Final Project

Adrianna Bassard

12/06/2020

Install Necessary Packages

```
# install packages
library(tidyverse)
library(lubridate)
library(janitor)
library(gridExtra)
```

Introduction

I have always been interested in social science and because we are in an election year, questions related to demographics, personal beliefs, and voting behaviors are especially salient. For this project, I plan to explore different relationships in the data that are driven by interest.

Data

Data Source

For my project I used a dataset obtained from the UCLA Social Science Data Archive. The dataset I used was most recently used and uploaded from the Comparative Study of Ethnic Minorities in California (M163V1) using data collected in 1984 and can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/M163V1 (1).

Cleaning the Data

The data was relatively clean overall. The data was saved as a '.txt' file and I used the read_delim() function so that the data could be stored as a tibble. When I looked at the data using print(), I realized some of the variable names had strings that made the overall data less tidy and it leaving it in this form would have also made it more difficult to name individual variables later in the data analysis. To fix this issue, I used the clean_names() function from the 'janitor' package to tidy up the column names (shown below).

```
# read in data
df <- read_delim('Data/processed/data_finalproject_UCLA.txt', delim = ',',col_types = cols())
print(df, n_extra = 1)
```

```
## # A tibble: 1,646 x 116
##    `\`CASE IDENTIF~ telephne  city county ethcode centract `HIGH PERCENTAG~
##               <dbl>    <dbl> <dbl>  <dbl> <chr>      <dbl> <chr>
```

```
##   1                  1   415992    85      81 H         601602 0
##   2                  2   213602    68      37 H           5540 0
##   3                  3   415668    85      75 N            154 0
##   4                  4   213566    56      37 N           2428 HIGH PERCENTAGE~
##   5                  5   415562    65       1 H           4322 0
##   6                  6   619697    80      73 N           3107 HIGH PERCENTAGE~
##   7                  7   415786    49       1 H           4373 0
##   8                  8   415793    49       1 H           4443 0
##   9                  9   408394    81      53 N            139 0
##  10                 10   213938    56      37 N           2162 HIGH PERCENTAGE~
## # ... with 1,636 more rows, and 109 more variables: `SAMPLING FRAME` <chr>, ...
```

```r
# clean names
df <- df %>%
  clean_names()
print(df, n_extra = 1)
```

```
## # A tibble: 1,646 x 116
##    case_identifica~ telephne  city county ethcode centract high_percentage~
##               <dbl>    <dbl> <dbl>  <dbl> <chr>      <dbl> <chr>
##   1                  1   415992    85      81 H         601602 0
##   2                  2   213602    68      37 H           5540 0
##   3                  3   415668    85      75 N            154 0
##   4                  4   213566    56      37 N           2428 HIGH PERCENTAGE~
##   5                  5   415562    65       1 H           4322 0
##   6                  6   619697    80      73 N           3107 HIGH PERCENTAGE~
##   7                  7   415786    49       1 H           4373 0
##   8                  8   415793    49       1 H           4443 0
##   9                  9   408394    81      53 N            139 0
##  10                 10   213938    56      37 N           2162 HIGH PERCENTAGE~
## # ... with 1,636 more rows, and 109 more variables: sampling_frame <chr>, ...
```

Data Analysis

Demographic Information I am first interested to see the distribution of respondent ages. When I looked at the data in .jmp (another program that was able to open the .sav file) I saw that responses of '100' in the 'yyrborn' column were to be considered missing. I saw in the documentation that all data collection was done in the year 1984. If the respondent was born before the year 1984, we can subtract 84 from the documented year of birth. However, if the respondent is born after 1984, we must add 100 to this age because the year is actually 1800 rather than 1900. This was only the case for one participant who has an age of 86 years old. I created a new variable 'age' that contains the correct information.

```r
df2 <- df %>%
  filter(yyrborn != 100) %>%
  mutate(age = ifelse(yyrborn < 84, (84 - yyrborn), ((84 - yyrborn) + 100)))
print(df2, n_extra = 1)
```

```
## # A tibble: 1,556 x 117
##    case_identifica~ telephne  city county ethcode centract high_percentage~
##               <dbl>    <dbl> <dbl>  <dbl> <chr>      <dbl> <chr>
##   1                  1   415992    85      81 H         601602 0
##   2                  2   213602    68      37 H           5540 0
```

```
## 3                 3   415668    85      75 N            154 0
## 4                 4   213566    56      37 N           2428 HIGH PERCENTAGE~
## 5                 5   415562    65       1 H           4322 0
## 6                 6   619697    80      73 N           3107 HIGH PERCENTAGE~
## 7                 7   415786    49       1 H           4373 0
## 8                 8   415793    49       1 H           4443 0
## 9                 9   408394    81      53 N            139 0
## 10               10   213938    56      37 N           2162 HIGH PERCENTAGE~
## # ... with 1,546 more rows, and 110 more variables: sampling_frame <chr>, ...
```

Overall, there were 1556 respondents in this dataset. I am interested in first looking at the distribution of 'age' as well as the distribution related to birth state, number of children, and income level. When looking at the distribution of age, I see the majority of the respondents are between the ages of 18 and ~40. The minimum age is 18 as that was the minimum age to participate in the study (note it is also the minimum age to participate in an election) and the mean age was 40 years old. In the next plot, I saw that the majority of respondents born in the U.S. are born in California where the data was collected. I thought this was an interesting measure to consider later when we consider voting habits. The final two plots allow us to look at some trends of income and number of children. Because NA/No answer were combined as a response for number of children, I looked to see how many respondents had children and then further how many children respondents had.

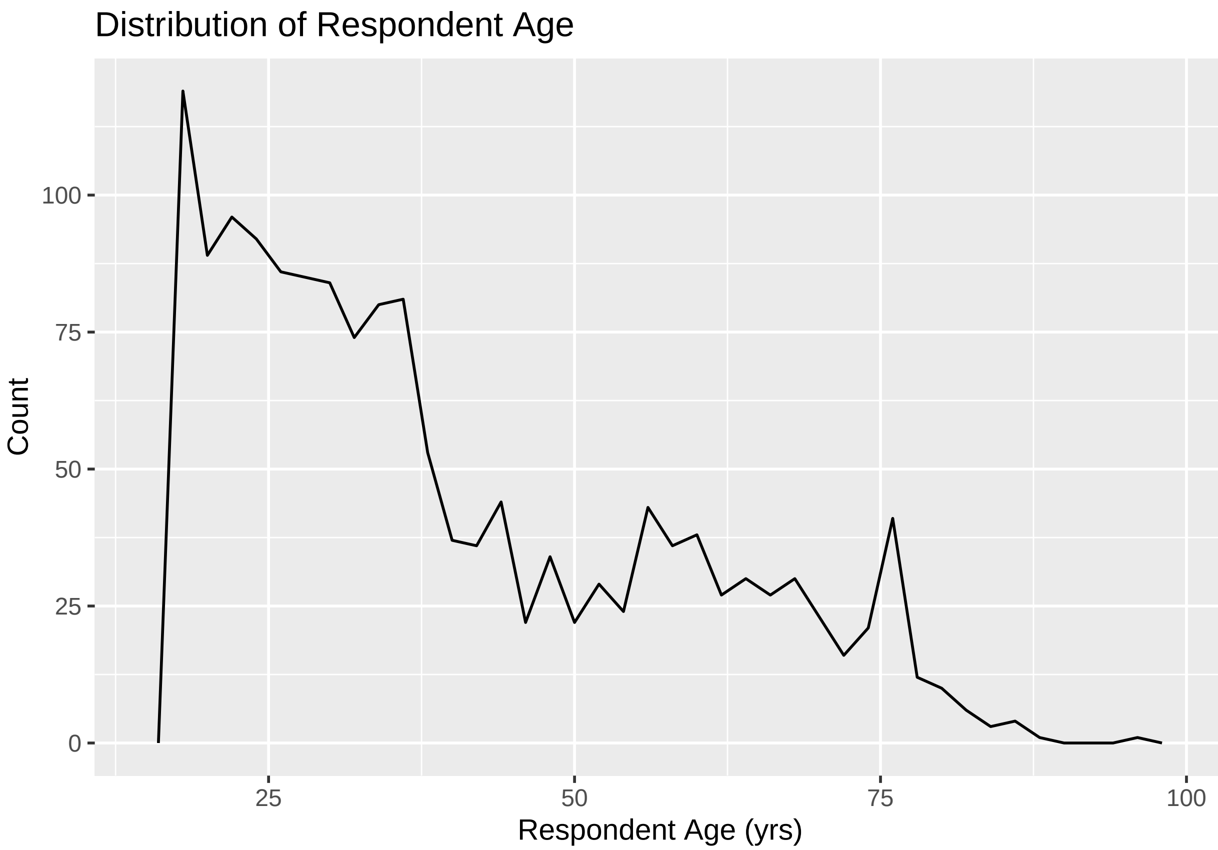
```
n <- length(df2$case_identification_number)
n
```

```
## [1] 1556
```

