STA304PS4JFinalRough1

```
# Load Packages
library(haven)
```

```
## Warning: package 'haven' was built under R version 3.6.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

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

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     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
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

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

```
library(labelled)
```

```
## Warning: package 'labelled' was built under R version 3.6.3
```

```
library(jtools)
```

```
## Warning: package 'jtools' was built under R version 3.6.3
```

```
library(boot)
```

```
## Warning: package 'boot' was built under R version 3.6.3
```


Abstract

2020 being an erratic year also marks the much-anticipated results of the United States Presidential election between the front-runners, Donald Trump and Joe Biden, on November 3rd, 2020. In this paper, we use both Nationscape survey and IPUMS data to conduct a regression analysis to determine the impact of social-demographic variables such as age, gender, language, region, and race of respondents on their choice of 2020 US presidential election. We discovered that age and race are strong predictors of respondents' presidential election choice and we expect Biden to win 52 percent of the national popular vote, with a 2 percent margin error. Our findings can not only predict the 2020 presidential election but also provide references for future presidential elections.

1 Introduction

In 2016, Donald Trump won the race for president of the United States which in turn rocked the world of statistical analysis as most pre-election predictions had suggested otherwise. This has led to an extremely heated political climate throughout the country that only grows in tension as election day approaches. Over the past four years there has been a steadily growing pool of citizens that hope to see a new leader take over control of the country. In 2020, this movement has seen a significant surge in support especially due to the numerous difficulties the country has faced this year. Through this study we will attempt to determine the possibility of this movement's success by providing a forecast for the winner of the 2020 US election.

In the US, a somewhat complex system is used to determine the next president which can make establishing a conclusive forecast very difficult. First, individuals cast their votes to elect a set of electors for their state, these electors then cast their final votes that actually decide on the winner of the election. As we only have data regarding these individual votes, we must base our forecast of the popular vote. While the popular and electoral votes often line up, the results of the 2016 election showed the exact opposite scenario and thus we must remain careful throughout this study to remember this.

In this paper we used multilevel regression modelling with post-stratification to analyze the effects of key characteristics on a citizen's voting decision and provide a final prediction on the outcome of the 2020 election. We have used two large survey datasets from _____ and _____ to accomplish this, applying our regression model to the smaller _____ survey data and post-stratifying using the second and much larger _____ dataset to generate our forecast. Our predictions suggest to us that, within a very small margin, Joe Biden will win the popular vote in the country primarily as a result of the effect a person's race has on their political preference. With the high amount of uncertainty not only in our model but in elections and voter preference as well, we cannot be completely certain of this forecast and must pay special attention to the factors likely to cause error in this prediction. Despite this, our paper is of great importance to voters as it provides an initial indication of the direction of the country for the next four years. While we may not be able to provide an indisputable forecast, this will still allow voters to begin making plans for the future rather than leave them completely in the dark.

In Section 2, we present the datasets used for our analysis as well as the modifications made to this data to allow our regression model to function. Section ___ provides a discussion of the model we have used for the study and examines the validity of its predictions. Additionally, we explain the multilevel modelling strategy

used to generate forecasts and its strengths and weaknesses in this section. The results of this model and the forecasts in produces are presented in Section _____, along with details on how to properly interpret these values.

Finally, in Section _____, we have summarized the main findings of this study and discussed their validity and the factors that may render them incorrect. At the end of this paper we have added an appendix which includes a link to where the code used for this study lives.

2 Data Discussion

Nationscape Survey Data

Democracy Fund plus UCLA Nationscape is a partnership between Democracy Fund Voter Study Group and University of California Los Angeles Political Scientists Chris Tausanovitch and Lynn Vavreck. It is one of the largest public opinion survey projects ever conducted — interviewing people in nearly every county, congressional district, and mid-sized U.S. city covering the U.S. 2020 campaign and election. Data Collection occurred through weekly surveys conducted by Nationscape. It began with the week of July 18, 2019, and concluded the week of December 26, 2019 (last interview on January 1, 2020). Phase 1 of the data, released in January of 2020, includes nearly 156,000 cases collected over 24 weeks. Phase 2 of the data, released in September of 2020, includes a re-release of Phase 1 data and new data from January 2020 to July 2020 (Phase 2 data) with the total cases being 318,697. Each data set is named after its first field date and is in the field for a week. The survey has been in the field since July 10, 2019, and it includes interviews in English with roughly 6,250 people per week. The interviews for the survey were conducted online in places where the respondents had access to the internet through a mobile device or computer. On average, 12% of the respondents who were asked to take the survey declined immediately while 5% did not complete the survey, and 8% of the respondents sped or straight-lined through the survey. According to the surveyors, respondents are considered to be speeding if they complete the survey in fewer than 6 minutes. Straight-lining through a survey occurs when respondents rush through the survey, clicking the same response every time. This usually happens when respondents are bored or impatient. In this survey, respondents were considered to be straight-lining if they selected the same response for every question in the three policy question batteries. This resulted in an average yield of roughly 75% of the original invited sample, depending on the wave.
There are roughly 200 variables in each weekly file. They are named to reflect the topic they measure. There is a rotating set of questions that vary over weeks of fielding, resulting in different numbers of variables in different weeks. Moreover, there were a number of changes made to the variables during different waves. For example, in wave 47 (2020-06-04, pilot-testing for general election vote questions began and the variable vote_2020_v1 was added, the question stem of which was changed later, with the variable being renamed as vote_2020 in Wave 49 and forward. Nationscape samples were provided by Lucid, a market research platform that runs an online exchange for survey respondents. The samples drawn from this exchange match a set of demographic quotas on age, gender, ethnicity, region, income, and education. Respondents were sent from Lucid directly to survey software operated by the Nationscape team. The sampling method used in this survey was convenience sampling based on demographic criteria. Convenience sampling is a non-probability sampling method in which researchers use the nearest and most conveniently available participants as the sample. This was essentially due to low response rates.

The survey, itself, in general, is a lengthy one comprising three main types of survey questions. Some ask respondents to report attitudes, such as whether they trust their neighbours. Other questions are about reported behaviours, such as voter turnout, which depend on respondents' recall. The third type of question asks people to report facts about their lives, such as how many children they have, whether they own their home, or how many cigarettes they have smoked in their lives.

The survey data was then weighted to be representative of the American population, based on the following factors: gender, the four major census regions (the four regions are Northeast, Midwest, West and South according to the US Census Bureau.), race, Hispanic ethnicity, household income, education, age, language spoken at home, nativity (U.S.- or foreign-born), 2016 presidential vote, and the urban-rural mix of the

respondent's ZIP code as well as the following interactions: Hispanic ethnicity by language spoken at home, education by gender, gender by race, race by Hispanic origin, race by education, and Hispanic origin by education. The weights were generated using a simple raking technique, generated for each week's survey. The targets to which Nationscape is weighted was derived from the adult population of the 2017 American Community Survey of the U.S. Census Bureau, with the exception of the 2016 vote, which is derived from the official election results released by the Federal Election Commission. Representativeness Assessment closely follows the Pew Research Center's evaluations of online non-probability samples in 2016 and 2018 as well as Pew's American Trends Panel (which was recruited using probability sampling) particularly for some smaller items. Pew compares the estimates generated by online vendors to estimates of the same quantities in census data and other high-quality sources, primarily the American Community Survey and supplements to the Current Population Study. Similar to Pew, the difference between the targets and the estimates across all the items is calculated in order to have the same amount of error as Pew(on average).

For our model, we use the survey of the week of June 25th, 2020, namely, the 20200625 dataset. Pertaining to the generic pattern used in this research, it is a lengthy survey with a variety of questions ranging from demographic variables to Covid-19 related ones as well as a collection of policy questions. Randomization was employed while conducting the survey, particularly for some policy questions. The policy questions were divided into two segments. Respondents were first asked two policy questions at random from abortion_any_time, abolish_priv_insurance, criminal_immigration, china_tariffs, egypt, saudi_arabia, and immigration_insurance. They were then asked 8 random policy questions from the remaining policy questions which comprised the second segment. Missing data is mainly coded to indicate the reason they are missing using the following codes: 888-Asked in this wave, but not asked of this respondent, 999-Not sure, don't know, " " - Respondent skipped. Our model uses vote_2020, age, gender, race_ethnicity, census_region and language.

In this study we will be examining the effects of a person's; age, gender, region, race, and household language on their intended vote in the 2020 US Election. To begin preparing our Nationscape survey data, using the select() function from the dplyr package, we selected only these variables of interest from our data and created a new dataset with them. Next, as the US election will most likely be a race between Joe Biden and Donald Trump, we filtered out all respondents that showed no interest in voting for either of these candidates.

Since we are using post-stratification in this study, this requires both of our datasets to have matching column names and response options. To accomplish this, we made a few more alterations to our dataset. We first classify our respondents based on their age group rather than their numerical age These were set up as breaks starting from 18 to 29 and followed a ten year age gap up until age 60 where we classify these respondents as 60 +. Grouping by age allows us to more easily create cells we can organize respondents into and generate overall forecasts from. To correct for the differences in responses in race across the data we organized the respondents into three primary groups based on the highest proportions we saw in the data: Black or African American, White, or Other. Lastly, to account for response inconsistencies between datasets in household language, much like our method for race, we organized respondents into three groups based on the proportion of languages used in the US: English, Spanish, or Other. English and Spanish can be considered the national languages of the USA as their proportions greatly outweigh all other language options which is why we chose to pay special attention to them in our analysis. For both gender, and participant region, these variables were already set up exactly as required by the Nationscape Data team and thus we had to perform no additional alterations to them. In addition to these, we also found many NA values that were provided by the survey data itself, likely due to cases of non-response. To remove these values we applied the na.omit() function from the base R package to our cleaned dataset so that we are left with only complete cases. By doing this, we allow our regression model to properly function with our data as it ensures that the number of observations from our model matches the size of our dataset.

```
raw_data <- read_dta("ns20200625.dta")

