     Murray
## 28    D000622 1968-03-12   94484     Tammy female       Female  Duckworth
## 32    C000127 1958-10-13   26137     Maria female       Female   Cantwell
## 34    F000062 1933-06-22   13061    Dianne female       Female  Feinstein
## 35    S000770 1950-04-29   45451    Debbie female       Female   Stabenow
## 36    B001230 1962-02-11   57884     Tammy female       Female    Baldwin
## 37    B001243 1952-06-06   31226    Marsha female       Female  Blackburn
## 44    H001042 1947-11-03   91216     Mazie female       Female     Hirono
## 45    G000555 1966-12-09 1022862   Kirsten female       Female Gillibrand
## 46    K000367 1960-05-25   83701       Amy female       Female  Klobuchar
## 53    S001191 1976-07-12   68489   Kyrsten female       Female     Sinema
## 54    W000817 1949-06-22 1023023 Elizabeth female       Female     Warren
## 57    F000463 1951-03-01 1034067       Deb female       Female    Fischer
## 66    C001035 1952-12-07   45738     Susan female       Female    Collins
## 75    S001181 1947-01-28   22850    Jeanne female       Female    Shaheen
##                                                                       link
## 4      https://www.govtrack.us/congress/members/lisa_murkowski/300075
## 5        https://www.govtrack.us/congress/members/patty_murray/300076
## 28    https://www.govtrack.us/congress/members/tammy_duckworth/412533
## 32     https://www.govtrack.us/congress/members/maria_cantwell/300018
## 34   https://www.govtrack.us/congress/members/dianne_feinstein/300043
## 35    https://www.govtrack.us/congress/members/debbie_stabenow/300093
## 36      https://www.govtrack.us/congress/members/tammy_baldwin/400013
## 37   https://www.govtrack.us/congress/members/marsha_blackburn/400032
## 44       https://www.govtrack.us/congress/members/mazie_hirono/412200
## 45 https://www.govtrack.us/congress/members/kirsten_gillibrand/412223
## 46      https://www.govtrack.us/congress/members/amy_klobuchar/412242
## 53     https://www.govtrack.us/congress/members/kyrsten_sinema/412509
## 54   https://www.govtrack.us/congress/members/elizabeth_warren/412542
## 57        https://www.govtrack.us/congress/members/deb_fischer/412556
## 66      https://www.govtrack.us/congress/members/susan_collins/300025
## 75     https://www.govtrack.us/congress/members/jeanne_shaheen/412323
##    middlename                          name namemod nickname      osid
pvsid
## 4          A.     Sen. Lisa Murkowski [R-AK]                  N00026050
15841
## 5                   Sen. Patty Murray [D-WA]                  N00007876
53358
```

```
## 28               Sen. Tammy Duckworth [D-IL]                   N00027860
57442
## 32                Sen. Maria Cantwell [D-WA]                   N00007836
27122
## 34              Sen. Dianne Feinstein [D-CA]                   N00007364
53273
## 35        Ann    Sen. Debbie Stabenow [D-MI]                   N00004118
515
## 36                 Sen. Tammy Baldwin [D-WI]                   N00004367
3470
## 37         W.   Sen. Marsha Blackburn [R-TN]                   N00003105
25186
## 44         K.       Sen. Mazie Hirono [D-HI]                   N00028139
1677
## 45         E. Sen. Kirsten Gillibrand [D-NY]                   N00027658
65147
## 46       Jean      Sen. Amy Klobuchar [D-MN]                   N00027500
65092
## 53                Sen. Kyrsten Sinema [D-AZ]                   N00033983
28338
## 54              Sen. Elizabeth Warren [D-MA]                   N00033492
141272
## 57                  Sen. Deb Fischer [R-NE]                   N00033443
41963
## 66         M.      Sen. Susan Collins [R-ME]                   N00000491
379
## 75                Sen. Jeanne Shaheen [D-NH]                   N00024790
1663
##                             sortname       twitterid           youtubeid
age
## 4      Murkowski, Lisa (Sen.) [R-AK]   LisaMurkowski    senatormurkowski
65
## 5        Murray, Patty (Sen.) [D-WA]     PattyMurray  SenatorPattyMurray
72
## 28    Duckworth, Tammy (Sen.) [D-IL]    SenDuckworth        repduckworth
54
## 32     Cantwell, Maria (Sen.) [D-WA] SenatorCantwell     SenatorCantwell
64
## 34   Feinstein, Dianne (Sen.) [D-CA]    SenFeinstein    SenatorFeinstein
89
## 35    Stabenow, Debbie (Sen.) [D-MI]     SenStabenow     senatorstabenow
72
## 36      Baldwin, Tammy (Sen.) [D-WI]  SenatorBaldwin      witammybaldwin
60
## 37   Blackburn, Marsha (Sen.) [R-TN] MarshaBlackburn  RepMarshaBlackburn
70
## 44       Hirono, Mazie (Sen.) [D-HI]     MazieHirono CongresswomanHirono
75
## 45 Gillibrand, Kirsten (Sen.) [D-NY]    GillibrandNY  KirstenEGillibrand
56
```

