Project1

Applied Data Science @ Columbia

Fall 2020

Project 1: A "data story" on how Americans vote

Project Description

This is the first and only *individual* (as opposed to *team*) this semester.

Term: Fall 2020

- Project title: How did features of voters in each racial group change?
- This project is conducted by Siran Qiu
- Project summary:First, I looked into the wordcloud of 2020 questionnaire question to decide my main topic–race and then I processed the time series data from 1968 to 2020 to see how the way of each racial group voting changed.

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(SnowballC)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(readtext)
library(haven)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(ggplot2)
library(labelled)
library(devtools)
```

```
## Loading required package: usethis
```

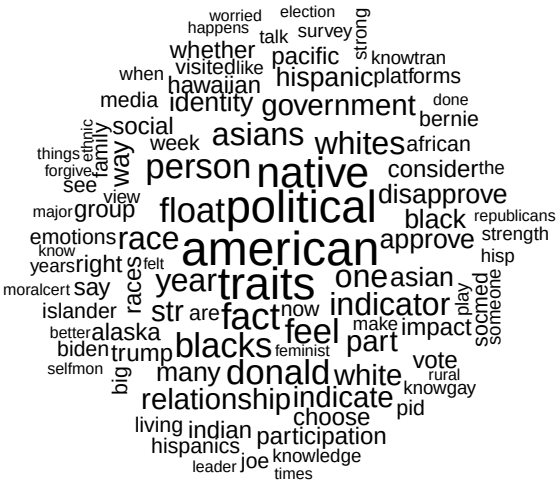
```
library(usethis)
```

Step 1: Wordcloud Word cloud of 2020 questionnaires-to find the theme of this project First,we are going to see what main words appears in 2020 questionnaires are and get some ideas about the information ANES tries to get from respondent.

```
## Warning in tm_map.SimpleCorpus(corp_2020, PlainTextDocument): transformation
## drops documents
```

```
## Warning in tm_map.SimpleCorpus(corp_2020, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corp_2020, removeWords, c("byte", "how", :
## transformation drops documents
```



From the wordcloud, the words relate to race (such as blacks,whites,asians,race) appear more time than other words and I am going to process some variables that may relate to this theme from the data of Time Series Cumulative Data from 1952 to 2016

Step2:Process the data

```
Election_years=as.character(seq(1968, 2016, 4))
anes_dat <-read_sav("~/Desktop/Github/Fall2020-Project1-siranq/data/anes_timeseries_cdf.sav")
anes_use<-anes_dat%>%
mutate(year=as_factor(VCF0004),turnout=as_factor(VCF0703),vote=as_factor(VCF0706),race=as_factor(VCF010
filter(year %in% Election_years)
anes_dat$VCF0102<-remove_labels(anes_dat$VCF0102)
anes_dat$VCF0102[anes_dat$VCF0102==c("1")]<-c("17-24")
anes_dat$VCF0102[anes_dat$VCF0102==c("2")]<-c("25-34")
anes_dat$VCF0102[anes_dat$VCF0102==c("3")]<-c("35-44")
anes_dat$VCF0102[anes_dat$VCF0102==c("4")]<-c("45-54")
anes_dat$VCF0102[anes_dat$VCF0102==c("5")]<-c("55-64")
anes_dat$VCF0102[anes_dat$VCF0102==c("6")]<-c("65-74")
anes_dat$VCF0102[anes_dat$VCF0102==c("7")]<-c("75-99 and over")
```

```
data.table(anes_use%>%
select(year, turnout, vote, race,age, gender,education)%>%
filter(!is.na(turnout))%>% sample_n(10))
```