# Add Labels
raw_data <- labelled::to_factor(raw_data)
```

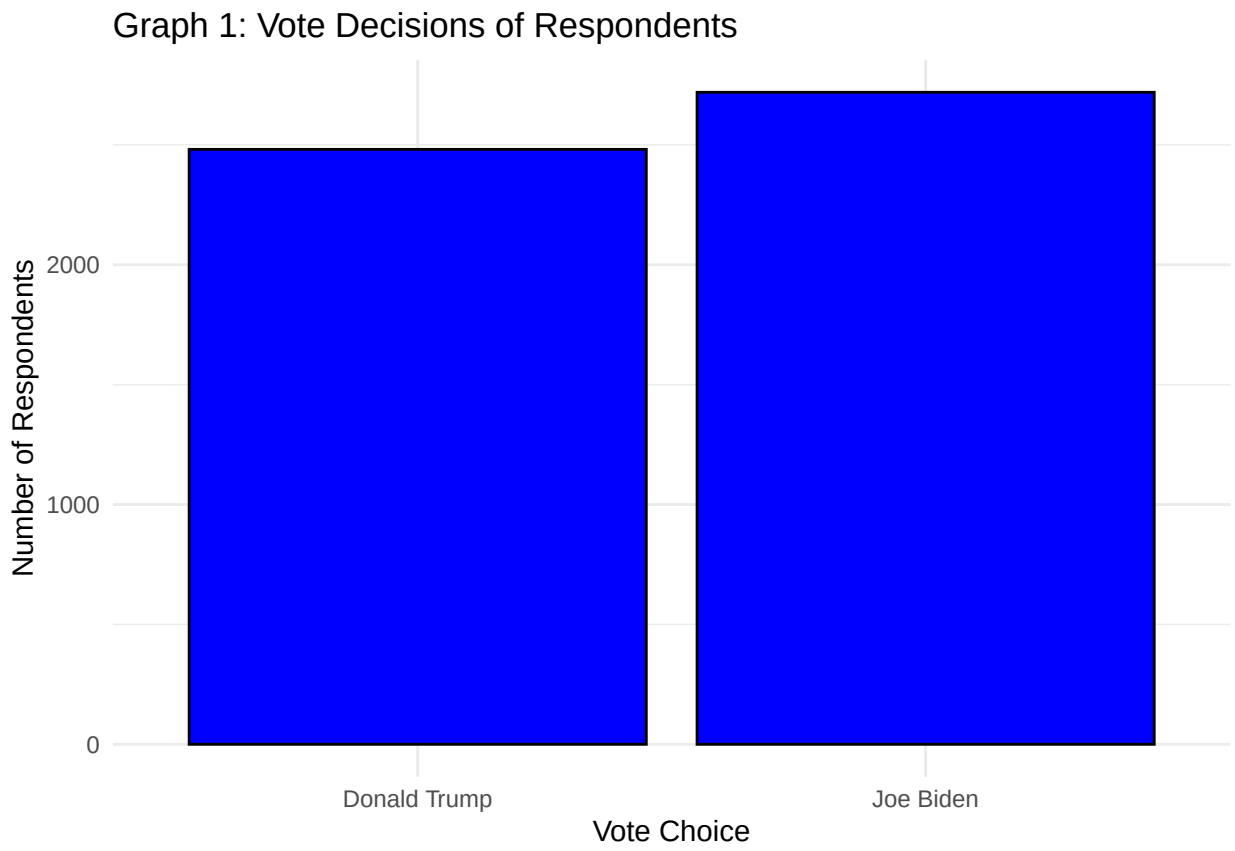
```r
select_data <- raw_data %>%
  dplyr::select(vote_2020,
                age,
                gender,
                race_ethnicity,
                census_region,
                language
  )
```

```r
clean_data <- select_data %>%
  na.omit() %>%
  filter(vote_2020 %in% c("Joe Biden", "Donald Trump"))
```

```r
clean_data <- clean_data %>%
  mutate(age_group = cut(age,
                         breaks = c(18, 30, 40, 50, 60, 100),
                         right = FALSE,
                         labels = c("18 to 29",
                                    "30 to 39",
                                    "40 to 49",
                                    "50 to 59",
                                    "60 +")
                         ),
         race_ethnicity = case_when(race_ethnicity == "Black, or African American" ~ "Black, or African
                                    race_ethnicity == "White" ~ "White",
                                    race_ethnicity != "Black, or African American" | race_ethnicity !=
         language = case_when(language == "Yes, we speak Spanish." ~ "Spanish",
                              language == "Yes, we speak a language other than Spanish or English." ~ "
                              language == "No, we speak only English." ~ "English")


  )
```

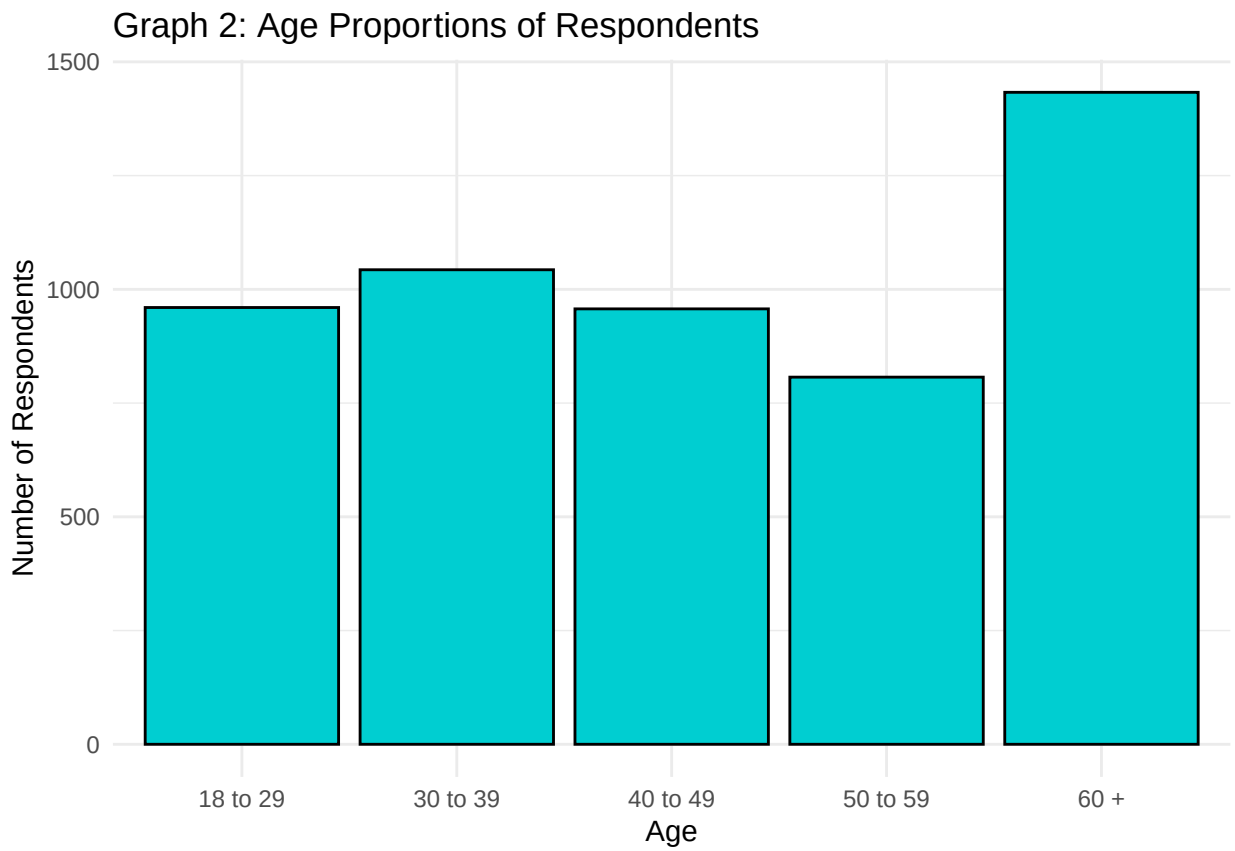
Display of Survey Data

```r
clean_data %>%
  ggplot(aes(x = vote_2020)) +
  geom_bar(colour="black", fill = 'blue') +
  labs(title = "Graph 1: Vote Decisions of Respondents",
       x = "Vote Choice",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



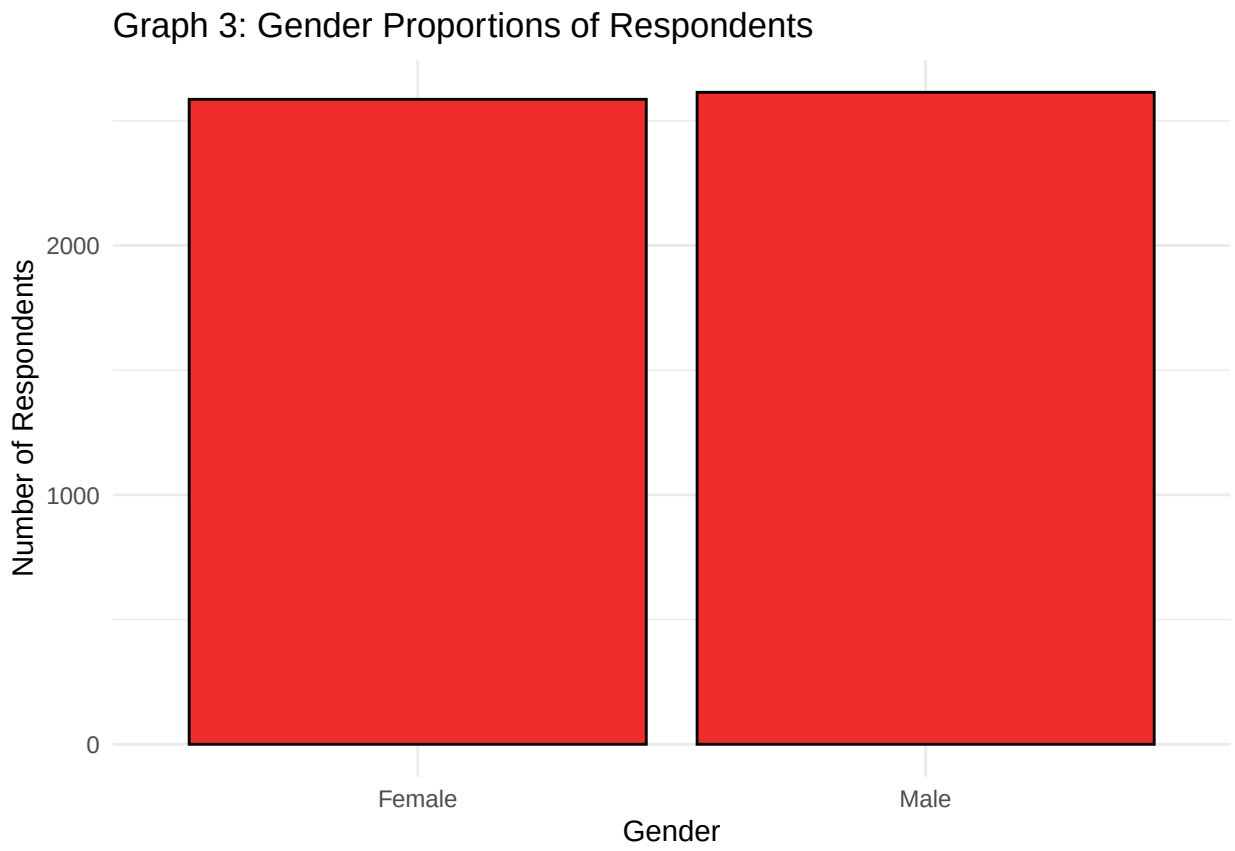
VOTE 2020: In our first plot, Graph 1, we display the results of which candidate respondents are likely to vote for in the 2020 election. We have isolated the data for this plot to display only Donald Trump or Joe Biden voters as these are the only two probable winners of the election and this is what we will base our forecasting on. As we can see, Biden voters slightly outweigh Trump voters, coming in with a proportion of 52% compared to 48% respectively. This can give us an initial idea of what party voters are currently favouring and how we can expect to see the popular vote split up. The small proportional difference between candidates suggests that the presidential race will not be easily won by either party and forecasting the outcome will require thorough analysis.

```
clean_data %>%
  ggplot(aes(x = age_group)) +
  geom_bar(colour = "black", fill = "darkturquoise") +
  labs(title = "Graph 2: Age Proportions of Respondents",
       x = "Age",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



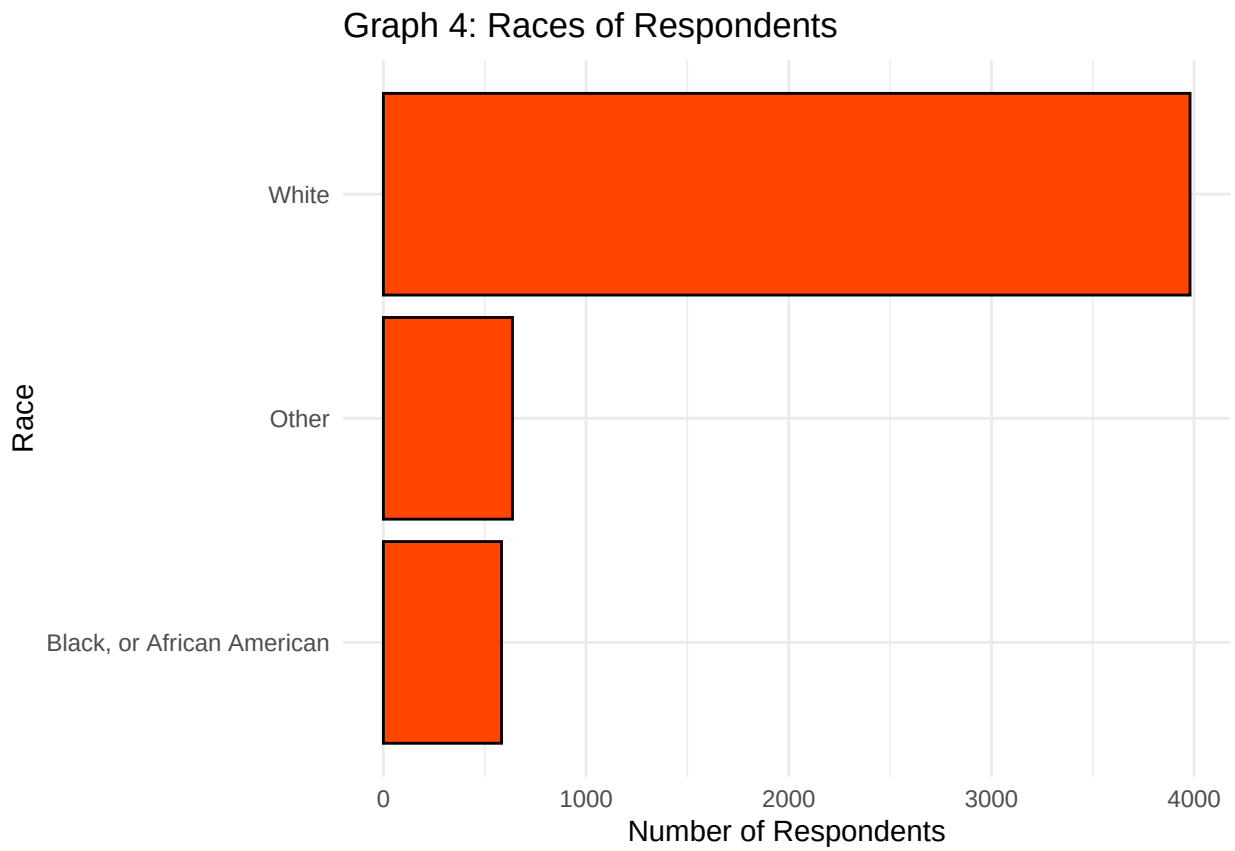
AGE: Graph 2 displays the age distribution of respondents to the survey. It appears that the highest number of respondents fall into the age group 60 + with the other groups holding similar proportions at a lower level. This is likely a result of how we decided to group our age data as we chose the cap to be 60 +, this leaves a much larger gap for ages than our other 10 year groups. Due to this, this result was to be expected however, it may possibly result in a high standard error in our prediction for this group as it may contain too high of a variation of voters.

