Yale's Climate Survey and Census data analysis

Yong Wu

12/9/2021

1.Set working directory and read the climate opinion and population data into the global environment

```
library(tidyverse)
data_loc = 'C:\\Users\\wuyon\\Desktop\\Sta9750'
setwd(data_loc)
require(data.table)
```

Loading required package: data.table

```
mydata1 = fread("yale_climate_cty_data.csv",
stringsAsFactors = F,
data.table = F,
colClasses=list(character=c(1)))
mydata2 = fread("ACS_16_5YR_DP05_with_ann.csv",
stringsAsFactors = F,
data.table = F)
mydata3 = fread("ACS_16_5YR_DP05_metadata.csv",
stringsAsFactors = F,
data.table = F)
```

2. Merge the two files together using the common ID method and the rename the new object

Renaming the cloumn GEO.id2 in mydata2 to cty_FIPS for merging.

```
mydata2 = rename(mydata2, cty_FIPS = GEO.id2)
```

Merging the two datasets by cty_FIPS column.

```
merged_data = inner_join(x = mydata1, y = mydata2, by='cty_FIPS')
```

3. From the newly merged object, rename variables to meaningful variable names

```
merged_data <- merged_data %>%
  rename('County ID' = cty_FIPS) %>%
  rename('Total Population' = HC01_VC03) %>%
  rename('White Population' = HC01_VC49) %>%
  rename('Black Population' = HC01_VC50) %>%
  rename('Hispanic Population' = HC01_VC88) %>%
  rename('Asian Population'= HC01_VC56) %>%
  rename('Median Age' = HC01_VC23) %>%
  rename('Female Population' = HC01_VC05) %>%
  rename('% of Respondents that believe global warming is Happening' = happening) %>%
  rename('% of Respondents that believe global warming is caused by human activities' = human) %>%
  rename('% of Respondent that are somewhat/very worried about global warming' = worried)
```

4. For each county, calculate the following percentages and store them in new columns: % White, % Black, % Hispanic, % Asian and % Female

Get rid of N/As from the dataset before calculations

```
merged_data = na.omit(merged_data)
```

Converting to columns to numeric for calculations

```
merged_data$`Total Population` = as.numeric(as.character(merged_data$`Total Population`))
merged_data$`White Population` = as.numeric(as.character(merged_data$`White Population`))
merged_data$`Black Population` = as.numeric(as.character(merged_data$`Black Population`))
merged_data$`Hispanic Population` = as.numeric(as.character(merged_data$`Hispanic Population`))
merged_data$`Asian Population` = as.numeric(as.character(merged_data$`Asian Population`))
merged_data$`Female Population` = as.numeric(as.character(merged_data$`Female Population`))
```

Adding new columns with the percentages of the races and female to the dataset with mutate

```
merged_data <- merged_data %>%
  mutate('% White' = (`White Population` / `Total Population`) * 100) %>%
  mutate('% Black' = (`Black Population` / `Total Population`) * 100) %>%
  mutate('% Hispanic' = (`Hispanic Population` / `Total Population`) * 100) %>%
  mutate('% Asian' = (`Asian Population` / `Total Population`) * 100) %>%
  mutate('% Female' = (`Female Population` / `Total Population`) * 100)
```

5. Viewing the summary stats write a short paragraph summarizing the distribution of the new percent variables.

```
summary(merged_data$`% White`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.903  76.917  89.909  83.382  95.416 100.000
```

```
summary(merged_data$`% Black`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6246  2.2135  9.0166 10.2691 86.1849
```

```
summary(merged_data$`% Hispanic`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.941   3.834   8.947   9.067  98.959
```

```
summary(merged_data$`% Asian`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2575  0.5692  1.3027  1.2348 42.8982
```

```
summary(merged_data$`% Female`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.51   49.45   50.42   49.94   51.13   58.50
```

From the summary of these new percent variables there are a few things that stick out to me. First for the Black, Hispanic and Asian percentages, the min. is 0 which is weird. This is likely to be that there is no data about their percentages in those areas or they recorded it wrong. The chances that none of the races lived in those areas is low. Secondly, the mean percentage of white people to any other race is overwhelmingly high while the others are really low (at less than 10%). This can mean that there are a lot of places in the U.S. where the diversity of the population is very low, or that when this data is taken there are not enough participants to represent the races in those areas. The third that stuck out to me from this summary is that for the percentage of white people, there are areas where the population is 100% white. This is surprising to me if this data is actually true because there should be some diversity in races all over the U.S. The last thing that stuck out to me was the min. of female population in certain areas. It is as low as 21.51% which is a very heavy bias towards Men in those areas. Other than that, everything else seems plausible.

6. Create a new state variable from the current county ID variable and merge the region variable (region.csv) to the working data frame (first 2 digits of ID represent the state).

Uploading the region.csv into R.

```
region = fread('Region.csv')
```

Using substr to take the first two digits from County ID and making it a new column called 'State.FIPS' also converting the substring to integer because 'State_FIPS' from region is in integer form.

```
merged_data = mutate(merged_data, 'State_FIPS' = as.integer(substr(`County ID`, 1, 2)))
```

Now merging region dataset with the merged_data dataset.

```
merged_data = inner_join(x = merged_data, y = region, by = 'State_FIPS')
```

7. What is, by State, the average % that believe global warming is occurring? Show the data in descending order

```
merged_data %>%
  group_by(State) %>%
  summarise('avg % that believe in global warming' = mean(`% of Respondents that believe global warming
  arrange(desc(`avg % that believe in global warming`))
```

```
## # A tibble: 51 x 2
##    State                ‘avg % that believe in global warming‘
##    <chr>                                                 <dbl>
##  1 District of Columbia                                   83.9
##  2 Hawaii                                                 78.3
##  3 Massachusetts                                          75.2
##  4 New Jersey                                             72.8
##  5 California                                             72.7
##  6 Vermont                                                72.0
##  7 Alaska                                                 72.0
##  8 Rhode Island                                           71.7
##  9 Connecticut                                            70.3
## 10 New Mexico                                             70.2
## # ... with 41 more rows
```

From this analysis we can tell that Washington D.C. has the most average percent of its residents that believe global warming is real. The top 10 are all Democratic states as well. The top 10 are also all states that are close to the water. The bottom 10 are mainly Republican states and are not near any bodies of water.

