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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

Voting in America

Every year millions of voters choose not to vote for various reasons. This analysis will try to identify reasons as to why someone would choose not to exercise this right. A more in depth analysis as well as the raw data can be found on fivethirtheight.

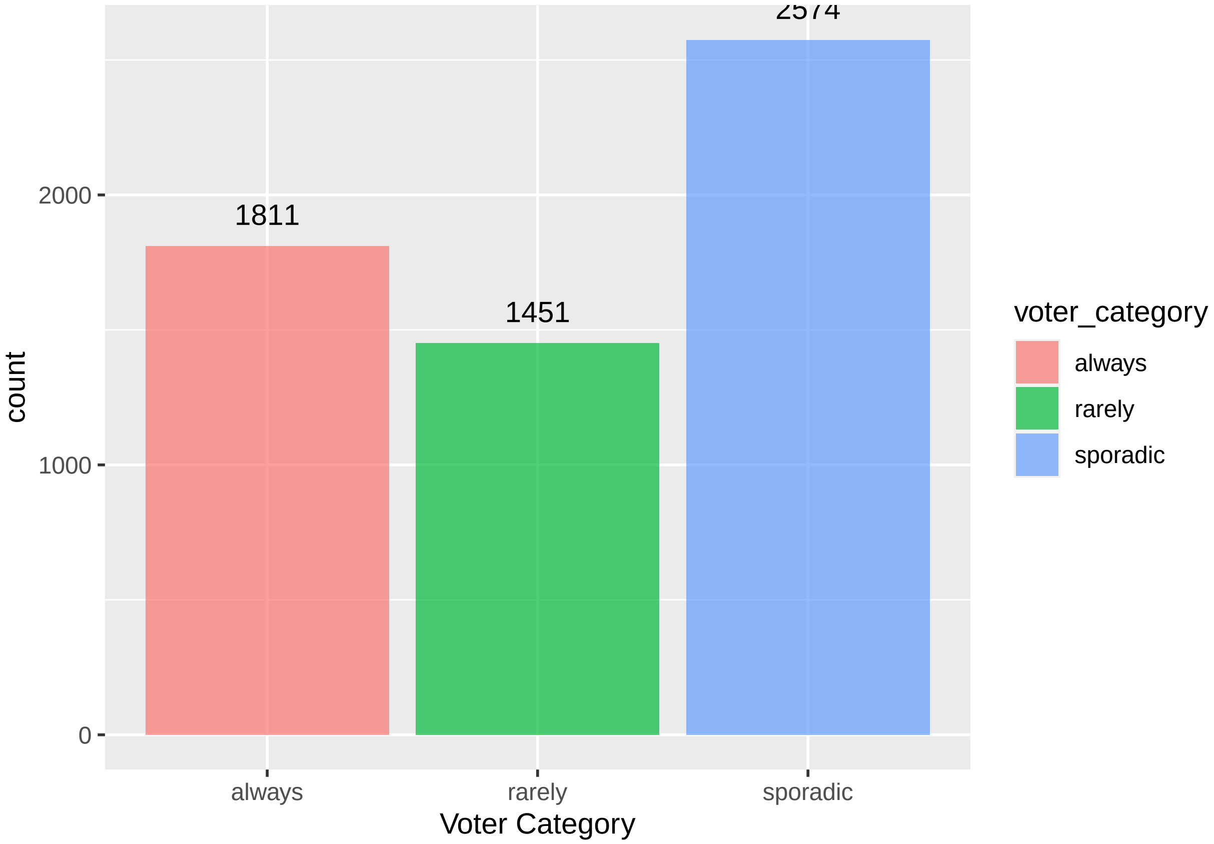
```
base_df <- read.csv(’https://raw.githubusercontent.com/ksooklall/CUNY-SPS-Masters-DS/main/DATA_607/non-v
df <- subset(base_df, select=c(’ppage’, ’educ’, ’race’, ’gender’, ’income_cat’, ’voter_category’))
df <- rename(df, age=ppage)

df <- df %>% mutate(income_cat=recode(income_cat,
                 "Less than $40k" = "low",
                  "$40-75k" = "lower_middle",
                  "$75-125k"="upper_middle",
                  "$125k or more"="high"),
                 voter_category=recode(voter_category,
                                       ’rarely/never’=’rarely’))
summary(df)
```

```
##       age                          educ                 race         gender
##  Min.   :22.00   College             :2330   Black      : 932   Female:2896
##  1st Qu.:36.00   High school or less:1796   Hispanic   : 813   Male  :2940
##  Median :54.00   Some college       :1710   Other/Mixed: 381
##  Mean   :51.69                               White      :3710
##  3rd Qu.:65.00
##  Max.   :94.00
##         income_cat    voter_category
##  high        :1394   always  :1811
##  lower_middle:1396   rarely  :1451
##  upper_middle:1628   sporadic:2574
##  low         :1418
##
##
```