```
clean_data %>%
  ggplot(aes(x = gender)) +
  geom_bar(colour = "black", fill = "firebrick2") +
  labs(title = "Graph 3: Gender Proportions of Respondents",
       x = "Gender",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



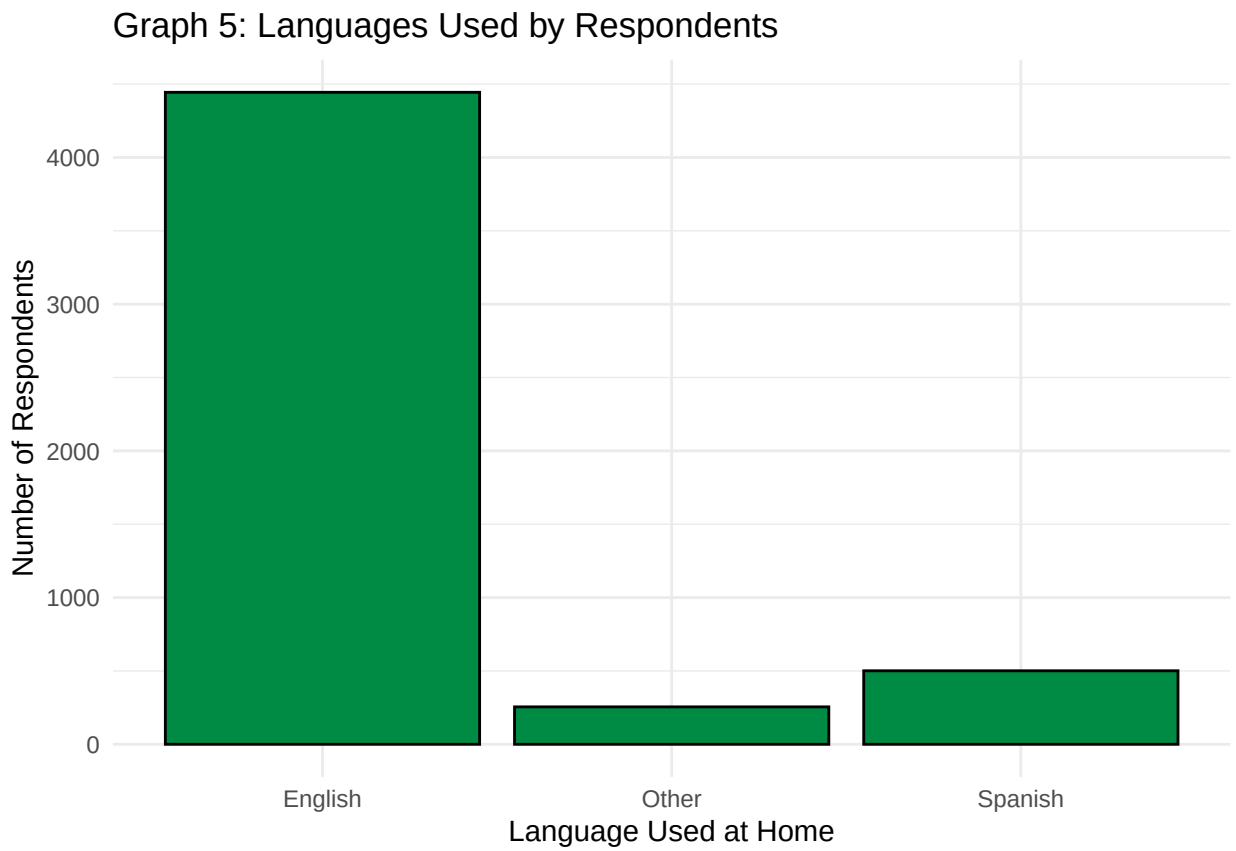
GENDER: From Graph 3, we can see an almost exactly even split between Female and Male respondents with proportions of approximately 50% each and a true difference of about 0.5%. This is not surprising at all considering the current gender proportions in the US were used for the weighting of the response data for this survey. As a result, our model and results should not be disturbed by this distribution.

```
clean_data %>%
  ggplot(aes(x = race_ethnicity)) +
  geom_bar(colour = "black", fill = "orangered1") +
  labs(title = "Graph 4: Races of Respondents",
       x = "Race",
       y = "Number of Respondents"
       ) +
  coord_flip() +
  theme_minimal()
```



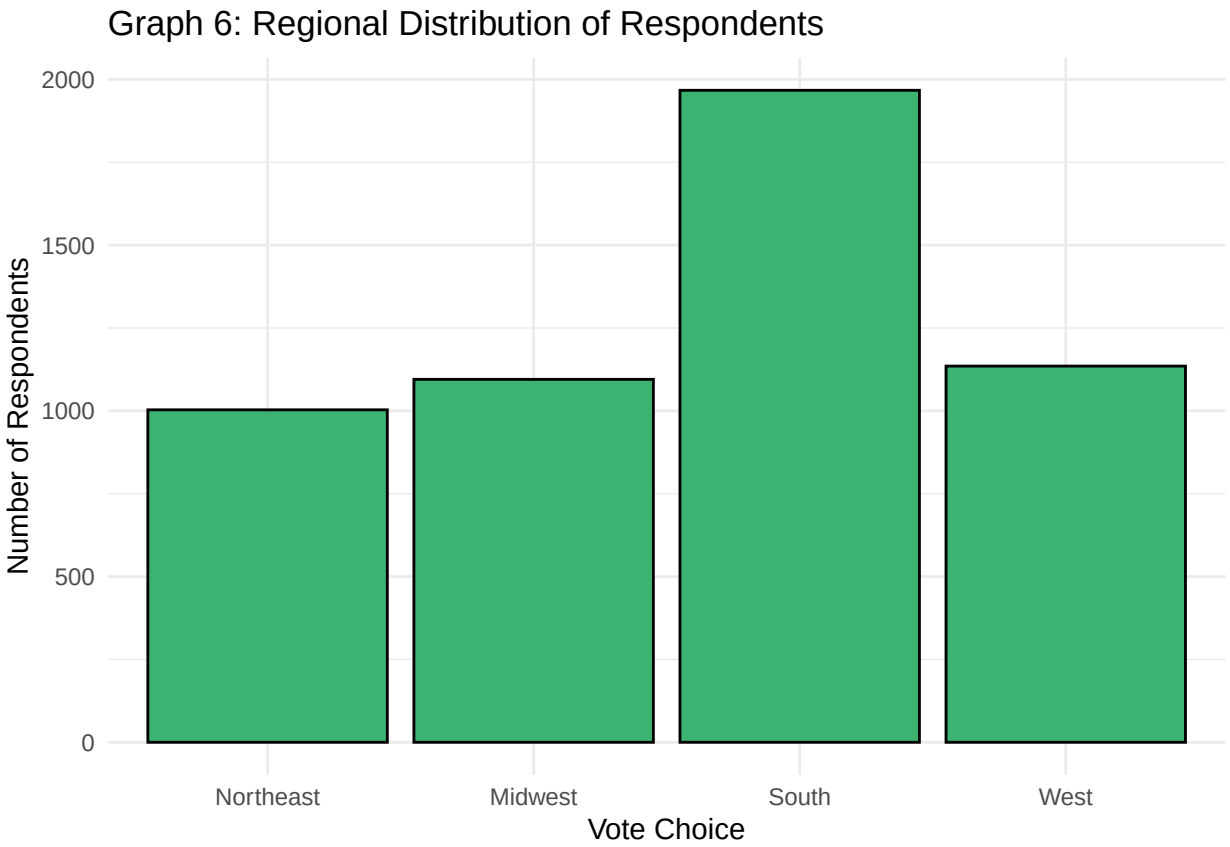
In Graph 4, we display the proportion of different races in our data. These races have been categorized as; White, Black or African American, or Other. These groups were made with consideration to the most popular races in the US. It is extremely clear that the proportion of white voters greatly outweighs all other options with a total proportion of almost 76%. Considering the current distribution of races in the US this is not surprising at all as the country has an extremely high population of white citizens compared to all other races.

```
clean_data %>%
  ggplot(aes(x = language)) +
  geom_bar(colour = "black", fill = "springgreen4") +
  labs(title = "Graph 5: Languages Used by Respondents",
       x = "Language Used at Home",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



Graph 5 presents the distribution of household languages used by respondents to the survey. Respondents are classified as English using, Spanish using, or users of an alternative language. As is evident, the highest proportion of respondents, by an extremely large margin, use English as their primary language making up more than 85% of our data. Much like the responses to the question of race, this result was entirely expected as English is often considered the national language of the US and most of North America for that matter. Moreover, the survey was conducted in English implying that people using English as their first language would be more likely to take the survey compared to people who speak other languages. Thus we expect that this should not hinder our model's predictive performance.

