

General profile of people voted

Introduction

This is a data story about the general profile of people who participated in the vote. I'm curious about things like gender, education level, age of people who participated in vote. Will the profile of participated people change over year? Or will it remain stable?

```
# 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
```

```
head(mydata, 5)
```

```
##              Version VCF0004 VCF0006 VCF0006a VCF0009x VCF0010x
## 1 ANES_CDF_VERSION:2019-Sep-10    1948    1001 19481001          1          1
## 2 ANES_CDF_VERSION:2019-Sep-10    1948    1002 19481002          1          1
## 3 ANES_CDF_VERSION:2019-Sep-10    1948    1003 19481003          1          1
## 4 ANES_CDF_VERSION:2019-Sep-10    1948    1004 19481004          1          1
## 5 ANES_CDF_VERSION:2019-Sep-10    1948    1005 19481005          1          1
##   VCF0011x VCF0009y VCF0010y VCF0011y VCF0009z VCF0010z VCF0011z VCF0012
## 1         1         1         1         1         1         1         1    <NA>
## 2         1         1         1         1         1         1         1    <NA>
## 3         1         1         1         1         1         1         1    <NA>
## 4         1         1         1         1         1         1         1    <NA>
## 5         1         1         1         1         1         1         1    <NA>
##              VCF0013
## 1 1. Post-election interview data present
## 2 1. Post-election interview data present
## 3 1. Post-election interview data present
## 4 1. Post-election interview data present
## 5 1. Post-election interview data present
##              VCF0014
## 1 1. Pre-election interview data present
## 2 1. Pre-election interview data present
```

```

## 3 1. Pre-election interview data present
## 4 1. Pre-election interview data present
## 5 1. Pre-election interview data present
##
##                                VCF0015a
## 1 0. Pre IW not abbreviated [1992:'Long' form Pre]
## 2 0. Pre IW not abbreviated [1992:'Long' form Pre]
## 3 0. Pre IW not abbreviated [1992:'Long' form Pre]
## 4 0. Pre IW not abbreviated [1992:'Long' form Pre]
## 5 0. Pre IW not abbreviated [1992:'Long' form Pre]
##
##                                VCF0015b                                VCF0016                                VCF0017
## 1 0. Post IW is not abbreviated 0. Fresh Cross case 0. All personal
## 2 0. Post IW is not abbreviated 0. Fresh Cross case 0. All personal
## 3 0. Post IW is not abbreviated 0. Fresh Cross case 0. All personal
## 4 0. Post IW is not abbreviated 0. Fresh Cross case 0. All personal
## 5 0. Post IW is not abbreviated 0. Fresh Cross case 0. All personal
##
##                                VCF0018a
## 1 0. IW conducted entirely in English; 2008,2012: beginning language
## 2 0. IW conducted entirely in English; 2008,2012: beginning language
## 3 0. IW conducted entirely in English; 2008,2012: beginning language
## 4 0. IW conducted entirely in English; 2008,2012: beginning language
## 5 0. IW conducted entirely in English; 2008,2012: beginning language
##
##                                VCF0018b VCF0019
## 1 0. IW conducted entirely in English; 2008,2012: beginning language <NA>
## 2 0. IW conducted entirely in English; 2008,2012: beginning language <NA>
## 3 0. IW conducted entirely in English; 2008,2012: beginning language <NA>
## 4 0. IW conducted entirely in English; 2008,2012: beginning language <NA>
## 5 0. IW conducted entirely in English; 2008,2012: beginning language <NA>
##
## VCF0050a VCF0050b VCF0070a VCF0070b VCF0071a VCF0071b VCF0071c VCF0071d
## 1 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 2 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 3 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 4 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 5 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##
## VCF0072a VCF0072b VCF0101 VCF0102 VCF0103 VCF0104
## 1 <NA> <NA> <NA> 3. 35 - 44 7. 1895 - 1910 1. Male
## 2 <NA> <NA> <NA> 3. 35 - 44 7. 1895 - 1910 2. Female
## 3 <NA> <NA> <NA> 2. 25 - 34 6. 1911 - 1926 2. Female
## 4 <NA> <NA> <NA> 3. 35 - 44 7. 1895 - 1910 2. Female
## 5 <NA> <NA> <NA> 2. 25 - 34 6. 1911 - 1926 1. Male
##
##                                VCF0105a                                VCF0105b                                VCF0106
## 1 1. White non-Hispanic (1948-2012) 1. White non-Hispanic 1. White non-Hispanic
## 2 1. White non-Hispanic (1948-2012) 1. White non-Hispanic 1. White non-Hispanic
## 3 1. White non-Hispanic (1948-2012) 1. White non-Hispanic 1. White non-Hispanic
## 4 1. White non-Hispanic (1948-2012) 1. White non-Hispanic 1. White non-Hispanic
## 5 1. White non-Hispanic (1948-2012) 1. White non-Hispanic 1. White non-Hispanic
##
## VCF0107 VCF0108 VCF0109                                VCF0110
## 1 <NA> <NA> <NA>                                1. Grade school or less (0-8 grades)
## 2 <NA> <NA> <NA>                                2. High school (12 grades or fewer, incl. non-college
## 3 <NA> <NA> <NA>                                2. High school (12 grades or fewer, incl. non-college
## 4 <NA> <NA> <NA>                                3. Some college (13 grades or more but no degree;
## 5 <NA> <NA> <NA>                                3. Some college (13 grades or more but no degree;
##
## VCF0111 VCF0112 VCF0113                                VCF0114 VCF0115 VCF0116
## 1 <NA> <NA> <NA>                                3. 34 to 67 percentile <NA> <NA>
## 2 <NA> <NA> <NA>                                5. 96 to 100 percentile <NA> <NA>

