Opinions on Global Warming and Census

Sai Pavani Cheruku

December 2021

PreReq : Loading the libraries required

```
library(data.table)
library(tidyverse)
require(data.table)
```

1 : Reading the climate opinion and population data into the global environment

```
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)
```

We have two files available for our analysis namely:

1) yale_climate_cty_data.csv - percentage of people who agree/disagree to certain parameters are listed by county. For example, percentages of people in Autauga couty who believe that global warming is happening, and percentage of people from the same county who do not believe in it.
2) ACS_16_5YR_DP05_with_ann.csv - percentages of population divided based on gender, age, race, voting age etc listed by county. For example, data from every county which gives total percentage of male, female, under 5 years etc under it.

2 : Merging two files to create MergedDataSet

Cleaning and preparing the data

```
#str(mydata1)
#str(mydata2)
mydata2 = mydata2 %>%
  rename(GeoName = 'GEO.display-label')
```

```
mergedDataSet = inner_join(x = mydata1, y=mydata2, by="GeoName")
#str(mergedDataSet)
```

- We can derive interesting insights like what percentage of certain race believe that global warming is happening, by merging the two datasets we have.
- The common column we have in both the datasets is the county details column.
- In file1 it goes by the name 'GeoName' and in file2 it is 'GEO.display-label'.
  - First step in merging the files would be to ensure the common column has the same name in both the files and thats what we do here by renaming the column in file2 as "GeoName"
  - Second step is using inner join by "GeoName" column to produce the merged data set that will consist of all data from both the files provided the GeoName colums have matching values

3 : Data preparation - Renaming the variables to meaningful names

```
#names(mergedDataSet)
mergedDataSet = mergedDataSet %>%
  rename(County_ID ='cty_FIPS')
mergedDataSet = mergedDataSet %>%
  rename(Total_Population ='HC01_VC03')
mergedDataSet = mergedDataSet %>%
  rename(White_Population ='HC01_VC49')
mergedDataSet = mergedDataSet %>%
  rename(Black_Population ='HC01_VC50')
mergedDataSet = mergedDataSet %>%
  rename(Hispanic_Population ='HC01_VC88')
mergedDataSet = mergedDataSet %>%
  rename(Asian_Population ='HC01_VC56')
mergedDataSet = mergedDataSet %>%
  rename(Median_Age ='HC01_VC23')
mergedDataSet = mergedDataSet %>%
  rename(Female_Population ='HC01_VC05')
mergedDataSet = mergedDataSet %>%
  rename(percentage_of_Respondents_that_believe_global_warming_is_Happening ='happening')
mergedDataSet = mergedDataSet %>%
  rename(percenatge_of_Respondents_that_believe_global_warming_is_caused_by_human_activities ='human')
mergedDataSet = mergedDataSet %>%
  rename(percentage_of_Respondent_that_are_somewhat_very_worried_about_global_warming = 'worried')
```

Adding meaningful column names is an important part of the data preparation for analysis. Often, the column names are not very descriptive (HC01_VC03). Although, we might have a metadata file that has the full length descriptions of what each columns mean, it would not be easy and productive to leave them as is. In this step, we choose the columns we are intrested and rename them to something that makes more sense, utilizing the data from the metadata file. For Ex, HC01_VC03 as Total_Population.

4 : For each county, what are the percentages of : % White, % Black, % Hispanic, % Asian and % Female

```
mergedDataSet$Total_Population =
  strtoi(mergedDataSet$Total_Population)
mergedDataSet$White_Population =
  strtoi(mergedDataSet$White_Population)
mergedDataSet$Black_Population =
  strtoi(mergedDataSet$Black_Population)
mergedDataSet$Hispanic_Population =
  strtoi(mergedDataSet$Hispanic_Population)
mergedDataSet$Asian_Population =
  strtoi(mergedDataSet$Asian_Population)
mergedDataSet$Female_Population =
  strtoi(mergedDataSet$Female_Population)


mergedDataSet =  mutate(mergedDataSet,
                        white_percentage = (White_Population/ Total_Population)*100,
                        black_percentage = (Black_Population / Total_Population)*100,
                        hispanic_percentage = (Hispanic_Population / Total_Population)*100,
                        asian_percentage = (Asian_Population / Total_Population)*100,
                        female_percentage = (Female_Population / Total_Population)*100)

#summary(mergedDataSet)
```

- The population estimate columns that we are interested in are in character format. To compute the percentage we had to convert the columns into integer data type.
- We added five new columns that are calculated by each estimate divided by the total population -
  - white_percentage - percentage of white population in the total population
  - black_percentage - percentage of black population in the total population
  - hispanic_percentage - percentage of hispanic population in the total population
  - asian_percentage - percentage of asian population in the total population
  - female_percentage - percentage of female population in the total population

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

```
summary(mergedDataSet$white_percentage)
```

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

```
summary(mergedDataSet$black_percentage)
```

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

```
summary(mergedDataSet$hispanic_percentage)
```

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

```
summary(mergedDataSet$asian_percentage)
```

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

```
summary(mergedDataSet$female_percentage)
```

