Cody Harris Statistical Inference

```
library(ggplot2)
library(tidyverse)
library(BSDA)
cbPalette <- c("#56B4E9", "#009E73", "#999999", "#E69F00",
               "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

Part A: Hurricanes and himmicanes

A 2014 paper published in PNAS was titled "Female hurricanes are deadlier than male hurricanes." The abstract:

> Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

The paper is here:
`http://www.pnas.org/content/111/24/8782.full`

The data can be downloaded here:
`http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1402786111/-/DCSupplemental/pnas.1402786111.sd01.xlsx`

The relevant data is in the "Archival Study" sheet; ignore the experimental results.

The data consists of observations of 92 hurricanes that made landfall in the U.S. from 1950 to 2012. Two key variables are:

- `Gender_MF`: 1 if the hurricane's name is considered "feminine," 0 if the hurricane's name is consider "masculine."
- `NDAM`: the normalized damage caused by the hurricane, adjusted for inflation, wealth, and population.

Various other variables are included in the data set; see the paper for discussion of these.

Questions

1. (5 points.) Draw plots to visualize similarities and differences between the distribution of damage caused by hurricanes with female names and the distribution of damage caused by hurricanes with male names.
2. (5 points.) Is there a meaningful difference between the distribution of damage caused by hurricanes with female names and the distribution of damage caused by hurricanes with male names? For this question, use the binary variable "Gender_MF" and the quantitative variable "NDAM," and no other explanatory variables.
3. (5 points.) Are there any other meaningful differences between hurricanes with female names and hurricanes with male names?

Part A: Hurricanes and himmicanes

We will first import our data into a dataframe so that we can analyze it. The data we are using comes from the archival study sheet in the xlsx file found on this page: https://www.pnas.org/content/suppl/2014/05/30/1402786111.DCSupplemental. For ease of import a csv file has been created that includes only this data.

```
hurr <- read.csv("hurricane.csv", header = TRUE)
summary(hurr)
```

```
##     ï..Year         Name               MasFem       MinPressure_before
##  Min.   :1950   Length:92          Min.   : 1.056   Min.   : 909.0
##  1st Qu.:1965   Class :character   1st Qu.: 2.667   1st Qu.: 950.0
##  Median :1985   Mode  :character   Median : 8.500   Median : 963.5
##  Mean   :1982                      Mean   : 6.781   Mean   : 964.9
##  3rd Qu.:1999                      3rd Qu.: 9.389   3rd Qu.: 983.0
##  Max.   :2012                      Max.   :10.444   Max.   :1002.0
##  Minpressure_Updated.2014   Gender_MF         Category       alldeaths
##  Min.   : 909.0           Min.   :0.0000   Min.   :1.000   Min.   :  0.00
##  1st Qu.: 950.0           1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:  2.00
##  Median : 964.0           Median :1.0000   Median :2.000   Median :  5.00
##  Mean   : 964.9           Mean   :0.6739   Mean   :2.087   Mean   : 20.65
##  3rd Qu.: 982.2           3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.: 20.25
##  Max.   :1003.0           Max.   :1.0000   Max.   :5.000   Max.   :256.00
##       NDAM        Elapsed.Yrs       Source            ZMasFem
##  Min.   :    1   Min.   : 1.00   Length:92          Min.   :-1.7740500
##  1st Qu.:  245   1st Qu.:14.00   Class :character   1st Qu.:-1.2748200
##  Median : 1650   Median :28.00   Mode  :character   Median : 0.5327200
##  Mean   : 7270   Mean   :30.91                      Mean   : 0.0000003
##  3rd Qu.: 8162   3rd Qu.:48.25                      3rd Qu.: 0.8081500
##  Max.   :75000   Max.   :63.00                      Max.   : 1.1352300
##  ZMinPressure_A          ZNDAM
##  Min.   :-2.8862200   Min.   :-0.561990
##  1st Qu.:-0.7694000   1st Qu.:-0.543125
##  Median :-0.0723950   Median :-0.434490
##  Mean   :-0.0000001   Mean   : 0.000001
##  3rd Qu.: 0.9343900   3rd Qu.: 0.069022
##  Max.   : 1.9153600   Max.   : 5.236570
```

What's in a name?

Before diving into the questions, let's just look at the distribution of female and male names.

