

STA/OPR 9750 Fall 2021 – Opinions on Global Warming and Census

Qiyu Liu & Jiayang Li, OPR/STA 9750, Final Project

12/16/2021

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

```
data_loc = "C:/Users/leejy/Desktop/school work/OPR 9750"
setwd(data_loc)
require(data.table)
```

Comment : Set up the direction of our dataset

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)
```

Comment : Through fread() function to read 3 csv files and assign them to mydata1, mydata2 and mydata3.

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

```
mydata2 <- mydata2[-1, ]
mydata2 <- mydata2 %>%rename("County ID"="GEO.id2")
mydata1 <- mydata1 %>%rename("County ID"="cty_FIPS")
mainDataSet = merge(mydata1, mydata2, by="County ID")
```

Comment : According to our observation, mydata2 column GEO.id2 and mydata1 column cty_FIPS have the same data. Therefore, we change the mydata2 column name from GEO.id2 to County ID and the mydata1 column name from cty_FIPS to County ID. We use the same column to merge yale_climate_cty_data and ACS_16_5YR_DP05_with_ann.

3) From the newly merged object, select the following variables and rename them to meaningful variable names:

```
mainDataSet1<-mainDataSet%>%
  rename("Total Population"="HC01_VC03",
        "White Population" = "HC01_VC49",
        "Black Population" = "HC01_VC50",
        "Hispanic Population" = "HC01_VC88",
        "Asian Population"= "HC01_VC56",
        "Median Age"= "HC01_VC23",
        "Female Population" = "HC01_VC05",
        "% of Respondents that believe global warming is Happening" = "happening",
        "% of Respondents that believe global warming is caused by human activities" = "human",
        "% of Respondent that are somewhat/very worried about global warming" ="worried")
newData<-mainDataSet1%>%
  select("County ID",
        "Total Population",
        "White Population",
        "Black Population",
        "Hispanic Population",
        "Asian Population",
        "Median Age",
        "Female Population",
        "% of Respondents that believe global warming is Happening",
        "% of Respondents that believe global warming is caused by human activities",
        "% of Respondent that are somewhat/very worried about global warming")
```

Comment : We rename columns HC01_VC03, HC01_VC49, HC01_VC50, HC01_VC88, HC01_VC56, HC01_VC23, HC01_VC05, happening, human and worried to “Total Population,” “White Population,” “Black Population,” “Hispanic Population,” “Asian Population,” “Median Age,” “Female Population,” “% of Respondents that believe global warming is Happening,” “% of Respondents that believe global warming is caused by human activities,” “% of Respondent that are somewhat/very worried about global warming”, and we select these new named columns to create a newData.

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

```
newData$`Total Population`<-as.numeric(newData$`Total Population`)
newData$`White Population` <-as.numeric(newData$`White Population`)
newData$`Black Population`<-as.numeric(newData$`Black Population`)
newData$`Hispanic Population`<-as.numeric(newData$`Hispanic Population`)
newData$`Asian Population`<-as.numeric(newData$`Asian Population`)
newData$`Female Population`<-as.numeric(newData$`Female Population`)

newData=newData%>%mutate( '% White'=(newData$`White Population`/newData$`Total Population`)*100)
newData=newData%>%mutate( '% Black'=(newData$`Black Population`/newData$`Total Population`)*100)
newData=newData%>%mutate( '% Hispanic'=(newData$`Hispanic Population`/newData$`Total Population`)*100)
newData=newData%>%mutate( '% Asian'=(newData$`Asian Population`/newData$`Total Population`)*100)
newData=newData%>%mutate( '% Female'=(newData$`Female Population`/newData$`Total Population`)*100)
newData%>%head()
```

```

## County ID Total Population White Population Black Population
## 1 01001 55049 42311 10196
## 2 01003 199510 172441 18594
## 3 01005 26614 12430 12756
## 4 01007 22572 17370 4788
## 5 01009 57704 55073 905
## 6 01011 10552 2301 7967
## Hispanic Population Asian Population Median Age Female Population
## 1 1416 444 37.8 28172
## 2 8712 1338 42.3 102139
## 3 1147 129 38.7 12375
## 4 502 14 40.2 10488
## 5 5036 92 40.8 29132
## 6 13 84 39.2 4833
## % of Respondents that believe global warming is Happening
## 1 59.476
## 2 60.062
## 3 67.862
## 4 58.232
## 5 52.956
## 6 73.924
## % of Respondents that believe global warming is caused by human activities
## 1 46.971
## 2 46.129
## 3 51.930
## 4 45.592
## 5 44.090
## 6 56.726
## % of Respondent that are somewhat/very worried about global warming % White
## 1 48.048 76.86062
## 2 48.790 86.43226
## 3 57.342 46.70474
## 4 47.417 76.95375
## 5 43.385 95.44052
## 6 64.653 21.80629
## % Black % Hispanic % Asian % Female
## 1 18.521681 2.5722538 0.80655416 51.17622
## 2 9.319834 4.3666984 0.67064308 51.19493
## 3 47.929661 4.3097618 0.48470730 46.49808
## 4 21.212121 2.2239943 0.06202375 46.46465
## 5 1.568349 8.7272979 0.15943435 50.48523
## 6 75.502274 0.1231994 0.79605762 45.80174

```

Comment : We noticed that the variables Total Population, White Population, Black Population, Hispanic Population, Asian Population, and Female Population from our data set are in categorical format. So for our further analysis, we change the name for these variables. In order to get the new percent variables, we use White Population divided Total Population, Black Population divided Total Population, Hispanic Population divided Total Population, Asian Population divided Total Population, and Female Population divided Total Population to get new columns White percentage, Black percentage, Hispanic percentage, Asian percentage, and Female percentage. We show the first 5 data by using head() to see if our new columns are created.

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

