Will characteristics of the voter change over time

Introduction

This is a data story about the general profile of voter. I’m curious about will the characteristics of voters change over time. If it will change over time, then when it change and how it change? In addition, this data story can tell us about what kind of people are voting. Will there be more old people or more young people participated in vote? And how about the gender and education level of people participated in vote?

Data import and prepossing

```
# import data
library(readstata13)
mydata<-read.dta13(’D:\\Applied data\\anes_timeseries_cdf_dta\\anes_timeseries_cdf.dta’)
mydata<-data.frame(lapply(mydata, as.character), stringsAsFactors=FALSE)
```

```
# check what data looks like
print(dim(mydata)[1])
```

```
## [1] 59944
```

```
print(dim(mydata)[2])
```

```
## [1] 1029
```

There are 1029 variables in data set. We only need some of it. I extracted five columns—-‘year’, ‘age’, ‘gender’, ‘race’ and ‘education’. In addition, I only use data after 1984, because questions in census changed frequently before 1984. Besides, there will be too many data if we use all the data since 1948.

```
# find useful variables
data<-mydata[,c(’VCF0004’, ’VCF0102’, ’VCF0104’, ’VCF0105b’, ’VCF0110’)]
data<-data[which(data$VCF0004>=1984),]
print(dim(data)[1])
```

```
## [1] 32764
```

```
print(dim(data)[2])
```

```
## [1] 5
```

```
head(data, 5)
```

```
##       VCF0004    VCF0102   VCF0104              VCF0105b
## 27181    1984 2. 25 - 34 2. Female 1. White non-Hispanic
## 27182    1984 5. 55 - 64 2. Female 1. White non-Hispanic
## 27183    1984 1. 17 - 24 2. Female 1. White non-Hispanic
## 27184    1984 6. 65 - 74 2. Female 2. Black non-Hispanic
## 27185    1984 6. 65 - 74   1. Male 1. White non-Hispanic
##                                                     VCF0110
## 27181 2. High school (12 grades or fewer, incl. non-college
## 27182         4. College or advanced degree (no cases 1948)
## 27183     3. Some college (13 grades or more but no degree;
## 27184 2. High school (12 grades or fewer, incl. non-college
## 27185     3. Some college (13 grades or more but no degree;
```

Now we have a dataframe of data we needed, we need to do some simple check about whether there are problems inside data. In this step, I check about missing data and repeated data.

```
# check data status
MissingValue<-c()
UniqueValue<-c()
MostFreqValue<-c()
MostFreqValueRate<-c()
for (i in colnames(data)){
  MissingValue<-c(MissingValue, sum(is.na(data[,i])))
  UniqueValue<-c(UniqueValue, round(length(unique(data[,i])),2))
  temp<-sort(table(data[,i], useNA = "ifany"), decreasing = TRUE)
  MostFreqValue<-c(MostFreqValue, names(temp)[1])
  MostFreqValueRate<-c(MostFreqValueRate, round(temp[1]/nrow(data),2))
}
df<-data.frame(MissingValue, UniqueValue, MostFreqValue, MostFreqValueRate)
rownames(df)<-c('year', 'age', 'gender', 'race', 'education')
df
```

```
##           MissingValue UniqueValue
## year                 0          14
## age                290           8
## gender              41           4
## race               242           5
## education          360           5
##                                                   MostFreqValue
## year                                                       2012
## age                                                  3. 35 - 44
## gender                                                2. Female
## race                                      1. White non-Hispanic
## education 2. High school (12 grades or fewer, incl. non-college
##           MostFreqValueRate
## year                   0.18
## age                    0.20
## gender                 0.54
## race                   0.71
## education              0.39
```

The data set has few missing value, and there are not many repeated value inside. Since there are not many missing value, I simply delete all the missing value.

```
# delete na data
newdata<-na.omit(data)
colnames(newdata)<-c('year', 'age', 'gender', 'race', 'education')
print(dim(newdata)[1])
```