```
table(hurr$Gender_MF)
```

```
##
##  0  1
## 30 62
```

We see that we have more than twice as many female named hurricanes, but this study does include what it calls the masculinity-femininity index. This helps neutralize some of the issue where some names could be considered male or female names such as: Alex, Charley, and Danny. The method that this index was

generated does not seem to be the best in my opinion. Nine people who were blind to the hypothesis rated the names on a basis of if they were masculine or feminine, 1 being masculine and 11 being feminine.

As the study seeks to show how the name of the hurricane effects the damage caused and loss of life, it would seem that the masculinity of a name should be based on the overall consensus of people in the year the hurricane made landfall. For example, Danny might have been exclusively a male name in 1954; whereas now 66 years later, there are women who go by Danny. So evaluating the names on masculinity in the 2010s when this paper was written, might not give the best indication of how the names were perceived when the hurricanes made landfall.

The only reason this explanation is given before answering the questions is because the data provided might give us an answer that could be misleading because the data itself was flawed. What is more, we will mostly be looking at the binary decision of if a name is feminine or masculine, when we know that in reality many names are not binary in this sense.

Data cleaning

Certain data cleaning might make it easier to visualize or analyze our data. Those changes will be done here.

First we will make a column that has the word "Female" or "Male" to describe the name of the hurricane instead of 1 or 0 respectively. This column will also be a factor.

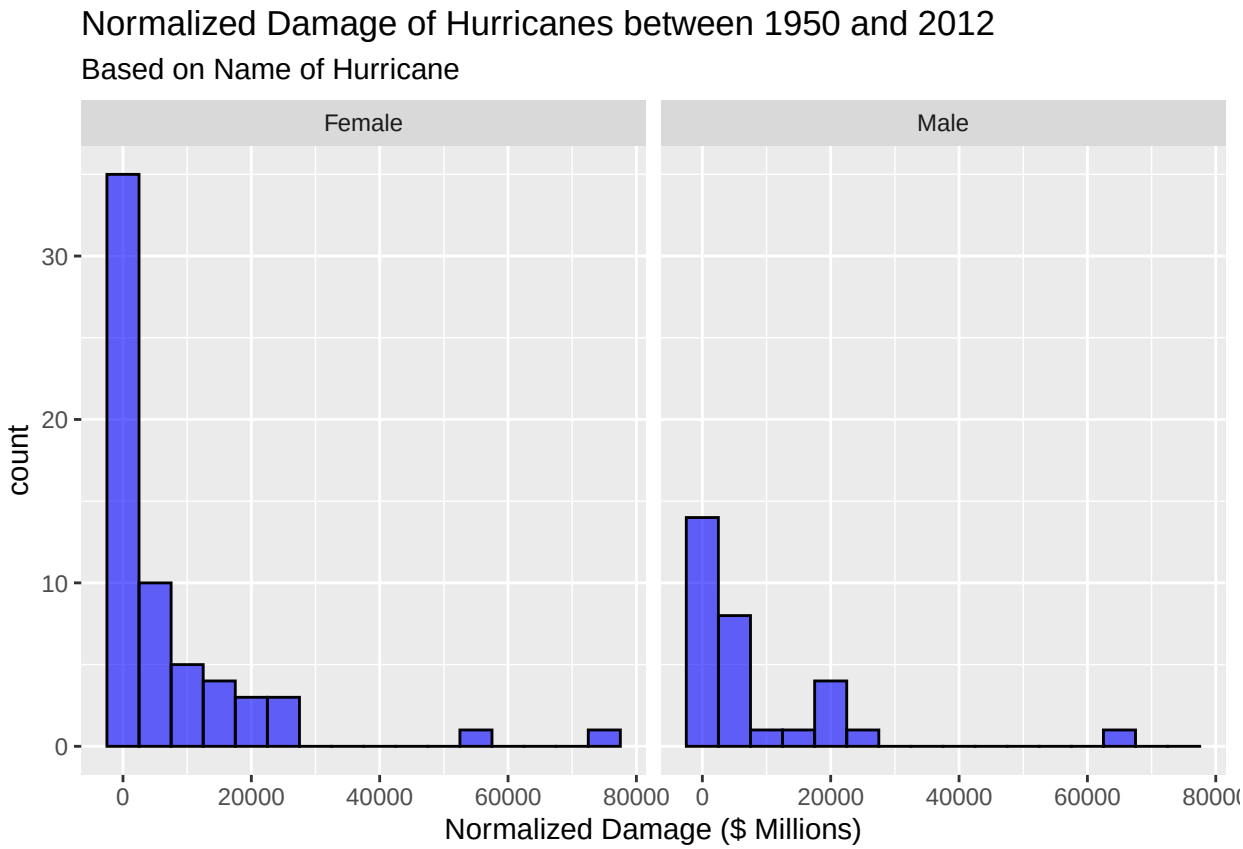
```
hurr$Gender <- as.factor(ifelse(hurr$Gender_MF == 1, "Female", "Male"))
summary(hurr$Gender)
```

