Writing:

Part I - Introduction:

The United States of America is founded on the principles of democracy, which emphasizes the use of government by the whole population. In order to have a government of this type, the people of the country have to elect representatives to make decisions for the country on their behalf. For a representative to be elected, they have to win the majority of the people's votes in their specific election, meaning every person's vote counts.

During voting periods, each person has to go to their assigned polling place in their county to cast their vote. This means that voting availability varies greatly by county based on the amount of polling places available and the amount of people wanting to vote. These factors, among others, play directly into how long each person has to wait in order to cast their vote.

Based on recent studies and collections of data, it has been found that these factors have played into 29% longer wait times and 74% more likely to spend more than 30 minutes waiting for entirely black neighborhoods compared to entirely white neighborhoods at their polling places (Chen). This is very important because these longer wait times for minorities could be pushing them away from voting at all because they simply don't have the time to do it. Not having a big chunk of voters from minorities voting can very easily skew election results and elect a candidate that really wasn't supported the most by the population.

Part II - How Long Do People Wait?:

	1 (not at all)	**2 (<10 min)**	**3 (10-30 min)**	**4 (31-60 min)**	**5 (>60 min)**	**6 (Don't know)**
Amount of responses	7090	5775	4748	2595	1277	77
Proportion	.329	.268	.220	.120	.059	.004

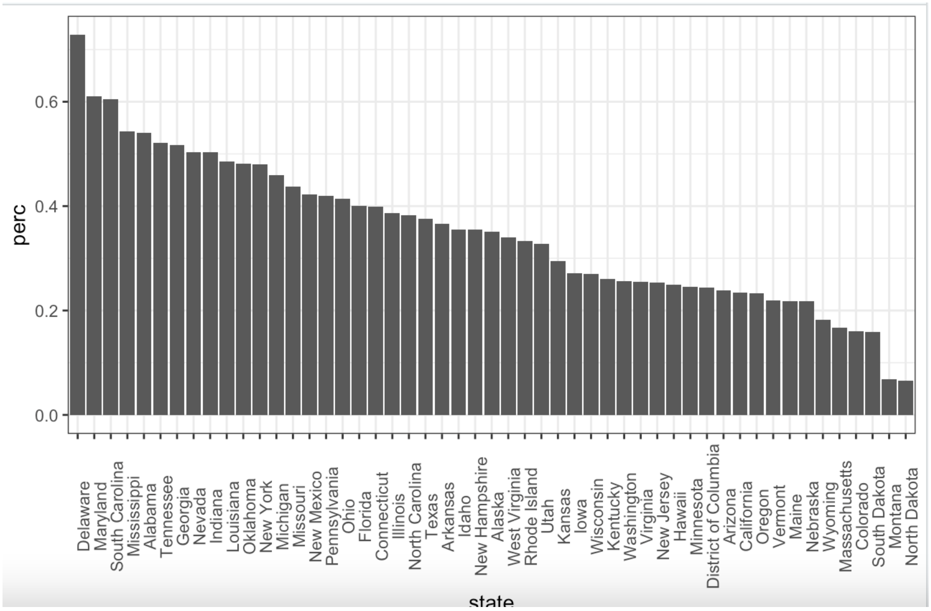
Mean wait response: 2.324	**Median wait response:** 2

To make the table above, we had to take two steps when coding in R. The first step we took was to get the number of responses of each number in our dataset. In order to do this, we used the table() command and put ces20$wait inside of it, which tells R to go to our ces20 dataset and look at the column labeled "wait". This command gave us an output which had the number of responses for each number in our wait column, which you see above in the "amount of responses" row. The second step we took was to turn those response numbers into proportions. In order to do that, we used the prop.table() function and put the table(ces20$wait) function we used before into that. What that does is just takes the table() function we did above and turns the number into proportions instead of just numbers. This is the data you see in the second row above.

The next two steps that we took in order to get the median and mean pieces of data for the wait response were pretty easy. To find the mean, we just used the mean() command and put the path to the wait column like we did before (ces20$wait) and that command enabled R to take the mean of that column and spit it out for us. To find the median, we did essentially the same thing. We used the median() command and put the path to the wait column inside of it, and R then gave us the median value.