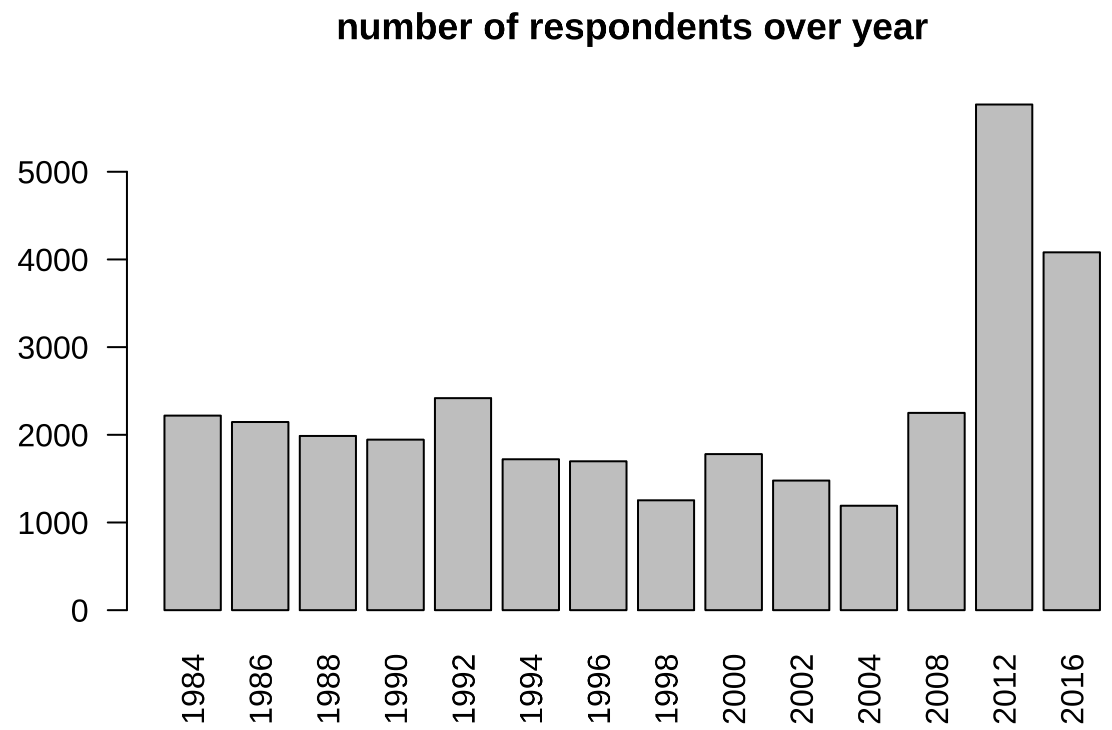
```
## [1] 31934
```

```
print(dim(newdata)[2])
```

```
## [1] 5
```