```
## Female   Male
##     62     30
```

Question 1.

We will first look at the difference in damage caused by hurricanes in the form of normalized damage in millions of dollars.

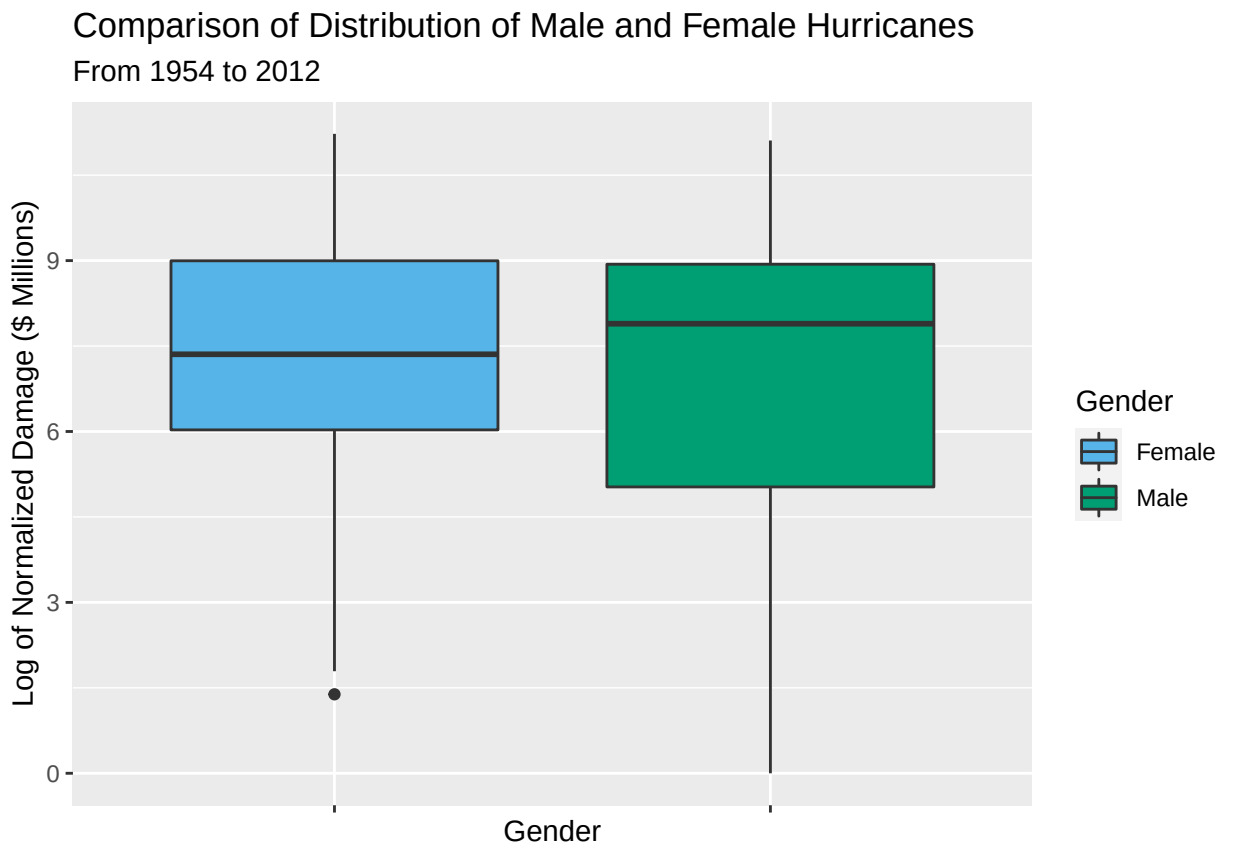
```
ggplot(hurr, aes(x = NDAM)) +
  geom_histogram(binwidth = 5000, fill = "blue", color = "black", alpha = 0.6) +
  facet_wrap(~Gender) +
  xlab("Normalized Damage ($ Millions)") +
  labs(title = "Normalized Damage of Hurricanes between 1950 and 2012",
       subtitle = "Based on Name of Hurricane")
```



We see that both male and female named hurricanes have right-skewed distributions with a majority of hurricanes causing $25 billion or less in damage. Remember that we have twice as many female named hurricanes as male, so although there are larger counts in the female histogram, the overall distributions are rather similar. We can see that the most destructive female hurricane, in terms of dollars of damage, comes from a female hurricane. The main difference between the distributions is that female hurricanes have a single peak at the low end of the damage range, whereas male hurricanes have a peak at the low end of the damage spectrum, but a dip in hurricanes that caused $10-$20 billion in damage and then a small peak at $20-$25 billion dollars of normalized damage.