```
summary(newData)
```

```
## County ID      Total Population  White Population  Black Population
## Length:3142    Min.      :    76    Min.      :    29    Min.      :    0.0
## Class :character 1st Qu.: 10996  1st Qu.:  9060  1st Qu.:   97.2
## Mode  :character Median : 25738  Median : 21934  Median :   784.0
##              Mean  : 101387  Mean  :  74366  Mean  : 12807.7
##              3rd Qu.: 67583  3rd Qu.: 57795  3rd Qu.: 5515.2
##              Max.   :10057155  Max.   :5283457  Max.   :1248948.0
## Hispanic Population Asian Population  Median Age      Female Population
## Min.      :    0    Min.      :    0.0    Length:3142    Min.      :    36
## 1st Qu.:   312    1st Qu.:   31.0    Class :character 1st Qu.:   5491
## Median :  1010    Median :   135.0    Mode  :character Median :  12935
## Mean   :  17568    Mean   :   5287.9                      Mean   :  51494
## 3rd Qu.:  4754    3rd Qu.:   706.8                      3rd Qu.:  34260
## Max.   :4861648    Max.   :1431361.0                    Max.   :5102539
## % of Respondents that believe global warming is Happening
## Min.      :49.08
## 1st Qu.:59.70
## Median :62.95
## Mean   :64.00
## 3rd Qu.:67.43
## Max.   :84.20
## % of Respondents that believe global warming is caused by human activities
## Min.      :39.36
## 1st Qu.:47.67
## Median :50.19
## Mean   :51.22
## 3rd Qu.:53.69
## Max.   :72.90
## % of Respondent that are somewhat/very worried about global warming
## Min.      :42.05
## 1st Qu.:49.64
## Median :52.84
## Mean   :54.18
## 3rd Qu.:57.52
## Max.   :78.77
## % White      % Black      % Hispanic      % Asian
## Min.      : 3.903    Min.      : 0.0000    Min.      : 0.000    Min.      : 0.0000
## 1st Qu.: 76.917    1st Qu.: 0.6246    1st Qu.: 1.941    1st Qu.: 0.2575
## Median : 89.909    Median : 2.2135    Median : 3.834    Median : 0.5692
## Mean   : 83.382    Mean   : 9.0166    Mean   : 8.947    Mean   : 1.3027
## 3rd Qu.: 95.416    3rd Qu.:10.2691    3rd Qu.: 9.067    3rd Qu.: 1.2348
## Max.   :100.000    Max.   :86.1849    Max.   :98.959    Max.   :42.8982
## % Female
## Min.      :21.51
## 1st Qu.:49.45
## Median :50.42
## Mean   :49.94
```

```
## 3rd Qu.:51.13
## Max. :58.50
```

Comment :According to our new variables, it shows that white people is the main part of the total population in the United States. The min and max of White population are 3.903% and 100%, and the mean of White population is 83.382%. The min and max of Black population are 0% and 86.1849%, and the mean of Black population is 9.0166%. The min and max of Hispanic population are 0% and 98.959%, and the mean of Hispanic population is 8.947%. The min and max of Asian population are 0% and 42.8982%, and the mean of Asian population is 1.3027%. White population has the highest min, median, mean, and max compare with other races, and Asian population has the lowest min, median, mean, and max compare with other races in this data. Meanwhile, the data shows that the min and max of Female are 21.51% and 58.50%, and the mean of Female is 49.94%. It means the population of men and women is almost equally divided.

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).

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

newData=newData%>%mutate('State_FIPS'=substr(newData$`County ID`, 1, 2))
newData$`State_FIPS`<-as.numeric(newData$`State_FIPS`)
newData=inner_join(x = newData, y = region)
str(newData)
```

```
## 'data.frame': 3142 obs. of 19 variables:
## $ County ID : chr "01001" "01003"
## $ Total Population : num 55049 199510 266
## $ White Population : num 42311 172441 124
## $ Black Population : num 10196 18594 1275
## $ Hispanic Population : num 1416 8712 1147 5
## $ Asian Population : num 444 1338 129 14
## $ Median Age : chr "37.8" "42.3" "3
## $ Female Population : num 28172 102139 123
## $ % of Respondents that believe global warming is Happening : num 59.5 60.1 67.9 5
## $ % of Respondents that believe global warming is caused by human activities: num 47 46.1 51.9 45.
## $ % of Respondent that are somewhat/very worried about global warming : num 48 48.8 57.3 47.
## $ % White : num 76.9 86.4 46.7 7
## $ % Black : num 18.52 9.32 47.93
## $ % Hispanic : num 2.57 4.37 4.31 2
## $ % Asian : num 0.807 0.671 0.48
## $ % Female : num 51.2 51.2 46.5 4
## $ State_FIPS : num 1 1 1 1 1 1 1
## $ Region : chr "South" "South"
## $ State : chr "Alabama" "Alabar
```

Comment : In order to read the the region variable from region.csv, we can use fread() function to read the region.csv file and name it as 'region'. We use substr method to slice the first 2 digits from variable County ID and use mutate function to create the new state variable 'State_FIPS'. Since the new variable that we created is a categorical variable, we change its data type to numeric. After that, we can use inner

join method to merge these two data sets by using the common variable 'State_FIPS'. Lastly, we use the `str()` function to see whether we merged these two data sets successfully.

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

```
newData %>% group_by(State) %>%
  summarize('the average % that believe global warming is occurring'
            = mean('% of Respondents that believe global warming is Happening'))%>%
  arrange(desc('the average % that believe global warming is occurring'))
```

```
## # A tibble: 51 x 2
##   State      'the average % that believe global warming is occurring'
##   <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
```

Comment : We group the state variable and create a new column 'the average % that believe global warming is occurring' which is the mean of '% of Respondents that believe global warming is Happening'. Meanwhile, we arrange the new variable 'the average % that believe global warming is occurring' by descending. Our data shows that District of Columbia has the highest average percentage that believe global warming is occurring. The second state is Hawaii, the third state is Massachusetts, the fourth state is New Jersey, and the fifth state is Vermont.