Data visualazation

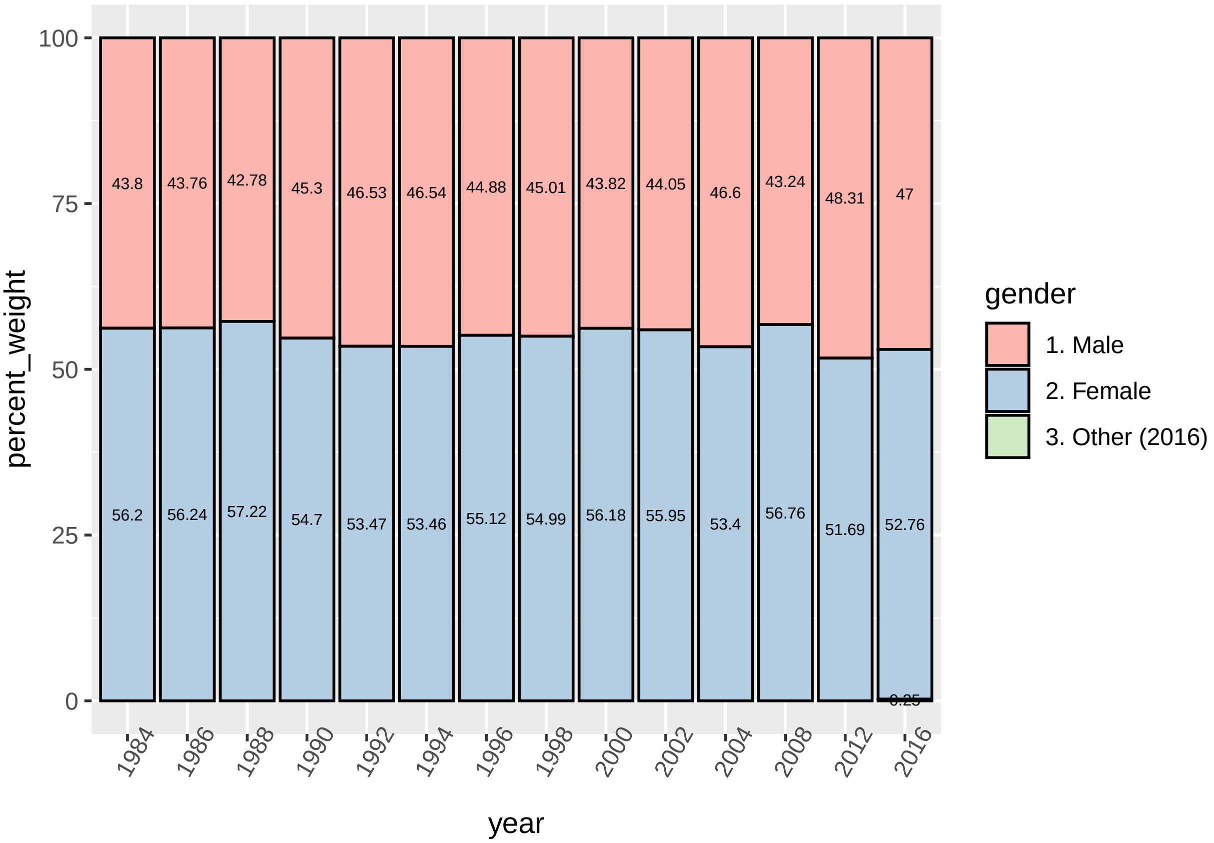
```
# number of respondents over year
barplot(table(newdata$year), las=2, main='number of respondents over year')
```



From above graph we can see that number of respondent is different over year. In order to eliminated the effect of number of people, I used percentage histogram to visualize each variables.