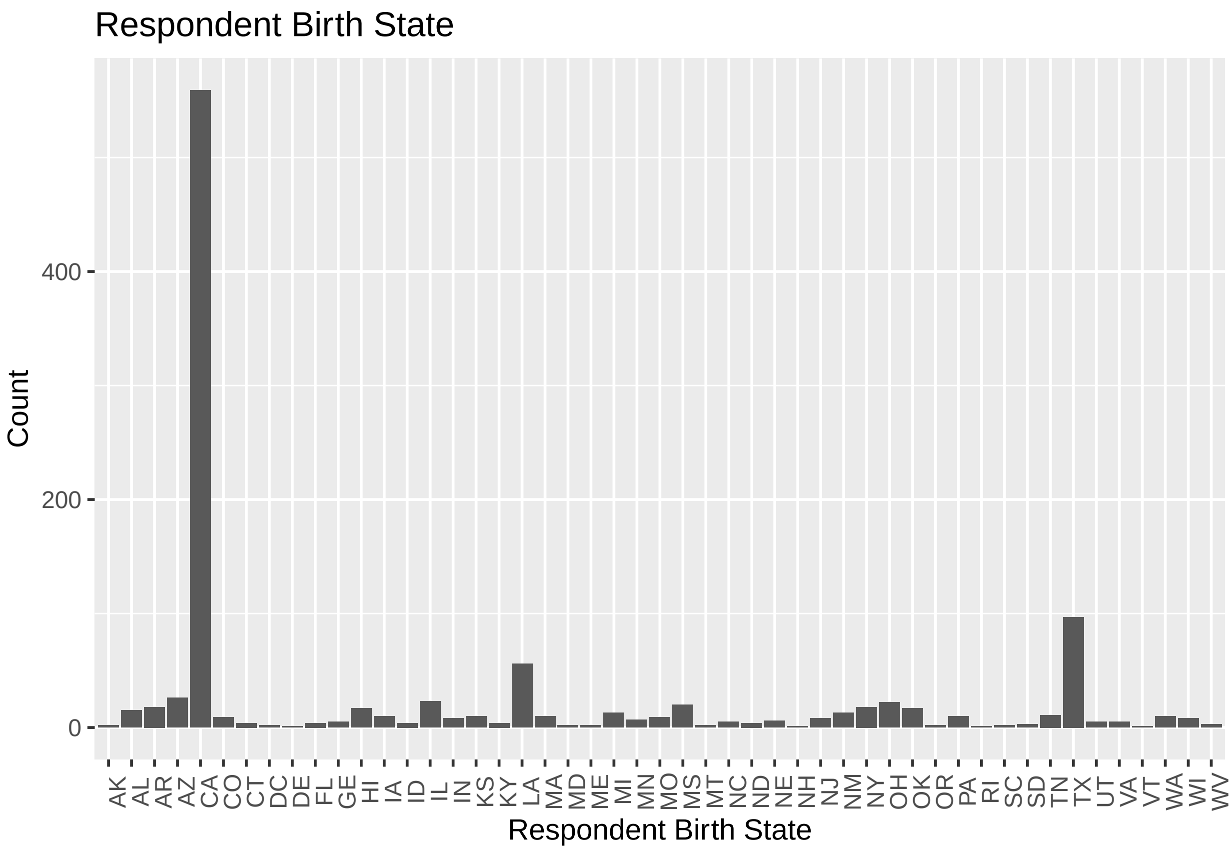
```
summary(df2$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      18      25      35      40      54      96
```

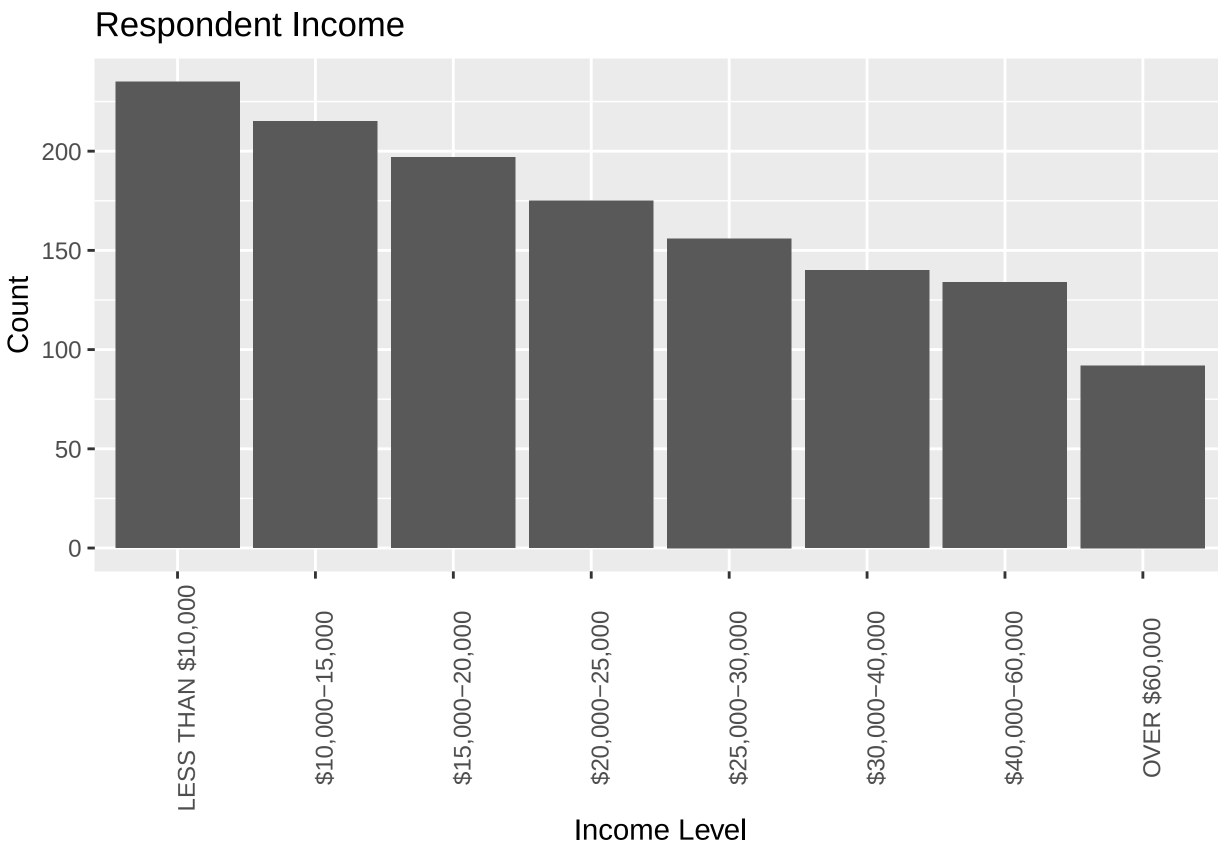
```
#plot distribution of age
df2 %>%
  ggplot(aes(x = age)) +
  geom_freqpoly(bins = 40) +
  labs(x = "Respondent Age (yrs)", y = "Count", title = "Distribution of Respondent Age")
```




```
#plot respondent birth state
df2 %>%
  filter(state_in_us_where_respondent_was_born != "ZZ") %>%
  ggplot(aes(x = state_in_us_where_respondent_was_born)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Respondent Birth State", y = "Count", title = "Respondent Birth State")
```