When looking at these numbers, we can obviously see that the majority of people who answered this survey did not have to wait very long, if at all, but it can also show that there are a good amount of people that do have to wait ridiculous amounts of time to vote. By looking at the mean wait response, we can see that the average person has to wait over 10 minutes to vote, which is not all that good in the grand scheme of things and could very well push some people away from trying to vote at all.

Part III - Long Waits by State:



In order to make the bar plot you see above, we had to find the percent of people in each state that waited over 10 minutes to cast their votes and arrange it in descending order. To do this, we had to take 6 different steps in our code.

The first step we had to take in our code was create a new column in our dataset that would say true or false if the voter had waited more than 10 minutes or not. In order to do this, we first had to add a new column to the dataset, which is possible by stating the dataset and the name of your new column put together by the $ symbol. Assigning “” to this makes this column empty, which means we can put our new stuff in it. What we end up with is ces20$more10 <- “”. The next thing we have to do is assign our true and false to these new empty rows. In R we can assign certain things to our data using criterias, which in this case is very helpful. When looking at the wait variable in part II, we can see that any response at 3 or above is considered to be over 10 minutes. Knowing this, we can add criteria to say false in the rows that the voters' wait time was under 3, true for over 3, and NA for 6 because that was described as “don't know”. This code looks like this:

```
ces20$more10[ces20$wait < 3] <- FALSE
ces20$more10[ces20$wait >= 3 & ces20$wait != 6] <- TRUE
ces20$more10[ces20$wait == 6] <- NA
```

As you can see, this is pretty straightforward as we just add brackets to state the criteria of looking at the number in the wait variable. Some things that you might not be aware of are that >= is greater than or equal to and != is does not equal.

After creating this new column in our dataset we have to take one more step before getting into making the bar plot. This step is to turn the trues and falses we just made in the previous step into 1s and 0s. This will make us be able to create our proportions for the bar plot. This step is very similar to the previous one, as we just create a new column and change the criteria inside the brackets to say if the more10 column is true to assign a 1, if it is false to assign a 0, and if it's NA to leave it. This new column is called more10int

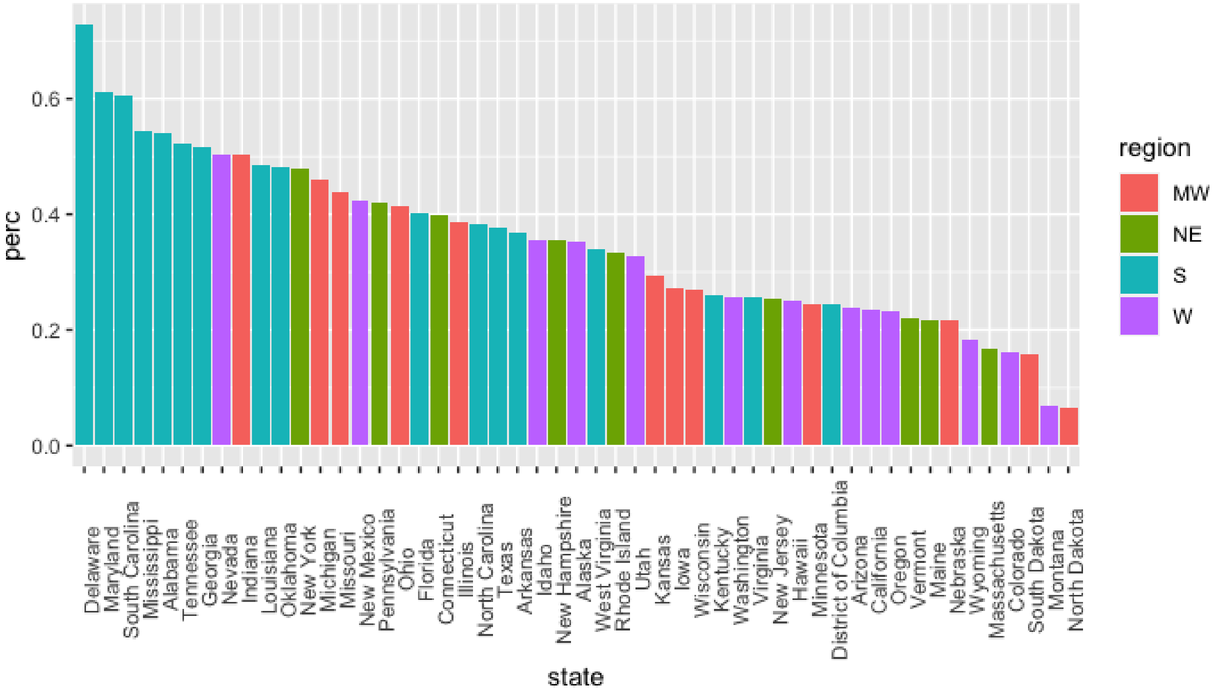
In order to make the bar plot, we first need to make a new dataset that pulls just the information we need from our big dataset. The information we will need in this case is the state and the percentage of people who waited more than 10 minutes. To start this we will name our new dataset (perc_more10) and start what we are assigning to it with an <-. The first thing we will assign to it is our big dataset (ces20) so we have somewhere to pull our data from. In making a new dataset we separate our commands with %>%. The next command we will use is the group_by() command. This is where we say what columns we want to take from ces20. In this case, we want to pull the state column, so we put state in the parentheses. The next command is how we get the percentage, which is summarize(). Inside this summarize we name the new column (perc =) and then say what we want R to calculate, which is the mean of the more10int column (mean(more10int, na.rm = TRUE)). The na.rm thing at the end makes it so R doesn't get confused with the NA values when calculating.