```
## 46      Klobuchar, Amy (Sen.) [D-MN] SenAmyKlobuchar    senatorklobuchar
62
## 53     Sinema, Kyrsten (Sen.) [D-AZ]   SenatorSinema           repsinema
46
## 54   Warren, Elizabeth (Sen.) [D-MA]       SenWarren  senelizabethwarren
73
## 57        Fischer, Deb (Sen.) [R-NE]  SenatorFischer   senatordebfischer
71
## 66      Collins, Susan (Sen.) [R-ME]  SenatorCollins SenatorSusanCollins
70
## 75     Shaheen, Jeanne (Sen.) [D-NH]  SenatorShaheen      senatorshaheen
75
```

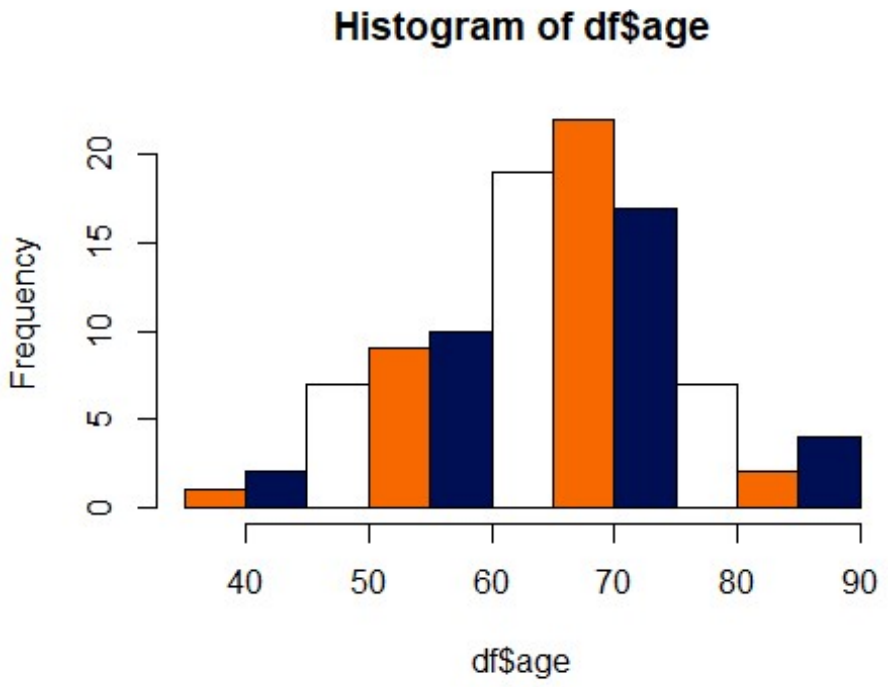
E. Make a histogram of the **age** of senators in **youtubeWomen**, and then another for the senetors in **df**. Add a comment describing the shape of the distributions.

```
hist(youtubewomen$age, col = c("#F76900","#000E54","#FFFFFF"))
```




```
#The shape shows a more older group of Women, mostly around the 60-80 Range.

hist(df$age, col = c("#F76900","#000E54","#FFFFFF"))
```




```
#Adding the men in here help balance out the chart with som botttom heavy
younger senators without Youtube.
```