Lab 1: Question 1

Your Names Here

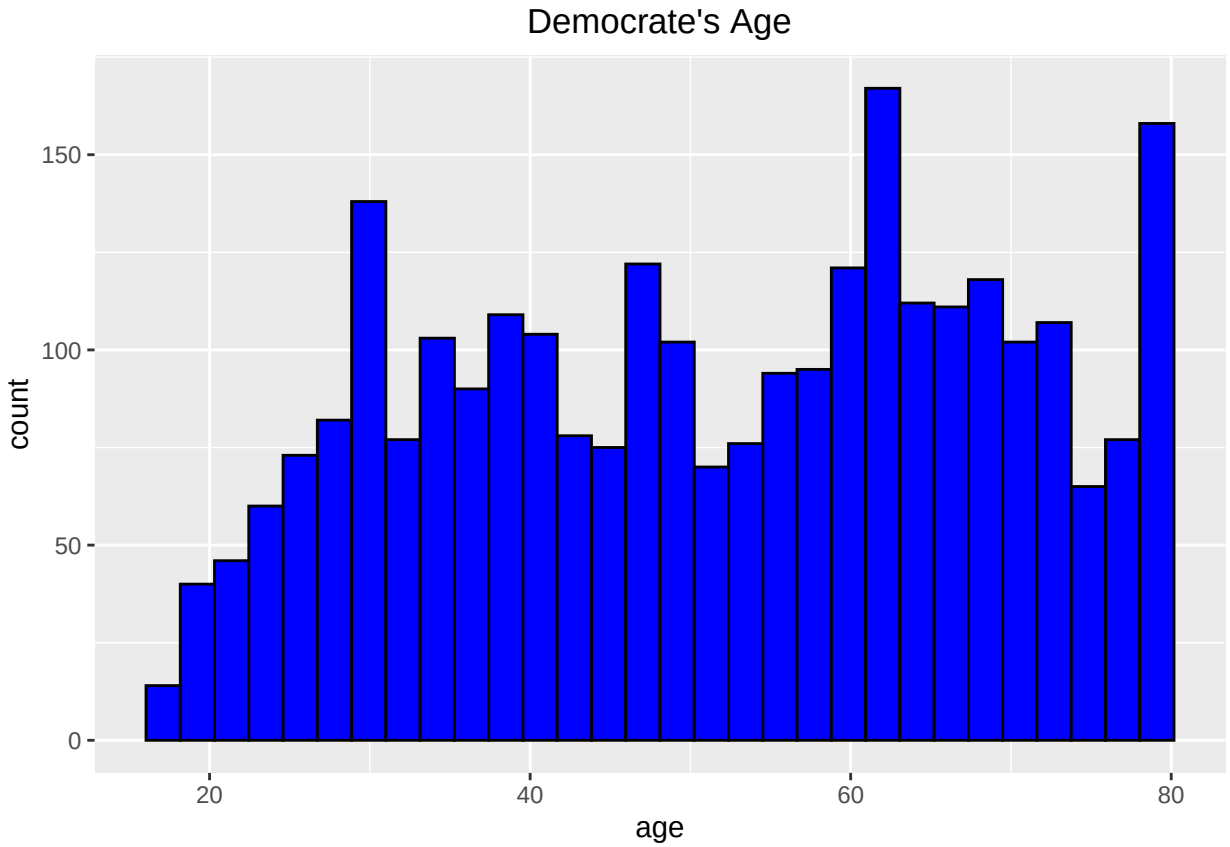
```
library(dplyr)
library(ggplot2)
# library(tidyverse) # if you want more, but still core, toolkit
```

