IT497 HW#3

Submit 2 files.

i. One showing the R Notebook (with R code) to answer the questions (***.Rmd**).
ii. One the knitted **HTML** or **Word** output from the R Notebook

Answer the following 5 questions completely.

1.

Below is a link to a file listing the number women software engineers at several Silicon Valley tech firms.

```
http://www.itk.ilstu.edu/faculty/jrwolf/wit.csv
```

a. Create the R code to find the number women software engineers at Reddit.

b. Using the same link, create the R code to find the total number women software engineers in the dataset.

2.

If you run the R code below, it will create a dataframe (named `pop`) with 139 observations and 4 variables.

```
### Start pop Code#######
install.packages("rvest")
library("rvest")
html <- read_html("https://www.ssa.gov/oact/babynames/numberUSbirths.html")
usBirthData <- html_table(html, fill=TRUE)
pop <- usBirthData[[1]]
pop$Male <- as.numeric(gsub(",", "", pop$Male))
pop$Female <- as.numeric(gsub(",", "", pop$Female))
pop$Total <- as.numeric(gsub(",", "", pop$Total))
names(pop) <- c("year", "male", "female", "total")
### End pop Code #######
```

The above data shows the number of applicants for a Social Security card by year of birth and sex. The number of such applicants is restricted to U. S. births where the year of birth and sex are known, and where the given name is at least 2 characters long.

This is a good proxy for the number of Americans born each year.

a. Create the R code to find the year with the most children born.

b. Using the same list, create the R code to find the year with the fewest children born.

3.

`https://en.wikipedia.org/wiki/List_of_cities_proper_by_population`

The above page contains a list of cities by population.

The following code will download the data from the wiki page above and create a dataframe named topCities.

```r
#### Start Top Cities Code ####
topCities <-
  read_html("https://en.wikipedia.org/wiki/List_of_cities_proper_by_population"
  )
topCities <- html_table(
  html_node(
    topCities, "table.sortable"
  ), header=TRUE, trim=F, fill = TRUE
)[-1,]

names(topCities) <- c(
  "City", "Country", "Image", "Population", "City.Def", "City.Population", "City.A
rea", "Metro.Pop",
  "Metro.Area", "Urban.Pop", "Urban.Area"
)
topCities$Image <- NULL

## Old String Substitute (gsub)

topCities$Population <- as.numeric(
  gsub(",", "", topCities$Population)

)



## tidyverse String Substitute (str_replace_all)

topCities$City<-str_replace_all(topCities$City, "\n", "")
topCities$Country<-str_replace_all(topCities$Country, "\n", "")
topCities$City.Def<-str_replace_all(topCities$City.Def, "\n", "")
topCities$City.Area<-str_replace_all(topCities$City.Area, "\n", "")
topCities$Metro.Pop<-str_replace_all(topCities$Metro.Pop, "\n", "")
topCities$Metro.Area<-str_replace_all(topCities$Metro.Area, "\n", "")
topCities$Urban.Pop<-str_replace_all(topCities$Urban.Pop, "\n", "")
topCities$Urban.Area<-str_replace_all(topCities$Urban.Area, "\n", "")

#### End Top Cities Code ####
```

a. Use the topCities dataframe and create the R code to find the 5 largest cities in India.

b. Use the topCities dataframe and create the R code to find the 5 largest cities in China.

4.

The Standard & Poor's 500, often abbreviated as the S&P 500, or just "the S&P",is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ.

The firms change often. Below is a link showing the firms and when they were added to the S&P 500.

```
https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
```

Below is the code to download the S&P 500 companies and the dates they were added.

```
########## SnP Code Start ##############
```

```r
url <- "https://en.wikipedia.org/wiki/List_of_S%26P_500_companies"

url %>%

  read_html() %>%

  html_nodes("#constituents") %>%

  html_table(header = T) -> sp500

sp500 <- sp500[[1]]

names(sp500) <- c(

  "Ticker", "Name", "SEC.Filing", "GICS.Sector",

  "GICS.SubIndustry", "Location.HQ", "Date.Added", "CIK", "Founded"

)

sp500$Date.Added <- as.numeric(substr(sp500$Date.Added,1,4))
```

```
########## SnP Code End ##############
```

a. Create the R code to find the names of firms added to the S&P 500 since 2010.
b. b. Create the R code to find the total number of firms in the table.

5. According to Wikipedia: "The 2015 Cricket World Cup was the 11th Cricket World Cup, jointly hosted by Australia and New Zealand. Fourteen teams played 49 matches in 14 venues.

```
https://en.wikipedia.org/wiki/2015_Cricket_World_Cup
```

The code below will download the data from Wikipedia and create a dataframe named venue.

```
### Start Venue R Code ####
cricket_world_cup <-
html("http://en.wikipedia.org/wiki/2015_Cricket_World_Cup")
venues = html_table(html_nodes(cricket_world_cup, "table")[[5]])
### End Venue R Code ####
```

a. Use the venue dataframe to find the name of the venue with the largest capacity.
b. Use the venue dataframe to find the name of the venue with the smallest capacity.

Note: You may need to convert the data