Lab 1: Question 1

Yao Chen, Jenny Conde, Satheesh Joseph, Paco Valdez, Yi Zhang

Importance and Context

The 2020 general election was very different from the ones that came before. It happened in the middle of a pandemic. It elected the first ever female VP. And it might have been the most polarized election in recent American history. What drove people apart, among other things, is their age.

Many people suspected that the Republicans have an older supporter base than the Democrats.

Is it just a myth or does it have some truth to it? That is what we're going to find out in this section, using the comprehensive 2020 Time Series Study from ANES (American National Election Studies).

Description of Data

Looking through the CodeBook, it looks like the variables that are useful for this question are: - `V201018`: PARTY OF REGISTRATION - `V201507x`: SUMMARY: RESPONDENT AGE

We removed the people who refused to answer the age question, and only left people who are registered as either Democratic or Republican for the test because we're not interested in other party affiliations.

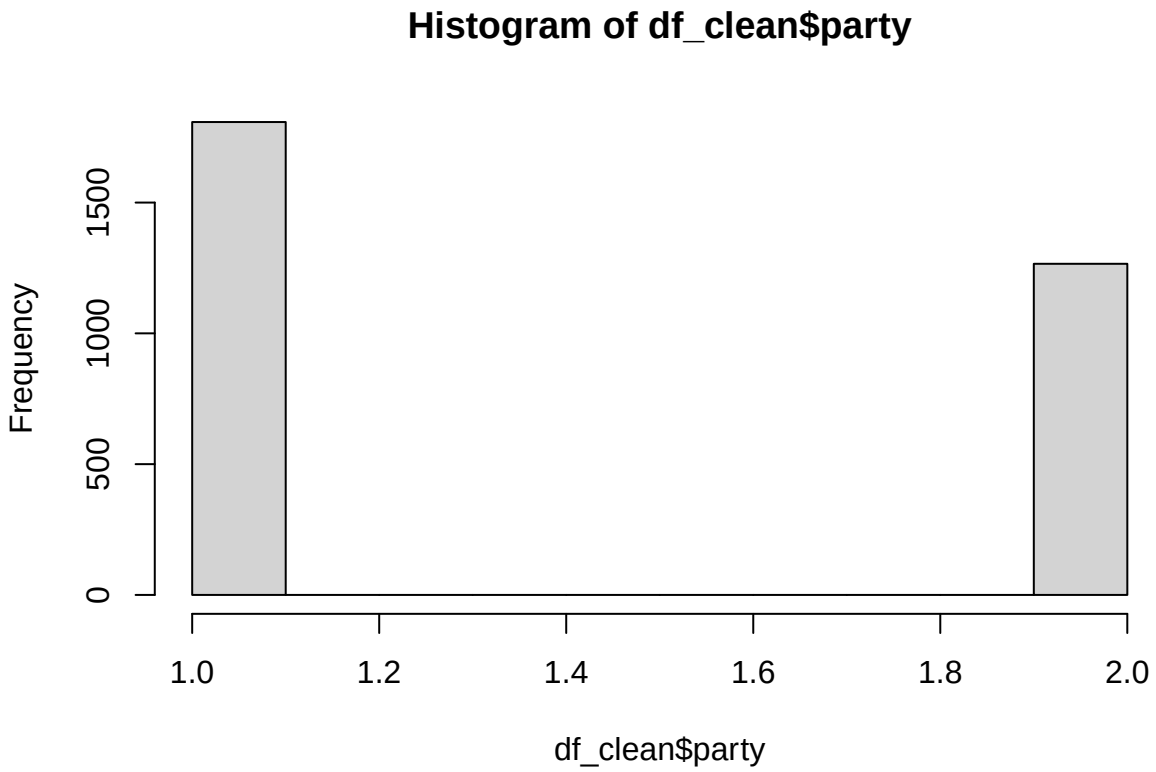
We should probably talk about the fact that the maximum is cut off at 80 in the survey.

Looking at the summary of the data:

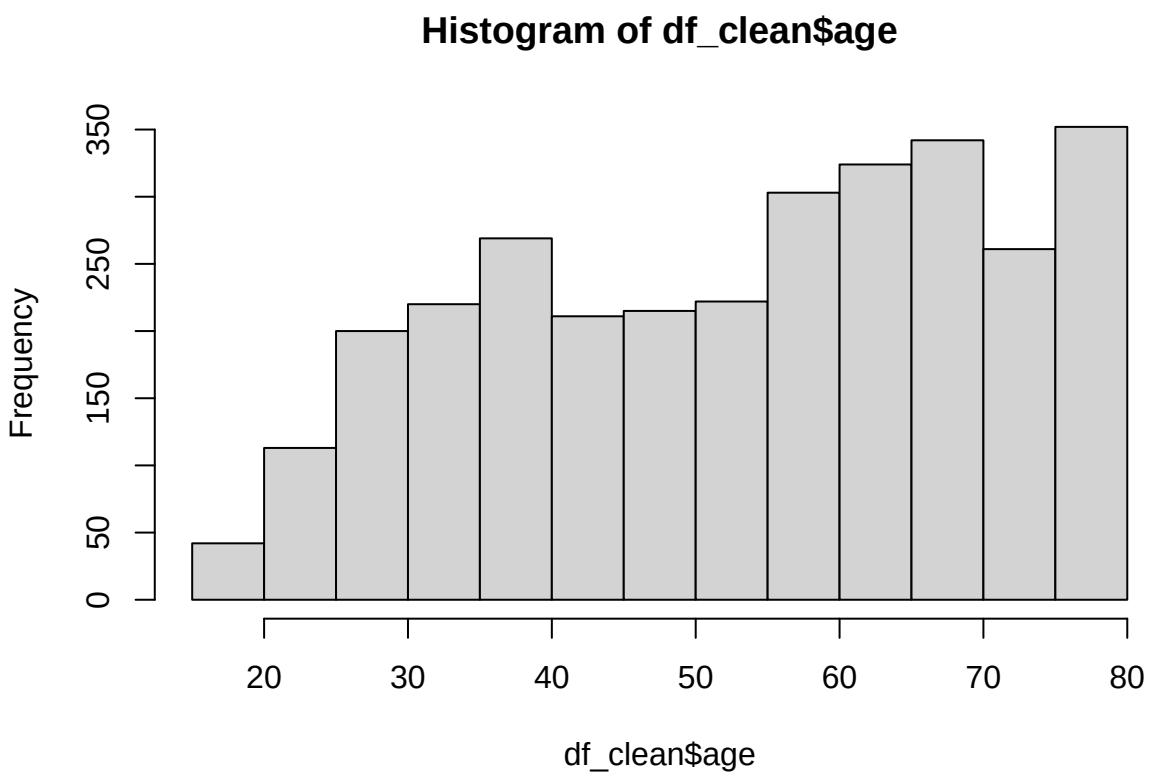
```
summary(df_clean)
```

```
##      party            age
##  Min.   :1.000   Min.   :18.00
##  1st Qu.:1.000   1st Qu.:39.00
##  Median :1.000   Median :56.00
##  Mean   :1.412   Mean   :53.91
##  3rd Qu.:2.000   3rd Qu.:68.00
##  Max.   :2.000   Max.   :80.00
```

```
hist(df_clean$party)
```




```
hist(df_clean$age)
```