```

[illegible]

[illegible]

5

[illegible]

[illegible]

## 3	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	2. Yes, voted
## 4	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	2. Yes, voted
## 5	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	2. Yes, voted
##	VCF0703	VCF0704	VCF0704a	VCF0705	VCF0706	VCF0707			
## 1	<NA>	1. Democrat	1. Democrat	1. Democrat	1. Democrat	<NA>			
## 2	<NA>	2. Republican	2. Republican	2. Republican	2. Republican	<NA>			
## 3	<NA>	1. Democrat	1. Democrat	1. Democrat	1. Democrat	<NA>			
## 4	<NA>	2. Republican	2. Republican	2. Republican	2. Republican	<NA>			
## 5	<NA>	1. Democrat	1. Democrat	1. Democrat	1. Democrat	<NA>			
##	VCF0708	VCF0709	VCF0710	VCF0711					
## 1	<NA>	<NA>	<NA>	<NA>					
## 2	<NA>	<NA>	<NA>	<NA>					
## 3	<NA>	<NA>	<NA>	<NA>					
## 4	<NA>	<NA>	<NA>	<NA>					
## 5	<NA>	<NA>	<NA>	<NA>					
##				VCF0712					
## 1	1.	Knew all along (incl.: always vote for same party;							
## 2	1.	Knew all along (incl.: always vote for same party;							
## 3	1.	Knew all along (incl.: always vote for same party;							
## 4		3. During conventions							
## 5		3. During conventions							
##				VCF0713	VCF0714	VCF0715			
## 1	1.	Democratic candidate (with or without qualifications,				<NA>	<NA>		
## 2	2.	Republican candidate (with or without qualifications,				<NA>	<NA>		
## 3		3. Undecided; DK (except 1964)				<NA>	<NA>		
## 4	2.	Republican candidate (with or without qualifications,				<NA>	<NA>		
## 5	1.	Democratic candidate (with or without qualifications,				<NA>	<NA>		
##		VCF0716	VCF0717	VCF0718	VCF0719	VCF0720	VCF0721	VCF0722	VCF0723
## 1	2.	Straight ticket		<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 2	2.	Straight ticket		<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 3	2.	Straight ticket		<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 4	2.	Straight ticket		<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 5	1.	Split-ticket		<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
##	VCF0723a	VCF0724	VCF0725	VCF0726	VCF0727	VCF0728	VCF0729	VCF0730	VCF0731
## 1	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 2	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 3	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 4	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 5	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
##	VCF0732	VCF0733			VCF0734	VCF0735	VCF0736		
## 1	<NA>	<NA>	1. INTENDED Democratic: voted Democratic				<NA>	<NA>	
## 2	<NA>	<NA>	9. INTENDED Republican: voted Republican				<NA>	<NA>	
## 3	<NA>	<NA>	2. INTENDED undecided: voted Democratic;				<NA>	<NA>	
## 4	<NA>	<NA>	9. INTENDED Republican: voted Republican				<NA>	<NA>	
## 5	<NA>	<NA>	1. INTENDED Democratic: voted Democratic				<NA>	<NA>	
##			VCF0737	VCF0738	VCF0738a	VCF0739	VCF0740		
## 1	2.	Yes (includes Rs who reported voting)		<NA>	<NA>	<NA>	<NA>		
## 2	2.	Yes (includes Rs who reported voting)		<NA>	<NA>	<NA>	<NA>		
## 3	2.	Yes (includes Rs who reported voting)		<NA>	<NA>	<NA>	<NA>		
## 4	2.	Yes (includes Rs who reported voting)		<NA>	<NA>	<NA>	<NA>		
## 5	2.	Yes (includes Rs who reported voting)		<NA>	<NA>	<NA>	<NA>		
##	VCF0741	VCF0742	VCF0743	VCF0744	VCF0745	VCF0746	VCF0747	VCF0748	VCF0749
## 1	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
## 2	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## 3    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 4    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 5    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##      VCF9282
## 1    <NA>
## 2    <NA>
## 3    <NA>
## 4    <NA>
## 5    <NA>
```

We need to extract some useful columns from this big dataset. 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.