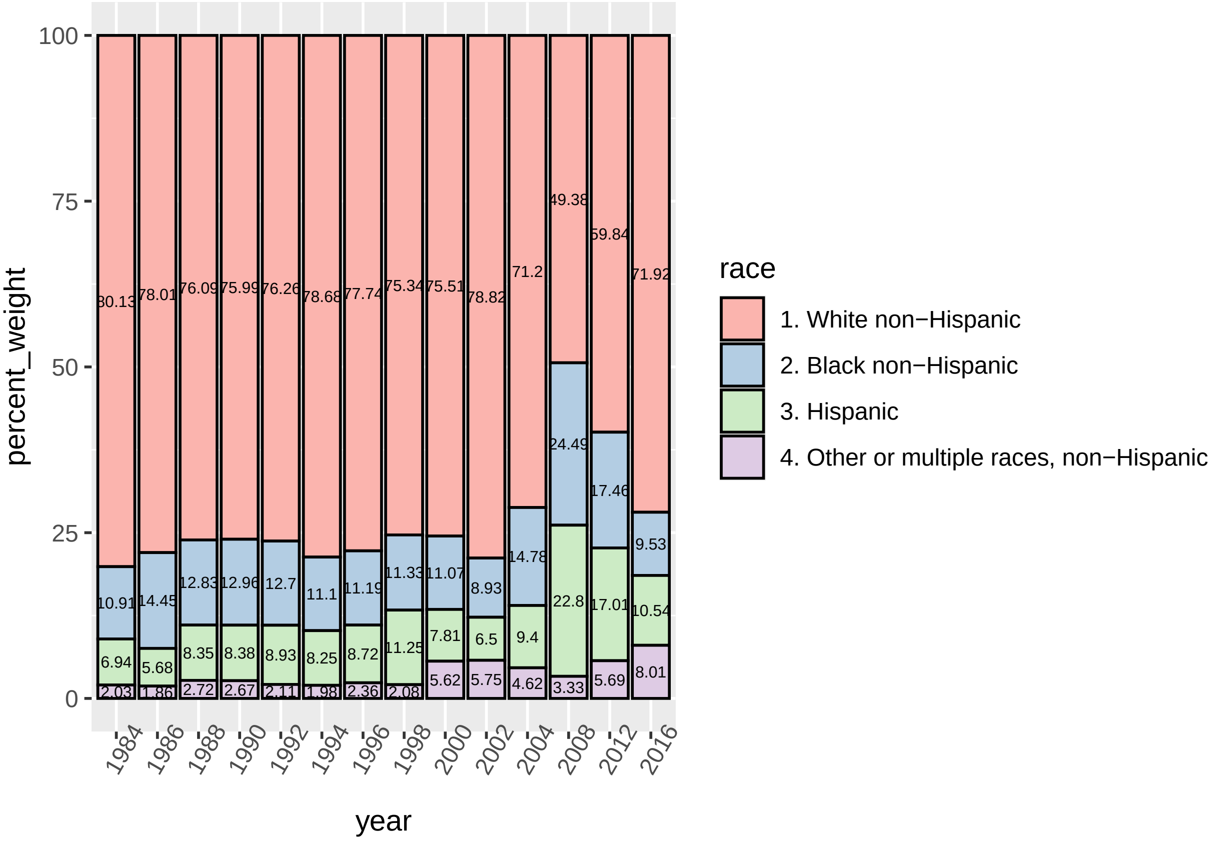
```
library(dplyr)
library(plyr)
library(ggplot2)
```

```r
# gender
a<-newdata %>%
  group_by(year, gender) %>%
  dplyr::summarize(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
ggplot(ce, aes(x = year, y = percent_weight, fill = gender)) +
  geom_bar(stat = "identity", colour = "black") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_text(aes(label = round(percent_weight,2), y = percent_weight), size = 2,
            position = position_stack(vjust = 0.5))+
  theme(axis.text.x = element_text(angle = 60))
```