```
clean_data %>%
  ggplot(aes(x = census_region)) +
  geom_bar(colour = "black", fill = "mediumseagreen") +
  labs(title = "Graph 6: Regional Distribution of Respondents",
       x = "Vote Choice",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



Lastly, in Graph 6 we display the regional distribution of respondents, dividing using the four main regions of the US; Northeast, Midwest, South, and West. There is a relatively similar proportion between the Northeast, Midwest, and West regions, each making up between 21% and 22% of our sample, however the Southern region heavily outweighs the others with a proportion of 38%. These results almost perfectly match real-world reports on US populations in each region, providing further evidence that the sample we have collected is properly representative of the US population.

ACS Survey Data

The ACS data we used in this paper were collected by U.S. Census Bureau back in 2018. The 2018 American Community Survey is a cross-sectional sampling survey of 3 million U.S. households conducted on a monthly rolling basis. This survey provides a wide range of important statistics about people and housing for every community in the United States. For example, it produces statistics for language, education, commuting, employment, mortgage status and rent, as well as income, poverty, and health insurance. Below we have displayed a preview of 2018 ACS Data.

The American Community Survey consists of two separate samples: housing unit (HU) addresses and residents of group quarters(GQ) facilities. There are two phases of HU address sampling for each area. During first-phase sampling, Bureau assign blocks to sampling strata, calculate sampling rates, and select the sample. During the second phase of sampling, they select a sample of nonresponding addresses for Computer Assisted Personal Interviewing (CAPI). The data were finally collected in continuous, 3-month cycles using a combination of mailout/mailback, Computer Assisted Telephone Interviewing(CATI), and Computer Assisted Personal Interviewing (CAPI) data collection modes. The sampling frames were derived from the Census Bureau's Master Address File (MAF). Since the ACS frame must be as complete as possible, the Census Bureau applies filtering rules during the creation of the 2018 ACS extracts to minimize both over-

coverage and undercoverage and to obtain an inclusive listing of addresses. For example, the 2018 ACS filter rules include units that represent new construction units, some of which may not exist yet. The 2018 ACS also includes other housing units that are not geocoded, which means that the address is one that has not been linked to a census tract and block yet. In addition, the 2018 ACS includes units that are "excluded from delivery statistics" (EDS); these units often are those under construction. Also, to protect individual confidentiality, geographic identifiers are currently restricted to the state level, and individual variables, such as income and housing values, are top coded.

In addition, the non-response problem was well handled. The ACS made every effort to minimize unit nonresponse, and thus, the potential for nonresponse error. First, the ACS used a combination of mail and CAPI data collection modes to maximize response. The mail phase included a series of three to four mailings to encourage housing units to return the questionnaire. Subsequently, a subsample of the mail nonrespondents were contacted by personal visit to attempt an interview. Combined, these efforts resulted in a very high overall response rate for the ACS.

However, most of the variables needed to be understood in the ACS dictionary, so we reduced the number of variables we were interested in down to 18. In this paper, we regrouped our interested variables to match with survey data, for example, we regrouped age variable into "18 to 29", "30 to 39", "40 to 49", "50 to 59","60 +" to match with survey data.(more detail in appendix)These are the variables we are going to focus on : age, sex, race, language and region and our objective is to determine the relationship between these five variables and voters' decision of the 2020 US presidential election.

To allow this data to enter our regression function we matched up the question and response options from this ACS dataset to those in our original survey Dataset. Thus, all groups for these graphs will appear the exact same except with slightly different distributions.

Upon examining our organized dataset, we see that there are many NA values, especially in the responses to age. This is largely due to the fact that there are over 500,000 respondents under the age of 18 that our age groups were not designed to include. As voting age in the US is 18 years and above, we are able to remove these respondents from our dataset without it harming the real-world interpretations of our forecasting.

```
raw_strat_data <- read_dta("usa_00002.dta")
```

```
raw_strat_data <- labelled::to_factor(raw_strat_data)
```

```
raw_select_strat_data <- raw_strat_data %>%
  select(age,
         sex,
         race,
         hispan,
         bpl,
         region,
         language,
         empstat

  )
```

```
strat_data <- raw_select_strat_data %>%
  na.omit()
```

```
clean_strat_data <-
  strat_data %>%
  select(region,
         sex,
         age,
```

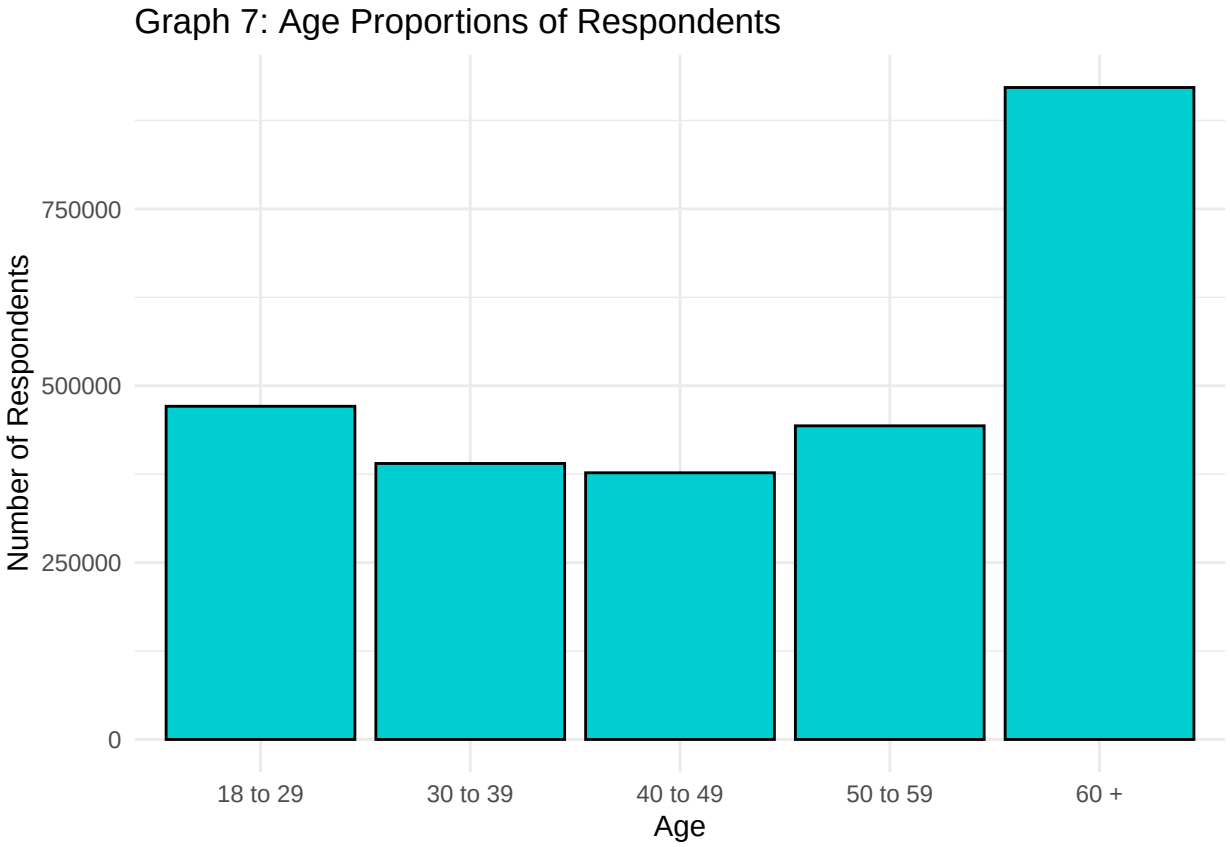
```
        race,
        bpl,
        language) %>%
  rename(gender = sex,
        race_ethnicity = race,
        census_region = region,
        foriegn_born = bpl)
```

```
clean_strat_data <- clean_strat_data %>%
  mutate(age_group = cut(as.numeric(age),
                         breaks = c(18, 30, 40, 50, 60, 100),
                         right = FALSE,
                         labels = c("18 to 29",
                                    "30 to 39",
                                    "40 to 49",
                                    "50 to 59",
                                    "60 +")
                         ),
         census_region = case_when(census_region == "new england division" ~ "Northeast",
                                   census_region == "middle atlantic division" ~ "Northeast",
                                   census_region == "east north central div" ~ "Midwest",
                                   census_region == "west north central div" ~ "Midwest",
                                   census_region == "south atlantic division" ~ "South",
                                   census_region == "east south central div" ~ "South",
                                   census_region == "west south central div" ~ "South",
                                   census_region == "mountain division" ~ "West",
                                   census_region == "pacific division" ~ "West"
                                   ),
         language = case_when(language == "english" ~ "English",
                              language == "spanish" ~ "Spanish",
                              language != "english" | language != "spanish" ~ "Other"
                              ),
         race_ethnicity = case_when(race_ethnicity == "black/african american/negro" ~ "Black, or Africa
                                    race_ethnicity == "white" ~ "White",
                                    race_ethnicity != "black/african american/negro" | race_ethnicity !=
                                    ),
         gender = case_when(gender == "male" ~ "Male",
                            gender == "female" ~ "Female")




  )
```

```
# Remove Data that doesn't fit the groups
clean_strat_data <- clean_strat_data %>%
  na.omit() %>%
  labelled::to_factor()
```