Perhaps a boxplot could further assert similarities in the distributions. As we have right-skewed data, a log transformation on the normalized damage can be done to look at the distribution of log normalized damage.

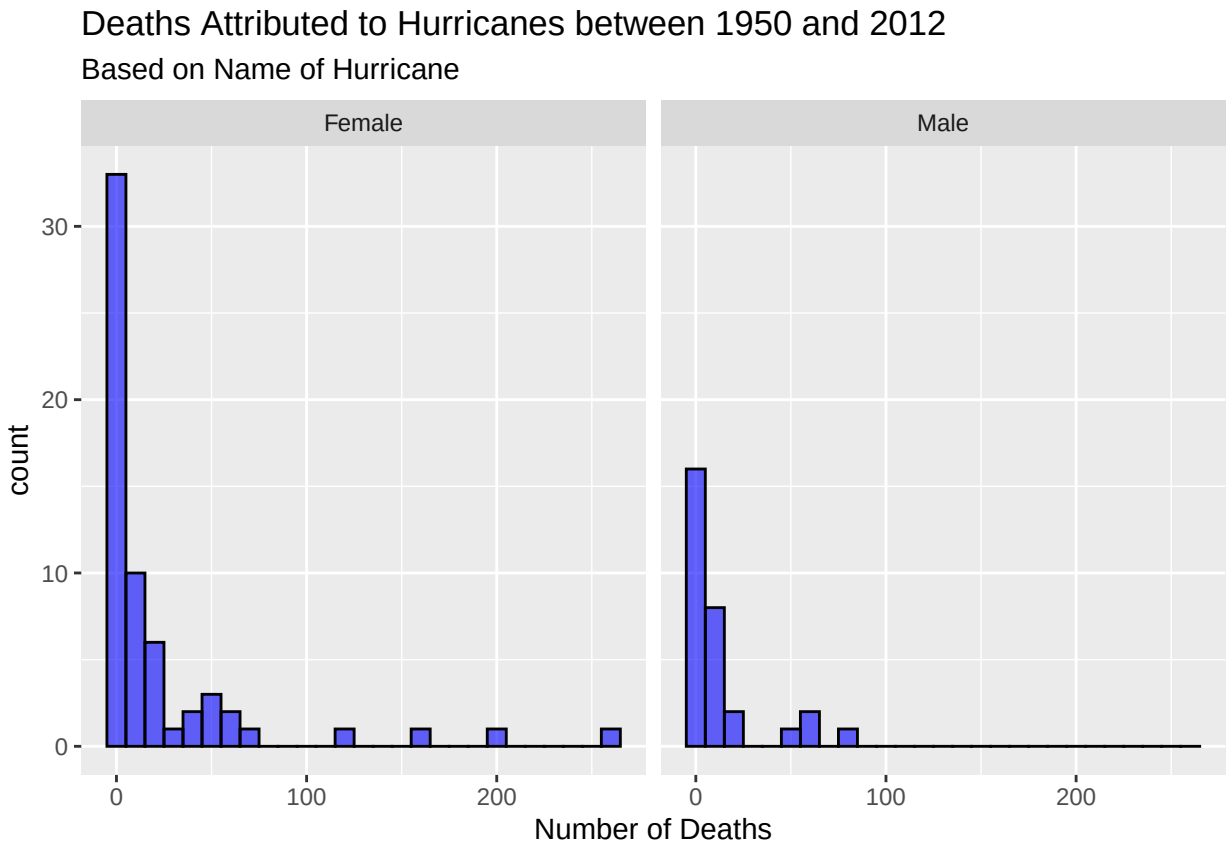
```
ggplot(hurr, aes(x = Gender, y = log(NDAM), fill = Gender)) +
  geom_boxplot() +
  scale_fill_manual(values = cbPalette) +
  ylab("Log of Normalized Damage ($ Millions)") +
  labs(title = "Comparison of Distribution of Male and Female Hurricanes",
       subtitle = "From 1954 to 2012") +
  theme(axis.text.x = element_blank())
```



When we look at the log of normalized damage, we see very similar distributions. Male hurricanes have a larger range and inter-quartile range, which suggests that there is a larger variance in the normalized damage caused by Male named hurricanes in comparison to female named hurricanes. Male hurricanes also have a slightly higher median log normalized damage. With all that said, the third quartile and max is almost equal. This suggests that both female and male named hurricanes in our data had similar potential when it comes to log normalized damage.

While economic damage is an important factor, loss of life is another measure of damage produced by a hurricane. We can look at the difference in distribution of deaths between male and female hurricanes.