The data set contains 5836 people who were polled and matched to their voting history. There are 6 columns in total. The first 5, age, education, race, gender and income_category will be analyzed against voter category

```
ggplot(df, aes(x=voter_category, fill=voter_category)) + geom_bar(alpha=0.7) +
geom_text(stat='count', aes(label=..count..), vjust=-1) + xlab('Voter Category')
```

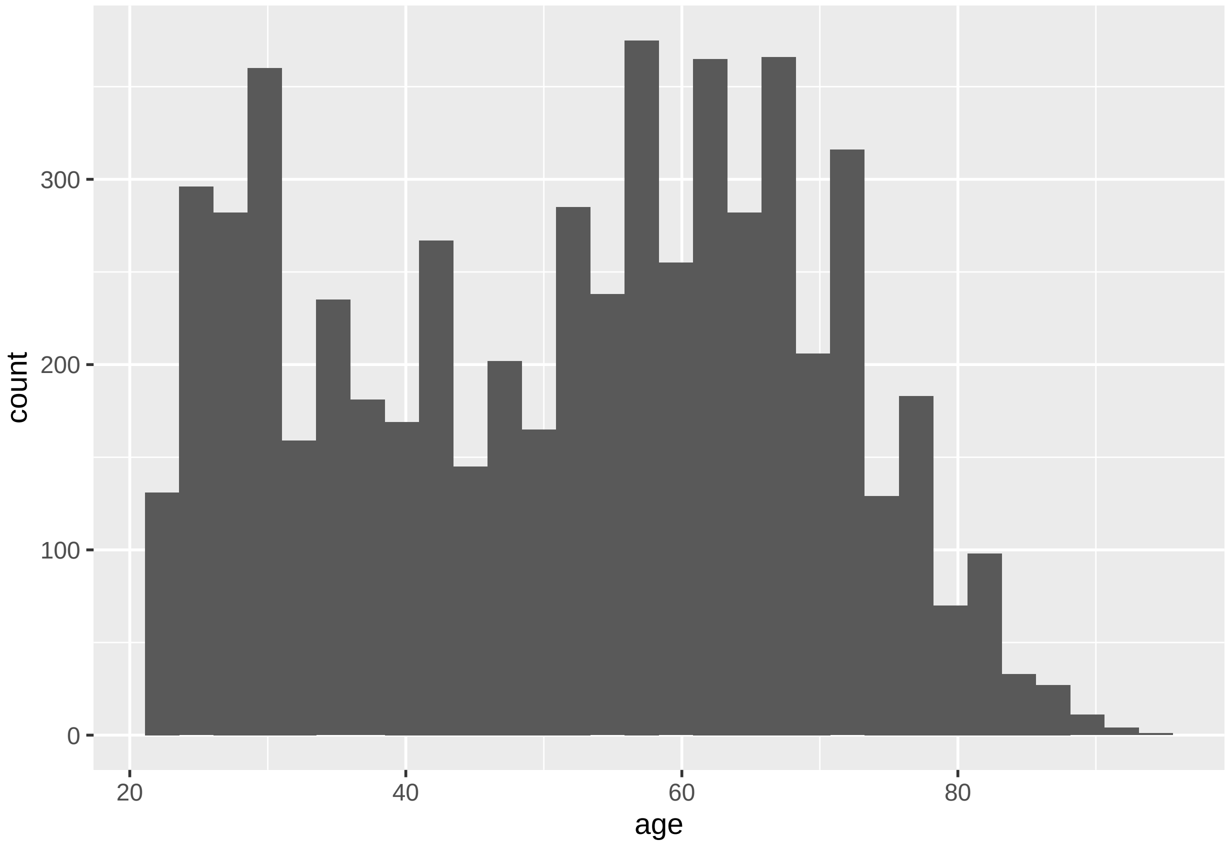


From the plot we can conclude that most voters are sporadic followed by those who always vote then those who rarely vote.

Age analysis

As we go older our interests change causing our willingness to participate in voting to change as well

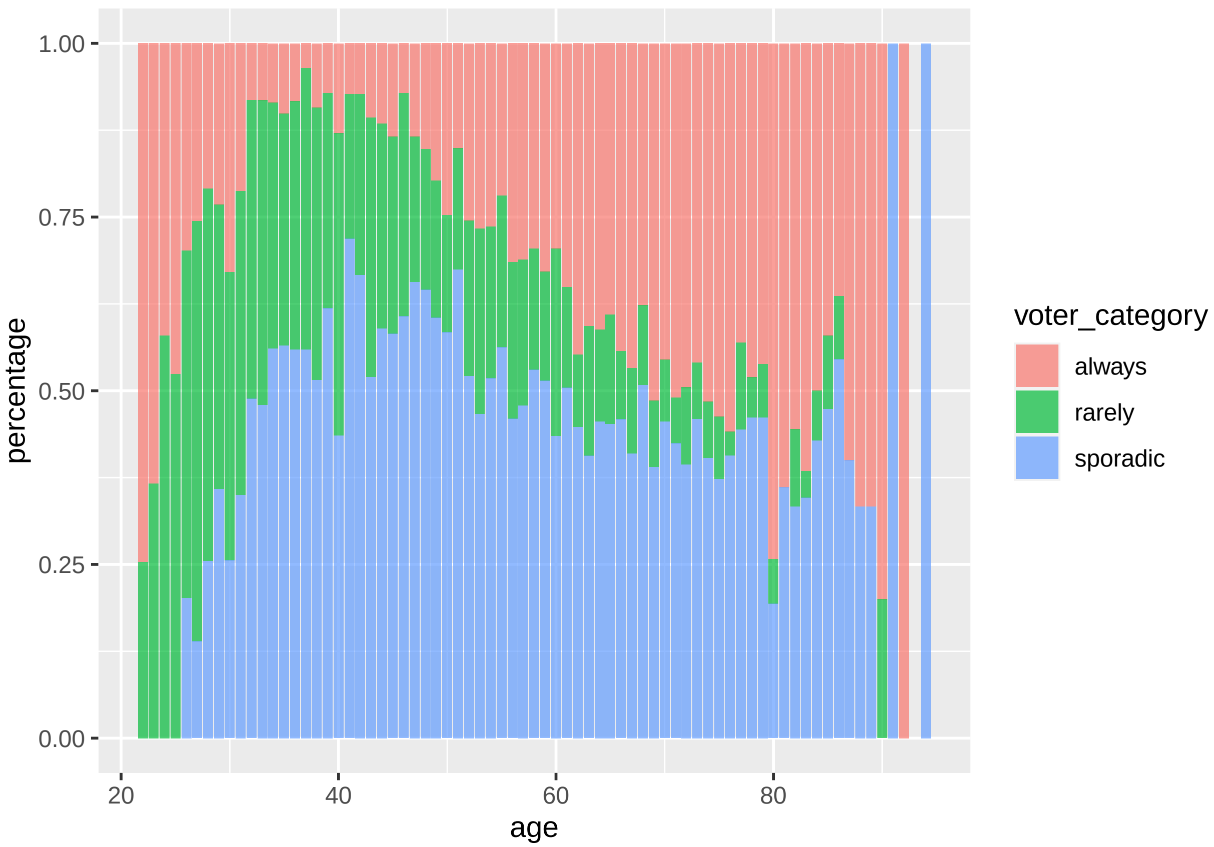
```
ggplot(df, aes(x=age)) + geom_histogram(bins=30)
```