```
##     year                                turnout
##  1: 2008                3. Voted (registered)
##  2: 1968                3. Voted (registered)
##  3: 1984                3. Voted (registered)
##  4: 1972     2. Registered, but did not vote
##  5: 1980                3. Voted (registered)
##  6: 1976                3. Voted (registered)
##  7: 2008                3. Voted (registered)
##  8: 1988 1. Not registered, and did not vote
##  9: 2008 1. Not registered, and did not vote
## 10: 2016                3. Voted (registered)
##                                                      vote
##  1:                                            1. Democrat
##  2:                                          2. Republican
##  3:                                            1. Democrat
##  4: 7. Did not vote or voted but not for president (exc.1972)
##  5:                                            1. Democrat
##  6:                                            1. Democrat
##  7:                                            1. Democrat
##  8: 7. Did not vote or voted but not for president (exc.1972)
##  9: 7. Did not vote or voted but not for president (exc.1972)
## 10:                                          2. Republican
##                                                       race        age    gender
##  1:                    2. Black non-Hispanic (1948-2012) 6. 65 - 74 2. Female
##  2:                    1. White non-Hispanic (1948-2012) 4. 45 - 54   1. Male
##  3:                    1. White non-Hispanic (1948-2012) 4. 45 - 54   1. Male
##  4:                    1. White non-Hispanic (1948-2012) 3. 35 - 44 2. Female
##  5:                    1. White non-Hispanic (1948-2012) 4. 45 - 54   1. Male
##  6:                    1. White non-Hispanic (1948-2012) 3. 35 - 44 2. Female
##  7:                    1. White non-Hispanic (1948-2012) 4. 45 - 54 2. Female
##  8:                    1. White non-Hispanic (1948-2012) 2. 25 - 34 2. Female
##  9:                                5. Hispanic (1966-2012) 2. 25 - 34 2. Female
## 10: 6. Other or multiple races, non-Hispanic (1968-2012) 4. 45 - 54   1. Male
##                                                    education
##  1: 2. High school (12 grades or fewer, incl. non-college
##  2:         4. College or advanced degree (no cases 1948)
##  3: 2. High school (12 grades or fewer, incl. non-college
##  4:     3. Some college (13 grades or more but no degree;
##  5:         4. College or advanced degree (no cases 1948)
##  6: 2. High school (12 grades or fewer, incl. non-college
##  7:     3. Some college (13 grades or more but no degree;
##  8:                                                 <NA>
##  9: 2. High school (12 grades or fewer, incl. non-college
## 10:     3. Some college (13 grades or more but no degree;
```

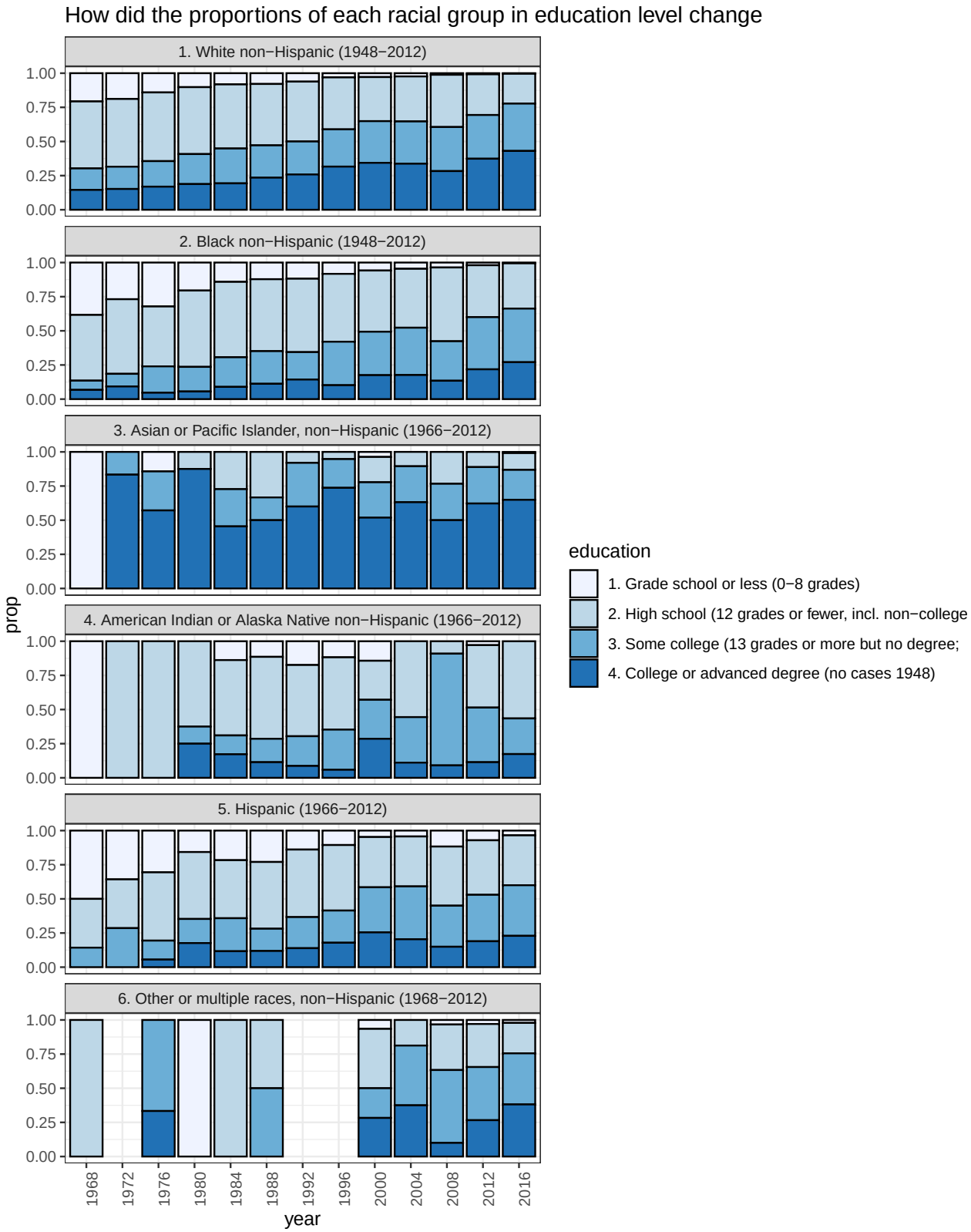
```
save(anes_use, file="~/Desktop/Github/Fall2020-Project1-siranq/output/data_use.RData")
```

Step3:Simple analysis

3.1 Education level How did the proportions of each racial group in education level change?