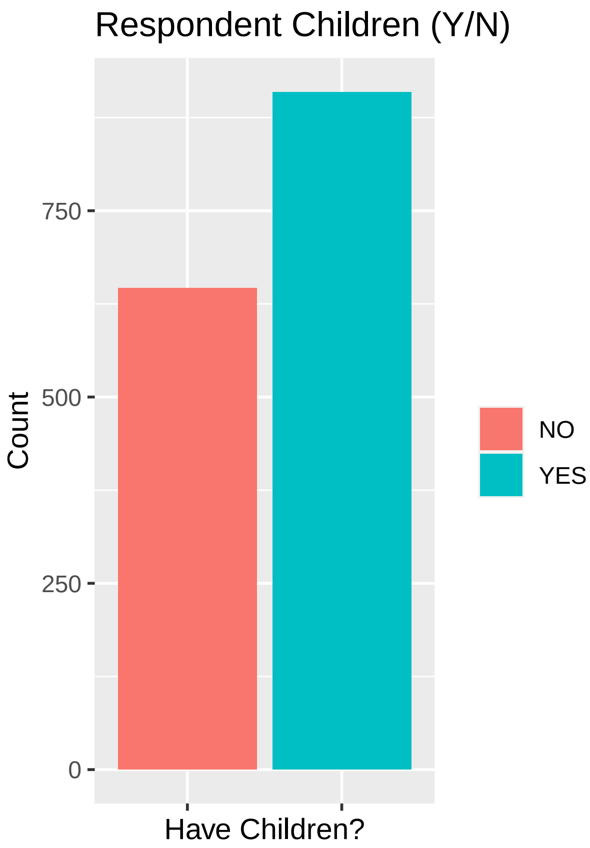

```
#plot city vs age including income level
income_f <- c("LESS THAN $10,000", "$10,000-15,000","$15,000-20,000", "$20,000-25,000", "$25,000-30,000
df2 %>%
  filter(income_level != "DK/NA") %>%
  mutate(income_level = factor(income_level, levels = income_f)) %>%
  ggplot(aes(x = income_level)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Income Level", y = "Count", title = "Respondent Income")
```

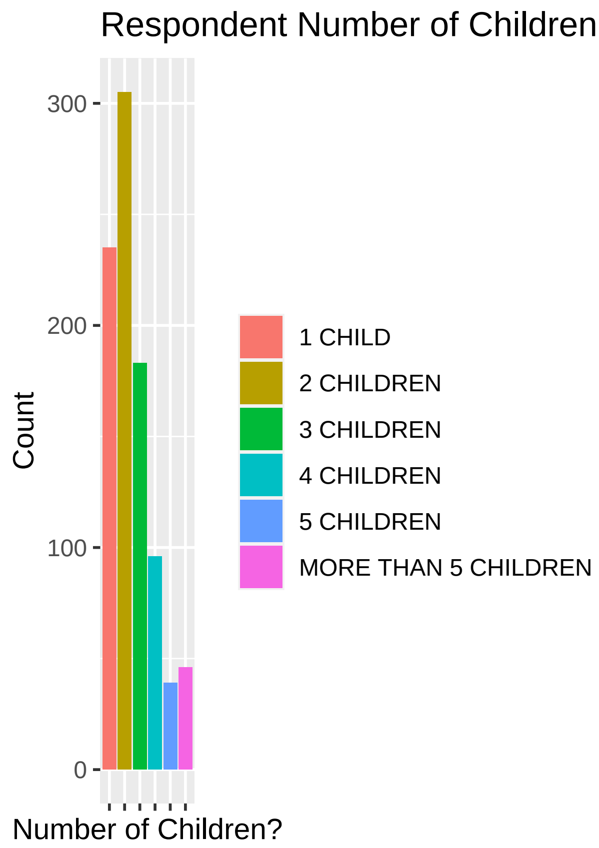



```
#have children?
have_children <- df2 %>%
  filter(have_children != ("DK/NA")) %>% #drop DK/NA response
  ggplot(aes(x = have_children, fill = have_children)) +
  geom_bar() +
  theme(axis.text.x = element_blank()) +
  labs(y = "Count", x = "Have Children?", title = "Respondent Children (Y/N)", fill = "")

#How many children?
num_children <- df2 %>%
  filter(number_of_children != "DK/NA") %>% #drop DK/NA response
  filter(number_of_children != "INAP") %>%
  ggplot(aes(x = number_of_children, fill = number_of_children)) +
  geom_bar() +
  theme(axis.text.x = element_blank()) +
  labs(y = "Count", x = "Number of Children?", title = "Respondent Number of Children", fill = "")

grid.arrange(have_children, num_children, ncol = 2)
```





Note* I removed levels of "DK/NA" to visualize more useful data.

Overall, respondents had a mean age of 40 with the majority being younger. I also saw the majority of respondents made less than 10,000 USD while the minority made over 60,000 USD. Further, most participants were parents and of these, most had two children followed by one child and three children.