Now that we have our new dataset with our percentages and states, we have to arrange it in descending order to make the bar plot look good and be easier to read. This is very easy as we just add another command to the new dataset called arrange() and inside we put desc(perc) which tells it we want it in descending order of the perc column. In order for R to recognize this new order of the data, we have to assign the same order to the state column as well, which is kindof tricky and weird as you have to do perc_more10$state <- factor(perc_more10$state, levels = perc_more10$state).

Now that we finally have the dataset done, we can finally make the bar plot from it. The first thing we have to do is name the plot (perc_more10_p). Once we do this we can start our assigning to it with the <-. The first command we use to assign to the plot is the ggplot() command, which tells R we are making a plot. Inside this command we will state the dataset we are pulling from (perc_more10) and then assign what we want on our y and z axis (state on x and perc on y). In order to assign our axes, we have to use the aes() command, which stands for aesthetic, and put x=state, y=perc inside. When separating our big commands in making a plot, we use the + sign. After the ggplot command, we use the geom_col() command to tell R we are putting this in a bar plot. We don't put anything in the parentheses here as the command is all R needs. The last two things we add to the plot are just to make it pretty, which are adding a black and white color theme with theme_bw() and making the state text on the x axis go up and down so we can read it with theme(axis.text.x = element_text(angle = 90)).

When looking at this new plot we have made, we can see that there is a great variety in the amount of time that states make their residents wait to vote. There are states where less than 10% of the residents have to wait more than 10 minutes and states where more than 60% of the residents have to wait 10 minutes. It can also be seen that there is not one number that a lot of states hover around and there is clearly a linear slope that goes down very slowly, which makes the percentages very diverse. One of the patterns that I notice about the states that have long wait times is that a lot of them are in the southeast. Out of the top 7 states in wait times, 5 of them are in the southeast part of the United States, which is very interesting.

Part IV - Waiting Times by State and Region:

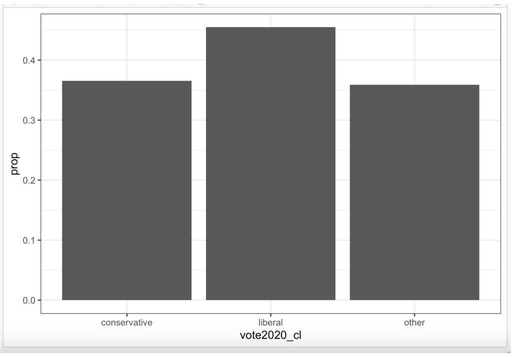


Making the new bar plot for this part is very easy as we just add one more thing to the plot from the previous part. This new thing we have to add is the region variable. In order to include it in our dataset so we can implement it in our plot, we have to go back to when we made the perc_more10 dataset. When we used the group_by() command before, we only put the state column in there, but now we have to add the region variable. This is really simple as you just add a comma and then “region”.