```
ggplot(hurr, aes(x = alldeaths)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black", alpha = 0.6) +
  facet_wrap(~Gender) +
  xlab("Number of Deaths") +
  labs(title = "Deaths Attributed to Hurricanes between 1950 and 2012",
       subtitle = "Based on Name of Hurricane")
```



As expected, we see another right skewed distribution for both male and female named hurricanes. Interestingly we see that none of the male hurricanes caused more than 100 deaths, whereas four female hurricanes killed more than 100 people. While there we don't have a huge number of data points, we do see something similar in both groups of hurricanes, there seems to be two peaks in the amount of deaths, the first peak being in the 0-10 deaths range and the second small peak being around 50 deaths. This could suggest that hurricanes are most likely not going to cause many deaths, but there is a subset of strong hurricanes that are more deadly. Then you have some outliers from especially strong hurricanes.

Question 2.

We know that the distribution of our data is not normal, and probably isn't drawn from a normal distribution based on our EDA so far. To determine if the distributions between male and female hurricane's damage, we can use a non-parametric test. The Wilcoxon Rank-Sum test could provide us with some evidence that the two distributions are different. The Wilcoxon rank-sum test is good test here as we do have some outliers and this test is less skewed by outliers as it uses ranks of each measure instead of the value of the variable in question. We will also do a two sided test as we just want to see if there is a difference and are not concerned with what direction the difference is in.