Display of ACS Survey Data

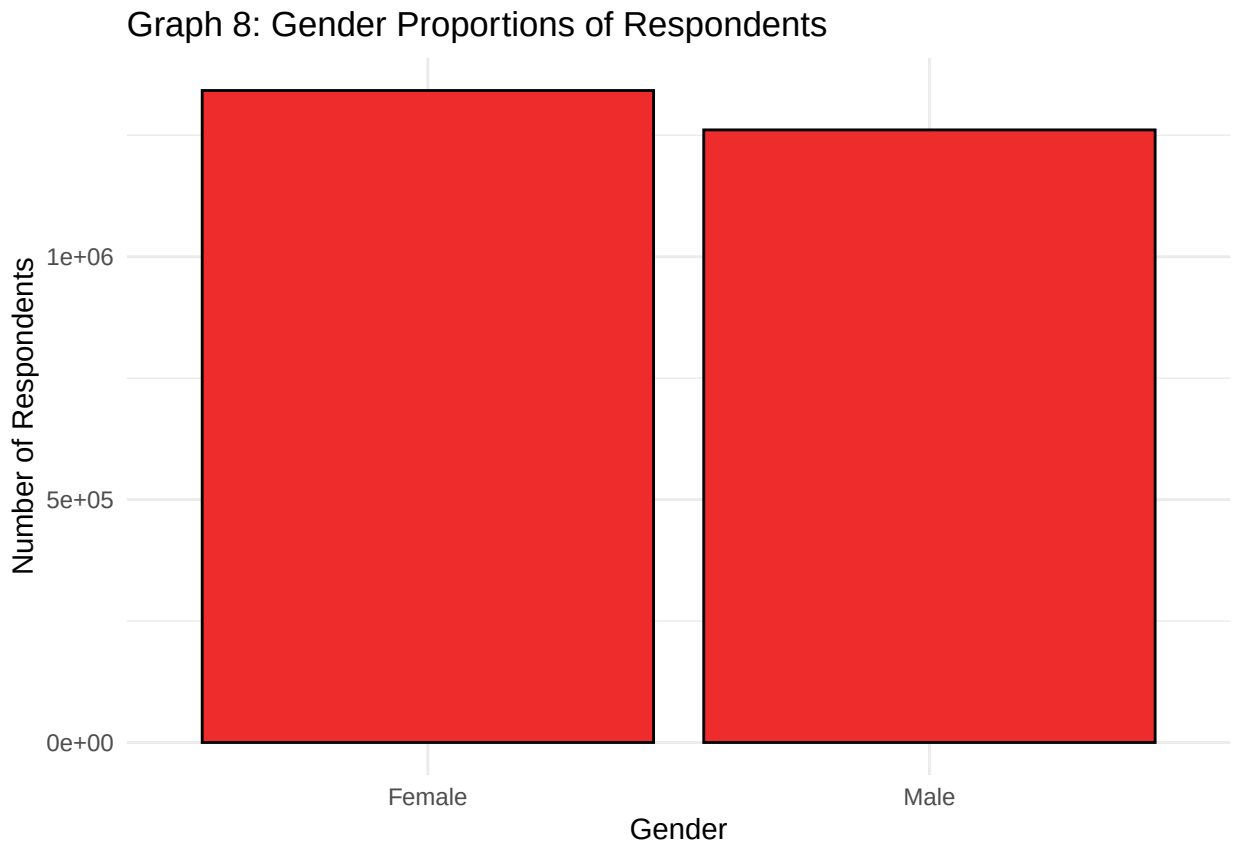
```
clean_strat_data %>%
  ggplot(aes(x = age_group)) +
  geom_bar(colour = "black", fill = "darkturquoise") +
  labs(title = "Graph 7: Age Proportions of Respondents",
       x = "Age",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



AGE: Graph 7 plots the distribution in age groups we see across our data. Much like our survey data we see an extremely large proportion of respondents fall into the 60 + age group totaling over 35%. We can see the remaining groups making up much smaller proportions between 14% and 18% each. For the same reasons mentioned previously, this result was to be expected and we expect this to have the same possible result of standard error in our predictions that we must pay attention to. However, according to the latest statistic, people over the age of 65 make up 16 percent of the total population in the United States, and our data show that people over the age of 60 make up about 35 percent. This may cause some disruptions to our model's strength as older respondents tend to vote for Biden. As a result of this, we may see a slight overestimation of the strengths of this predictor.

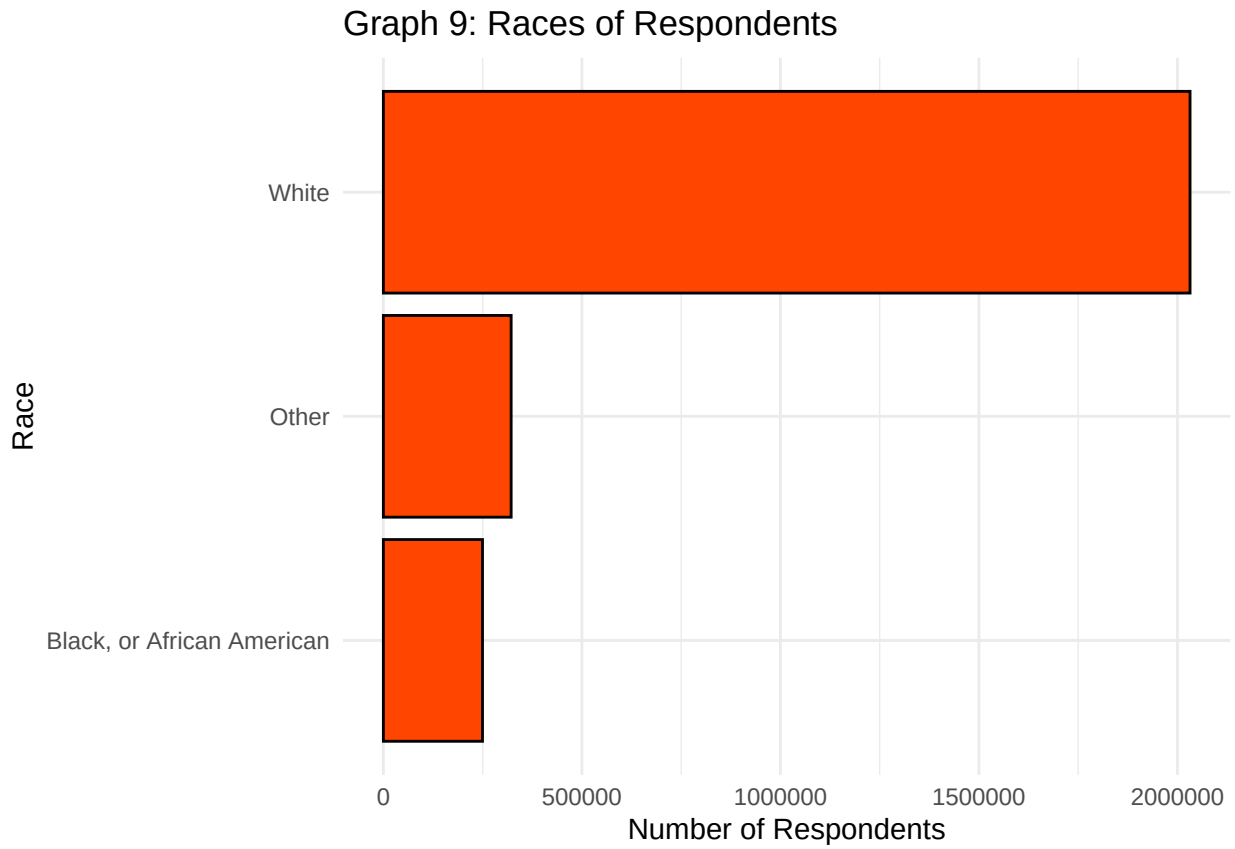
```
clean_strat_data %>%
  ggplot(aes(x = gender)) +
  geom_bar(colour = "black", fill = "firebrick2") +
  labs(title = "Graph 8: Gender Proportions of Respondents",
```

```
      x = "Gender",
      y = "Number of Respondents"
      ) +
  theme_minimal()
```



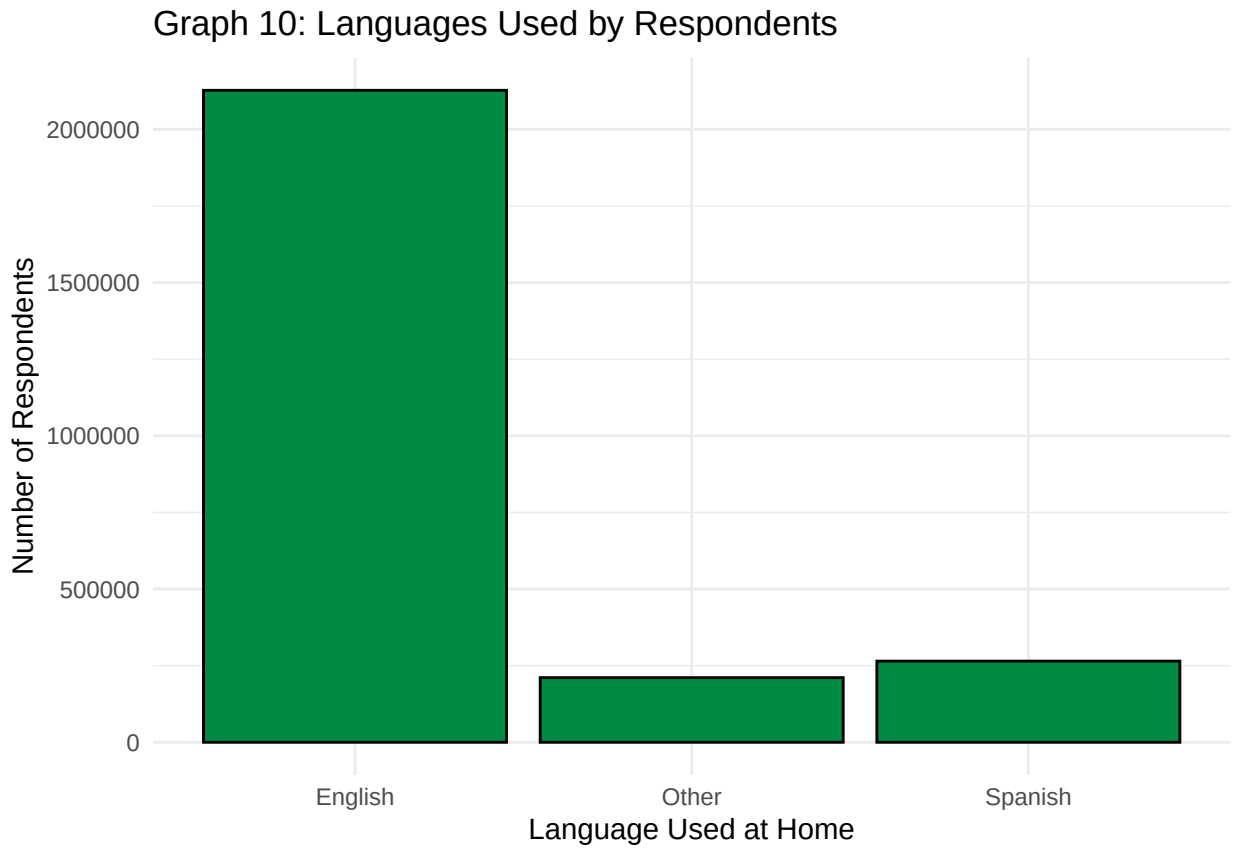
GENDER: In Graph 8, we display the proportion of each gender in our dataset. Again, we see an extremely close difference between Female and Male respondents, making up 52% and 48% of our dataset, respectively. The slightly higher proportional difference we see is most likely due simply to a variation in individuals contacted in this survey compared to our original data. As a result, we do not believe that this should cause any problem with our model's performance.

```
clean_strat_data %>%
  ggplot(aes(x = race_ethnicity)) +
  geom_bar(colour = "black", fill = "orangered1") +
  labs(title = "Graph 9: Races of Respondents",
       x = "Race",
       y = "Number of Respondents"
       ) +
  coord_flip() +
  theme_minimal()
```



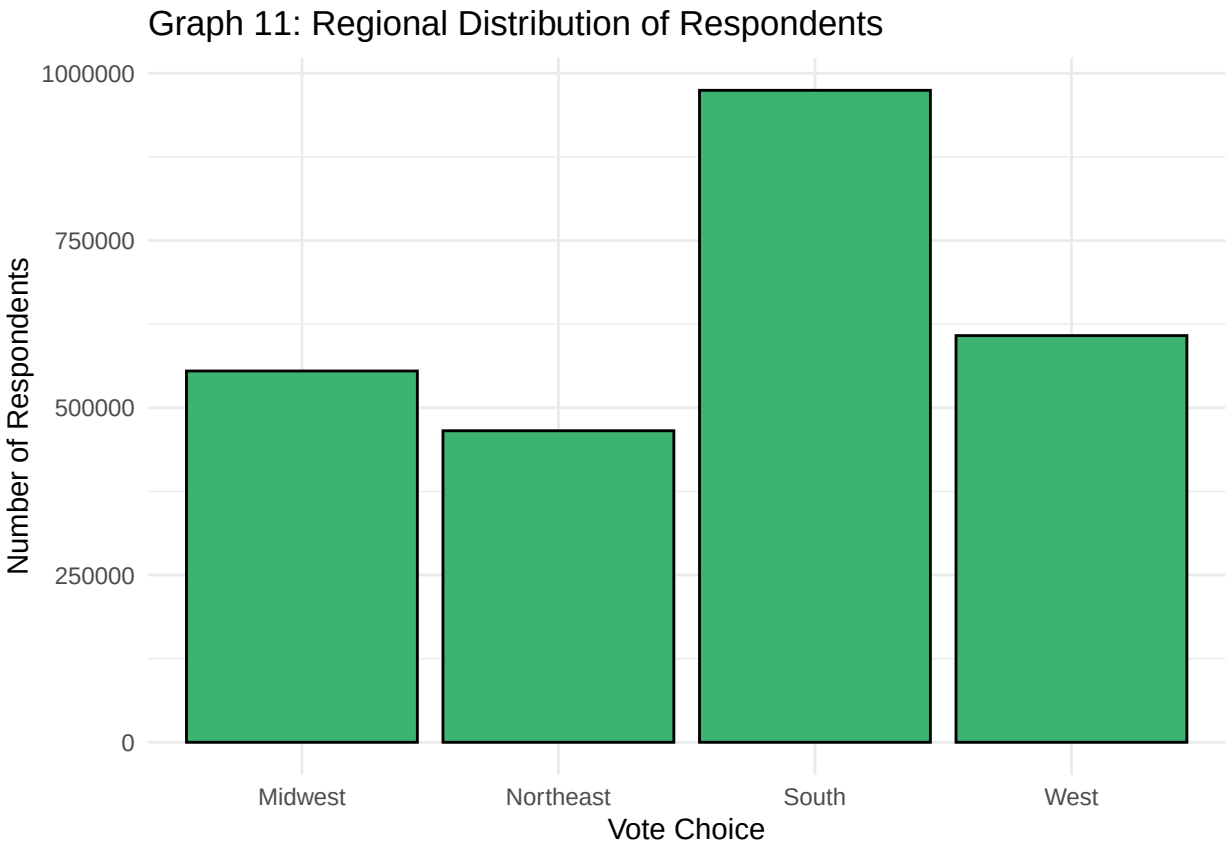
RACES: From Graph 9, we see an extremely similar result in the race distribution of respondents to that of our original data. With White respondents making up over 78% of our data we are not surprised at all by this result given the real distribution of races in the US today and therefore do not believe that it will create any problems with our regression analysis.