Now that we have the region variable in our dataset, we just have to add it into the plot. Because we want the bars to be different colors for each region, we have to go to the aes() command in the ggplot() command to change the way the graph looks (aesthetics). In the aes() command before, we just had x=state, y=perc, but in order to change the fill color based on region, we need to add another comma and say “fill = region”. Adding this makes the plot look like it does above.

When looking at this new color coded bar plot, we can see that the observation I made above is correct in regards to how the southern part of the United States is heavily represented in the higher wait times and not represented much at all in the lower wait times. I wold say that the rest of the United States is pretty well spread out in how long their wait times are in regards to their region, but you can definitely see the disparity in the southern states. I think that this is definitely an interesting piece of information because southern states stereotypically vote republican, which in theory could be impacted by the wait times we see above.

Part V - Waiting times by Prior Vote:



Now let's take it a step further, to the individual wait times by ideology in the 2020 vote. The two major parties that were factors in the 2020 vote were the Conservative and Liberal parties, both of these parties took up a vast majority of the vote in the election. Firstly, by taking out those who waited less than ten minutes to vote by their political ideology we can better gauge how wait times affected these parties. Also, by putting aside those who voted for other parties in a separate category called “other” we can look at these three categories with more objective and simplistic understanding. Then by taking the proportions of each respective parties voting wait times that took over ten minutes, a bar chart can be produced to visually represent these findings. In the 2020 election, proportionately, those who identified as liberal voters on average had longer wait times to vote than both conservative voters and other voters. The results, given in proportions, are as such, 45.5% of liberal voters waited over ten minutes to vote, 36.6% of conservative voters waited over ten minutes to vote, and 35.9% of other voters waited over ten minutes to vote. Interestingly, both the other and conservative categories showed very similar results in their above ten minutes wait time proportions. So why did the those who voted liberal on average have more voters waiting over ten minutes? The main factor that played into these results was the location of where the vote was placed.

The location of where a vote is being placed as it concerns to voting wait time is a very important factor in this study. The locations of where the votes for liberal voter's vs conservative voters took place, it was more common to see that liberal voters voted in more densely populated counties like large cities, as opposed to conservative voters who more commonly vote in less densely populated counties. Within the issue of voting location there are two differing issues that stem from this over encompassing issue of voting location.

The first being the shear overflow of voting done in densely populated districts. Districts that include cities and other larger metropolitan areas see a high influx of voters on average. Statistically, people with a more liberal ideology reside in these more densely populated districts as many of the United States larger cities in general, and by state, are home to more liberal ideology and progressive ideals. On the other hand, more conservate voting is seen in smaller cities and more rural areas that do not see as many people on average voting, therefore less wait times at the polls. Yet, that alone is not substantive enough to come to this conclusion that is evident in this study.

Other than population density being a factor in these findings, racial and socioeconomic factors are evident in these conclusions. While at this point in the study those factors have not been examined yet, it is important to recognize their importance to this data in foresight. Both of these factors play a role in voting wait times, as minority voters are more likely to have a liberal ideology and on the spectrum of socioeconomic status, those that live in cities often have a more liberal leaning viewpoint on political matters.

This bar chart is but a small representation of the voting wait times seen in the 2020 election. It only examines the wait times by ideology and is not influenced by any other external factors. Those external factors, however, are very important to keep in mind when examining a study like this, and in the next couple sections those other factors will be examined for their influence on voting wait times in 2020.