```
# 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 dataset has few missing value, and there are not many repeated value in the dataset. 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
```

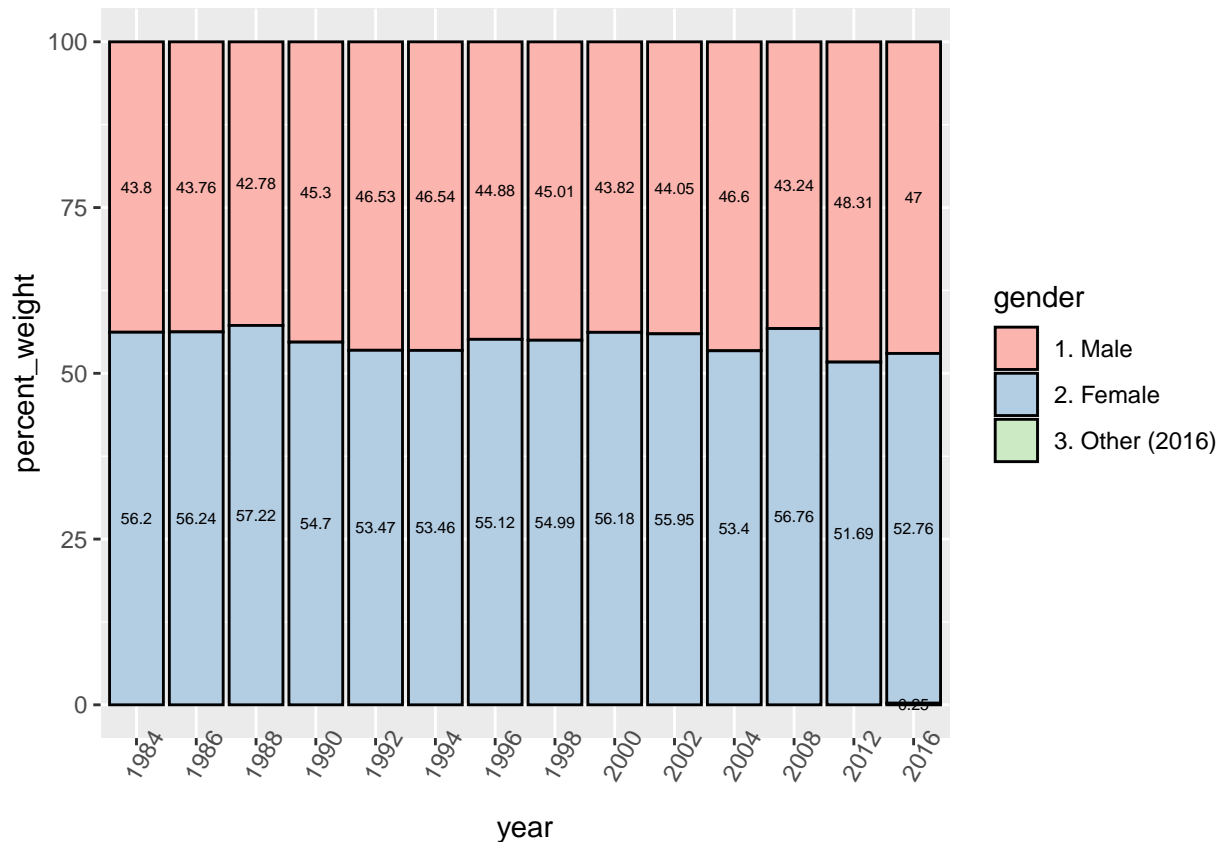
I used percentage histogram to visualize each variables.