```
clean_strat_data %>%
  ggplot(aes(x = language)) +
  geom_bar(colour = "black", fill = "springgreen4") +
  labs(title = "Graph 10: Languages Used by Respondents",
       x = "Language Used at Home",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



LANGUAGES: Much like the results seen in our original data, Graph 10 displays the proportions of household languages used by our respondents and signifies an extremely large proportion of English users. Making up almost 82% of our data this result is as exactly as expected. Examining Graph 9 again, we can see that the low proportions of other races seen in the US are the likely cause of this significant difference.

```
clean_strat_data %>%
  ggplot(aes(x = census_region)) +
  geom_bar(colour = "black", fill = "mediumseagreen") +
  labs(title = "Graph 11: Regional Distribution of Respondents",
       x = "Vote Choice",
       y = "Number of Respondents"
       ) +
  theme_minimal()
```



REGION: Finally, in Graph 11, we have displayed the regional distribution of our respondents. As we can see, the large proportion of southern voters seen in our original survey dataset is present here as well with the other regions making up similarly smaller proportions. Again, Southern voters make up a total of 37% of this data exactly as expected given the regional distribution seen today.

Now that we have viewed all our our predictor variables and confirmed that there were no troubling observations in the data, we are ready to move on to developing our model and begin to make forecasts for the election based on this.

Model Discussion and Development

Regression Model

For the purposes of this study, we will be using a binomial logistic regression model to both analyze the Nationscape survey data and forecast the election results using the ACS data. To accomplish these tasks we will be using R to carry out our analyses. A binomial logistic regression is a type of generalized linear model of the binomial family. Generalized linear regression models are generalizations of linear models that don't include the same harsh requirements such as a normally distributed response variable. This model allows us to link our response variable to a specific distribution function known as the model's family. In this case, the binomial distribution, being a series of Bernoulli distributions, requires a response variable with binary outcomes; 0 or 1, heads or tails, true or false, etc.

Logistic Regression models are primarily used to predict the odds of an event occurring, given the inputs to the model's predictor variables. Due to the logistic model's link function logit, these odds are interpreted as the log-odds of seeing our event occur, expressed as the log ratio of the probability of success to the

probability of failure. As an example, we can use logistic regression to determine the odds of a person having a heart attack, given factors like their age, gender, and medical history.

We made this choice largely because our response variable, the respondent's intended vote in the 2020 election, is an option between two candidates, which a linear regression model is not equipped to handle as this requires numerical and continuous response data. In addition, the predictor variables we have chosen are categorical and do not follow any of the strict criterion of linear models, thus making it an incompatible option for this study. In general, linear regression models are often ineffective for real-world analysis as their requirements such as homoscedasticity and a normally distributed response variable are rather unrealistic. As this is the case with the data we are studying here, a generalized linear regression model appears to be our best option. Moving to the decision on the distribution family, this choice depends primarily on the format and distribution of our response variable. Since the 2020 election will likely be won by one of two parties, our response variable has been set up as an option between Donald Trump and Joe Biden. As this can easily be converted into a Bernoulli-like outcome, with 1 representing a vote for Trump and 0 representing a vote for Biden, the binomial distribution stands out as the clear winner for the family of our regression model.

```
survey_glm <- glm(as.numeric(vote_2020 == "Donald Trump") ~
                    age_group +
                    gender +
                    census_region +
                    race_ethnicity +
                    language,
                  family = 'binomial',
                  data = clean_data
)
```

```
survey_glm_plot <- glm(vote_2020 ~
                    age_group +
                    gender +
                    census_region +
                    race_ethnicity +
                    language,
                  family = 'binomial',
                  data = clean_data
)
```

Model Validation

Using a 10-fold cross validation analysis, we find a cross validation estimate of prediction error of 0.22. More simply, when we split our survey data into 10 groups and cross validate 10 times using each of these subsamples as a testing set, we compute an average prediction error 0.22. Taking into account that we are on a log-odds scale, this value means that our model's predictive strength will not be extremely strong as errors can seriously sway final results. THere are a number of possible causes for this such as a conservative number of explanatory variables and likely the generally complex set of factors that determine an individual's political preference. With future analysis this value can be brought down to allow for more precise predictions, however, our analysis will still provide important information regarding a person's voting choice and the likely outcome of the 2020 US election.

```
test_data_MSPE <- clean_data %>%
  mutate(id = row_number())

test_data_msres <- clean_data %>%
  mutate(id = row_number())
```

```r
# Splicing Data into Testing and Training Set
set.seed(5799893)
glm_test_MSPE <- sample_n(test_data_MSPE, 1295)
glm_test_msres <- test_data_msres %>%
  anti_join(test_data_MSPE)
```

```
## Joining, by = c("vote_2020", "age", "gender", "race_ethnicity", "census_region", "language", "age_gr
```

```r
#Model validation, comparing MSPE and MSres
# fitting sample dataset to our model
glm_test1 <- glm(as.numeric(vote_2020 == "Donald Trump") ~
                   age_group +
                   gender +
                   census_region +
                   race_ethnicity +
                   language,
                 family = 'binomial',
                 data = test_data_msres)

glm_test2 <- glm(as.numeric(vote_2020 == "Donald Trump") ~
                   age_group +
                   gender +
                   census_region +
                   race_ethnicity +
                   language,
                 family = 'binomial',
                 data = test_data_MSPE)

#calculation of the mspe for sample dataset
mspe = sum(resid(glm_test2)^2)/ length(test_data_MSPE$vote_2020)
#calculation of the MSres for our main dataset
# 13 represent the degrees of freedom in our model
msres = sum(resid(glm_test1)^2)/(length(test_data_msres$vote_2020) - 13)
mspe
```

```
## [1] 1.260798
```

```r
msres
```

```
## [1] 1.263957
```

Next to further validate our model we tested to find the MSPE and MSRes of our model. We decided that the 20% of the data set should be to calculate the MSPE, which was 1040 observations and the remaining 4160 were used to calculate the MSRes. Through running calculations in R, we found the MSPE to be 1.26, when compared to the MSres from our model, at 1.26 as well, we find that true the difference is smaller than 0.04. An MSPE value that is close to the MSRes value based on the regression fit to the model-building data set, indicates a high predictive ability of our model. Thus, this incredibly small difference helps to prove the validity of our model selection being a binomial regression and its strength of prediction.

Multilevel Modelling

Multilevel Modelling With Post-Stratification Benefits: With multilevel modelling we can make use of non probability sampling by using something like a market for respondents such as _____ which was used by the

Nationscape team. By doing this, we can cut out the higher cost of probability sampling and use smaller sample sizes to explain the results of larger data with conducting additional surveys. In addition, the use of post-stratification allows for us to gain information on underrepresented groups. If we had a subpopulation that were underrepresented such as a specific city, we can use post stratification to gain an understanding of how the overall country feels and use this as a way of predicting the responses in these underrepresented groups. Cons: On the other hand, this requires us to have well defined cells with sufficient proportions in each so that we can make accurate forecasts. As is often the case with real world data, we run into situations where our cell structures may result in severely underrepresented groups that we must account for.

Multilevel modeling is a method used in regression analysis to estimate the outcome of a target population's response through a regression model developed using data from a sample population. The process is as follows: First we use our sample data to generate a regression model that can estimate the relationship between our response and predictor variables. We then post-stratify our target population data by grouping the individual responses into cells. Cells are essentially broad categorical identifiers we can use to group our data. Referring back to our heart attack example, we would first group people by their genders, and then as a subcategory we would group them by their age groups. This would leave us with the counts of; males 18 to 29 years old, females 18 to 29 years old, males 30 to 39 years old, etc. and each of these groups are referred to as 'cells'. These cells are mutually exclusive and organized to represent the entire population studied. Using the responses from each of these cells, we plug these values into our regression model which in turn provides us log odds of these groups voting for Trump.

With a sufficiently significant regression model based on sample data, this can allow us to make predictions and forecasts for our target population without the need for us to conduct additional surveys.

```
summary(survey_glm)
```

```
##
## Call:
## glm(formula = as.numeric(vote_2020 == "Donald Trump") ~ age_group +
##     gender + census_region + race_ethnicity + language, family = "binomial",
##     data = clean_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5721  -1.1342  -0.4239   1.0698   2.5235
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.81474    0.16246 -17.326  < 2e-16 ***
## age_group30 to 39     0.46524    0.09867   4.715 2.42e-06 ***
## age_group40 to 49     0.65329    0.10122   6.454 1.09e-10 ***
## age_group50 to 59     0.60524    0.10560   5.731 9.96e-09 ***
## age_group60 +         0.44729    0.09417   4.750 2.03e-06 ***
## genderMale            0.42799    0.05983   7.153 8.50e-13 ***
## census_regionMidwest  0.10854    0.09246   1.174 0.240432
## census_regionSouth    0.45023    0.08342   5.397 6.77e-08 ***
## census_regionWest     0.03343    0.09170   0.365 0.715460
## race_ethnicityOther   1.57878    0.15921   9.916  < 2e-16 ***
## race_ethnicityWhite   2.17563    0.13463  16.159  < 2e-16 ***
## languageOther        -0.17940    0.14665  -1.223 0.221215
## languageSpanish      -0.36031    0.10577  -3.406 0.000658 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 7197.8  on 5199  degrees of freedom
## Residual deviance: 6556.1  on 5187  degrees of freedom
## AIC: 6582.1
##
## Number of Fisher Scoring iterations: 4
```

```
summ(model = survey_glm)
```

```
## MODEL INFO:
## Observations: 5200
## Dependent Variable: as.numeric(vote_2020 == "Donald Trump")
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>²(12) = 641.69, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.15
## Pseudo-R² (McFadden) = 0.09
## AIC = 6582.15, BIC = 6667.38
##
## Standard errors: MLE
## ------------------------------------------------------------
##                              Est.   S.E.   z val.      p
## ------------------------- ------- ------ -------- ------
## (Intercept)                -2.81   0.16   -17.33   0.00
## age_group30 to 39           0.47   0.10     4.71   0.00
## age_group40 to 49           0.65   0.10     6.45   0.00
## age_group50 to 59           0.61   0.11     5.73   0.00
## age_group60 +               0.45   0.09     4.75   0.00
## genderMale                  0.43   0.06     7.15   0.00
## census_regionMidwest        0.11   0.09     1.17   0.24
## census_regionSouth          0.45   0.08     5.40   0.00
## census_regionWest           0.03   0.09     0.36   0.72
## race_ethnicityOther         1.58   0.16     9.92   0.00
## race_ethnicityWhite         2.18   0.13    16.16   0.00
## languageOther              -0.18   0.15    -1.22   0.22
## languageSpanish            -0.36   0.11    -3.41   0.00
## ------------------------------------------------------------
```

From this output, we can determine the formula for our final regression model:

log(Y/(1-Y)) = -2.81 + 0.47A1 + 0.65A2 + 0.61A3 + 0.45A4 + 0.43G1 + 0.11C1 + 0.45C2 + 0.03C3 + 1.58R1 + 2.18R2 - 0.18L1 - 0.36L2

Where:

Y represents the probability of an individual casting their vote for Trump in the upcoming election and -2.81 represents our intercept value.