Age is slightly right skewed as expected since the median age is 54. Two major age group dominate the data set, those who are in their late 20s and those in their late 50s and early 60s.

```
adf <- df %>% group_by(age, voter_category) %>% summarise(count=n(), .groups='drop')
adf$percentage <- (adf %>% group_by(age) %>% summarise(norm=count / sum(count), .groups='drop'))$norm

ggplot(adf, aes(x=age, y=percentage, fill=voter_category)) + geom_col(alpha=0.7)
```



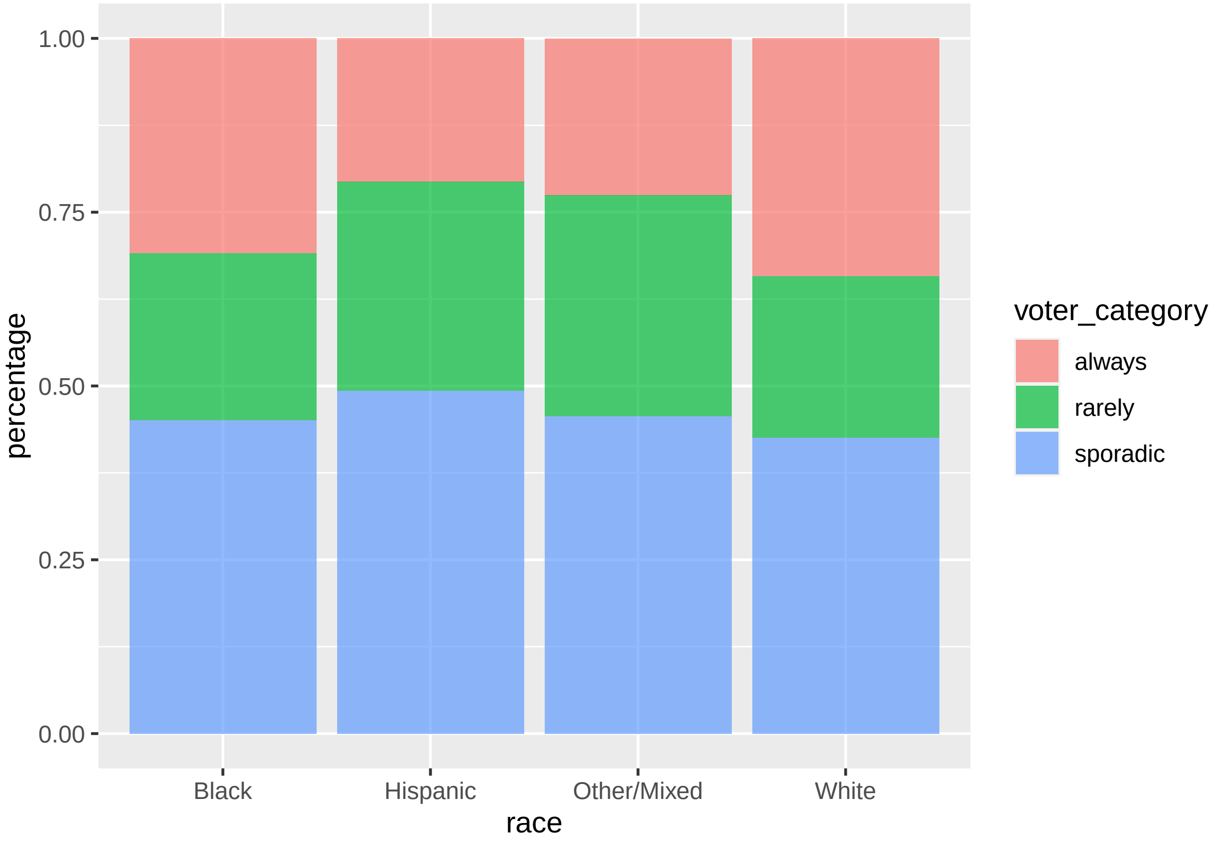
The thickness of the *always* bock grows in length from left to right, thus it seems as voters get older they exercise their right to vote more often. After 80 years old most voters are either *always* voting or *sporadically* voting.