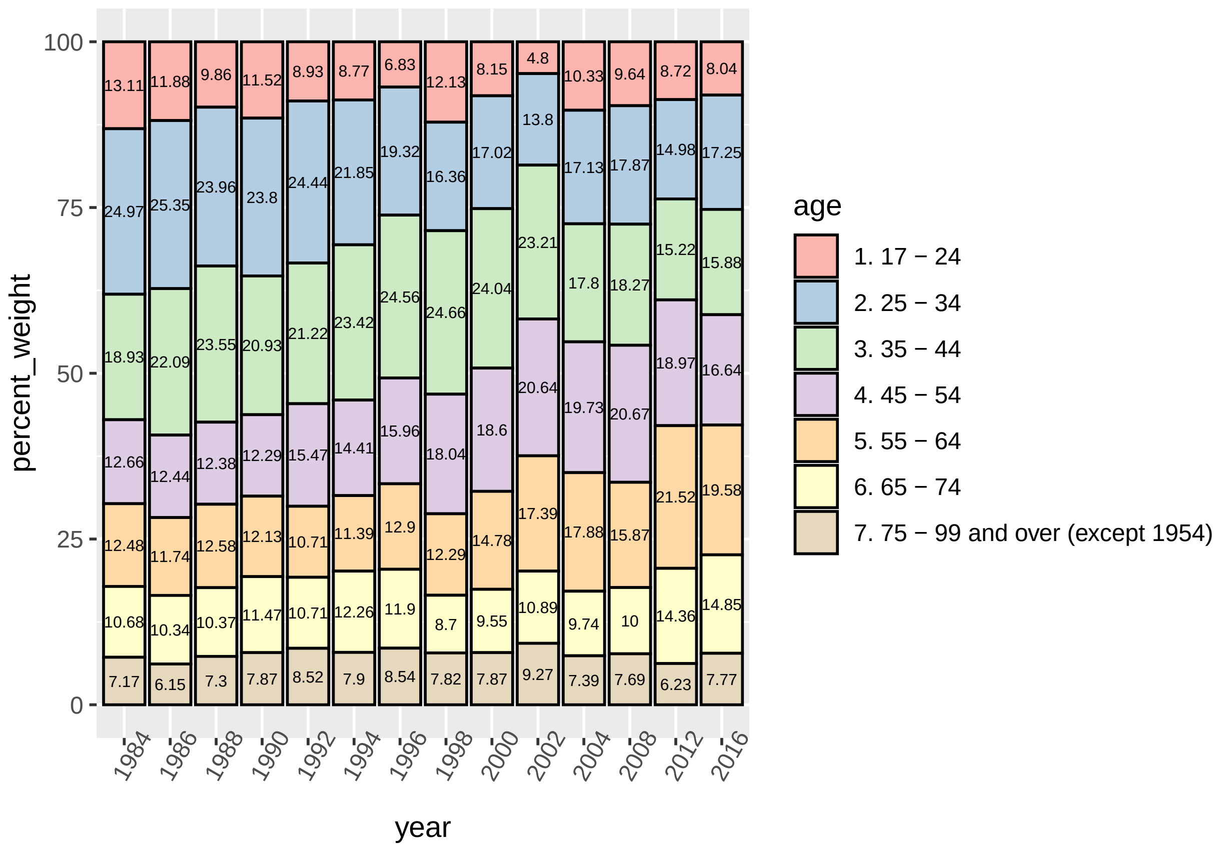
The proportion of male and female is stable over year. Female is slightly fewer then male. I checked the total population in the United States by gender and found that the number of female is slightly more than male since 1960. This maybe imply that male are more willing to participated in vote.

```r
# race
a<-newdata %>%
  group_by(year, race) %>%
  dplyr::summarise(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
ggplot(ce, aes(x = year, y = percent_weight, fill = race)) +
  geom_bar(stat = "identity", colour = "black") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_text(aes(label = round(percent_weight,2), y = percent_weight), size = 2,
            position = position_stack(vjust = 0.5))+
  theme(axis.text.x = element_text(angle = 60))
```