```

library(dplyr)
library(plyr)
library(ggplot2)
# gender
a<-newdata %>%
  group_by(year, gender) %>%
  dplyr::summarize(n=n())

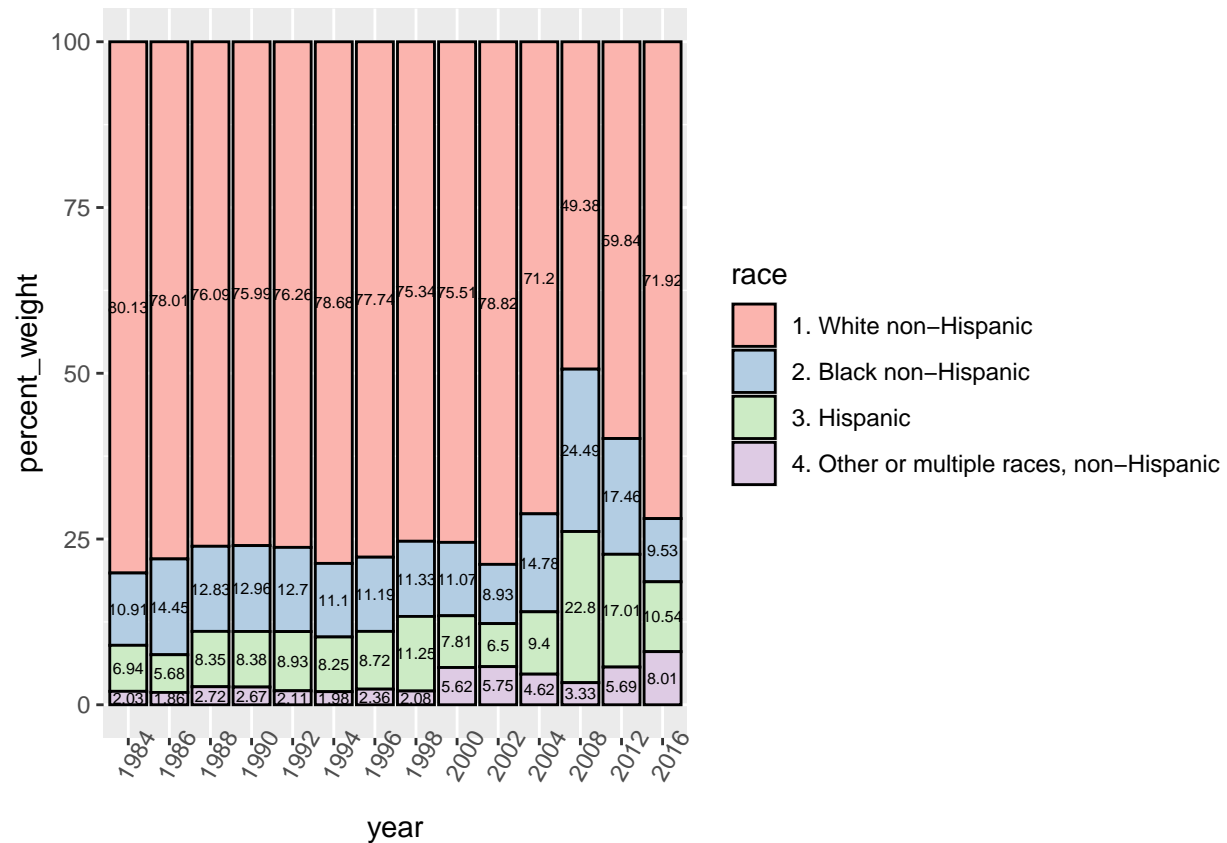
```

```
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 = "Pastell1")+