Race analysis

In the recent years race has played a larger role in media for various reasons and it's importance in voting is just another role.

```r
rdf <- df %>% group_by(race, voter_category) %>% summarise(count=n(), .groups=’drop’)
rdf$percentage <- (rdf %>% group_by(race) %>% summarise(norm=count / sum(count), .groups=’drop’))$norm

ggplot(rdf, aes(x=race, y=percentage, fill=voter_category)) + geom_col(alpha=0.7)
```

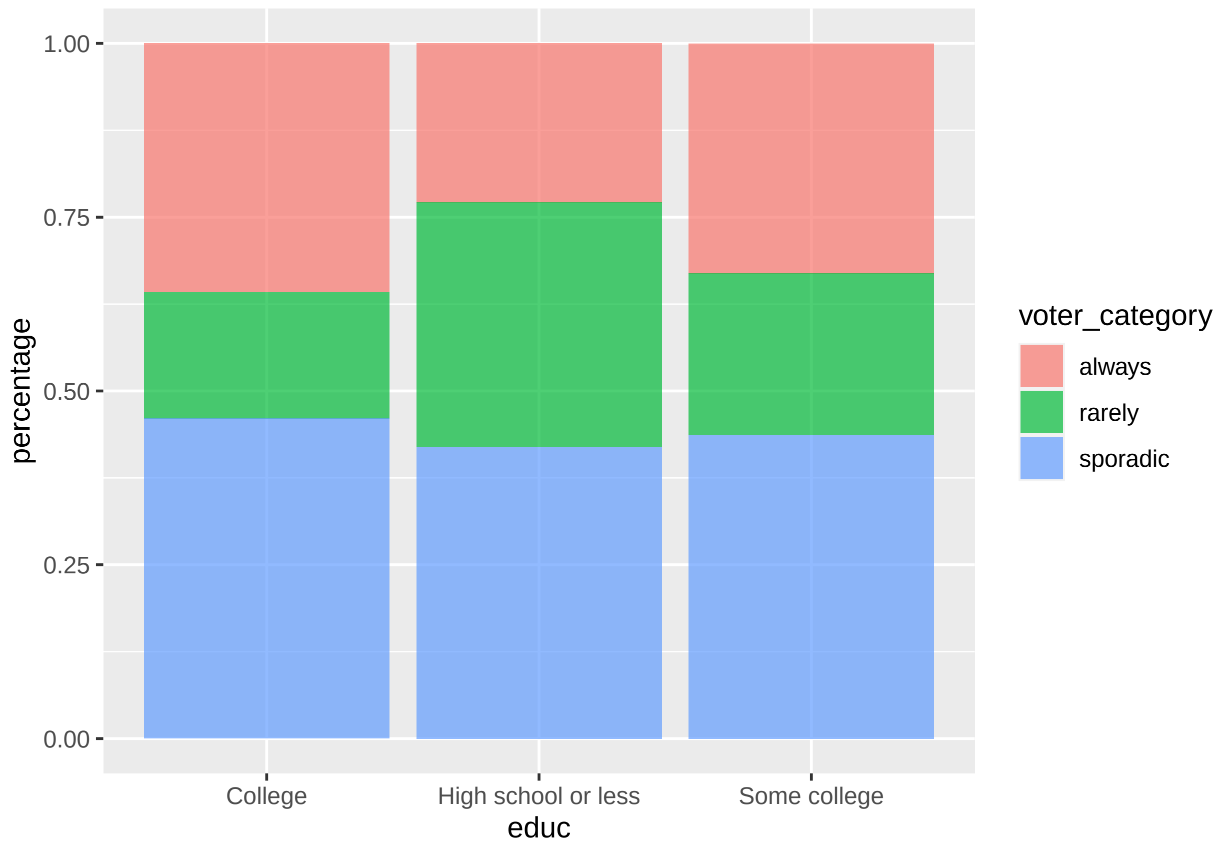


Hispanic and Other/Mixed race vote sporadically comapre to black and white voters. White voters vote the most with the smallest *rarely* block and largest *always* block.