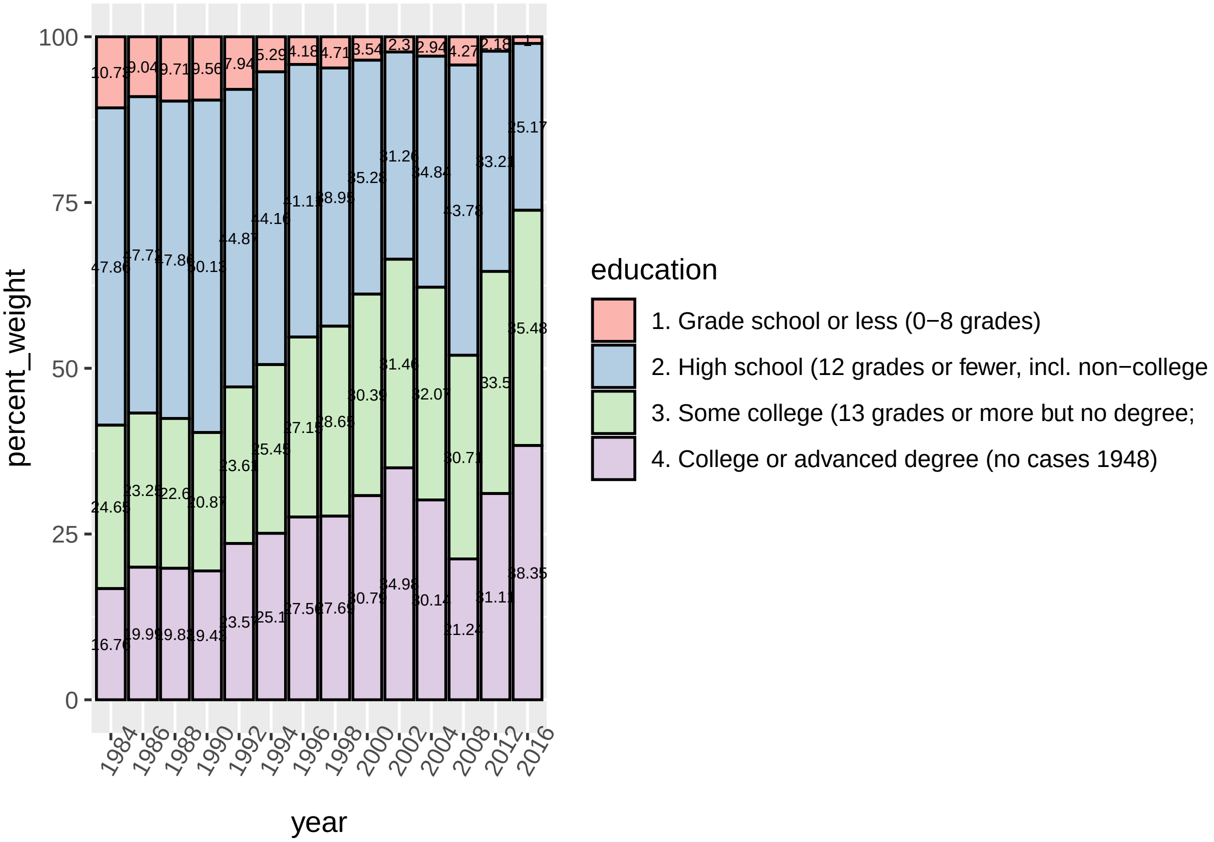
The porpotion of different race of people participated in vote is stable until 2000. Maybe this is because survey sample change.

```
# age
a<-newdata %>%
  group_by(year, age) %>%
  dplyr::summarise(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
ggplot(ce, aes(x = year, y = percent_weight, fill = age)) +
  geom_bar(stat = "identity", colour = "black") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_text(aes(label = round(percent_weight,2), y = percent_weight), size = 2,
            position = position_stack(vjust = 0.5))+
  theme(axis.text.x = element_text(angle = 60))
```