  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))
```



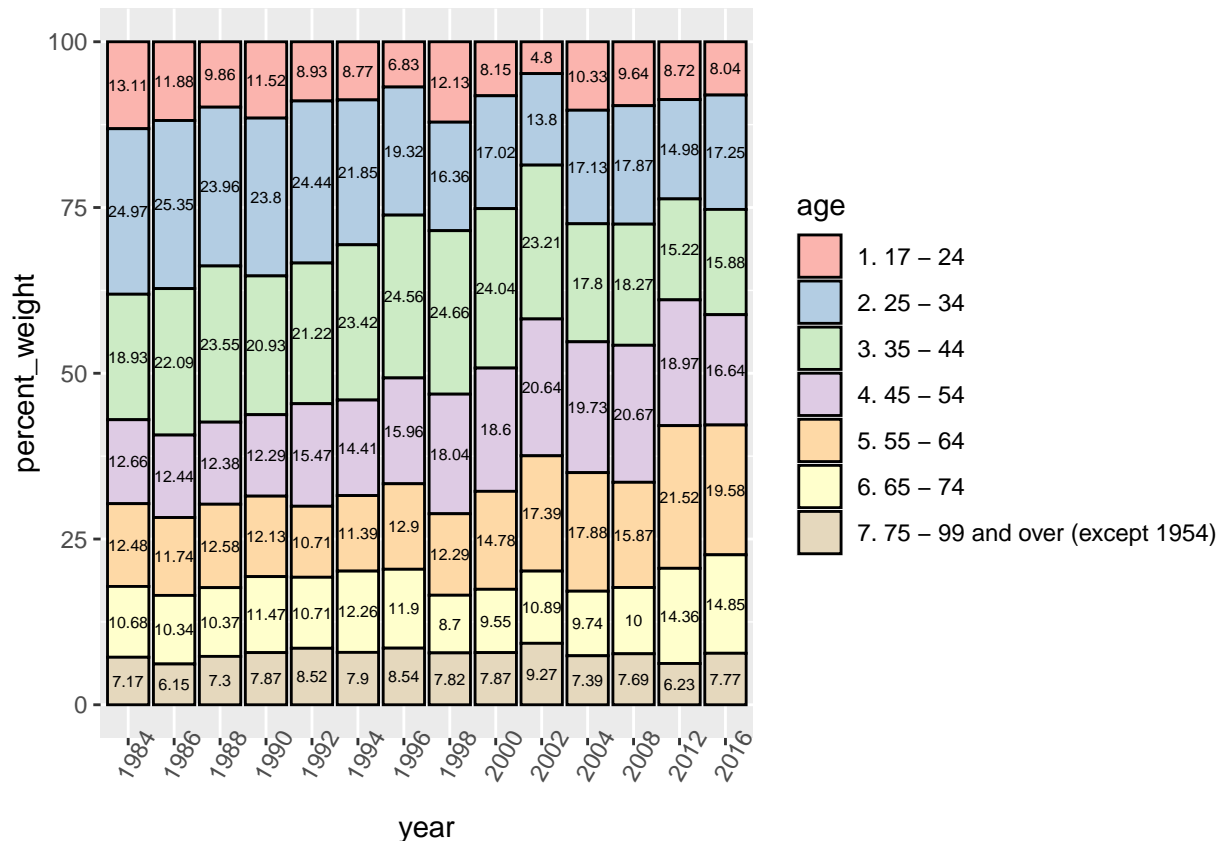
The porpotion of male and female is stable over year. Female is slightly fewer male.