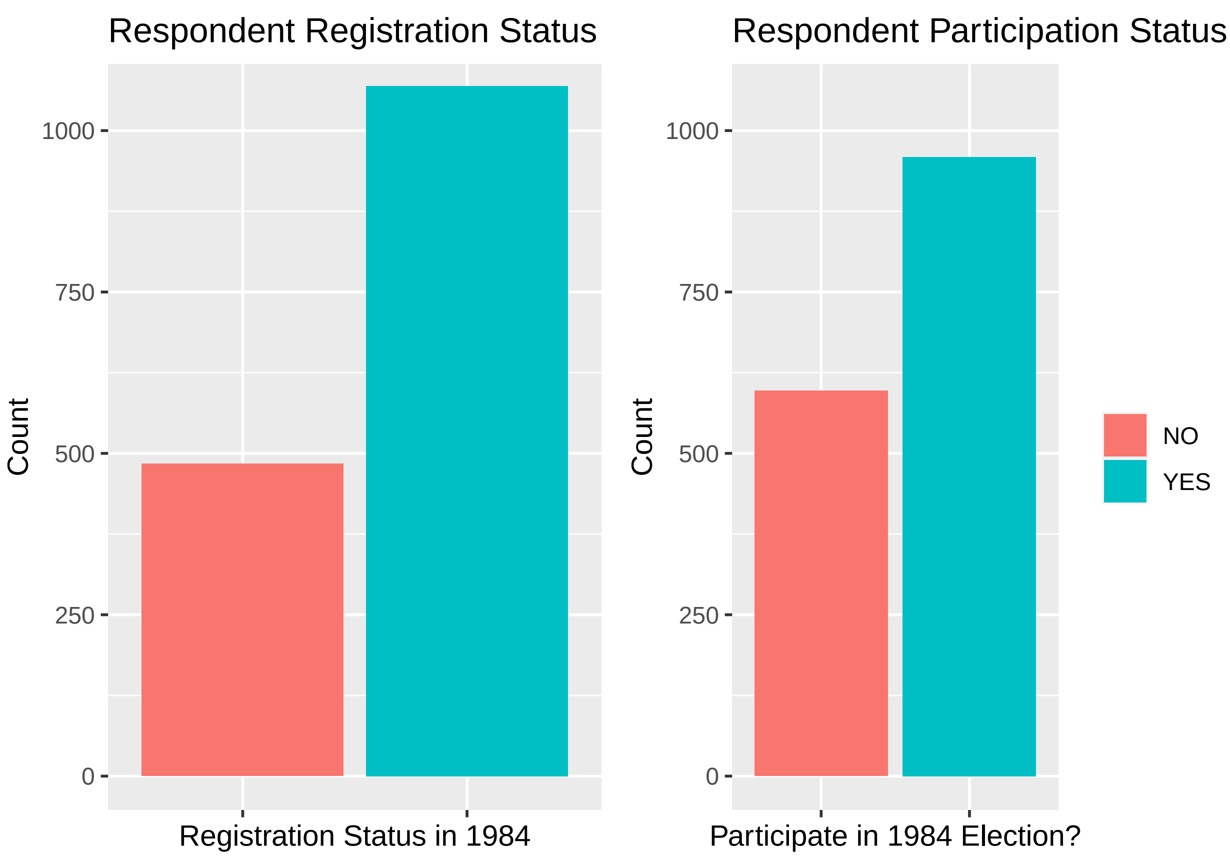
Respondents Political Views Now that I have visualized some of the data I'm interested in, I next going to use this information to guide some further questions. First, I am interested to see the breakdown of participation in the 1984 election by age. I first created a barplot of registration and participation in the election. Here we can see that over 1000 respondents reported being registered. Of these, the majority reported voting in the 1984 election.

```r
#barplot of registration status
registration <- df2 %>%
  filter(registration_status_in_1984 != "DK/NA") %>%
  ggplot(aes(x = registration_status_in_1984, fill = registration_status_in_1984)) +
  geom_bar(show.legend = FALSE) +
  coord_cartesian(ylim = c(0, 1050)) + #make ylim match other plot
  theme(axis.text.x = element_blank()) +
  labs(y = "Count", x = "Registration Status in 1984", fill = "", title = "Respondent Registration Statu

#barplot of participation in election
participation <- df2 %>%
  ggplot(aes(x = participate_in_1984_election, fill = participate_in_1984_election)) +
  geom_bar() +
  coord_cartesian(ylim = c(0, 1050)) +
  theme(axis.text.x = element_blank()) +
```

```
  labs(y = "Count", x = "Participate in 1984 Election?", fill = "", title = "Respondent Participation S

grid.arrange(registration, participation, ncol = 2)
```



Next, I wanted to see if age could predict whether a respondent participated in the 1984 election or not. To do so, I compared the frequency plot of respondent age for those who did and those who did not participate in the election. Here we see an interesting trend: younger respondents reported not participating in the election while older participants were more likely to indicate that they had participated (below). This graph represents what is a common trend found in voter habits: older people are more likely to participate in elections.

```
#participate in election by age
df2 %>%
  ggplot(aes(x = age, color = participate_in_1984_election)) +
  geom_freqpoly(bins = 40) +
  labs(y = "Count", x = "Age of Respondent (yrs)", color = "Participate in Election?", title = "Voting 
```

I was next interested in looking at other external factors that drive possible voter habits. First, I wanted to explore what respondents thought were the biggest issues facing America and how the responses differed depending on ethnicity. I was interested in this because I think it is likely to imagine that for a diverse sample, the responses may be related to ethnicity. I filtered the data only to look reflect the most popular choices. To do so, I determined which were the top 10 issues overall based on number of endorsements. Upon first glance, it was interesting to see that the majority of the most chosen issues are related to the state of the economy and financial burdens such as taxes. I will use these top 10 issues for all further analyses (see below).

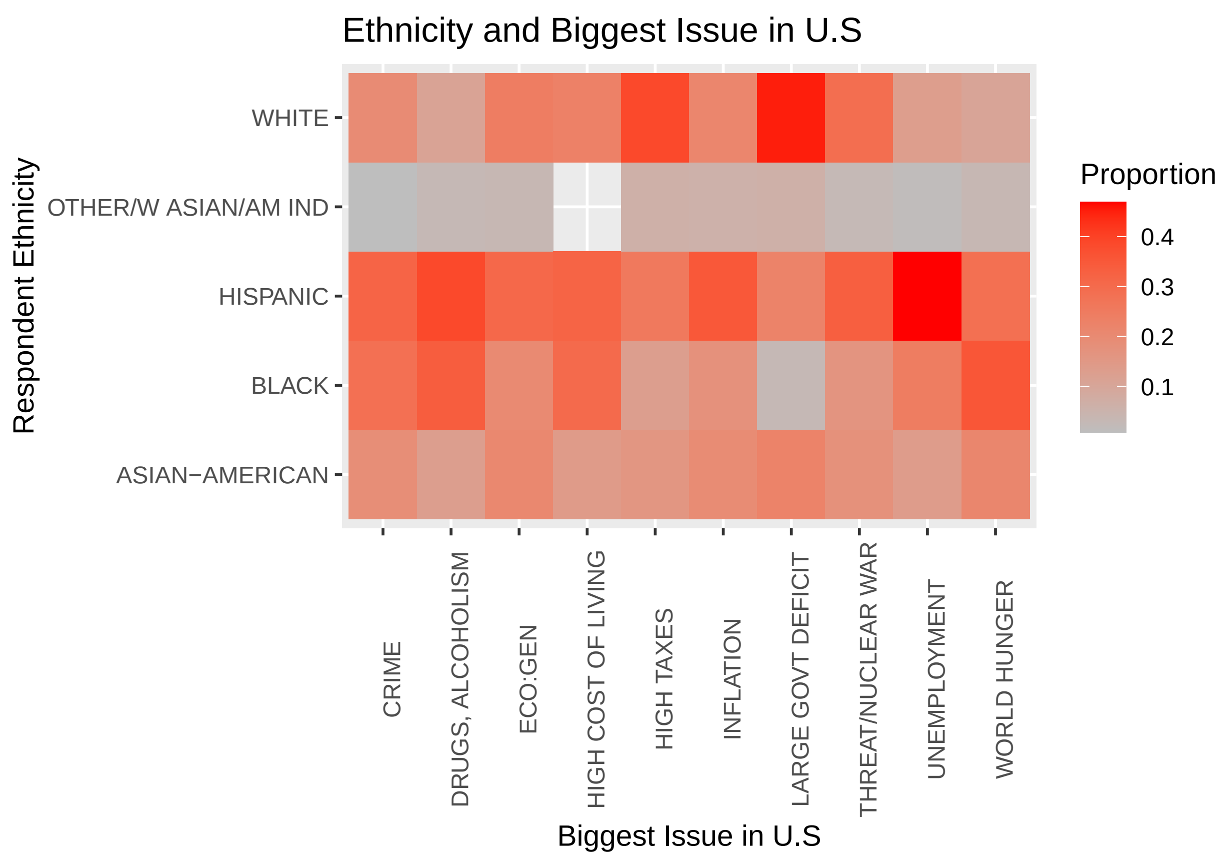
From this graph based on the proportion of endorsements for political issues by ethnicity, I found that Hispanic respondents were more likely to state that unemployment was the biggest issue facing America while White respondents chose large government deficit followed by taxes. Black and Asian American voters were more variable in what they chose as being the top issue in the U.S.

```
#determine top 10 biggest problems based on count
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  count(biggest_problem_of_usa) %>%
  arrange(desc(n))
```

```
## # A tibble: 46 x 2
##    biggest_problem_of_usa      n
##    <chr>                   <int>
##  1 UNEMPLOYMENT              314
##  2 THREAT/NUCLEAR WAR        179
##  3 ECO:GEN                   147
##  4 CRIME                     122
```