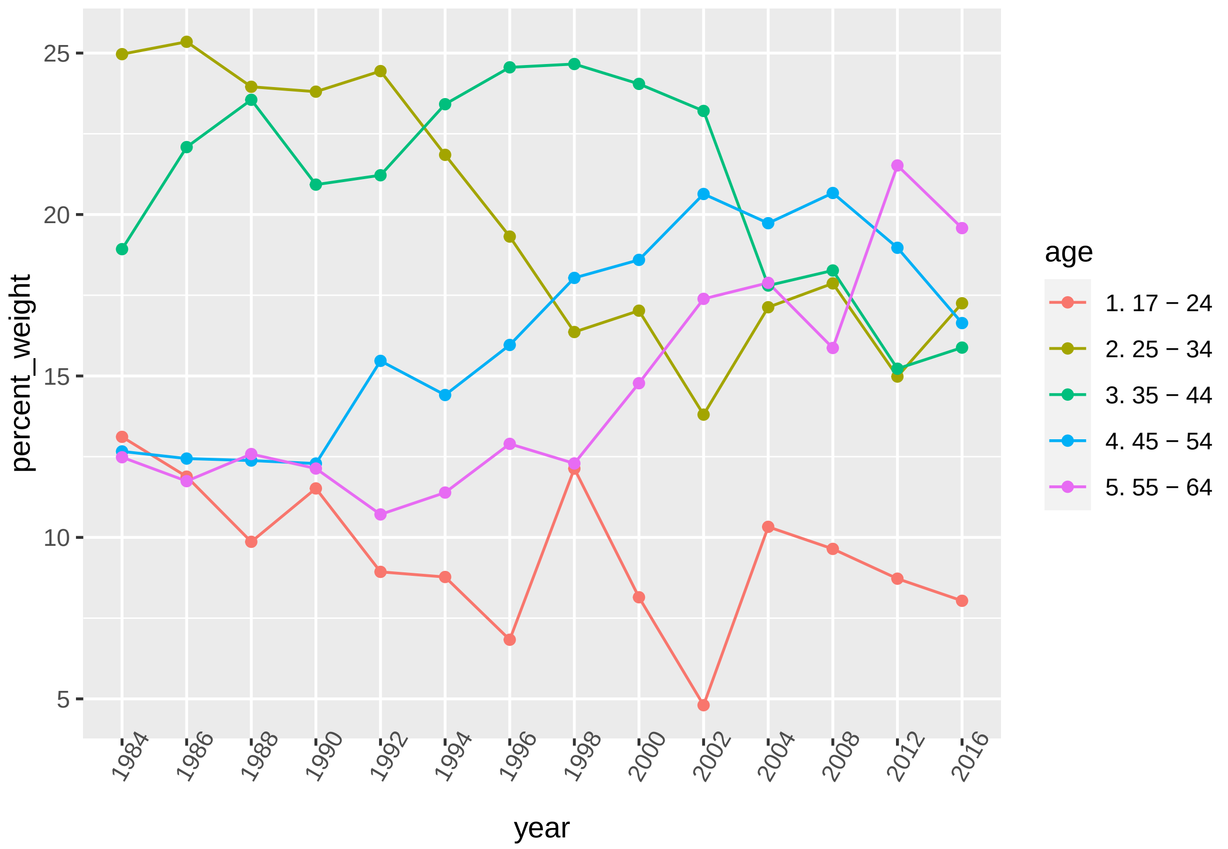
In recent year, people in group 4 and 5 are major voter. However, the proportion of different age of people participated in vote is not stable. From the graph we can see number of people in age group 1 going down first then going up and then going down. People in age group 2 keep going down. People in age group 3 and 4 going up then going down. people in age group 5 keep going up. In order to visualize the pattern, I will produce a line chart later.

```r
# education
a<-newdata %>%
  group_by(year, education) %>%
  dplyr::summarise(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
ggplot(ce, aes(x = year, y = percent_weight, fill = education)) +
  geom_bar(stat = "identity", colour = "black") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_text(aes(label = round(percent_weight,2), y = percent_weight), size = 2,
            position = position_stack(vjust = 0.5))+
  theme(axis.text.x = element_text(angle = 60))
```