```
# 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 = "Pastell1")+
  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))
```



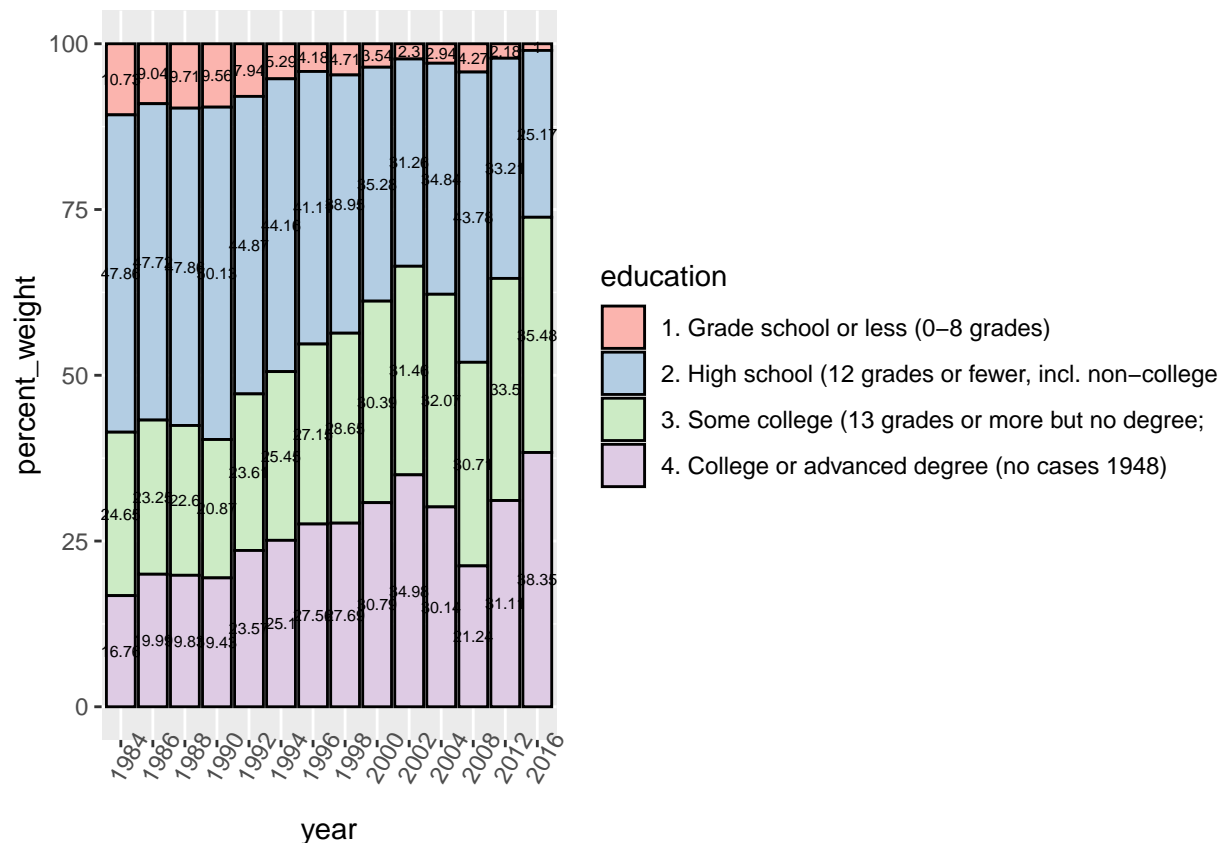
The proportion of different race of people participated in vote is stable until 2000. Maybe this is because survey sample change.