8. What is, by Region, the average % that believe global warming is caused by humans? Report the data in descending order

```
merged_data %>%
  group_by(Region) %>%
  summarise('avg % that believe global warming is caused by humans' = mean(`% of Respondents that belie
  arrange(desc(`avg % that believe global warming is caused by humans`))
```

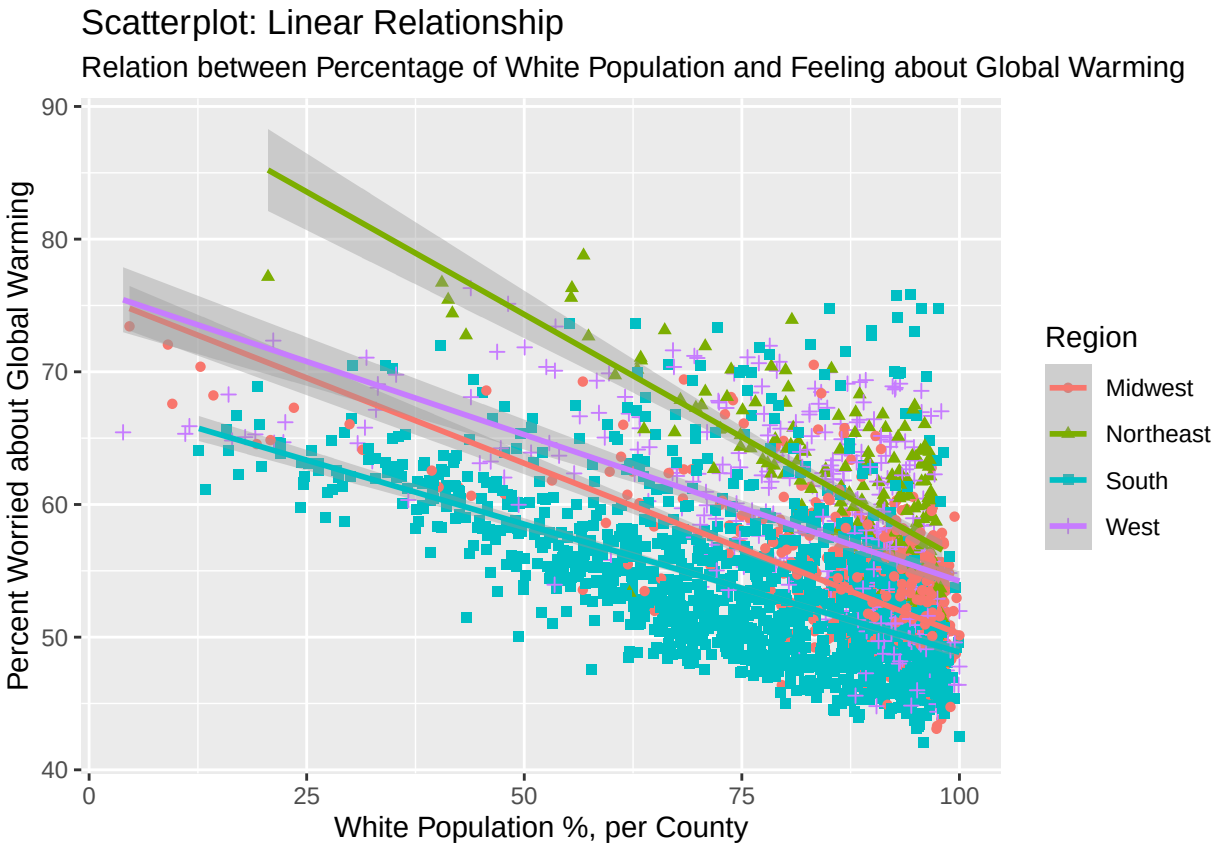
```
## # A tibble: 4 x 2
##   Region    `avg % that believe global warming is caused by humans`
##   <chr>                                                       <dbl>
## 1 Northeast                                                    55.9
## 2 West                                                         53.9
## 3 Midwest                                                      50.7
## 4 South                                                        50.0
```

This analysis tells us that people who believe that global warming is man made tend to be in the Northeast and West which is where all the Democratic states are. Perhaps political affiliation may have something to do with how people perceive global warming. This could be a potential topic to perform an analysis on in the future.

9. Plot the relationship between the % of white population and the % of the respondents that are worried about global warming, separating the 4 different regions.

```
ggplot(merged_data, aes(x = `% White`, y =`% of Respondent that are somewhat/very worried about global w
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From this graph we can see that there is an inverse relationship between percent of white population per county and percent of the population that are worried about global warming. This graph also shows that

people in the Northeast and West tend to worry more about the global warming when there is less white people within the county. This can be due to many reasons like political affiliation. This graph also shows some outliers where some Southern counties with high percentage of white population have around 70% to 75% of their population be worried about global warming. These outliers can be further researched on to find out if this is an error or something in those counties are causing this phenomenon.