The reason these groups appear to be missing one category is due to the way they were interpreted by our regression model. For each group in a predictor variable, their estimate value is computed in relation to a conditional group selected by our model where this value states the difference in log odds between these two groups. Thus, these predictor values take on a 0 or a 1 depending on which group a respondent is in and when all predictors are set to 0 this represents a respondent in the conditional group.

As such, Ai is an indicator variable representing the age group of a respondent, conditioned on the age group 18 to 29: A1 corresponds to those aged 30 to 39, A2 to those aged 40 to 49, A3 to those aged 50 to 59, and A4 representing those aged 60 and over.

Gi is an indicator variable representing the gender of a respondent, conditioned on female respondents: G1 simply corresponds to respondents that are male.

Ci is an indicator variable representing the region of a respondent, conditioned on the Northeastern region: C1 corresponds to those in the Midwestern region, C2 to those in the Southern region, and C3 to those in the Western region of the United States.

Ri is an indicator variable representing the race of a respondent, conditioned on black or African American individuals: R1 corresponds to those of races other than black or white, and C2 corresponds to white respondents.

Lastly, Li is an indicator variable representing the home language used by a respondent, conditioned on those using English: L1 corresponds to those using a language other than Spanish or English, and L2 corresponds to those using Spanish as their primary language.

Results

Regression Model Results

Observations In Table ***, we have displayed a table of summary statistics using*** _____ which includes the predictor's; estimate, standard error, z-value, and p-value. Utilizing these values, we can determine the strength and significance of our predictor variables on _____. The estimate value tells us the change in the log-odds of the respondent's vote going toward Donald Trump, given their response to the question and the standard error will tell us the expected error we will see in this estimate value. The z-value and p-value work together to tell us whether or not we can reject the null hypothesis that our estimate value truly zero, this allows us to determine the significance of our predictions_____. To conform with a 95% confidence interval, we are looking for z-values with a magnitude greater than 1.96, and p-values with values smaller than 0.05. Interpret Rules To interpret the estimate values we must realize that they are set up as categorical variables and not like our usual continuous or count data. This means that for each predictor, the estimates are conditioned on a specific response and their values indicate the difference we expect to see from this conditional response. Agee_group Estimates For this variable, the response we use as conditional is the age group 18 to 29. As we can see, the values of all estimates are positive which means that the younger voters in the 18 to 29 year age group are the least likely to vote for Trump. It appears that the age group 40 to 49 has the highest log odds of voting for Trump with an estimate of _____. Respondents aged 50 to 59 have a slightly lower probability of voting for Trump with an estimate of _____. Lastly, respondents that are 30 to 39 or 60 + have considerably lower likelihoods of voting for Trump with estimates of _____ and _____ respectively. Each of these values, however, are well above 0 suggesting that respondents aged 18 to 29 have a much lower likelihood of voting for Trump in the upcoming election. P and z-values In all categories we see our requirements met by the z and p values of this predictor thus enabling us to reject the null hypothesis and confirm their significance to our model. Gender Estiamtes For this variable the conditional group is female respondents . With a value of _____ we can determine that male voters are much more likely to vote for Trump in the election than female respondents P and z-values With values of _____ and _____, the p value and z value both satisfy our requirements meaning that this predictor is significant in determining the intended vote of the respondent. Region Estimates The census region predictor variable uses the Northeast region of the US as conditional. As shown, all categories have positive estimates suggesting that respondents living in the Northeastern region of USA have the lowest odds of giving a vote to Trump. Respondents living in the Southern region displayed the highest probability of casting a vote for Trump by a wide margin with an estimate of _____. Comparatively, voters in the Midwestern and Western regions had estimates of _____ and _____ respectively, suggesting that a sizeable group of these individuals will still lean toward

a Biden vote in the election. P and z values For the Southern region, both the z and p values satisfy our requirements with a 95% confidence interval which allows us to reject the null hypothesis and confirm its significance. However, both the Midwestern and Western regions fail to satisfy these tests, thus preventing us from confirming their significance with certainty.

Race Estimates The conditional group used for race is Black or African American. From the estimate values we can determine that Black or African American voters have a considerably lower probability of voting for Trump than respondents of other races. We can see that White voters have the highest probability of casting a vote for Trump with an estimate of _____ and those of other races have a slightly lower probability at an estimate of _____

P and z-values The p and z values determined for each category by our model satisfy both requirements necessary and allow us to confirm that their estimated values are significant.

Language Estimates In our language variable, the conditional group we have used are respondents that speak primarily English in their households. With both Spanish users, and users of an alternative language, their estimate values are negative implying that those who use English as their main language are the most likely to cast a vote for Trump out of these categories. Spanish users come in with an estimate of _____ with alternative language users have an estimate of _____. This suggests that those who use Spanish as a main language are the least likely to end up voting for Trump in the coming election.

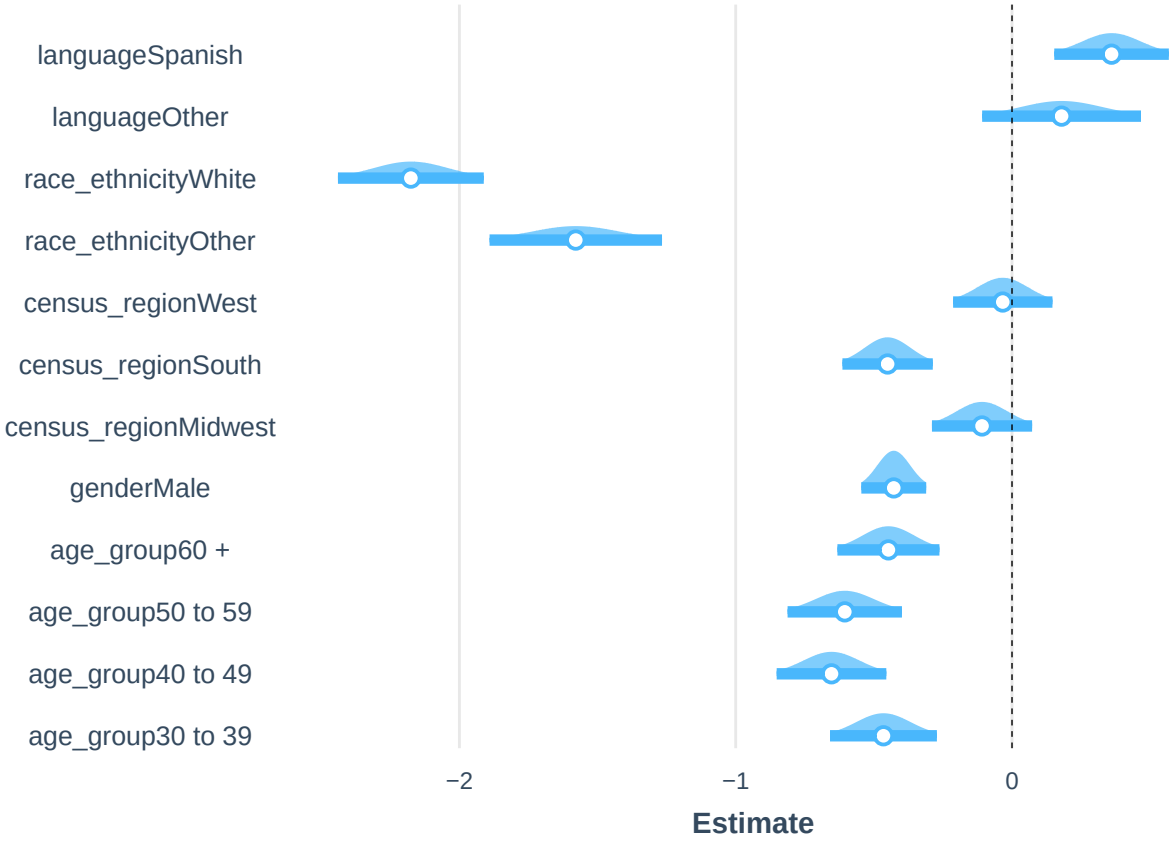
P and z values As we can see, for Spanish speakers, the p-value and z-value both pass our tests and let us confirm this category's significance to our model. On the other hand, the ____ values for the group of other languages did not satisfy either of these requirements and thus prevent us from rejecting its null hypothesis.

```
plot_summs(survey_glm_plot, plot.distributions = TRUE, inner_ci_level = .95)
```

```
## Loading required namespace: broom.mixed
```

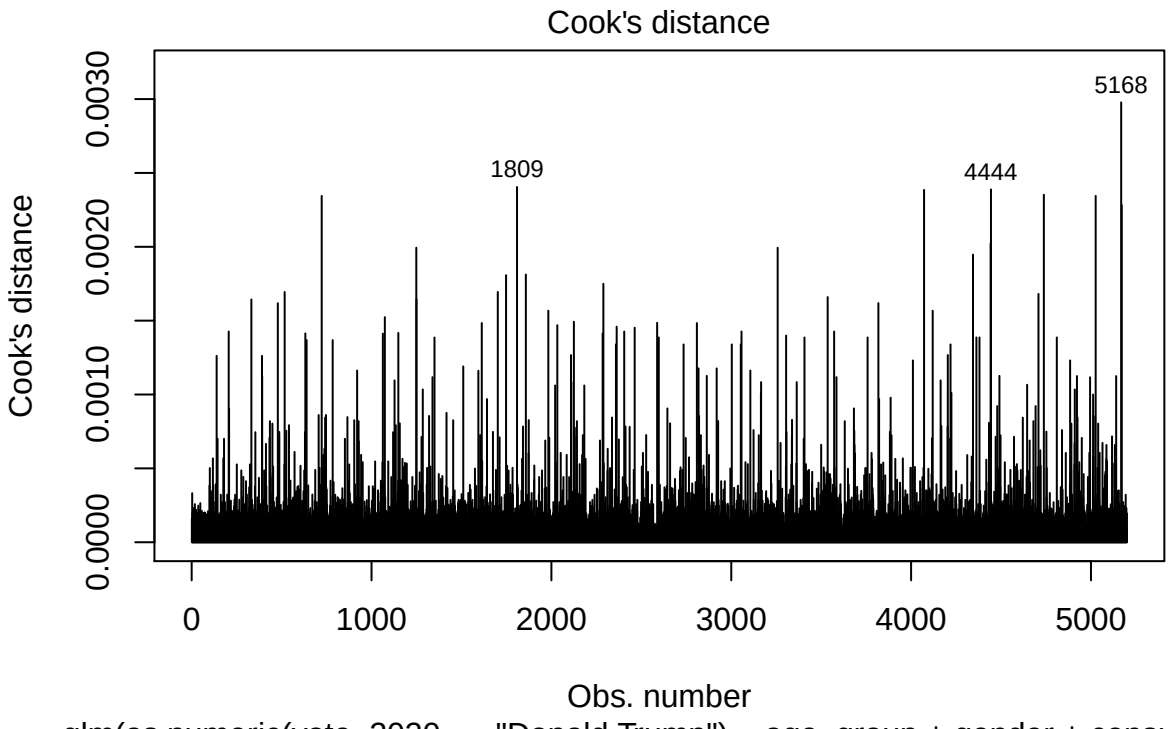
```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.2.18
## Current Matrix version is 1.2.17
## Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRAN for a
```

```
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```



Graph _____ provides us with a visualization of these estimates as well as the distribution we expect around these estimates and their 95% confidence intervals. The plot shows the log-odds estimates of each groups likelihood of voting for Biden. As we can see, just like our regression model suggested, white voters have an extremely low probability of voting for Biden in the election compared to black voters and those of other races have as well seem to heavily favor Trump. This helps to express the extreme significance a person's race has on their voting preference in the US.