```
library(coin)
```

```
## Loading required package: survival
```

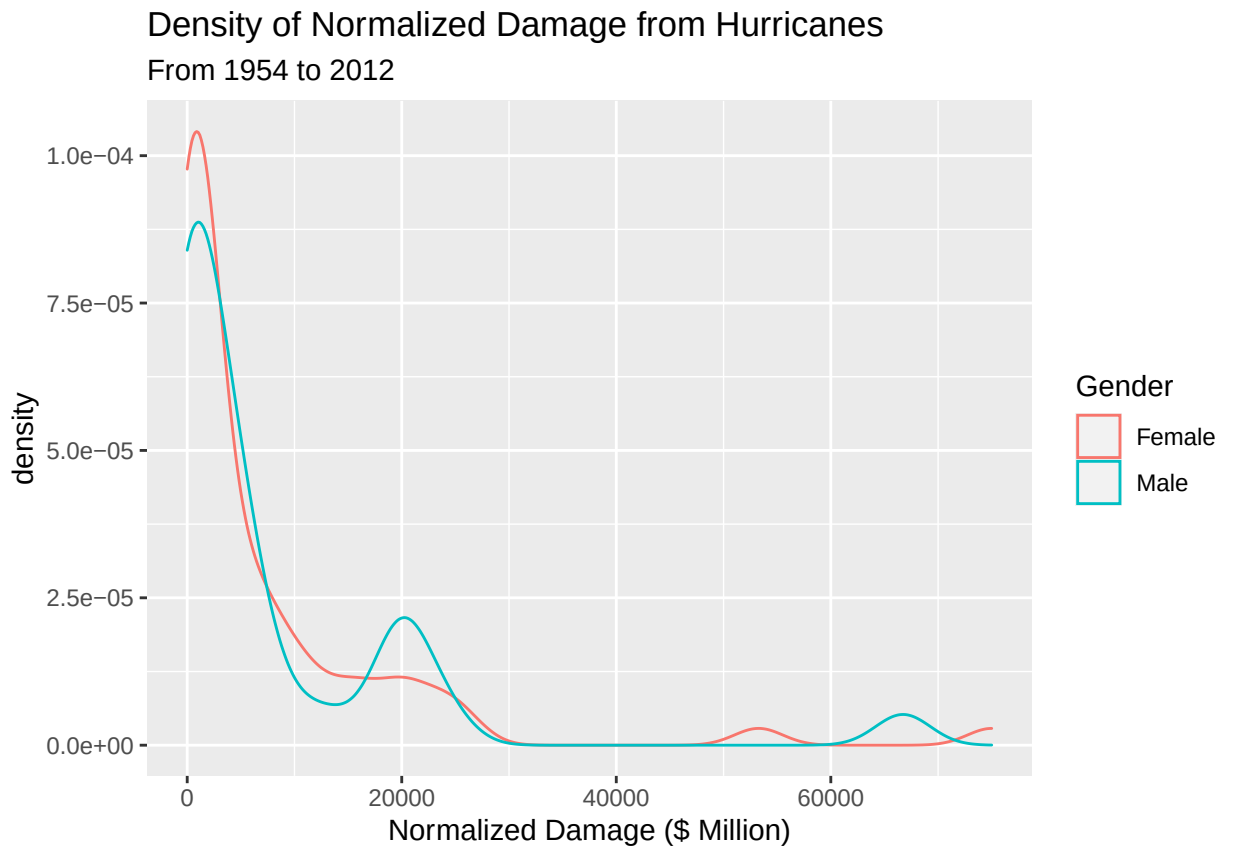
```
wilcox_test(hurr$NDAM ~ hurr$Gender, dist = "exact")
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  hurr$NDAM by hurr$Gender (Female, Male)
## Z = 0.51641, p-value = 0.6089
## alternative hypothesis: true mu is not equal to 0
```

We get a P-Value that is high by any standard which indicates that we fail to reject the null hypothesis that there is no difference between the groups. This is some evidence that there really isn't any difference between the hurricanes that are labeled female compared to the hurricanes labeled male when it comes to normalized damage.

Let's take a look at the density plots.