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

- The range of percentages of White, Black and Hispanic races is approximately 0 to 90, in contrast to the Asian and female percentages that range from approximately 0 to 50.
  - In the White percentage distribution, there is remarkable difference between the min value and the first quartile point. However, we do not observe that between the 3 quartiles and the max value.
  - In the Black, Hispanic and Asian percentage distribution, there is remarkable difference between the max value and the third quartile point. However, we do not observe that between the 3 quartiles and the min value.
  - In the female percentage distribution, there is no remarkable difference between the min, max values and the three quartiles.
- The White % distribution is centered at the highest of 89% followed by female % distribution at 50%. Black, Hispanic and Asian % distributions are centered at a much lower level like 1~3%. We can say that first half of these three percentage distributions are below 3%.
- Occupancy scenarios in counties:
  - There are not more than 43% of Asians in any given county, and there could be counties with 0% of Asians as well.
  - There are not more than 58% of Female population in any given county which makes sense because usually there are no places where only female population live.
  - There are counties where there are no Black or Hispanic population. However there are even counties where the Hispanics and Blacks in majority and occupy 80 to 90% of the total population in the county.
  - Whites are the only race that has complete occupancy in any county (max value of the distribution is 100).

6 : Merging the Region data file

```
mydata4 = fread("Region.csv",
                stringsAsFactors = F,
                data.table = F)

mergedDataSet$State_FIPS = substr(mergedDataSet$County_ID, 0, 2)

#str(mydata4)
#str(mergedDataSet$State_FIPS)

mergedDataSet$State_FIPS =
  strtoi(mergedDataSet$State_FIPS)
```

```
#glimpse(mergedDataSet)

finalMergedDataSet = inner_join(x = mergedDataSet, y=mydata4, by="State_FIPS")
#glimpse(finalMergedDataSet)
```

- So far we have analysed various parameters with respect to counties. Now we have a new file called Region.csv that can be merged with our data set, which gives the regional distribution values for each of the county listed.
- The first two digits in the County_ID column represent the state ID. This is present in the region.csv file as State_FIPS code which has the respective region details mentioned for each state.
- We split the County_ID and took the first two digits and created that as the new State_FIPS column in our mergedDataSet.
- We then use this common column to merge the merrgedDataSet to the Region.csv file as required.
- With this new information we can derive insights based on regions, states in addition to counties.

7 : Average % of population that believe global warming is occuring by State

```
byState =
  finalMergedDataSet %>%
  group_by(State) %>%
  summarize(avg_percent = mean(percentage_of_Respondents_that_believe_global_warming_is_Happening))

arrange(byState, desc(avg_percent))
```

```
## # A tibble: 49 x 2
##    State                avg_percent
##    <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 New Mexico                  70.2
## 10 New Hampshire               69.4
## # ... with 39 more rows
```

- District of Columbia (84%), Hawaii (78%), Massachusetts (75), New Jersey (73) and California (73) are the top 5 states in the US that believe global warming is occurring.
- 84% of the total population District of Columbia believe that climatic change is a serious concern and according to an article in Wikipedia the state is implementing a ClimateReady DC plan. It has also mandated 50% of renewable energy by 2032.
- 78% of Hawaii's population are alarmed with the changes occurring in the climate and believe that the global warming is a concern that we need to deal with. The effect of global warming is seen in Hawaii as the rainfall decreases over the years, coral bleaching, rising sea levels etc, and it makes sense why the people of Hawaii are worried.

8 : Average % of population that believe global warming is occuring by State

```
byRegion =
  finalMergedDataSet %>%
  group_by(Region) %>%
  summarize(avg_percent = mean(percenatge_of_Respondents_that_believe_global_warming_is_caused_by_human

(byRegion = arrange(byRegion, desc(avg_percent)))
```

```
## # A tibble: 4 x 2
##   Region    avg_percent
##   <chr>           <dbl>
## 1 Northeast        55.8
## 2 West             53.8
## 3 Midwest          50.7
## 4 South            50.0
```

- Further in the analysis, we wanted to see how the percentage of people who believe that global warming is caused by humans, is distributed between the four regions in the US.
- We group the data by Region and compute the average % of population who believe that we humans cause global warming.
- Interestingly, the data shows us that almost half of the population in all 4 regions believe in this claim and Northeast region tops the list with 56%.
- Although Northeast region has the highest percentage, given the current conditions its astonishing to see that such low proportions of population are actually worried.