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

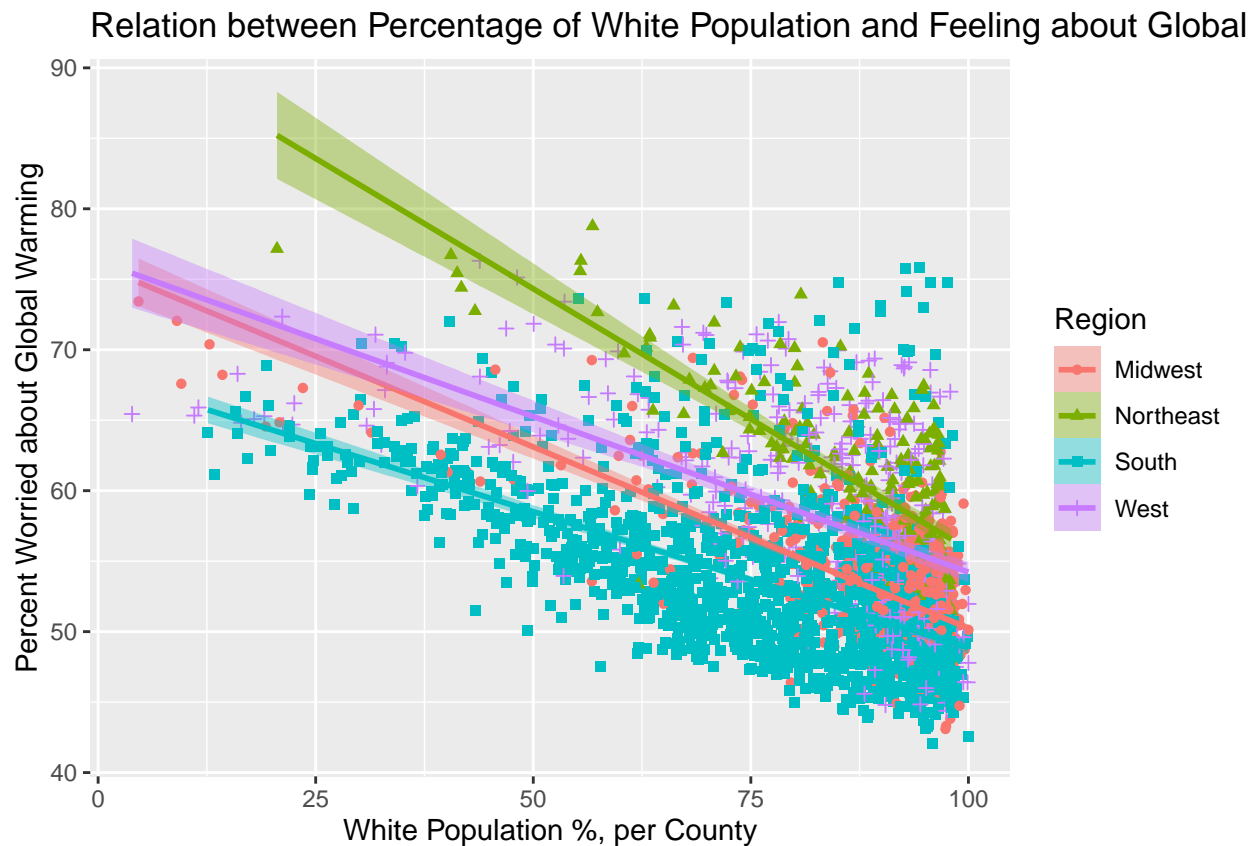
```
newData %>% group_by(Region) %>%
  summarize('the average % that believe global warming is caused by humans'
            = mean('% of Respondents that believe global warming is caused by human activities'))%>%
  arrange(desc('the average % that believe global warming is caused by humans'))
```

```
## # A tibble: 4 x 2
##   Region      'the average % 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
```

Comment : We group the Region variable and create a new column 'the average % that believe global warming is caused by humans' which is the mean of '% of Respondents that believe global warming is caused by human activities'. Meanwhile, we arrange the new variable 'the average % that believe global warming is caused by humans' by descending order. Our data shows that Northeast has the highest average percentage that believe global warming is caused by humans. The second region is West, the third region is Midwest, and the fourth region is South. These regions all have the percentage greater and equal 50%, and it means more than half of the population believe global warming is caused by humans.

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. You can do 4 “facet” plots or overlay each relationship on the same graph. You do NOT have to make the colors as fancy as in the example below.

```
qplot(x=~% White`,
      y=~% of Respondent that are somewhat/very worried about global warming`,
      data=newData,fill=Region,shape=Region,color=Region)+
  geom_smooth(method='lm')+
  ylab('Percent Worried about Global Warming')+
  xlab('White Population %, per County')+
  ggtitle('Relation between Percentage of White Population and Feeling about Global Warming')
```