```
ggplot(hurr, aes(x = NDAM, group = Gender, color = Gender)) +
  geom_density() +
  xlab("Normalized Damage ($ Million)") +
  labs(title = "Density of Normalized Damage from Hurricanes",
       subtitle = "From 1954 to 2012")
```



It seems like the distributions are similar when plotted on top of each other, but do not seem like one is a shift of another. With that said, the distributions look similar enough to each other that we could look to see if one statistically is just a shift in mean when compared to the other.

```
wilcox_test(hurr$NDAM ~ hurr$Gender, dist = "exact", conf.int = TRUE)
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  hurr$NDAM by hurr$Gender (Female, Male)
## Z = 0.51641, p-value = 0.6089
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
##  -1950  1140
## sample estimates:
## difference in location
##                    135
```

Our confidence interval for the true mu of these distributions is (-1950, 1140). We see that a 95% confidence interval for the shift parameter between the distributions encapsulates 0. As this confidence interval crosses 0, it gives evidence that we are not really sure if one distribution really is a shift one way or the other from each other. This gives more evidence to the fact that the distributions of normalized damage are exchangeable and not statistically different from one another.

As we might be looking for more evidence that the models are not different, we can look at the Hodges-Lehmann estimate of the shift parameter.

```
NDAM.m <- hurr$NDAM[which(hurr$Gender == "Male")]
NDAM.f <- hurr$NDAM[which(hurr$Gender == "Female")]
median(outer(NDAM.m, NDAM.f, '-'))
```