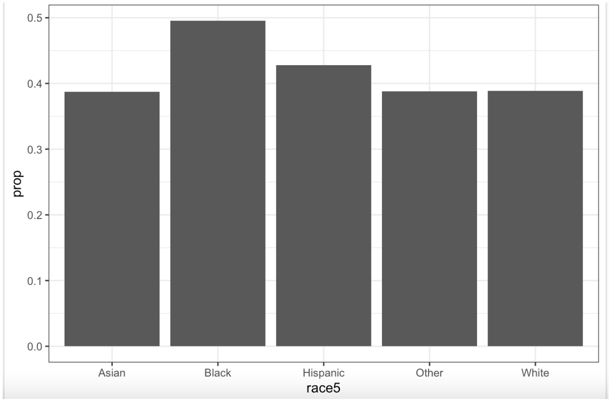
Part VI - Waiting times by Race:

Now we are in essence zooming in and getting more specific on what variables contribute to longer or shorter waiting times when compared person to person. The first variable we wish to look at is how a voter's race affects their wait time. With the dataset provided to us we do have a race column that identifies each person's race on a scale from 1-8:

1: White
2: Black
3: Hispanic
4: Asian
5: Native American
6: Mixed
7: Other

8: Middle Eastern
However, to make the dataset easier to understand and group we created a new variable called “race5”. We essentially used the same scaling except two changes were made:
In the dataset, in plain english the person’s race is specifically written. Moreover, individuals who identified as Mixed, Other, or Middle Eastern were all categorized as “Other”. Next to see who waited more than 10 minutes or more in line. Once again, we created another column in the dataset called “more10” by using the $ command in R. Our dataset already had its own column called “wait” that tells us the wait time each individual reported with a scaling of 1-6 to represent if the individual waited more than 10 minutes in line.
1 Not at all
2 Less than 10 minutes
3 10-30 minutes
4 31 minutes – 1 hour
5 More than 1 hour
6 Don’t know
Then by using the operator functions in R such as the “greater than or equal to” or “less than” we were able to tell R to fill in our “more10” column by checking each individual’s value in the “wait” column and checking if its greater than or equal to 3. If it was not, we assigned the value in that person’s row False. The great thing about R is that False will be read as a value of 0 and read True as a value of 1 so later statistical analysis will have easier math. For individuals that had a 3 or more we assigned them a value of True. Although, since 6 means “Don’t know” that doesn’t add any useful information, so we used the” not equal to” operator which essentially tells R to not look individuals who had a 6 in the “wait column.” To fully make sure individuals who reported Don’t know doesn’t affect our proportions we told R that if an individual had a 6 in the “wait” column to assign it the value NA because this tells R that it’s a missing value and to not include it in any future calculations. Now making the bar graph is the easy part as we used a similar method from previous sections. We told R by using tidyverse commands to tell R to make a new dataset of information we want to look at from our big dataset. We named our new dataset bargraph_wait_race and assigned it the value of our big dataset and used the %>% to let R know there’s more lines of code to look at. We then grouped by race and told R to summarize the grouping by proportions which will equal the mean of our more10 column. The we once again used the ggplot command to plot our dataset and set the x axis to our race5 column and our y to the proportions we told R to find.

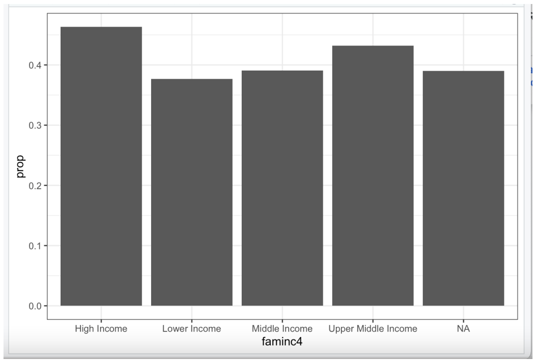
When looking at the results of our bar graph, it’s shows that Black people were the majority of people who waited 10 minutes or more in voting lines. While Hispanic people were 2nd in the bar graph. Interestingly Asian people, White people, and people who were classified as Other had about the same average wait time in line.