Education analysis Education is important in almost every aspect of ones life. A proper education has helped many families move out of poverty and even up in social classes. It has been theorized the more educated an individual is the more likely they are to get involved in government issues.

```
edf <- df %>% group_by(educ, voter_category) %>% summarise(count=n(), .groups=’drop’)
edf$percentage <- (edf %>% group_by(educ) %>% summarise(norm=count / sum(count), .groups=’drop’))$norm

ggplot(edf, aes(x=educ, y=percentage, fill=voter_category)) + geom_col(alpha=0.7)
```



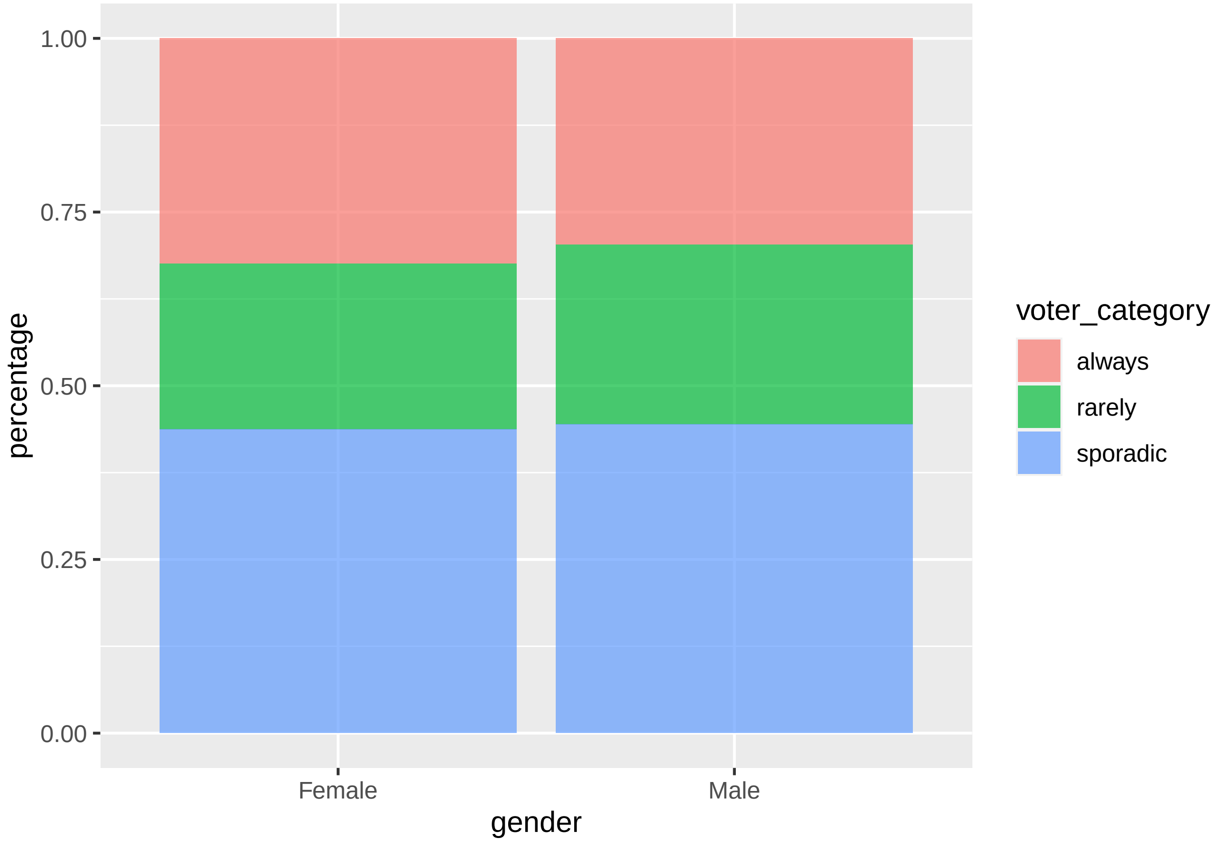
Those with high school or less education are least likely to vote while those with some college - college education are more likely to vote. Those with a college education have the largest block for always voting.