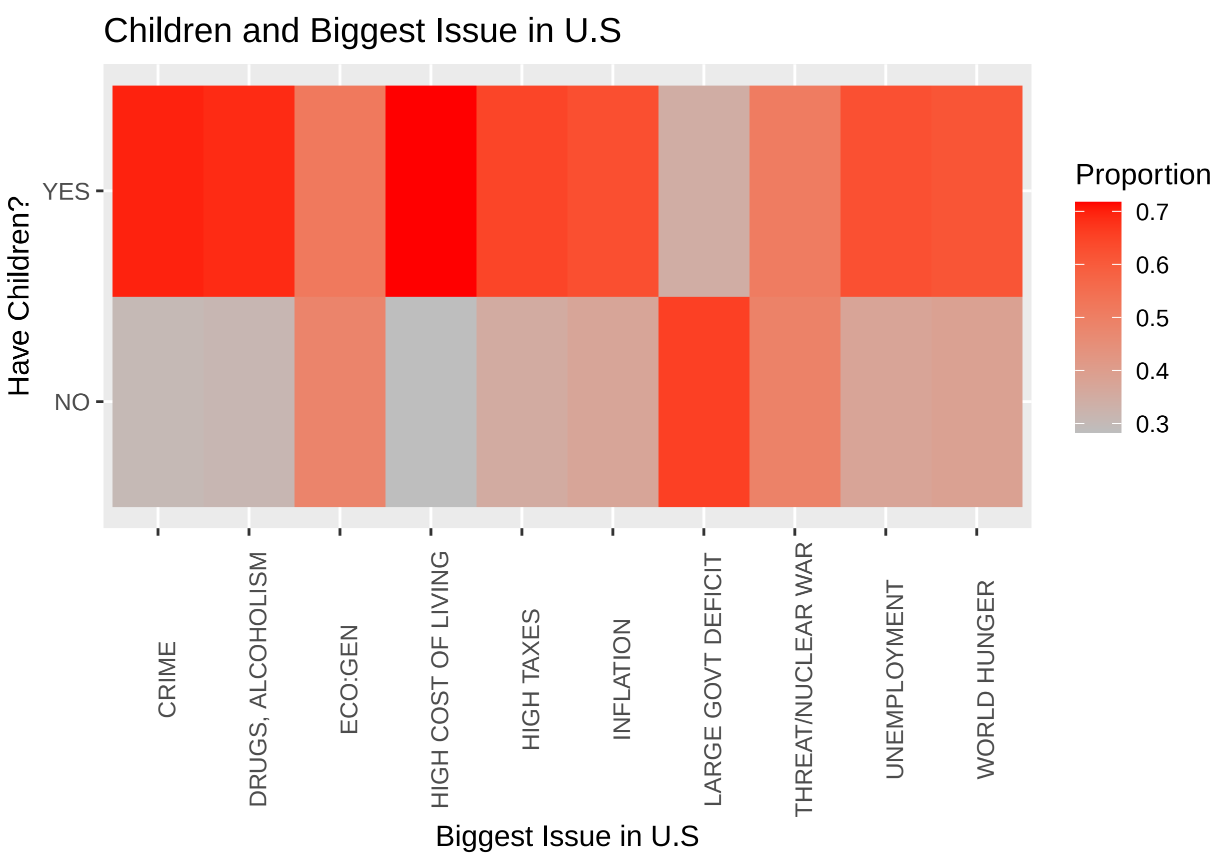
```
##   5 DRUGS, ALCOHOLISM         64
##   6 HIGH COST OF LIVING       60
##   7 INFLATION                 51
##   8 HIGH TAXES                34
##   9 LARGE GOVT DEFICIT        32
##  10 WORLD HUNGER              31
## # ... with 36 more rows
```

```
#break down top 10 problems by ethnicity
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  group_by(biggest_problem_of_usa) %>%
  filter(n() > 30) %>%
  ungroup() %>%
  filter(ethnic_identification != "NOT ASCERTAINED") %>%
  count(biggest_problem_of_usa, ethnic_identification) %>%
  group_by(biggest_problem_of_usa) %>%
  mutate(biggest_problem_prop = n / sum(n)) %>%
  ggplot(mapping = aes(x = biggest_problem_of_usa, y = ethnic_identification)) +
  geom_tile(mapping = aes(fill = biggest_problem_prop)) +
  scale_fill_continuous(low = "grey", high = "red") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Biggest Issue in U.S", y = "Respondent Ethnicity", fill = "Proportion", title = "Ethnicity a
```