```
# 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 = "Pastell1")+
  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))
```



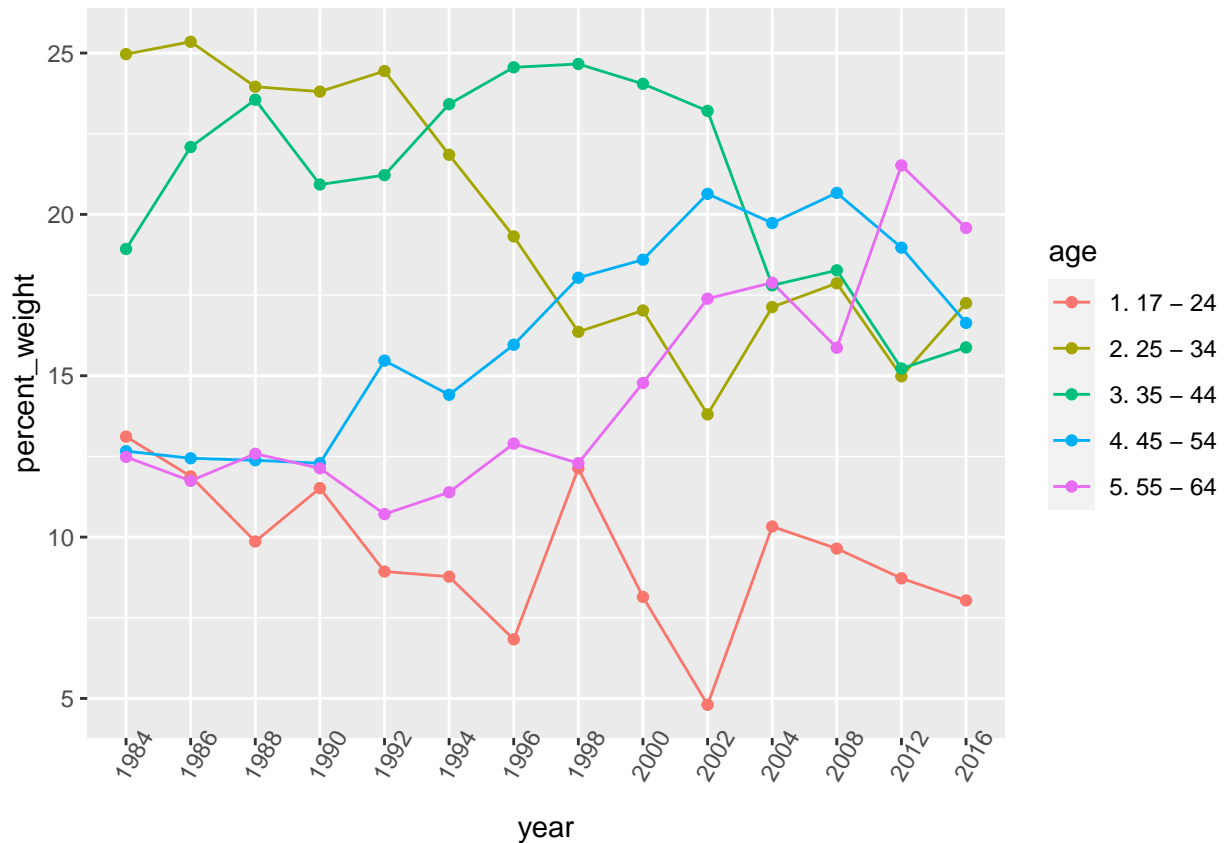
The proportion of different age of people participated in vote is not stable. From the graph we can see number of people in age group 1 going down first then going up and then going down. People in age group 2 keeping going down. People in age group 3 and 4 going up then going down. people in age group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