```
plot(survey_glm, which = 4, id.n = 3)
```



Graph ____ plots the cook's distance score of all observation seen in our dataset by the model. Cook's distance is an estimate of the influence of outlying observations in a regression analysis. When data points result in overly large errors this can negatively affect our model's predictions as they are considered to influence the overall estimate computed. To evaluate these results we look for observations with a Cook's distance greater than 0.0008. As we can see from our plot there are a large number of values above this level with upper levels reaching almost 0.003. As a result, we are likely to run into some issues with our model's accuracy here. _ This is not surprising however when we consider the wide range of personal factors that influence a person's vote. As we have only selected a few of these factors, having multiple influential outliers is something we completely expected thus we must be careful to remember this when analyzing the results of our forecast.

Forecasting Results

```
# Create Cells
cell_counts <- clean_strat_data %>%
  group_by(age_group, gender, census_region, race_ethnicity, language) %>%
  count() %>%
  mutate(proportion = n/2603150
         )
```

```
# Add Predictions to cell counts
cell_counts2 <- cbind(cell_counts, predictions$fit, predictions$se.fit)
```

```
## New names:
```

```
## * NA -> ...8
## * NA -> ...9
```

```
# First Forecasting Result
vote_pred <- cell_counts2$proportion * cell_counts2$...8
sum(vote_pred)
```

```
## [1] 0.4764565
```

```
lower_pred = cell_counts2$...8 - cell_counts2$...9
upper_pred = cell_counts2$...8 + cell_counts2$...9
cell_counts2 <- cbind(cell_counts2, lower_pred, upper_pred)
```

```
## New names:
## * NA -> ...10
## * NA -> ...11
```

```
cell_counts2 <- cell_counts2 %>% rename(
  prediction = ...8,
  StdError = ...9,
  LowerBound = ...10,
  UpperBound = ...11
)
```

```
# Errors
lower_vote_pred <- cell_counts2$proportion * cell_counts2$LowerBound
sum(lower_vote_pred)
```

```
## [1] 0.4527774
```

```
upper_vote_pred <- cell_counts2$proportion * cell_counts2$UpperBound
sum(upper_vote_pred)
```

```
## [1] 0.5001355
```

After applying our regression function to our post-stratified cells and evaluating the predictions that are made, we find a final value of _____ 0.48 with a standard error of _____ 0.02. This means that we predict, with a 2% error, that 48% of our studied survey respondents will vote for Trump in the upcoming election. If this prediction is correct, we will see Joe Biden taking the remaining 52% of our respondents. As we have discussed previously, the 2020 presidential election is most likely going to be a competition between only Donald Trump and Joe Biden due to the serious lack of votes that all other parties receive. Thus, our prediction suggests that Joe Biden will likely win the popular vote in the country. However, if we examine our standard error of 2% we can see that the race to win the popular vote will actually be quite close, giving Trump a lowest prediction of 46% of votes and upper bound of just over 50%.
Each Cell Individually Examining the predictions of each cell we can begin to gain an understanding of what groups each candidate will take in the most votes from. With the highest prediction estimate at 0.71, white male voters aged 40 to 49 living in southern USA show the highest likelihood of casting a vote for Trump. Overall we can see that all cells with a prediction above 0.5, that is, that have a higher proportion of Trump voters than Biden voters, are either white or some other race with black voters coming nowhere near this level. The highest prediction for a cell of black voters is at 0.22 coming from males aged 40 to 49 in the

southern region. Given that the only difference between this cell and the one with the highest estimate is that their race is black or African American, rather than white, we can understand just how strong of an effect race has on a person's political preference. Continuing with this method of comparative statics we find further evidence supporting race's strength as a predictor as altering any other factors from our highest estimate cell does not result in nearly the same decline in vote estimates. We can do the same thing again if it is female voters and then shift to black vs white

On the other hand, cells with the lowest estimates suggest a strong tendency to vote for Biden in the election. Examining these cells suggests that Biden has a great deal of support from black or African American voters as these groups all have estimates below 0.25. When we look at the proportions of these cells however, we can see that many of these groups are not very well represented which will result in a less impactful increase in votes as they are outweighed by the more populated groups. The cells with the highest populations appear to have estimates much closer to 0.5 with the majority of them leaning slightly in favour of Trump. This suggests these smaller groups may in fact provide enough of a push for Biden to take office as the heavily populated groups remain relatively divided. The first group that Biden appears to gain a significant amount of votes from are white females between 18 and 29 living in midwestern states. With an estimate of 0.37 and total size of 35887 respondents this group can provide a solid boost to Biden's chances of winning the popular vote. Furthermore, similar females living in other regions of the US also show significantly lower odds of voting for Trump in the election and can greatly benefit Biden's chances.

Discussion

Preamble

Predicting the results of an election is an extremely difficult task given the wide variety of factors that influence an individual's vote. There are many subtle characteristics that affect someone's political preference which cannot be accounted for when performing simple analyses on limited survey data. This, coupled with the complex system that the US uses to determine a winner, means that coming to a conclusive answer is almost impossible using purely statistical methods. Regardless, we have made a thorough attempt at forecasting the results of the 2020 election with the use of multilevel regression modelling in R.

Regression Model Discussion

Applying our regression model to our original dataset provided us with key results that explain some of the effects of a person's characteristics on their voting choice.

We found that out of all age groups, individuals between 18 and 29 years old have the lowest likelihood of voting for Trump whereas those aged 40 to 49 are the most likely to vote for him.

With Joe Biden proposing an active education plan that would increase funding for schools in low-income areas, help teachers pay off student loans and double the number of health professionals working in schools, it is no wonder that a majority of the youth intend to vote for him rather than Trump. In October 2019, Biden unveiled a plan that would cut student loan debt obligations, waiving $10,000 per year, for up to five years, for those in public service work, like teachers or members of the military. While Trump, as President, has promised to fix student loan debt, his administration has been seen rescinding a number of Obama-era policies, including those that promoted racial diversity in schools and protections for transgender students in public schools that let them use bathrooms and other facilities corresponding to their gender identities, not to mention it has repeatedly proposed ending a student loan forgiveness program for public workers. It has also rolled back two rules that were intended to hold for-profit colleges accountable. Thus, it is natural for the younger generation, who are more likely to be affected by these policies to cast their vote towards the candidate that draws a policy benefiting them.

Comparing between genders in our study, it appears that male voters favor Trump much more than female voters. This is unsurprising given the lack of incentives females have been given to vote for Trump over the past four years. For example, when Donald Trump was elected, one of the biggest promises he made during his September 2016 Pennsylvania speech was to ensure six weeks of paid leave for employed mothers whose employers were not providing them with such. Although a legislation was signed, providing 12 weeks paid paternal leave, a vast majority of women did not receive this guarantee.

Across the four regions of the United States, Trump stands to gain a sizeable amount of votes as the respondents in the Southern region showed the highest probabilities of casting their vote for him. When examining the distribution of voters across these regions in Graph ____, we can see that being the most populated region, gaining the Southern region's voters will be extremely beneficial to his odds of taking office for a second term.

Race and home language also appear to have a significant effect on a voter's decision. By a wide margin, black or African American voters have a much lower probability of voting for Trump than all other races while white voters seem to have the highest odds of favouring him.

This may primarily be due to the surge of the "Black Lives Matter" movement in mid 2020 and Trump's ineptitude approach towards it. The movement resulted in a "Defund the Police" call from the public. Although Joe Biden does not support this call, he has promised to back proposals that increase spending on social programs separate from local police budgets as well as more funding for police reforms such as body cameras and training on community policing approaches.

Referring back to the distribution of of races we again see that Donald Trump can receive many votes for this as the white population heavily outweighs the other races.

Similarly in the home language used by respondents, those using languages other than English appeared to have lower odds of a Trump vote compared to those that use English primarily. The distribution of languages used in the US once again suggests that Trump can take in many votes here as English voters showed significantly higher proportions.

Forecasting Discussion

In the end, our forecast suggests that Joe Biden will win the popular vote by only a very small margin. Most often, electoral votes in each state align with the popular vote due to the method with which electors are chosen. Thus, if there are not any inconsistencies with this belief, we will likely see Biden taking office for the next four years. Factoring in the error we expect on this forecast, we can see that the presidential race will be exceptionally close as Trump is suggested to win on the upper bound of this error. As a result, we make this forecast with great caution and with the intention that further work be carried out to find a stronger prediction to either side. Unsurprisingly, Trump seems to gain the most amount of votes from white males above the age of 30 across all regions but primarily in the southern region. Conversely, younger female voters of all races appear to heavily favor Biden. There are many areas that we believe can have a significant effect on making our forecast inaccurate. The distribution of our predictor variables and the heavily outweighed groups we see can cause problems with our models as errors in the estimates of highly populated groups will create a much larger shift in the number of votes each candidate gets. Furthermore, the underrepresented groups can be problematic as our regression model has less data to work with here, leaving us with less overall confidence in our predictions for these groups.

Weaknesses and Future Work

The Nationscape survey is extremely long and as previously mentioned there was the obvious problem of speeding and straight-lining through it that resulted in elimination of data. Since respondents were given the option to not answer certain questions, there is the usual problem of missing values which further shrinks our testable sample population. However, one of the biggest drawbacks of the survey is that it was conducted only

in English after Phase 1. Initially, it was offered in both English and Spanish, but the Spanish version was rescinded in Phase 2. As the dataset we use is from Phase 2, we can obviously expect an under-representation of individuals who speak languages other than English.

The most prominent limitation of our study was a lack of explanatory variables. Restricting ourselves to only a person's age, gender, race, region, and home language gives us only a small window through which we can analyze their preferences. Additionally, the limited amount of data we had to work with seriously constrained the options we had to choose from. As the ACS dataset contained no questions on things like political policies or politicians, we were prevented from analyzing these responses seen in the survey data. A person's vote choice is often not purely defined based on physical characteristics and so this leaves a lot of room for error in our predictions as we simply ignore these other options.

While our work provides important information about voting preferences and the outcome of the US election, there are many areas for future work where our estimates and predictions can become stronger. As mentioned in section ____, a good first place to start is by including additional variables to the regression model. Adding factors like a person's; income level, education level, birthplace, and state, would give our model more strength, allowing it to make more precise predictions about an individual's voting preferences. There are just a few key elements we believe would have a significant effect on someone's vote. Making use of additional surveys and data would allow for even more variables to be added into our model and improve its predictive capability. Since the US election is based on electoral votes out of each state, including state into our model would allow us to make a prediction specifically on what candidates were likely to win each state. This would result in a more realistic estimate of the election winner as the winning popular vote is not enough for a candidate to take office.

Appendix

Code for this study can be found at:

References