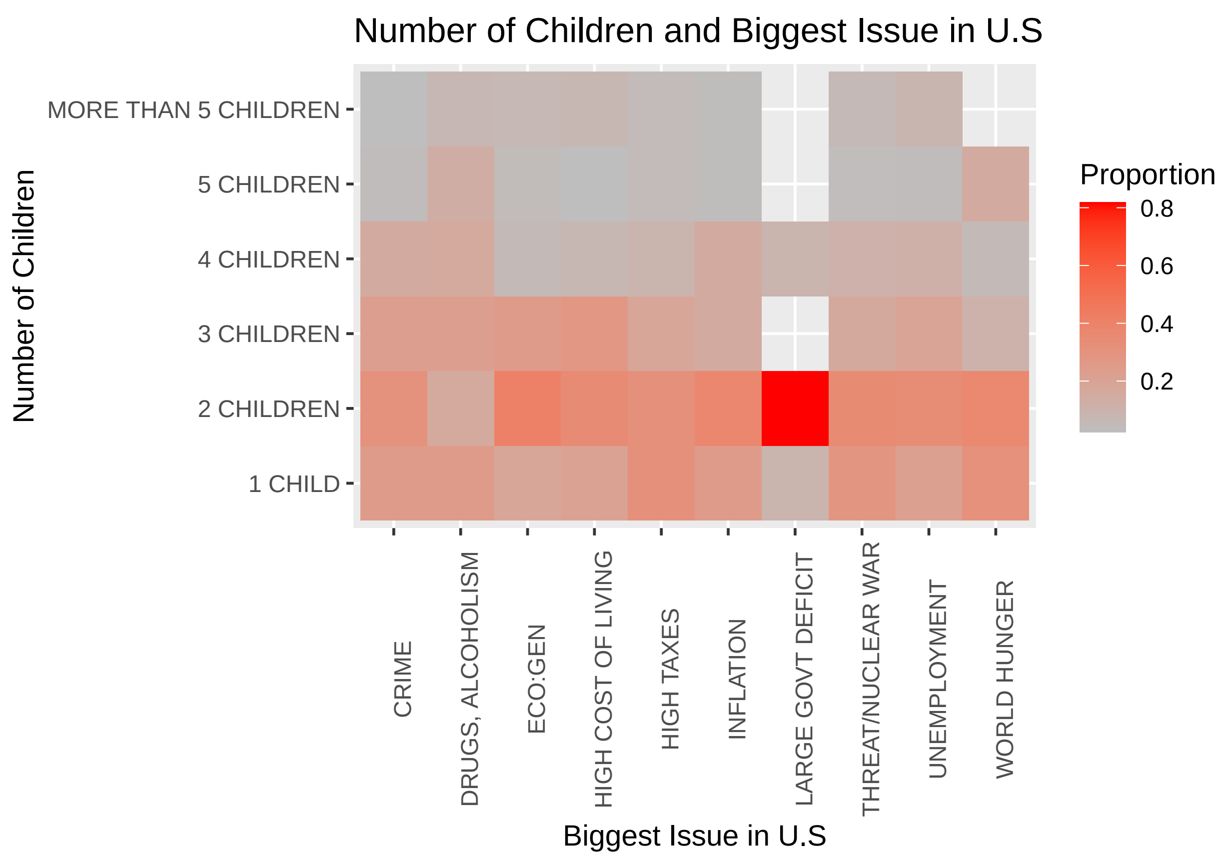


I next looked to see if there were any patterns to be found when relating issues to children. First, I made a plot to determine which issues respondents tended to endorse depending on whether they had children or not. I found there was indeed a pattern that stood out with respondents who were not parents choosing the large government deficit while those with children tended to be relatively spread among the top 10 issues but interestingly, tended not to choose the government deficit as the most important issue. I decided to look further and see if the number of children participants had affected issue choice. For this analysis, we see those with two children (which, as we saw before, make up the majority of the respondents) the large government deficit was chosen as the biggest threat with other groups being more variable in their responses. This is interesting as it shows that the previous result may have been driven primarily by respondents who do not have two children.

```
#issue based on have children?
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  group_by(biggest_problem_of_usa) %>%
  filter(n() > 30) %>%
  ungroup() %>%
  filter(have_children != "INAP") %>%
  filter(have_children != "DK/NA") %>%
  count(biggest_problem_of_usa, have_children) %>%
  group_by(biggest_problem_of_usa) %>%
  mutate(biggest_problem_prop = n / sum(n)) %>%
  ggplot(mapping = aes(x = biggest_problem_of_usa, y = have_children)) +
  geom_tile(mapping = aes(fill = biggest_problem_prop)) +
  scale_fill_continuous(low = "grey", high = "red") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Biggest Issue in U.S", y = "Have Children?", fill = "Proportion", title = "Children and Bigg
```



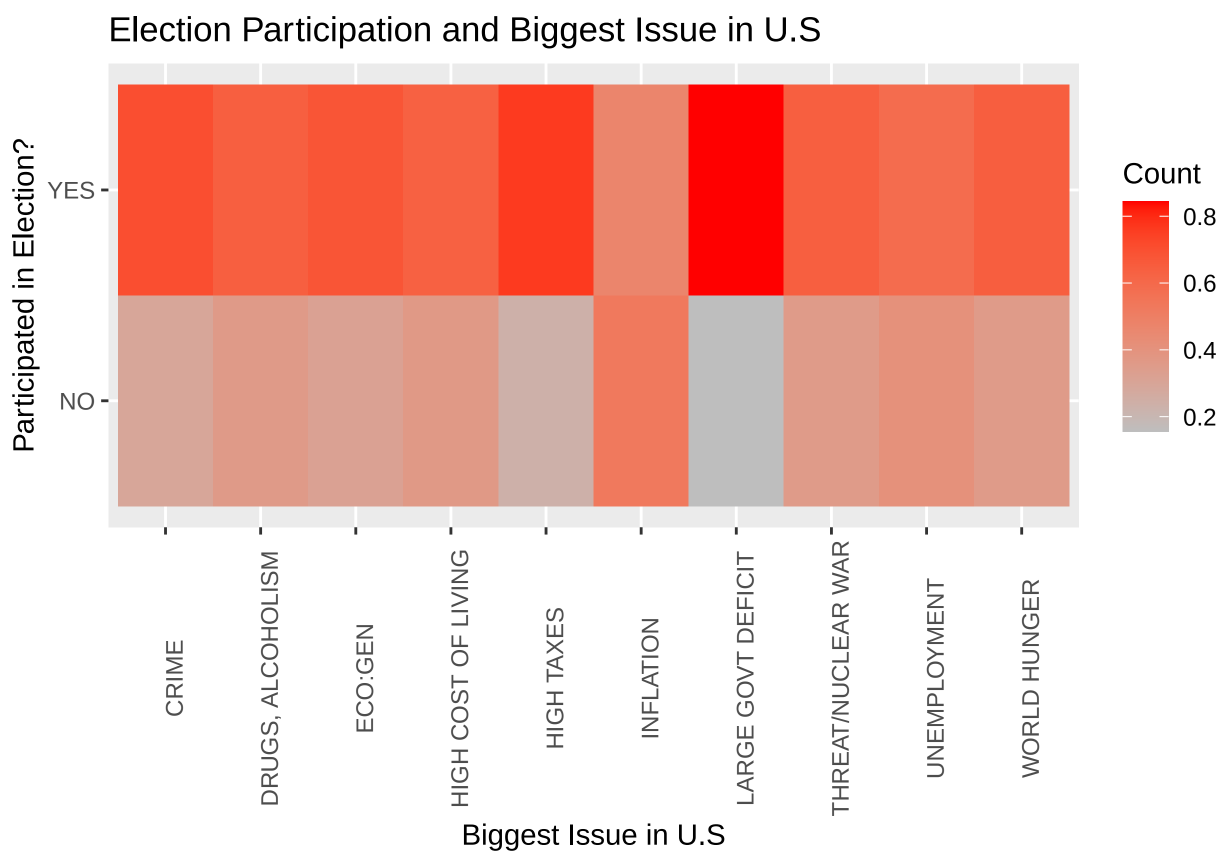
```
#issue based on number of children
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  group_by(biggest_problem_of_usa) %>%
  filter(n() > 30) %>%
  ungroup() %>%
  filter(number_of_children != "INAP") %>%
  filter(number_of_children != "DK/NA") %>%
  count(biggest_problem_of_usa, number_of_children) %>%
  group_by(biggest_problem_of_usa) %>%
  mutate(biggest_problem_prop = n / sum(n)) %>%
  ggplot(mapping = aes(x = biggest_problem_of_usa, y = number_of_children)) +
  geom_tile(mapping = aes(fill = biggest_problem_prop)) +
  scale_fill_continuous(low = "grey", high = "red") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Biggest Issue in U.S", y = "Number of Children", fill = "Proportion", title = "Number of Chi
```



Next I looked to see if respondents biggest issue choice was related to whether they had participated in the election or not. Each of the issues had relatively high proportion of endorsements from respondents who participated in the election with the threat of nuclear war and taxes seeming to stand out. This result follows as we are looking at the top 10 issues most commonly chosen but it is interesting to see that overall, for these respondents, each of these issues seems nearly equally as important among participants of the 1984 election.

```
#issue by participated in election?
df2 %>%
```