Gender analysis

Males and Females tends to have different prioritize when it comes to voting, for example a females is more concerned with equal pay than males.

```
gdf <- df %>% group_by(gender, voter_category) %>% summarise(count=n(), .groups='drop')
gdf$percentage <- (gdf %>% group_by(gender) %>% summarise(norm=count / sum(count), .groups='drop'))$norm

ggplot(gdf, aes(x=gender, y=percentage, fill=voter_category)) + geom_col(alpha=0.7)
```



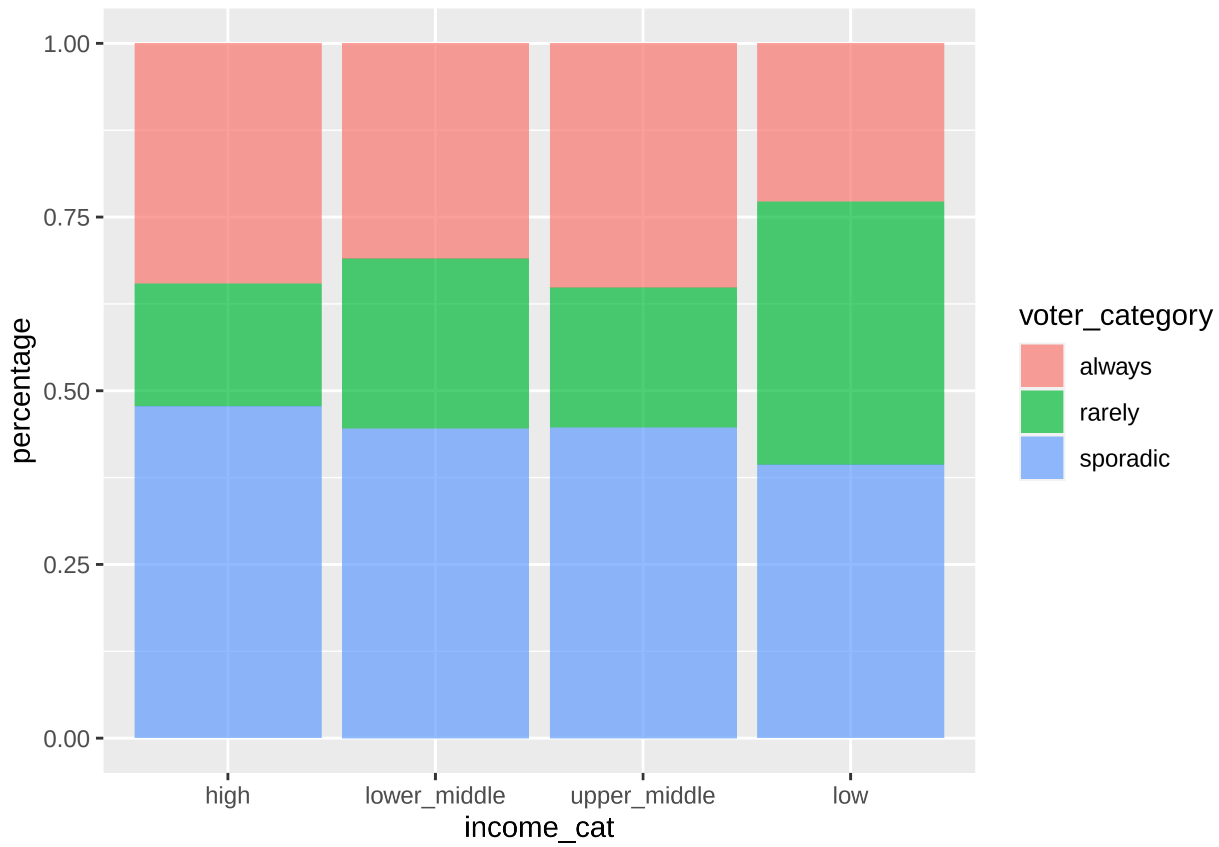
The different between males and female here is very small. Males vote sporadically more often than females, while some females will always vote.