```r
load(file="~/Desktop/Github/Fall2020-Project1-siranq/output/data_use.RData")
anes_to_race_edu = anes_use %>%
filter(!is.na(race) & !is.na(education) &!is.na(turnout))%>%
group_by(year,education,race)%>%
count(education)%>%
group_by(year,race)%>%
mutate(prop=n/sum(n))

ggplot(anes_to_race_edu,aes(x=year,y=prop,fill=education),rep="best")+
geom_bar(stat="identity", colour="black")+
facet_wrap(~race, ncol=1)+
theme_bw()+
theme(axis.text.x=element_text(angle=90))+
scale_fill_brewer(palette="Blues")+
labs(title="How did the proportions of each racial group in education level change")
```



Story 1: From the graph above, I discover that both the proportions of high education level of White non-Hispanic and Black non-Hispanic respondent grew from 1968 to 2016 but the interesting thing is that the proportion of high education level of Asians or Pacific Islander non-Hispanic varies from 1968 to 2016. It absolutely makes sense that in 1968 all the Asian respondent is grade school or less because at that time

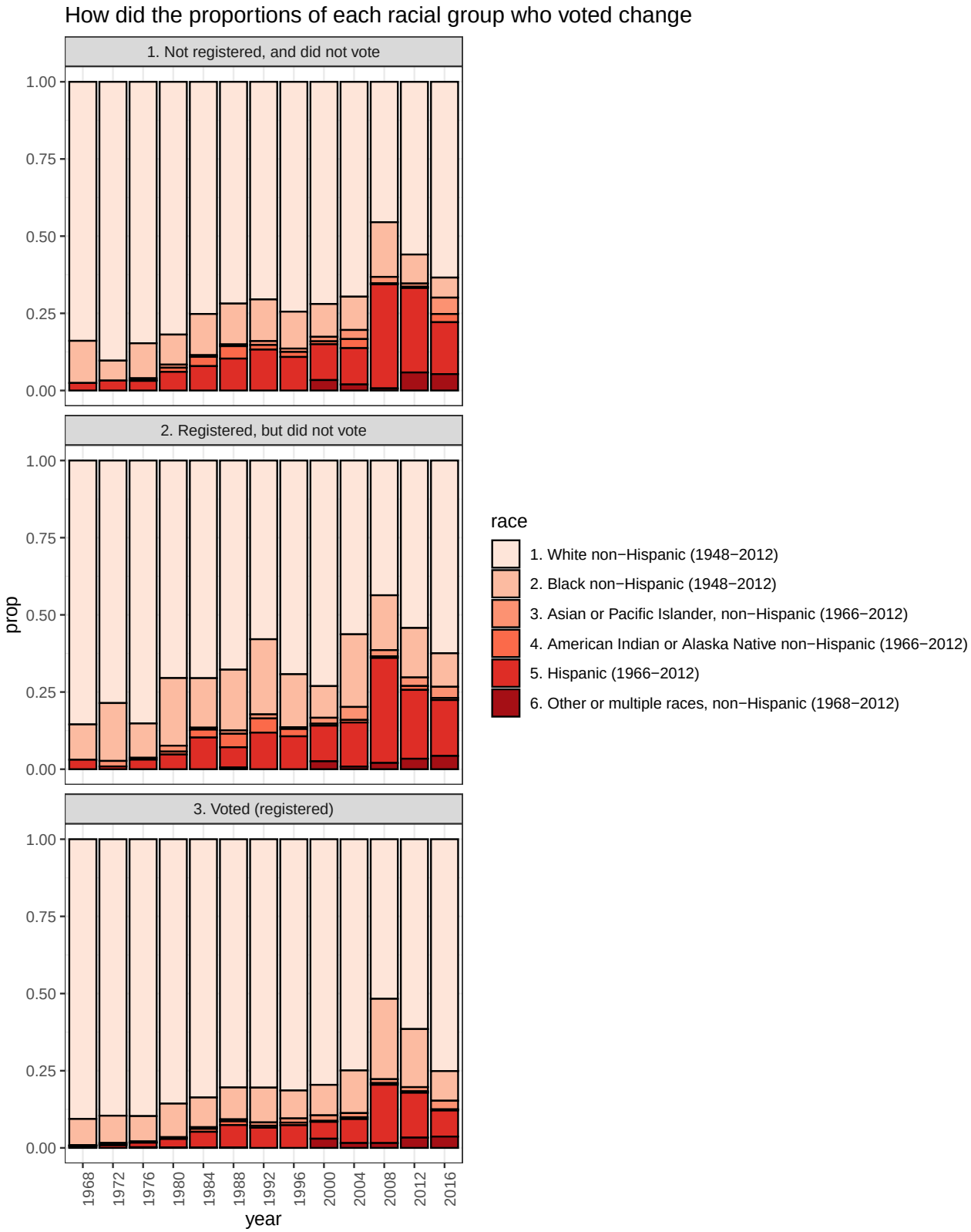
most of them were immigrants. My guess of reason of this variation is that the number of Asian respondent is not as much as White and Blacks.

Wait!Let's see an old news!

3.2 Who voted? How did the proportions of each racial group who voted change?

```
anes_to_race_year = anes_use %>%
  filter(!is.na(race) & !is.na(turnout))%>%
  group_by(year, turnout,race)%>%
  count(race)%>%
  group_by(year,turnout)%>%
  mutate(
    prop=n/sum(n)
  )