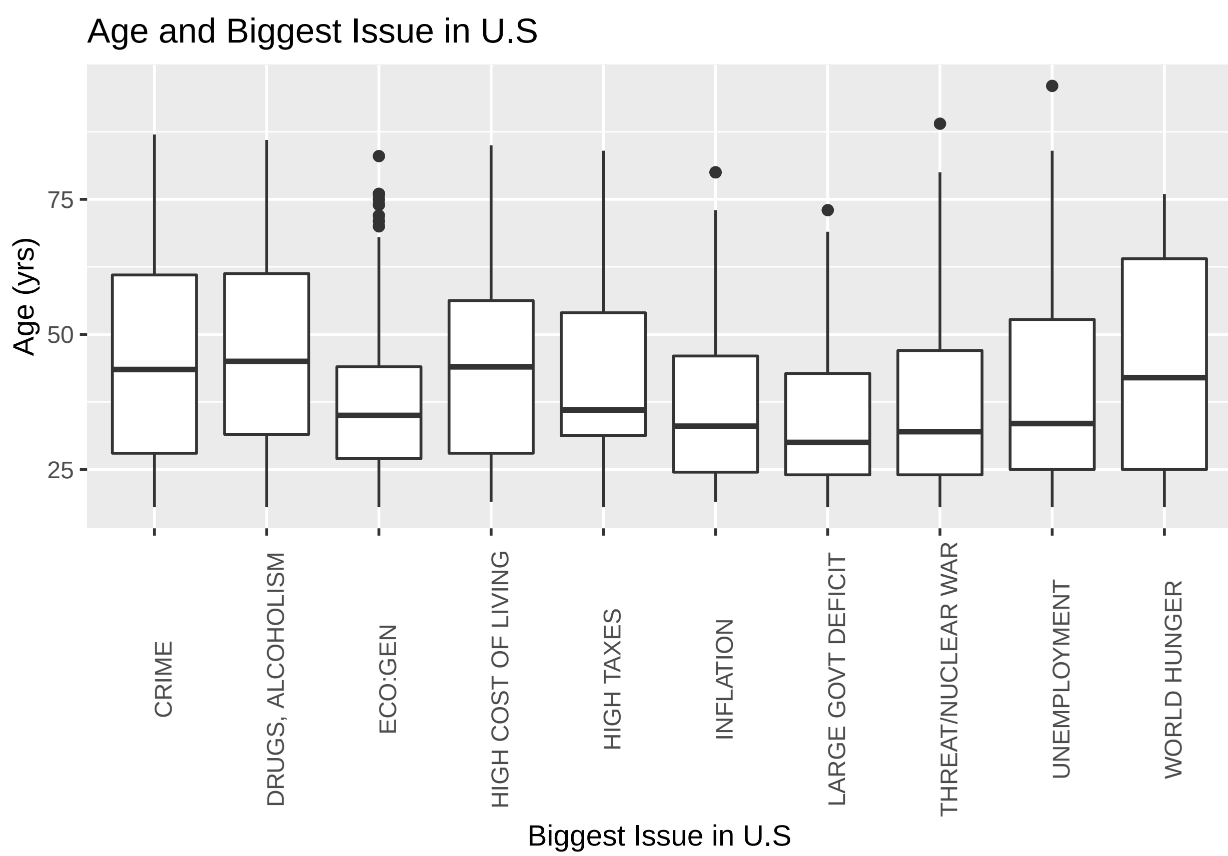
```
filter(biggest_problem_of_usa != "DK/NA") %>%
group_by(biggest_problem_of_usa) %>%
filter(n() > 30) %>%
ungroup() %>%
filter(participate_in_1984_election != "DK/NA") %>%
count(biggest_problem_of_usa, participate_in_1984_election) %>%
group_by(biggest_problem_of_usa) %>%
mutate(biggest_problem_prop = n / sum(n)) %>%
ggplot(mapping = aes(x = biggest_problem_of_usa, y = participate_in_1984_election)) +
geom_tile(mapping = aes(fill = biggest_problem_prop)) +
scale_fill_continuous(low = "grey", high = "red") +
theme(axis.text.x = element_text(angle = 90)) +
labs(x = "Biggest Issue in U.S", y = "Participated in Election?", fill = "Count", title = "Election Pa
```



Presidential Choice It is also important to see which presidential candidate respondents preferred and how that was affected by factors such as age and U.S. issues. I first wanted to see how age broke down by biggest issue facing the USA. I found that the government deficit has the youngest average age of endorsement while older respondents tended to choose drugs and alcoholism. Next, I looked at how respondents rated issues depending on who they preferred as the presidential candidate. Here we see that those who preferred Mondale thought unemployment and nuclear war was the biggest issues facing the U.S but were more spread generally while those who preferred Reagan thought the government deficit was the most important issue.

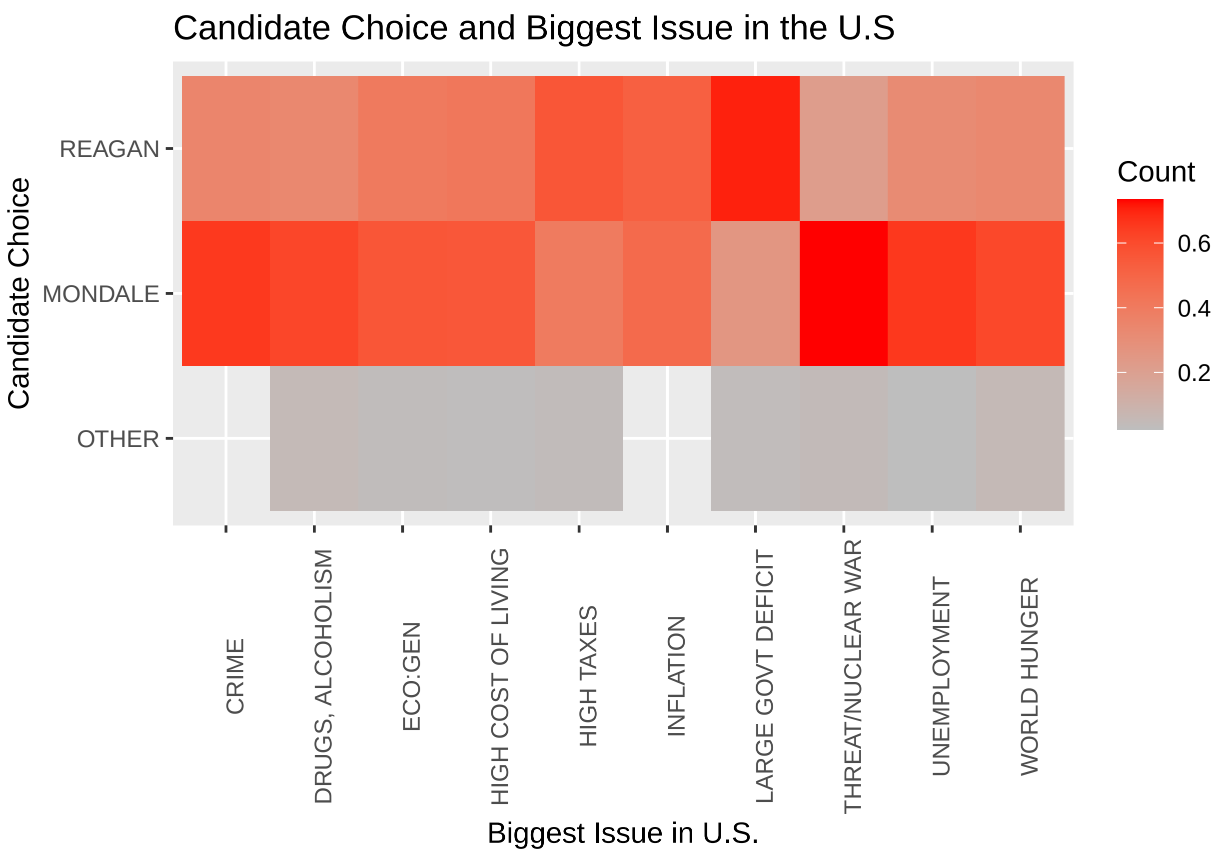
To tie this together, I was interested to see how age affected voter choice. I will focus on Reagan and Mondale for this analysis but the ‘other’ option is shown for comparison. Here we see that Reagan had the largest range in age of voter support an average age just under 40. Mondale had a similar range in age of voter support but the highest average age.

```r
#age vs biggest issue
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  group_by(biggest_problem_of_usa) %>%
  filter(n() > 30) %>%
  ungroup() %>%
  ggplot(mapping = aes(x = biggest_problem_of_usa, y = age)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Biggest Issue in U.S", y = "Age (yrs)", title = "Age and Biggest Issue in U.S")
```