```
# 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))
```



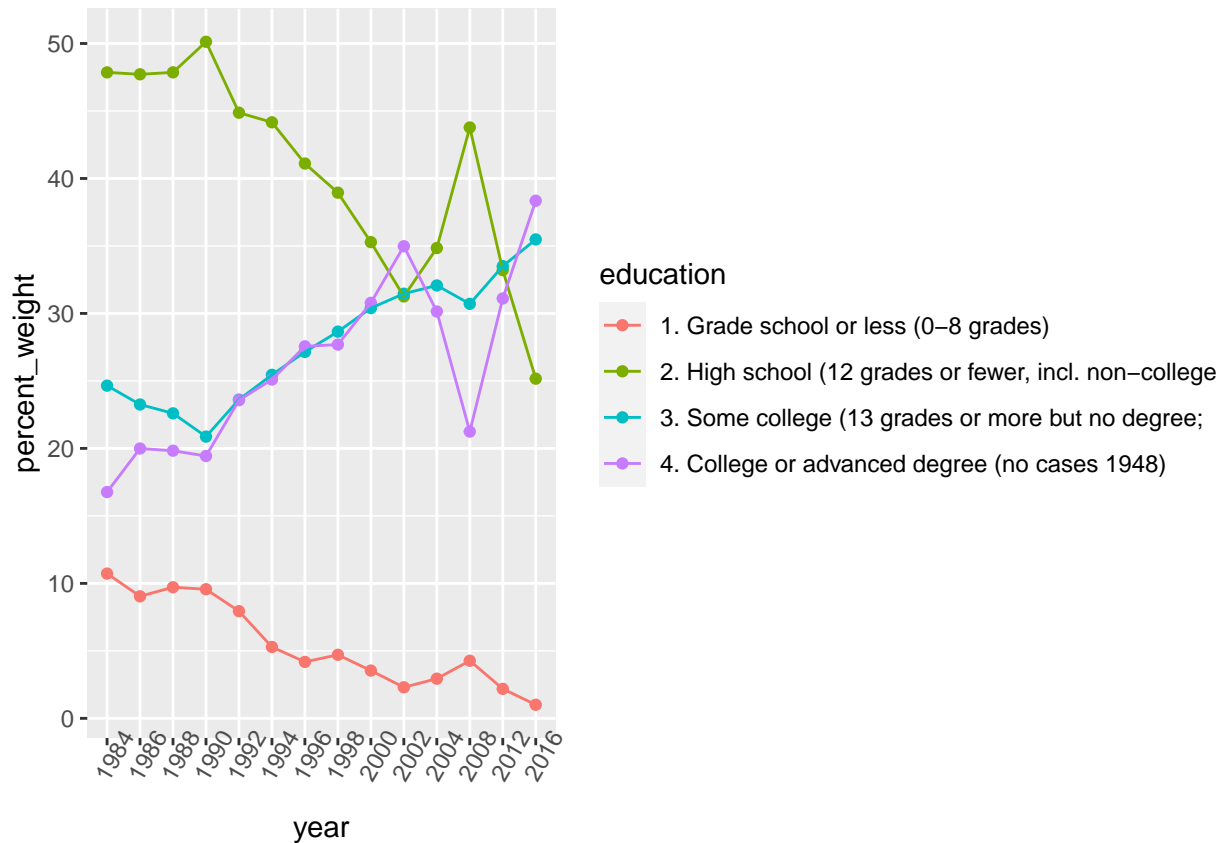
The proportion of different education level of people participated in vote is not stable. From the graph we can see number of people in group 1 and 2 keep going down except for 2008. People in group 3 and 4 keep going up except for 2008. I think 2008 year should be exclude from analysis.

```
# 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.

```
# 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.

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.