ggplot(anes_to_race_year,
       aes(x=year, y=prop, fill=race)) +
  geom_bar(stat="identity", colour="black") + facet_wrap(~turnout, ncol=1) + theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Reds")+
  labs(title="How did the proportions of each racial group who voted change")
```



Story 2: From the graph above, we can see that although from past to present the respondents were mainly White non-Hispanic, the races of respondents has had more diversity from 1968 to 2016. Especially, during 2008, the proportions of all the racial groups other than Whites enlarged. Guess why? Yes,it is because of him.

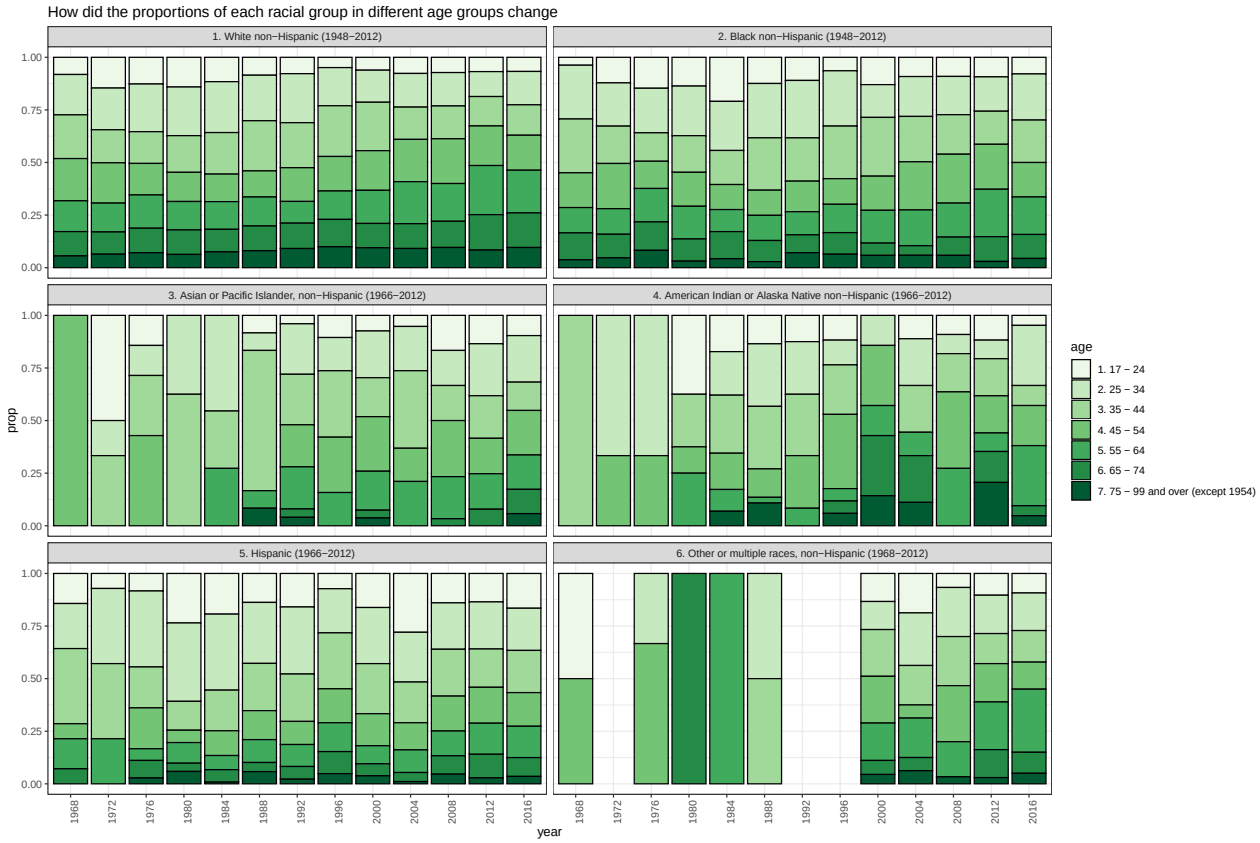
The first Black President in U.S.A. Hence, people may pay more attention to vote.

3.3 Age

How did the proportions of each racial group in different age groups change?

```
anes_to_race_edu = anes_use %>%
filter(!is.na(race) & !is.na(age) &!is.na(turnout))%>%
group_by(year,age,race)%>%
count(age)%>%
group_by(year,race)%>%
mutate(prop=n/sum(n))

ggplot(anes_to_race_edu,aes(x=year,y=prop,fill=age))+
geom_bar(stat="identity", colour="black")+
facet_wrap(~race, ncol=2)+
theme_bw()+
theme(axis.text.x=element_text(angle=90))+
scale_fill_brewer(palette="Greens")+
labs(title="How did the proportions of each racial group in different age groups change")
```



Story 3: From the graph above, we can see that the respondents of Black and Whites are from every age group and mainly concentrated at 25-64. However,the respondents of Asians and Hispanics before 1988 are all from young age group. My guess is that maybe similar to the reason of Story 2, the Asian and Hispanic respondents in that day might be immigrants so they came to U.S.A at young age or they might be the second generations of Asian or Hipanic immigrants. As time flew, after 1988, the respondent who attended the interview became older and the diversity of the respondent's age group became large.