Most appropriate test

We have large sample, i.i.d. data, age is an interval variable so we can do t-test. Because people are not registered as Democrat and Republican at the same time, there doesn't seem to be a natural pairing going on. So we should do unpaired t-test.

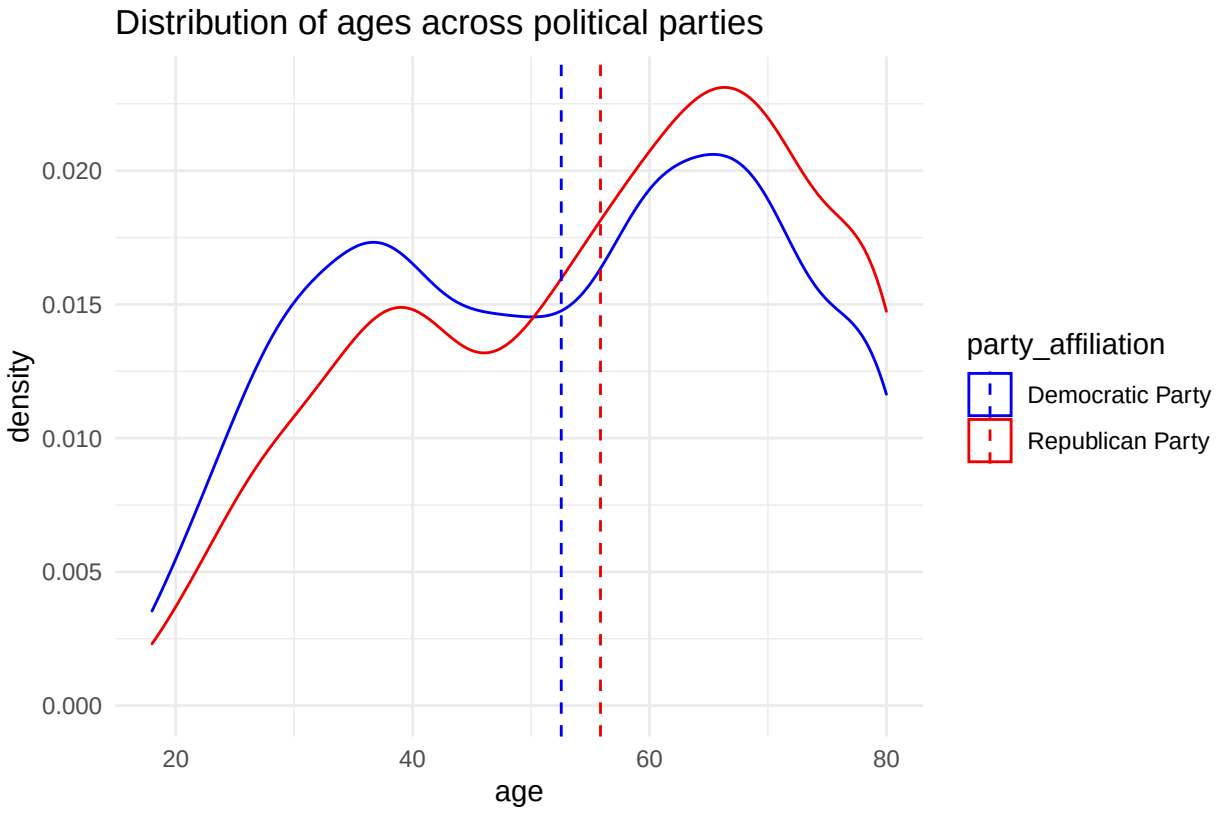
Test, results and interpretation

Null hypothesis is the average age of Democrats and Republicans are the same. The alternative hypothesis is that they're not. So this should be a two tailed test.

```
t.test(df_clean$age ~ df_clean$party)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_clean$age by df_clean$party
## t = -5.3376, df = 2781.1, p-value = 1.017e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.531263 -2.096511
## sample estimates:
## mean in group 1 mean in group 2
##        52.54867        55.86256
```

We have very small p-value, we can reject the null and say the average ages are indeed different.



The dotted lines represent the average age in each respective political party.