Assignment 5 MSDS 6306 Spring 2019

Part 1 Data Munging

* Import yob2016.txt, a text file that contains a list of names (Male and Female) born in 2016 along with the frequency count of the name being chosen.
  + df <- read.table("Folder\_Path\\FileName", sep=";")   
      
    sep defines the delimiter being using in the file
* Find the dimension, summary and structure of the data frame df once the text file import has been completed.
  + dim(df)   
      
    to find dimension of the data frame
  + summary(df)   
      
    summary of the data frame
  + str(df)   
      
    the data structure of the data frame
* Find the misspelled name that has 3 y’s at the end of name using a regex and display the name.
  + grep('y{3}',df$V1)   
      
    To match the string that has 3 Y’s in column 1 of the data frame
* Remove the row that contains the misspelled name and save the data frame df into y2016.
  + y2016 <-subset(x=df,df$V1!="Fionayyy")   
      
    subset function to remove the desire row

Part 2 Data Merging

* Import yob2015.txt, a text file that contains a list of names (Male and Female) born in 2015 along with the frequency count of the name being chosen.
  + yob2015 <- read.table("Folder\_Path\\FileName", sep=",")   
      
    sep defines the delimiter being using in the file
* Show the last 10 rows from the data frame y2015
  + tail(y2015,10)   
      
    show only the last 10 row in the df
* Merge y2016 and y2015 by the name column then assign to data frame final. Remove all NA values.
  + final<-merge(y2016, y2015, by="V1")   
      
    merge y2016 and y2015 by name (“V1”)

Part 3 Data Summary

* Add a column Total in the final data frame to sum up the count in 2015 and 2016
  + total<-data.frame(final$V3.x+final$V3.y)   
      
    add the count column from 2016 and count column from 2015, save to total
  + final<-cbind(final$V1,final$V2.x,total)   
      
    binds the V1 ,V2.x (gender) and total column into final data frame
  + names(final) <- c("Name", "Gender", "Total\_Count")   
      
    rename the headers in final into human readable form
* Sort the data frame by total
  + head(arrange(final,desc(Total\_Count)),10)   
      
    head(object,10) find the first 10 row from the object. Arrange(object, desc(variable\_to\_be\_sorted) function to arrange the object based on the specified variable in descending order
* Filter out all the male names and then find the top 10 female names.
  + GirlNames<-final%>%filter(Gender == "F")   
      
    filter the data frame based on a criteria
* Write the top 10 girl names to a CSV file. Remove the Gender column before exporting to CSV.
  + write.csv(object,”filename.csv”)   
      
    export the object into a CSV format with a designated filename

Part 4 Github Upload

* Upload at assignment 5 + a code book to a repo