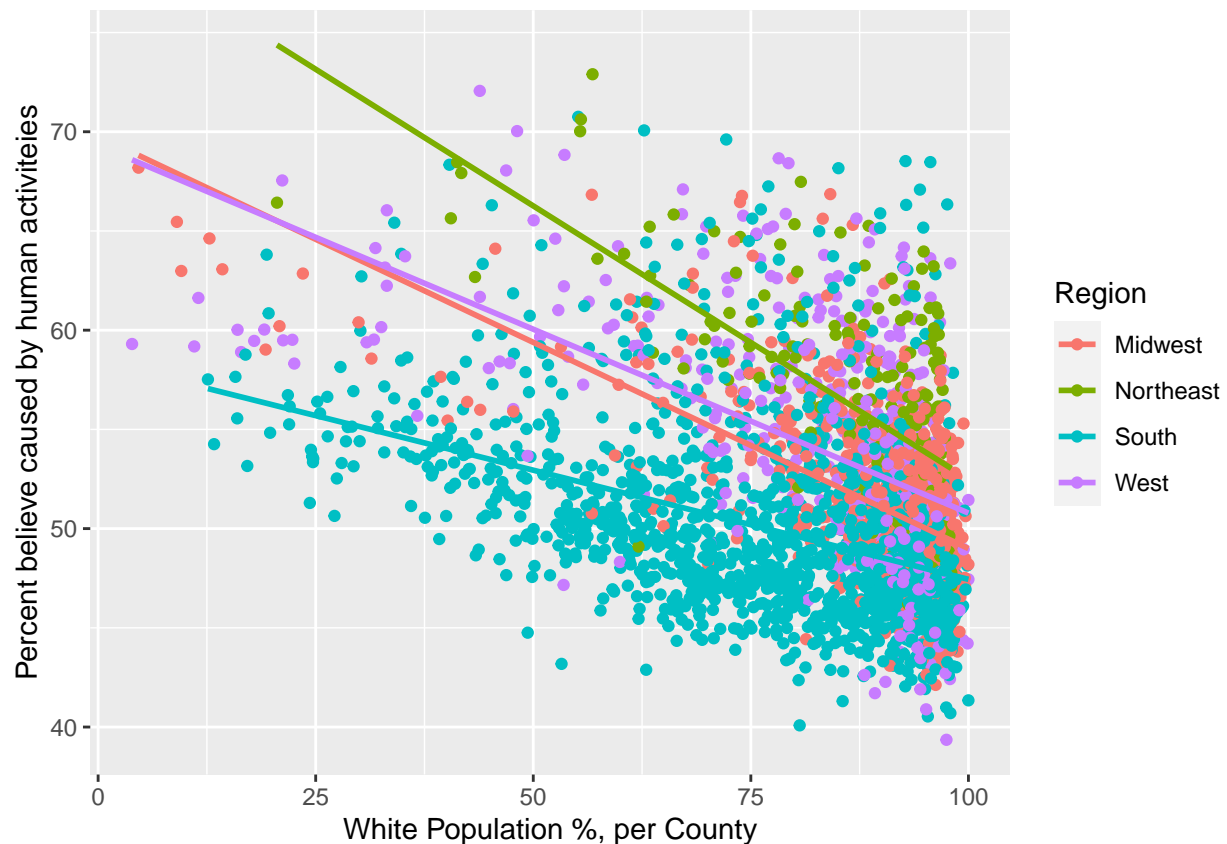


Comment : Based on the output above, it shows there is a negative correlation relationship between the White Population %, per County. The Northeast has the highest Percent Worried about Global Warming when these four regions have the same White Population %, per County, and the Northeast has the highest

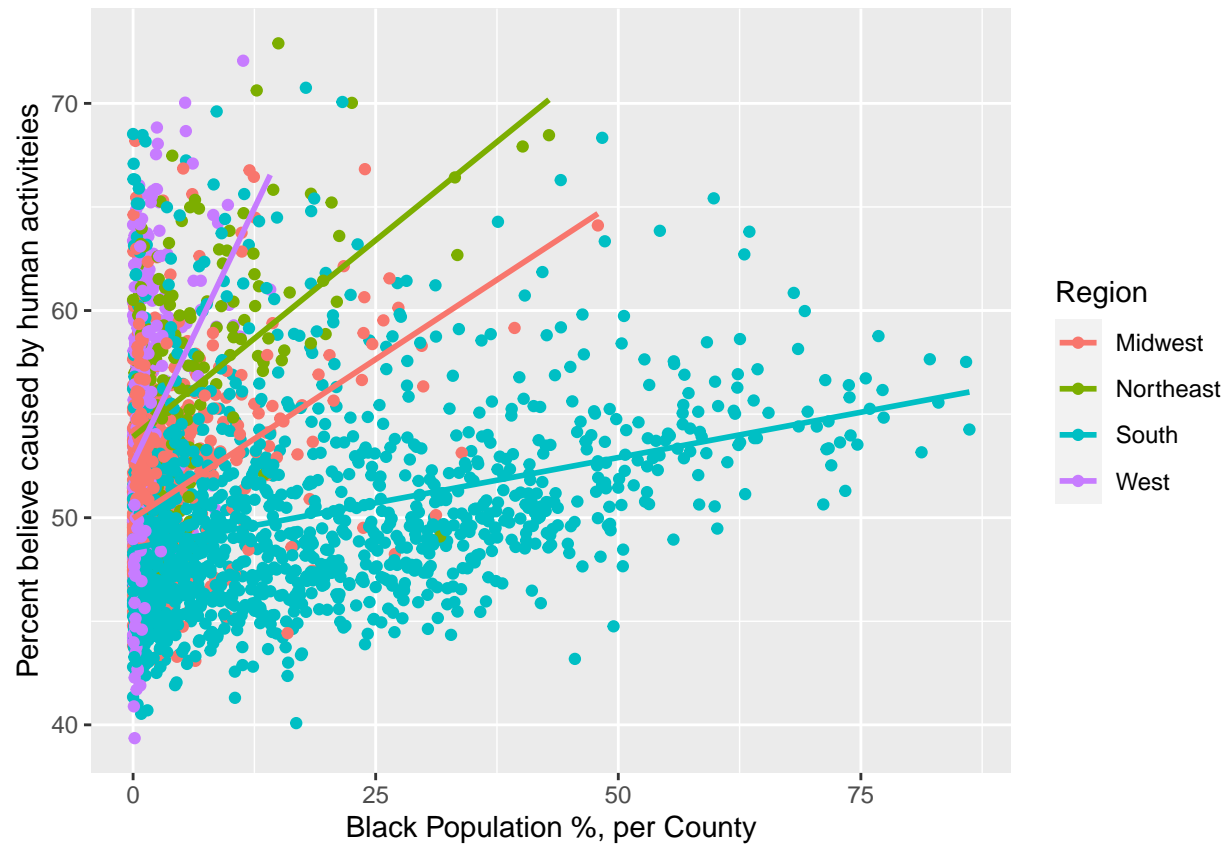
White Population %, per County when these four regions have the same Percent Worried about Global Warming. It means the Percent Worried about Global Warming will decrease as the White Population %, per County increases. In addition, the people from the Northeast seem have most concern about the global warming, the South region has the least percentage of people who are worried about the global warming among all four regions.