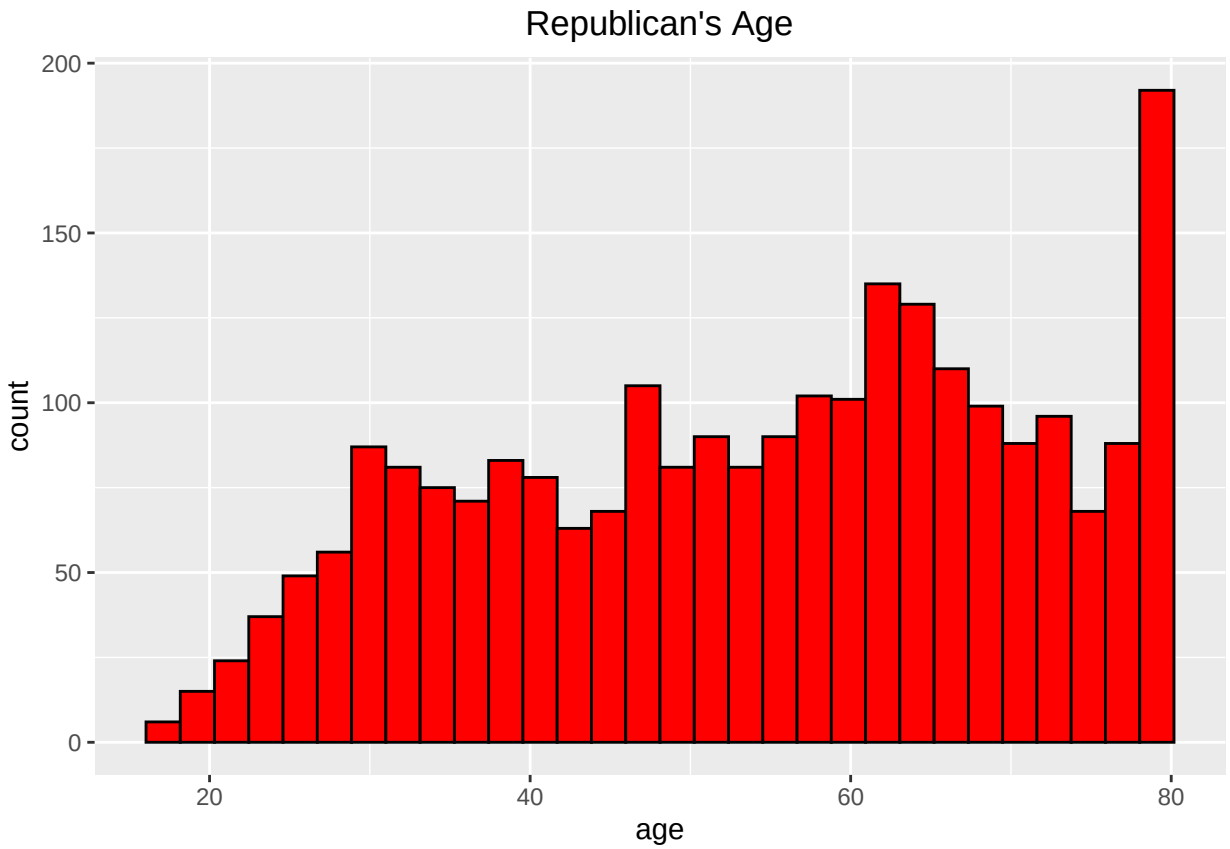
Importance and Context

The United States has been on a two party political system since the 1850's. These two parties have been the republican and democratic party. Each party has had its ups and downs throughout the years. They have also appealed to different demographics over the years. It has long been the goal of each party to gain the vote of the younger generation. We would like to look at a recent survey to see if the age demographic between the two parties is different. This survey was conducted on a cross-sectional sample on the USPS computerized delivery sequence file, which includes residential addresses across the 50 states and Washington DC from August 18, 2020 and November 3, 2020.

Description of Data

The two main variables for this analysis are party affiliation and age. We used code V201228 to determine what party each subject affiliated with. The question on the survey asked the subject what party they self identified with. We also used code V201507x to determine the subjects age. The subject could refuse to submit an age, so they were excluded from the data set. After the data clean up we had 2,786 Democratic subjects and 2,448 Republican candidates. You can see the distribution of ages for each party below. As you can see both parties have similar distributions, but one thing to note is that the republicans have a large portion at 80 years old. This is because the survey had a category of 80 +. This means that the republican subjects could actually be much older than we are showing, but because of the nature of the survey we are missing that data and truncating it at 80. This will be important to acount for when looking at the practical significance of the results.


Democrate's Age
150
100
50
0
count
20
40
60
80
age



Most appropriate test

We believe that the most appropriate test for our question is an independent two tail t test. The two samples that we are comparing are completely independent of each other, because one person can either be a Democrat or a Republican, these two are not linked. We also don't have a very strong assumption of the data leaning one way or another so we need to use the two tailed test. There are three main assumptions needed for this test. first that the data is numeric. The age data that we are using is numeric between 18 and 80+. Second the sample needs to independent and identically distributed (iid). The data is being pulled from only one individual per house hold. So there should not be subjects that are impacting each others results. So the data is independent. Also the data is being collected in a very short period of time so the population is not changing, meaning the data is also identical. Lastly the data should have no major distributions in normality, considering the size of the sample. As seen in the charts above the data is not normally distributed, but our sample sizes are 2,500 or greater. This is well above 30 so the central limit theorem will kick in and we will have no problem meeting this assumption.

Test, results and interpretation

Below you can see the results of our independent two tail t test. The t value for this test is -5.75 this is well outside of our 95% confidence interval (-3.63 - -1.78). This results in a p-value of 9.366e-09. Therefore we have enough statistical significance to reject the null hypothesis. On the other hand if you look at the two means they are only showing a difference of 3 years for the groups. This isn't a lot of practical significance. So although we can reject the null hypothesis the actual age difference might not be very significant, but it does show that Democrats lean younger than Republicans. Another important note is that Republicans have a higher percentage of people in the 80+ group this means that we could possibly be truncating older people from the republican group. This could cause a larger spread in our means, giving us a more practical significance.

```r
t.test(df_dem$age, df_rep$age, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_dem$age and df_rep$age
## t = -5.7512, df = 5185, p-value = 9.366e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.630179 -1.784487
## sample estimates:
## mean of x mean of y
##  51.61701  54.32435
```