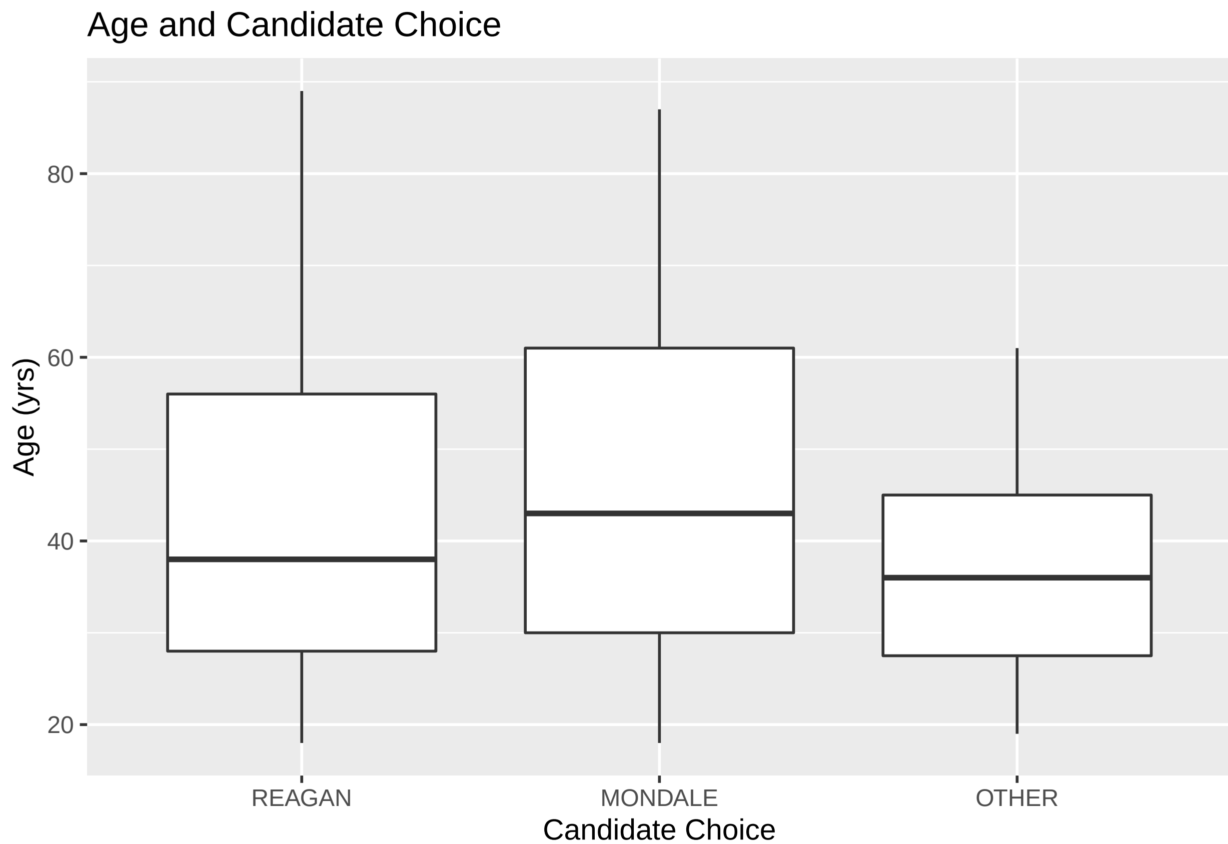



```r
#Presidential Vote Choice vs biggest problem of USA
df2 %>%
  filter(biggest_problem_of_usa != "DK/NA") %>%
  group_by(biggest_problem_of_usa) %>%
  filter(n() > 30) %>%
  ungroup() %>%
  filter(presidential_vote_choice_84 != "DK/NA") %>%
  filter(presidential_vote_choice_84 != "INAP") %>%
  mutate(presidential_vote_choice_84 = factor(presidential_vote_choice_84, levels = c("OTHER", "MONDALE
  count(biggest_problem_of_usa, presidential_vote_choice_84) %>%
  group_by(biggest_problem_of_usa) %>%
  mutate(biggest_problem_prop = n / sum(n)) %>%
  ggplot(mapping = aes(x = biggest_problem_of_usa, y = presidential_vote_choice_84)) +
  geom_tile(mapping = aes(fill = biggest_problem_prop)) +
  scale_fill_continuous(low = "grey", high = "red") +
```

```
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Biggest Issue in U.S.", y = "Candidate Choice", fill = "Count", title = "Candidate Choice ar
```




```
#age vs vote choice
df2 %>%
  filter(presidential_vote_choice_84 != "DK/NA") %>%
  filter(presidential_vote_choice_84 != "INAP") %>%
  mutate(presidential_vote_choice_84 = factor(presidential_vote_choice_84, levels = c("REAGAN", "MONDALE
  ggplot(mapping = aes(x = presidential_vote_choice_84, y = age)) +
  geom_boxplot() +
  labs(x = "Candidate Choice", y = "Age (yrs)", title = "Age and Candidate Choice")
```



Discussion and Conclusion

Discussion Overall, there are some findings that stand out. First, related to demographics, I found that the majority of respondents had an annual income of less than ~$30,000. The majority of respondents had children with most having two children. Further, most who are born in the United States were born in California where the data was collected.

Getting further into voting behaviors, we see that most respondents who reported being registered to vote did indeed vote. We also saw that older respondents were more likely to have voted which supports what is often observed in election results. Finally, we were able to determine the top 10 most endorsed issues facing the U.S. being crime, drugs, the economy, cost of living, taxes, inflation, large government deficit, threat of nuclear war, unemployment, and world hunger.

Conclusion This project has allowed me to dive deeper into the political climate and voter beliefs in the 1984 election of Reagan vs. Mondale. These analyses are especially interesting considering a presidential election took place this year. The first thing that stands out is the trend of older people being more likely to participate in the election. This is a common trend that we see continues to persist today.

Today, however, we do see that the biggest issues surrounding the U.S and election are much different. During the 1984 election, issues related to economics were the most endorsed by participants. We can see that these issues are related to the candidate who won the election: Reagan as he is known for “Reaganomics” - his policies and ideas surrounding taxes and free-market activity (2). Today, however, we see a rise of issues that weren’t endorsed by respondents in 1984 including immigration, government leadership, healthcare, racism, unifying the country, and climate change (3). We have seen this shift today following President Trump’s focus on issues of immigration in 2016 and the recent attention given to the Black Lives Matter movement following focus on police brutality has given increased attention to issues of race. Further, the COVID-19 pandemic has called further attention the U.S. healthcare system. The 2020 election is probably one like no other in U.S. history and comparing it with respondents views surrounding the 1984 presidential election highlight these differences even further. Though it is interesting to compare the current political climate with the past, I am hopeful of the future and interested to see how we will tackle the issues we face today.

References

(1) Cain, Bruce; Kiewiet, D. Roderick; Uhlaner, Carole J., 2017, "Comparative Study of Ethnic Minorities in California (M163V1)", https://doi.org/10.7910/DVN/0KTV57, Harvard Dataverse, V1

(2) Lexico powered by Oxford Dictionary. "Definition of Reaganomics." 2010. https://www.lexico.com/en/definition/reaganomics

(3) Statista. "The Most Important Issue Facing the U.S. Today." Martin Armstrong. 2019. https://www.statista.com/chart/10278/the-most-important-issues-facing-the-us-today/