10) Plot the relationship between the % of different group of population and the % of Respondents that believe global warming is caused by human activities, separating the 4 different regions.

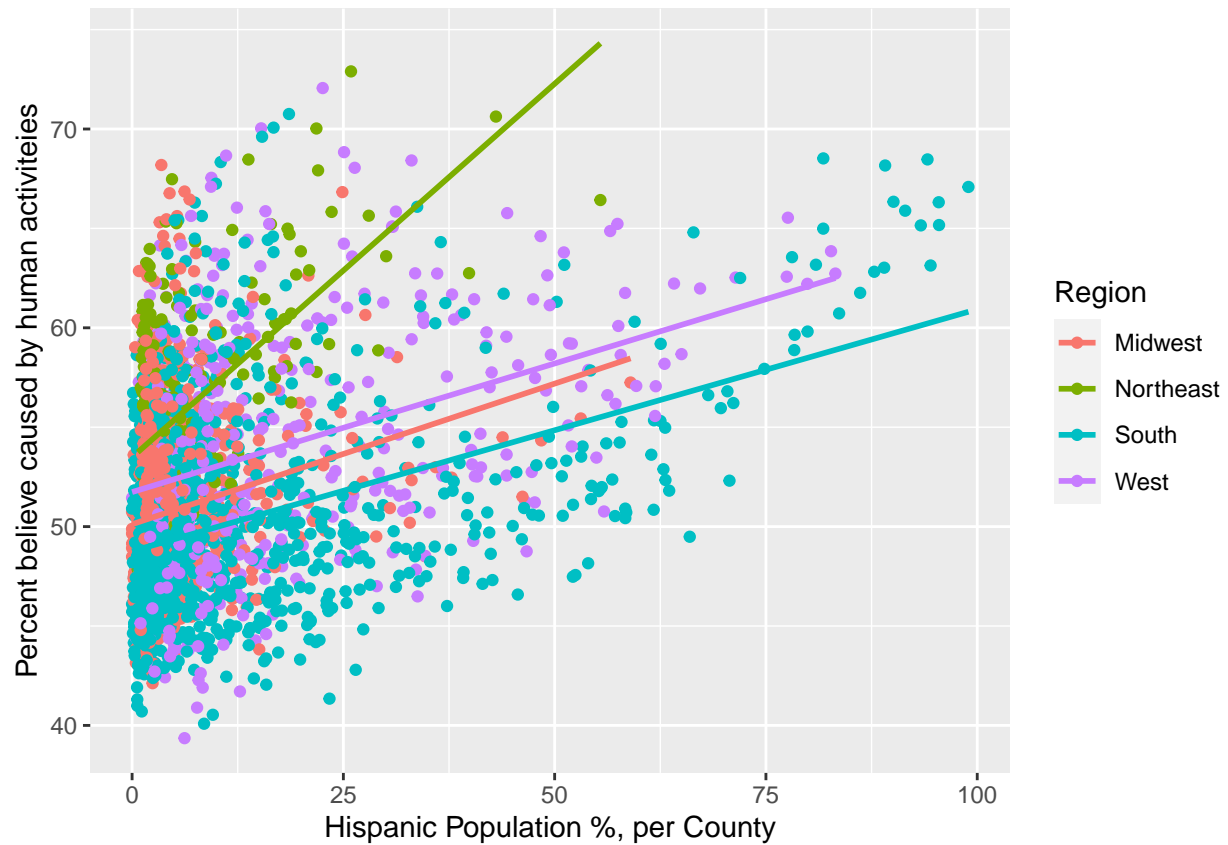
```
par(mfrow=c(2,2))
ggplot(newData,aes(`% White`, `% of Respondents that believe global warming is caused by human activities`))
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe caused by human activiteies')+
  xlab('White Population %, per County')
```



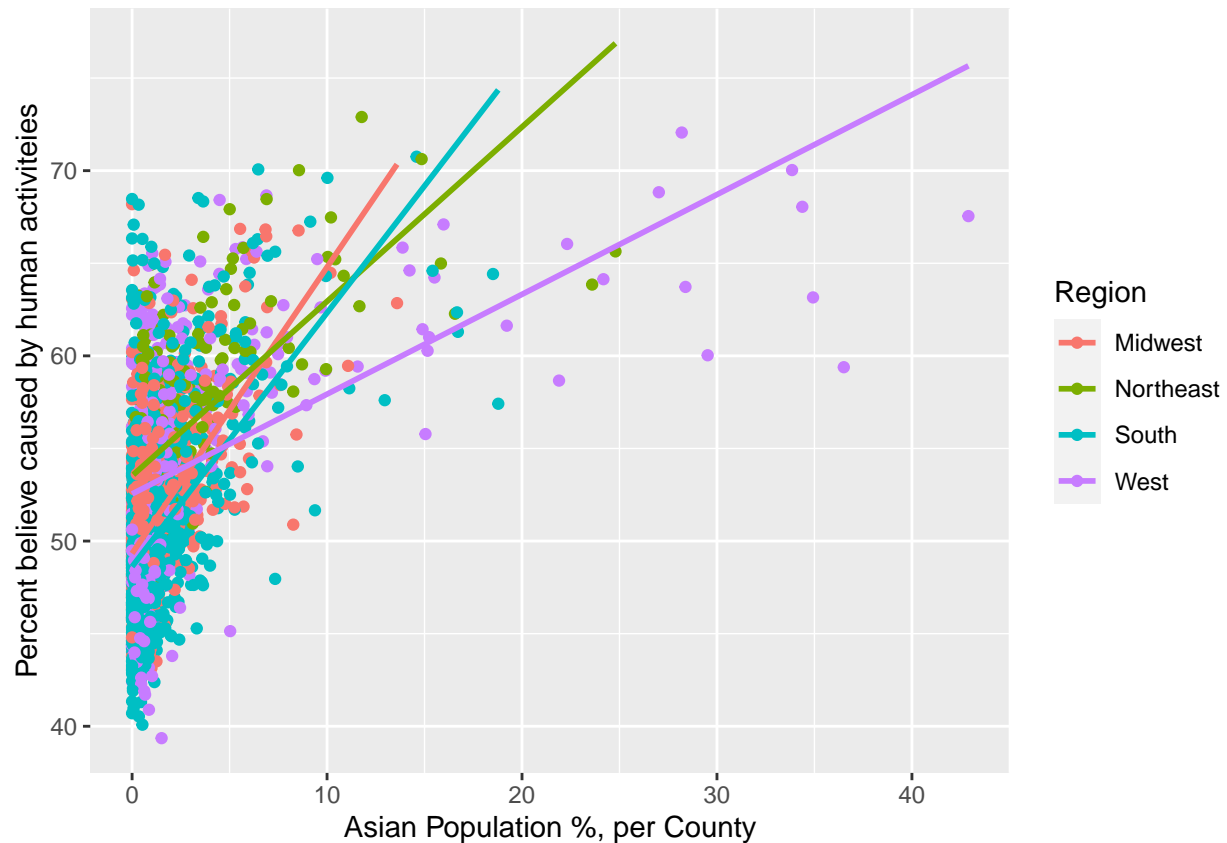
```
ggplot(newData,aes(`% Black`, `% of Respondents that believe global warming is caused by human activities`))
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe caused by human activiteies')+
  xlab('Black Population %, per County')
```

```
ggplot(newData,aes(`% Hispanic`, ` % of Respondents that believe global warming is caused by human activi
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe caused by human activiteies')+
  xlab('Hispanic Population %, per County')
```