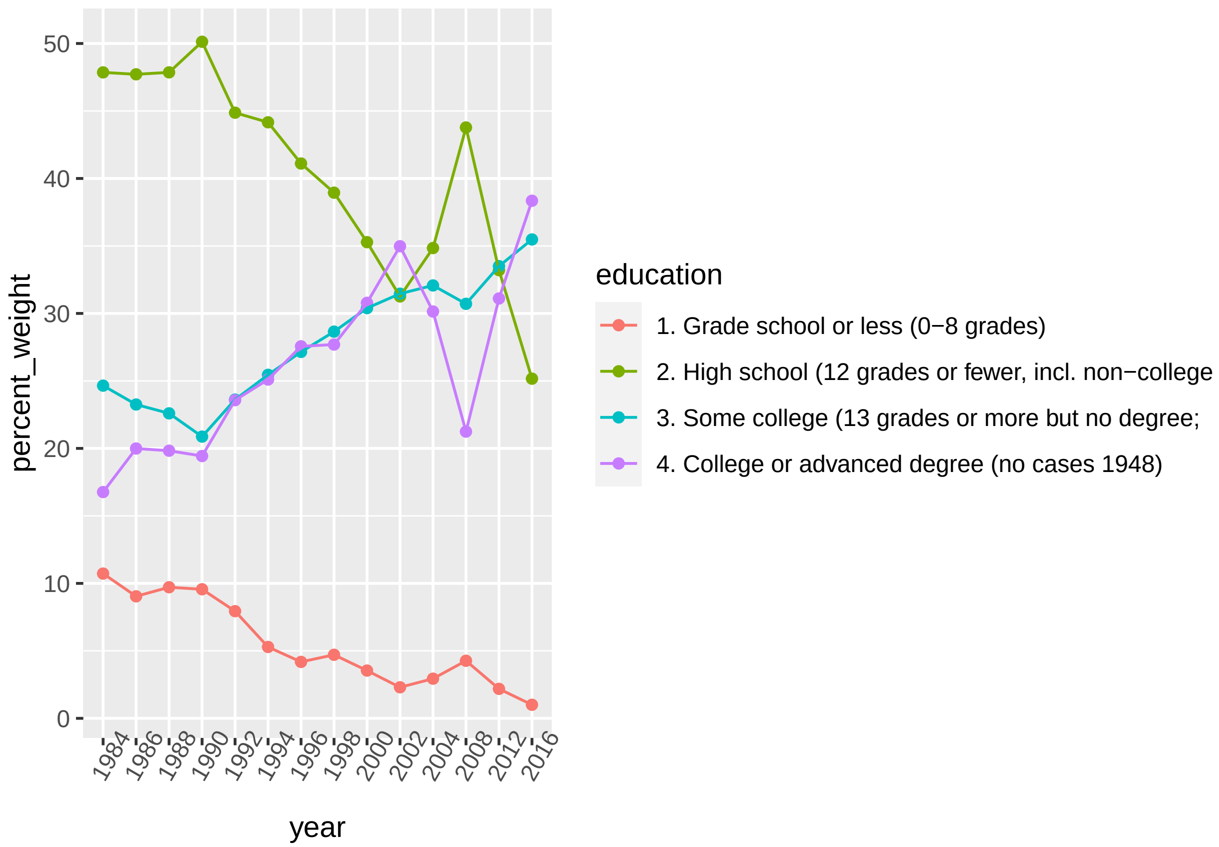
In rencent year, people of group 3 and 4 are major voter. However, the proportion of different education level of people participated in vote is not stable. From the graph we can see number of people in group 1 and 2 keep going down except for 2008. People in group 3 and 4 keep going up except for 2008. I think 2008 year should be exclude from analysis.

```
# line chart for age
a<-newdata %>%
  group_by(year, age) %>%
  dplyr::summarise(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
temp<-ce[which(ce$age != ’6. 65 - 74’),]
temp<-temp[which(temp$age != ’7. 75 - 99 and over (except 1954)’),]
ggplot(temp, aes(x=year, y = percent_weight, colour = age, group = age)) +
  geom_line()+
  geom_point()+
  theme(axis.text.x = element_text(angle = 60))
```



From the graph we can see that group 2 number of people participated in vote peak at approximately year 1984. Group 3 peak at approximately 1996. Group 4 peak at approximately 2004, group 5 at 2012. The peak point of each groups are approximately 10 year apart, which is exactly the age difference between each groups. Maybe only the same group of people are willing to vote, young people are not very willing to participate. Age of people participate in vote tend to become elder and elder. There are many papers and articles on the internet have discuss about the power of senior voter.

```
# line chart for education
a<-newdata %>%
  group_by(year, education) %>%
  dplyr::summarise(n=n())
ce<-ddply(a, "year", transform, percent_weight = n / sum(n) * 100)
ggplot(ce, aes(x = year, y = percent_weight, colour = education, group=education)) +
  geom_point()+
  geom_line()+
  theme(axis.text.x = element_text(angle = 60))
```



From line chart above we can clearly see that number of people in group 1 and 2 keep decreasing, number of people from group 3 and 4 keep increasing. More and more people with higher level of education participate in vote. This maybe because elder people are more willing to vote.

Summary

White non Hispanic and male people are main voter. Age of people participate in vote tend to become elder and elder. Educational level of people participate in vote tend to become higher and higher.