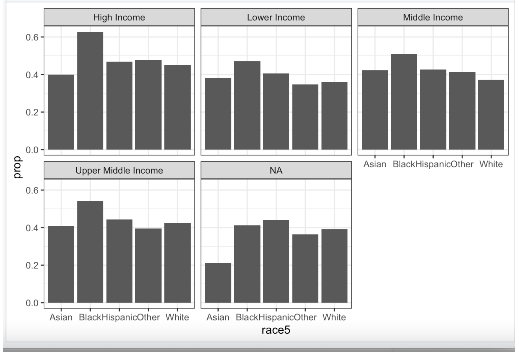
Part VII - Waiting times by Income:

Another variable we wished to observe was how an individual’s income affected their wait times in voting lines. The process we used to understand this relationship was like how we analyzed race vs wait times. Again, we start off by making a new column in our big dataset and named it “faminc4”. In our dataset, individuals reported income is saved under the column named “faminc” and had a scaling of 1-16 and if someone preferred not to say their income, they were given the value of 97. We told R to fill the new column by using its greater than or equal to / less than or equal to operators and told it to check everyone’s value they reported in the faminc column. If their faminc value was less than or equal to 4, in plain English on the new faminc4 column it would read “Lower income.” If an individual’s value in the faminc column was greater than or equal to 5 but less than or equal to 8, it would read “Middle Income.” If an individual’s value in the faminc column was greater than or equal to 9 but less than or equal to 12, it would read “Upper Middle Income.” If an individual’s value in the faminc column was greater than or equal to 13 but less than or equal to 16, it would read “High Income.” Also, to ensure that the people who did not report their income does not skew our results we told R if an individual had a value of 97 in the faminc column to assign it NA in the famninc4 column so that R would understand that they are missing values and will not be included in further calculations. To create the table we used the prop.table command which creates a table for a set of values its given, so for our case it would make a table for the faminc4 column we just made. Now, since we want a bar graph on how income affects wait times we used a similar process for when we made a bar graph on how race affects wait time. We told R by using tidyverse commands to tell R to make a new dataset of information we want to look at from our big dataset. We named our new dataset bargraph_wait_income and assigned it the value of our big dataset and used the %>% to let R know there’s more lines of code to look at. We then grouped by race and told R to summarize the grouping by proportions which will equal the mean of our more10 column. Then we once again used the ggplot command to plot our dataset except now our x-axis for the bar graph is faminc4 while the y axis is still prop.

When looking at the results of this figure, we can see that as income goes up, the wait times go up. This finding clashes with what we found in part VI in how the black population has the longest wait times because now we are saying that high income people have the longest wait times as well, which we know is not made up of mostly the black population. If we just looked at the results from section VI without producing these results, it would not be crazy to infer that lower income people have longer wait times than higher income people because of the distribution between races that we saw with minorities being the highest. With the data we see now with this figure though we see this is not true, which can make us question our results in both parts. This next section, though, will hopefully clear up some of this confusion.



Part VIII - Using subclassification to Account for the Effect of Income:



In order to make the subclassified graphs that you see above we took 2 steps in our coding. The first thing we had to do was make changes to the bargraph_raceinc dataset that we used in the previous part. The changes we had to make in this dataset was to include the race5 variable so that we have the race of each respondent to separate them within each income level. In order to do this, we just had to add our race5 column to our group_by() command along with faminc4.

The next and last step was to make our new bargraphs. To do this we just had to take the code from the previous part and make a few changes. First, obviously, we had to change the name of it (pt_8_bargraph). We then copy and pasted the same ggplot command from the previous part, but instead this time we are using our new race5 data on our x-axis since we are separating the graphs by the faminc4 variable. All we had to do to do this was go into our aes() command and change our x= to x=race5. The last change we had to make to the code was to add a new command after the ggplot() command. This new command we used is called facet_wrap(). What this command does is it will take the plot you already made and separate it into different plots based on the variable you tell it to use. In our case, we want to separate by our faminc4 variable, so inside the command we just added “~faminc4” and it did the job.

Looking at these figures, we can see that there is a slow constant increase in wait times as we go up in income groups. In each separate bar graph except the NA one, we also can see that the wait time for Blacks is substantially higher than the wait times for all of the other races. This realization supports the conclusion we came to in part VI of the paper that blacks have to wait substantially longer than any other race to vote and it also clears up some of the confusion we had above in part VII. The confounder that we account for in making these subclassified graphs is that there could have been more of one race in a certain group, which could have been skewing the results in Part VII, but that was not the case. This is not the case because in each income level you can see the same pattern across the races in terms of their wait times and it just gets bigger and bigger as you go up. A confounder that remains even with this use of subclassification is that the people in the NA division who didn’t want to state their income level could be part of higher incomes, which would bring the higher income levels’ wait times down if they were accounted for.