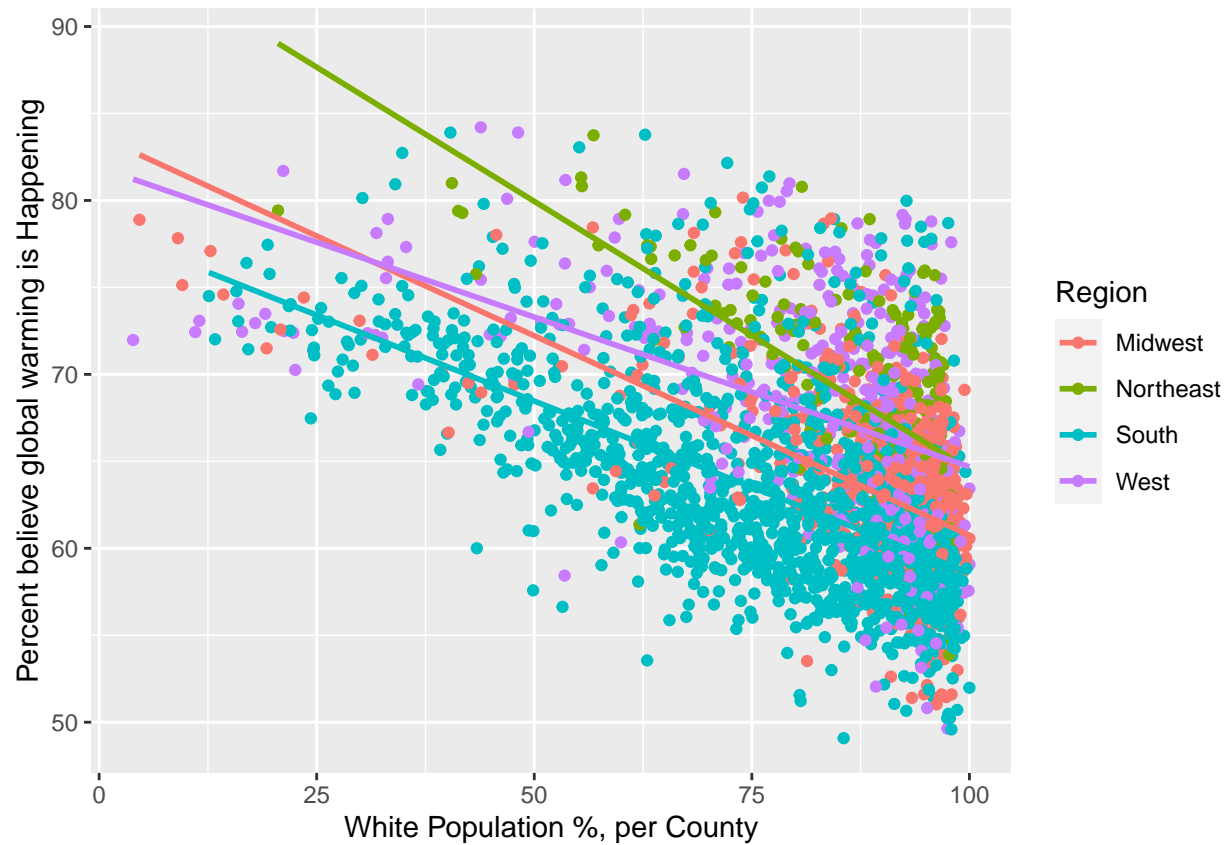
```
ggplot(newData,aes(`% Asian`, ` % of Respondents that believe global warming is caused by human activities`)) +
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe caused by human activiteies')+
  xlab('Asian Population %, per County')
```