Income analysis

Income and education have a very strong correlation as one would expect.

```
idf <- df %>% group_by(income_cat, voter_category) %>% summarise(count=n(), .groups='drop')
idf$percentage <- (idf %>% group_by(income_cat) %>% summarise(norm=count / sum(count), .groups='drop'))$

ggplot(idf, aes(x=income_cat, y=percentage, fill=voter_category)) + geom_col(alpha=0.7)
```



High and upper_middle income individuals vote the most while low incomes individuals vote the least. This result is expected based on the education analysis.

Always vs rarely

```
summary(df %>% filter(voter_category == 'always'))
```

```
##       age                          educ               race         gender
##  Min.   :22.00   College             :834   Black      : 288   Female:939
##  1st Qu.:44.00   High school or less:411   Hispanic   : 168   Male  :872
##  Median :62.00   Some college       :566   Other/Mixed:  86
##  Mean   :56.69                             White      :1269
##  3rd Qu.:70.00
##  Max.   :92.00
##         income_cat    voter_category
##  high        :482   always  :1811
##  lower_middle:433   rarely  :   0
##  upper_middle:573   sporadic:   0
##  low         :323
##
##
```

White females who are around 56 years old with a college education and have high income are most likely to vote

```r
summary(df %>% filter(voter_category == 'rarely'))
```

```
##       age                            educ              race         gender
##  Min.   :22.00   College            :423   Black      :224   Female:690
##  1st Qu.:29.00   High school or less:631   Hispanic   :244   Male  :761
##  Median :38.00   Some college       :397   Other/Mixed:121
##  Mean   :42.33                             White      :862
##  3rd Qu.:55.00
##  Max.   :90.00
##         income_cat   voter_category
##  high        :246   always  :   0
##  lower_middle:341   rarely  :1451
##  upper_middle:327   sporadic:   0
##  low         :537
##
##
```

White males who are around 42 years old with a high school or less and low income are least likely to vote

Conclusion

Many factors goes into someones choice to vote or not vote and this analysis was just scratching the surface. When it comes to the variables analyzed here some we have more control over like education and consequently incomes as opposed to race and gender. Age is the one variable that we have no control of as time never stops. The more educated an individual is the more likely they will vote as the other variables fall in place, ie income/age. As we move to the next election it should be empathized that everyone regradless of age, race ,gender, education and income should always vote.