Part IX - Using Regression to Account for Multiple Confounders Simultaneously:

	model1	model2
Black	0.284	0.318
	0.029	0.031
Hispanic	0.076	0.049
	0.033	0.035
Asian	0.041	0.017
	0.064	0.068
Other	0.011	-0.006
	0.039	0.042
faminc_reg		0.026

		0.003
income_county		0.001
		0.001
density		0.009
		0.002
Intercept	2.279	2.037
	0.009	0.035
N	21,480	19,233
R^2	0.005	0.013

Within a study like this one there are many factors, or confounding variables, that can impact the findings and outcomes of what is being studied. Confounding variables play a crucial part in the shaping and understanding of data and results of studies like this one of the 2020 election. In this study's case, there are multiple confounding variables that need to be introduced into the study to garner more concise conclusions. Aside from race, the factors of socioeconomic status, income, and population density need to be examined through the data. Although we have already looked at income as a factor, socioeconomic status and population density are new to the study.

The two new factors of socioeconomic status and population density are seen as confounding factors because they shape how the results are presented to the vote / wait time study. When these two variables are added to the study, noticeable trends of voting wait times are seen congruent with the scales of both socioeconomic status and population density of locations. The race / wait time relationship is impacted by both of these factors in a couple of ways.

Part X - Conclusion:

Voting is a vital element that helps America's democracy run as smoothly as possible. Ideally it should be a streamline process. A simple check on a ballot holds much weight on who you want to represent you and this great power all Americans have should be taken with pride.

However, through this experiment, it was made apparent that long wait times are plaguing the efficiency of voting. But the thing that is the most shocking is that long wait times do not affect Americans on a consistent level.

From this study on the 2020 election, there are many disparities that exist in our society that inhibit equal voting opportunities for all American citizens. Long voting wait times are prevalent across the United States, that exist from factors of socioeconomic disparity, racial divisions, and population density, inhibiting all American citizens from an easy voting experience. In essence creating barriers from people exercising one of the most important of political civic duties that exists in our society. While wait times are a common inconvenience that exist across the country, there are numerous demographics that unfortunately experience more troublesome voting wait times and issues than others. With disparity like this existing in one of our governments most important and fundamental processes, there is much need for reform to even the playing field for all American citizens to solve issues of long wait times and incentivize better voter participation. While this study reflected those fundamental issues in our voting system in 2020, there is a future where these many issues can be, and hopefully will be, recognized and fixed.

One possible policy counties or even states could implement to reduce long wait times many Americans face is an introduction to a voting appointment online based system. With a voting appointment system, instead of long lines throughout the day or even week to vote, certain amount of time slots will allow only an X amount of people (which can be decided by the state or county) to be at the polling place at a time reducing long lines to a more manageable amount. Furthermore, an appointment based system will aid in remedying one of the variables we discovered affects wait times. In the experiment we discovered that Black and Hispanic people wait the most in lines and also found out that a good proportion of Black and Hispanic people when compared to White and Asian people have occupations in construction, maintenance, service, production, transportation etc which can be argued require longer hours of the day and to work more days in comparison to other occupations. Implementing a system where you can make an appointment online to vote will enable Americans who work those types of jobs listed above to fit it into their schedule to not only vote earlier but to also wait a significantly less amount of time in line.

Part XI - Individual Contributions

Works Cited

Chen, M. Keith, et al. “Racial Disparities in Voting Wait Times: Evidence from Smartphone Data.” MIT Press, MIT Press, 14 Nov. 2022, https://direct.mit.edu/rest/article-abstract/104/6/1341/97747/Racial-Disparities-in-Voting-Wait-Times-Evidence?redirectedFrom=fulltext.

Presentation: Group 3, Section VI

Poli 281 Pres - Group 3 (Part VI)