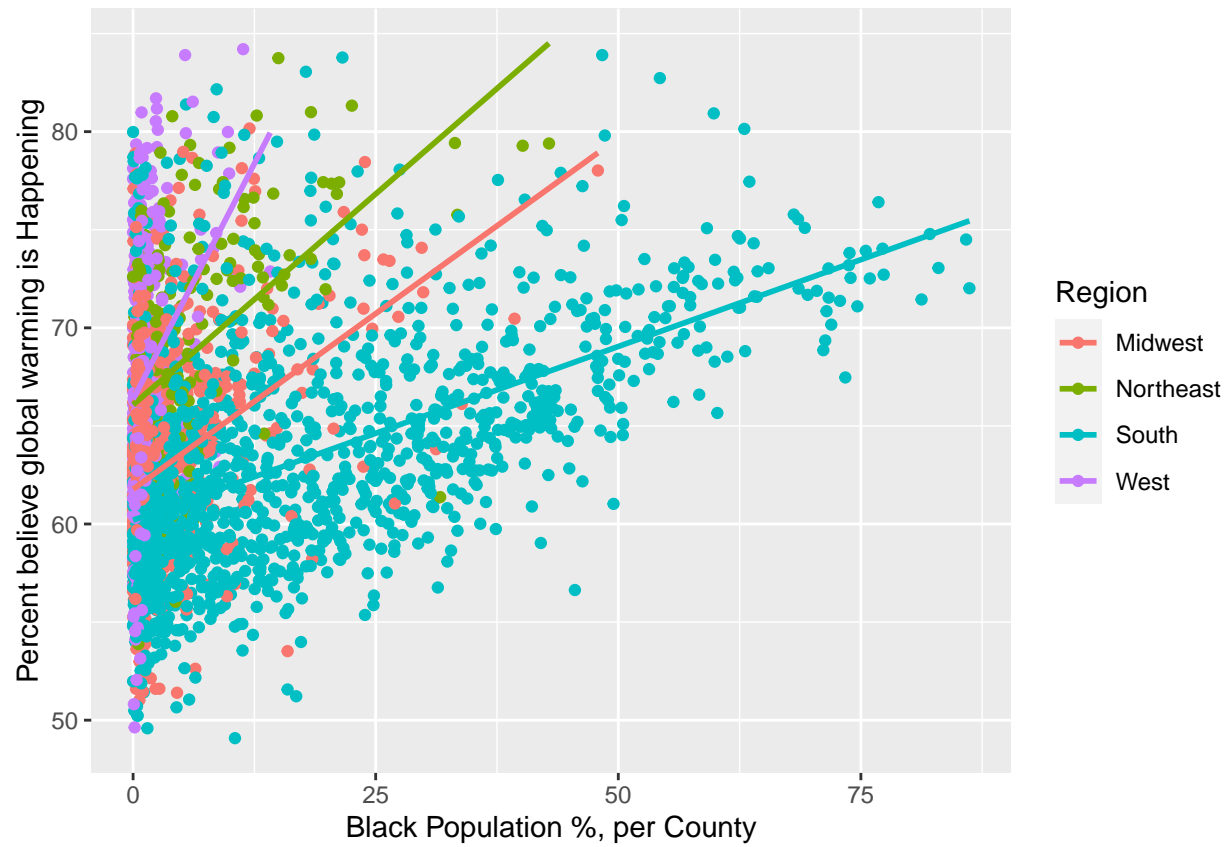
Comment : Based on the output above, it shows there is a negative correlation relationship between the % of white people and % believe warming caused by human activities. There are positive correlation relationship between % of other group of population and % believe warming caused by human activities. That means except white population, other group of people who believe that global warming is linked to humanity increase with their % of population increase.

11) Plot the relationship between the % of different group of population and the % of Respondents that believe global warming is Happening, separating the 4 different regions.

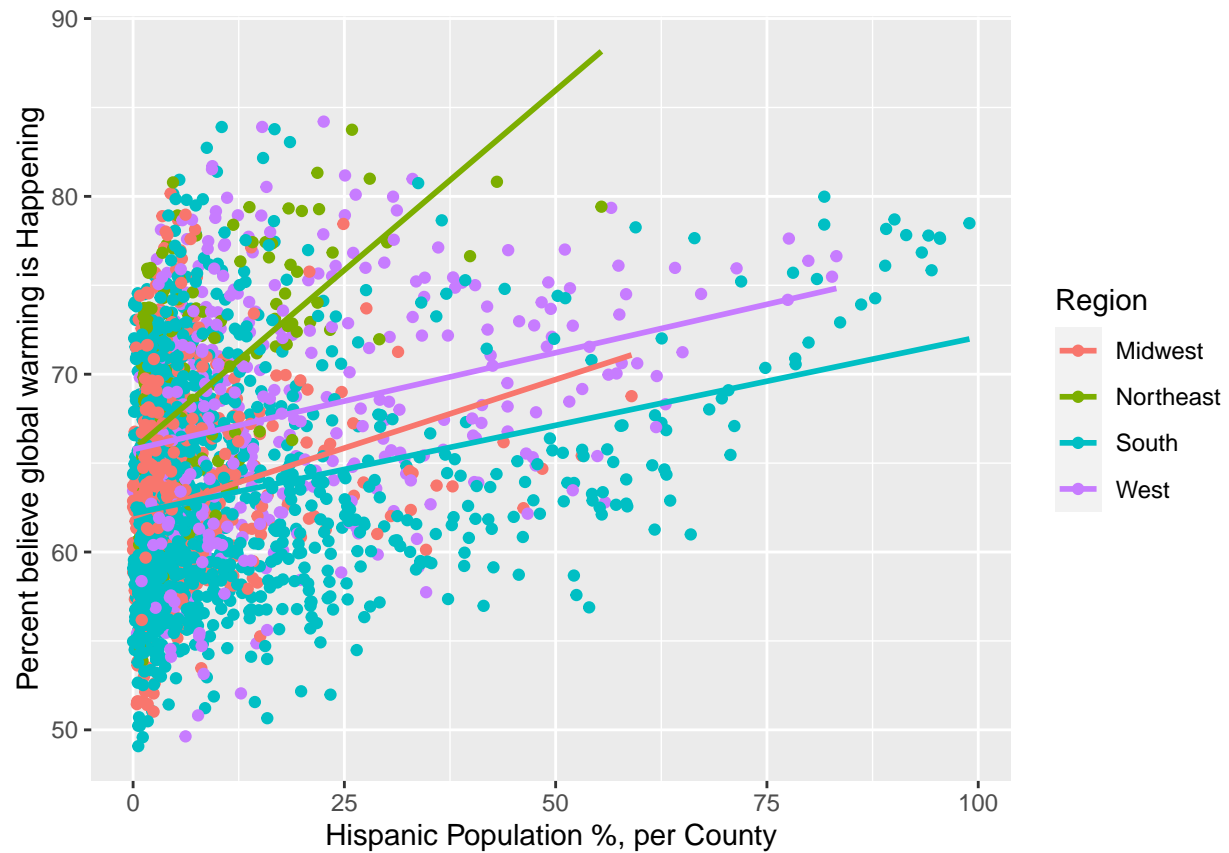
```
par(mfrow=c(2,2))
ggplot(newData,aes(`% White`, ` % of Respondents that believe global warming is Happening`,color=Region))
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe global warming is Happening')+
  xlab('White Population %, per County')
```



