Mustapha Brima - DACSS 601 Assignment 5

1. The variables that I used in this Assignment are the 'age' and 'Duartion..in.Seconds' Variables of the covid_responses_tab data set. The data collection for the 'age' variable had been done regarding only the years of birth of the participants, but not the actual ages of the participants up to the time of the study. Also, the 'Duration..in.Seconds' variable had shown the time spent by each participant when taking the study. But the name of the variable to represent this data 'Duration..in.Seconds' seemed too cumbersome of a name to represent it.

2.In order to fix these issues, I used different functions for each of the variables in the data set. For the 'age' variable, I changed any missing values ('NA' values), to 0. Then I converted the entries for the 'age' variable to integer values in order to use a function to change the ages. I then used the mutate function to convert the years of birth into ages by using the function 2020 - 'age'. Finally, I used the filter function to remove the entries of those who had not put down the years of their birth when they took the survey.

For the 'Duration..in.Seconds.' Variable, I used the rename function to instead name the variable as 'Duration' to make it easier to reference the entries of the column.

3.The recoded data now includes the newly named 'Duration' column as well as the 'age' column with the ages of all the participants who entered their years of birth when taking the survey.

4.I originally believed that the entries of the variables would suggest that the might be a negative trend in regards to the data in which, the younger the participant (or the 'age'), the faster they would complete the survey since they might become exposed and accustomed to newer kinds of technology compared to older participants and thus, the shorter amount of time spent (or the 'Duration') of their completion of the survey. After reorganizing the data and plotting it, i had found that this was somewhat false given that most of the participants of the surey aged 30 to 60 (and above) tend to have an even amount of participants taking similar short instances of time to complete the survey. In fact, the longest durations of the survey belonged to more of the participants closer to the age of 30 than those that were closer to the age of 60.

library(readxl) covid_responses_tab <- read_excel("Assignment 5/covid-responses.tab.xlsx") View(covid_responses_tab)

install.packages("tidyverse")

R Markdown

```
Note that the 'echo = FALSE' parameter was added to the code chunk to prevent printing of the R code tha
## To convert the 'age' column entries to intergers.

  covid_responses_tab$age <- as.integer(covid_responses_tab$age)

## To replace the empty entries for the 'age' variables with a value of 0 for each one.

  covid_responses_tab$age[is.na(covid_responses_tab$age)] <- 0

## Code for cleaning up the entries for the 'age' and 'Duration..in.seconds.' columns.

Plot:
```

```r
#install.packages("ggplot2")
library(readxl)
#library(datasets)
#library(ggplot2)
# library(tidyverse)
#library(lubridate)
#library(magrittr)
covid_responses_tab <- read_excel("/Users/Mustapha/Documents/SENIOR YEAR SIXTH SEMESTER/DACSS 601/R Cod

#`df` is `CPOD_Time` saved as `df<-as.data.frame(CPOD_Time)`
#covid_responses_tab<-as.data.frame(covid_responses_tab)
#load(covid_responses_tab)
covid_responses_tab %>%
  rename(Duration = Duration..in.seconds.)%>%
  select(Duration, age)%>%
  mutate(age = 2020 - age)%>%
  filter(age < 2020)%>%
  arrange(age)%>%
  ggplot(aes(x = age, y = Duration)) + geom_point()
```