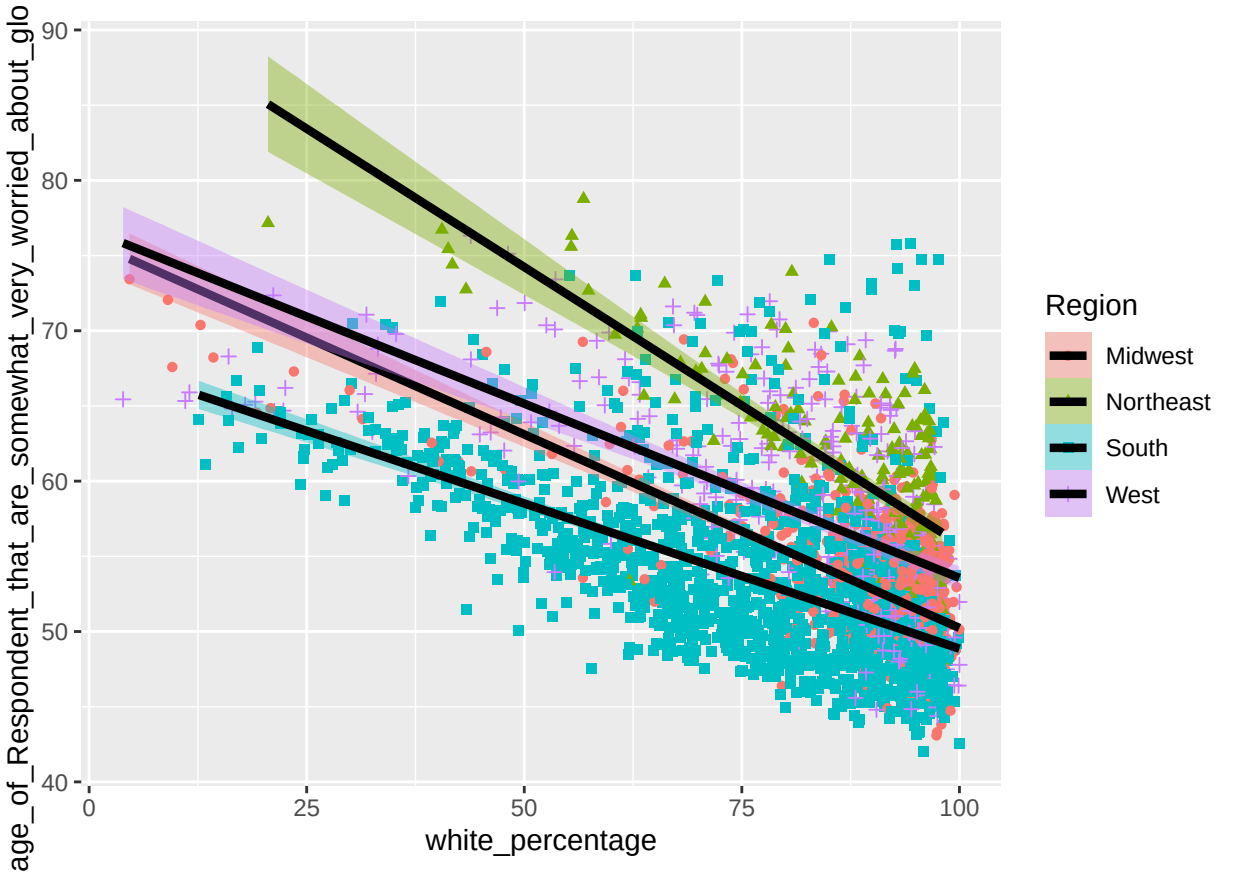
9 : Relationship plot between the %white population and %of respondents that are worried about global warming

```
finalMergedDataSet$white_percentage =  as.double(finalMergedDataSet$white_percentage)
#str(finalMergedDataSet$white_percentage)

stop_dist_model = lm(white_percentage ~ percentage_of_Respondent_that_are_somewhat_very_worried_about_gl

ggplot(finalMergedDataSet, aes(y = percentage_of_Respondent_that_are_somewhat_very_worried_about_global_
  geom_point() +
  geom_smooth(method="lm",color="black",size=1.5,aes(fill=Region))
```

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




```
#facet_wrap(~Region)
```

- Now that we have analysed what percentage of people believe in global warming by county and by state, we wanted to further visualize the relationship between the %white population by county against the %of respondents who are worried about global warming.
- We did a scatter plot using the two parameters and used color, shape to create a visual distinction between the four regions.
- The data is densely populated around 50 to 100% white population (on x axis) and 45 to 65% of people worried about global warming category (on y axis).
- We added a linear model regressing the two parameters and included the best fit lines along with their confidence intervals in the graph.
  - Adding these lines shows us the trend in the relationship between these two parameters.
  - The confidence intervals are tighter around the South region line, however they are spread out in the other regions when compared to South
- From the plot we can say that as the counties with lower % white population seems to have higher % of people who are worried about the global warming. In other words, the counties with higher % white population seems to have lower % of people who are worried about the global warming.

Conclusion:

Global warming is a concern for our planet even though the predicted damage is well ahead in the future. The repercussions of our actions are visible in the slow changes that we are observing off late. The temperatures are soaring high, heat waves are becoming more common, arctic ice slowly melting leading to sea level rising,

rainfall is varying, ecosystem balance is going off, marine life being effected, number of natural disasters like floods, cyclones becoming too frequent, health issues arising from climate changes, etc all point towards one thing; Global warming. There is a pressing need for people to realize what is happening and how things could get worse if we don't take necessary actions. In our analysis so far we see only about approx 50% of the total population in the US that actually are worried about Global Warming and this number seems pretty alarming. More poeple need to be made aware of the grave danger that we are putting the planet in and work towards making things better.