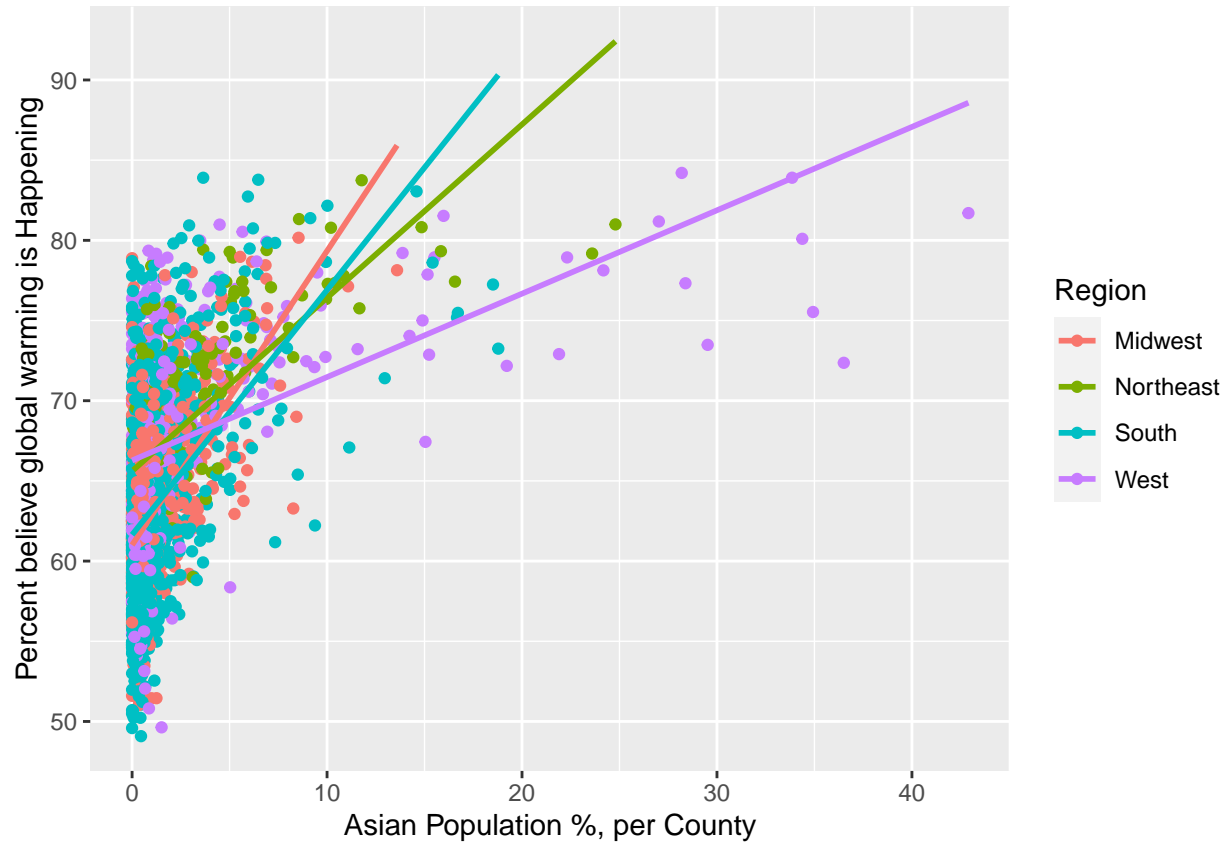
```
ggplot(newData,aes(`% Black`, ` % of Respondents that believe global warming is Happening`,color=Region))+
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe global warming is Happening')+
  xlab('Black Population %, per County')
```



```
ggplot(newData,aes(`% Hispanic`, ` % of Respondents that believe global warming is Happening`,color=Region))
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe global warming is Happening')+
  xlab('Hispanic Population %, per County')
```



```
ggplot(newData,aes(`% Asian`, ` % of Respondents that believe global warming is Happening`,color=Region))
  geom_point(show.legend = TRUE)+
  geom_smooth(method = 'lm',se = FALSE)+
  ylab('Percent believe global warming is Happening')+
  xlab('Asian Population %, per County')
```



Comment : Based on the output above, it shows there is a negative correlation relationship between the % of white people and % of Respondents that believe global warming is Happening. There are positive correlation relationship between % of other group of population and % of Respondents that believe global warming is Happening. That means except white population, other group of people who believe that global warming is happening increase with their % of population increase.

According to the analytics of the the relationships between different group of population and their believes of global warming in the United States, it shows that white population has different point of views with other group of population. As their % of the population increases per county, the % of believe global warming, the % of believe global warming happening, and % of believe global warming is caused by human activities are all decrease. Comparing with Black population, Hispanic population and Asian population, as their % of the population increases per county, the % of believe global warming, the % of believe global warming happening, and % of believe global warming is caused by human activities are all increase.