```
## [1] -135
```

While the Hodges-Lehmann estimate is not equal to 0, the shift is rather small in comparison to the size of many of the normalized damage numbers. As this shift seems insignificant in the grand scheme of the overall normalized damage amounts, it would lead me to believe that the two distributions aren't really a shift of one another and the fact that the Hogdes-Lehmann estimate isn't equal to zero is just due to random variances between the two distributions.

In conclusion, when only using a hurricanes name's gender to explain normalized damage, we see that there really is no meaningful statistical difference between the two distributions.

Question 3.

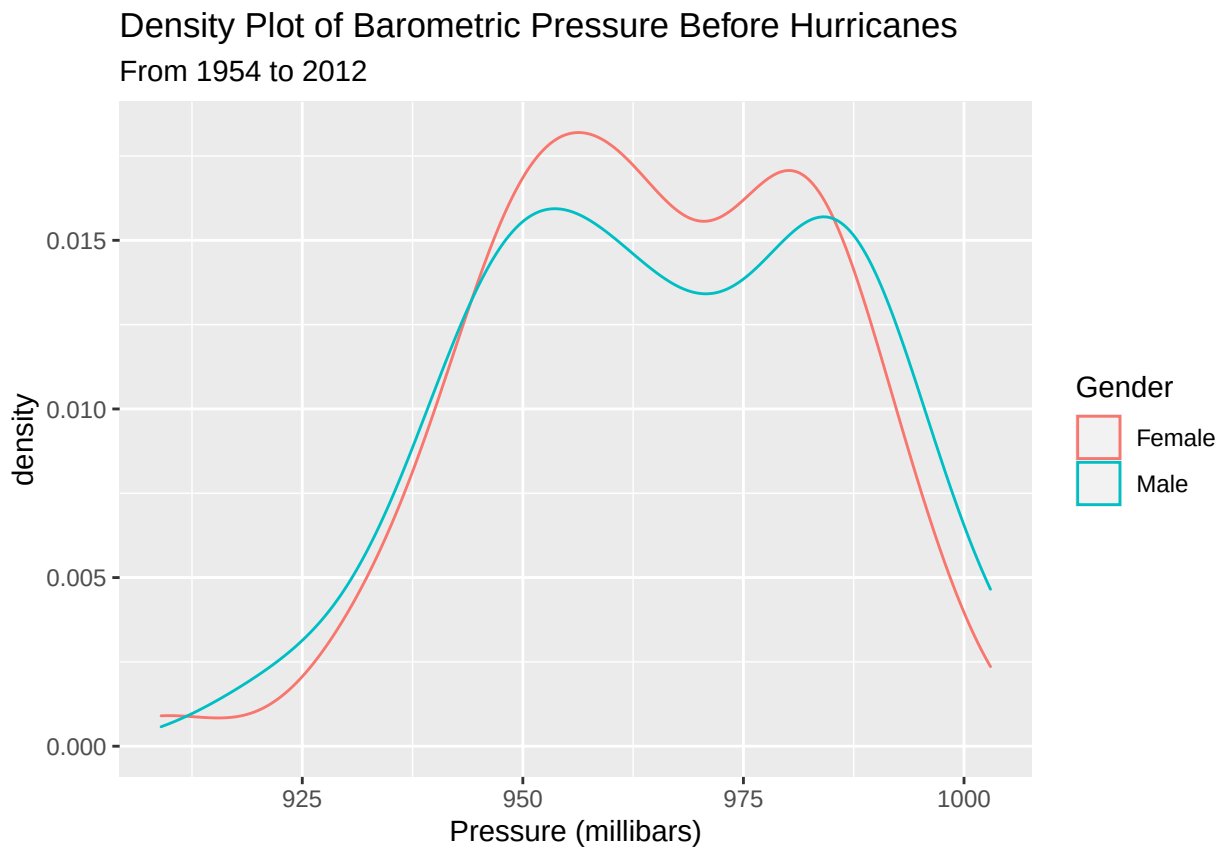
When it comes to the data we are given, we have the following other variables we can look at to see if there is a difference between female and male named hurricanes: min pressure before, and category.

We can first analyze the minimum barometric pressure from before the hurricanes. Usually the lower the pressure, the stronger a hurricanes winds are. While precipitation amount does effect the destructive power of a hurricane, the wind speeds also account for a lot of the destructive power of a hurricane.

There are two columns for the minimum barometric pressure before a hurricane, one is an adjusted number from 2014 probably based on new data. We will examine the adjusted numbers, so this means looking at the column named "Minpressure_Updated 2014"

Let's look at the density plots of the hurricanes in our dataset.

```
ggplot(hurr, aes(x = Minpressure_Updated.2014, color = Gender, group = Gender)) +
  geom_density() +
  xlab("Pressure (millibars)") +
  labs(title = "Density Plot of Barometric Pressure Before Hurricanes",
       subtitle = "From 1954 to 2012")
```



We see similar density plots between male and female hurricanes. The only difference seems to be that male named hurricanes have a slightly larger spread of barometric pressures which results in slightly less density at each pressure.

Again we could use a Wilcoxon rank-sum test to see if there is a difference in the barometric pressure distributions between male and female named hurricanes.

We can also just start with looking at the two means.

```
pres.m <- hurr$Minpressure_Updated.2014[which(hurr$Gender == "Male")]
pres.f <- hurr$Minpressure_Updated.2014[which(hurr$Gender == "Female")]
```

```
mean(pres.m)
```

```
## [1] 965.9667
```

```
mean(pres.f)
```

```
## [1] 964.4032
```

The means are very similar between the two groups.

```
wilcox_test(hurr$Minpressure_Updated.2014 ~ hurr$Gender, dist = "exact", conf.int = TRUE)
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  hurr$Minpressure_Updated.2014 by hurr$Gender (Female, Male)
## Z = -0.41662, p-value = 0.6801
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
##  -10   8
## sample estimates:
## difference in location
##                     -2
```

Based on the Wilcoxon rank-sum test, we get a large P-Value of about 0.68, which suggests that the two distributions are not statistically different. Also the confidence interval encapsulates 0 which also gives more evidence that there is no statistical shift between male and female named hurricanes when it comes to barometric pressure.

Next we can look at the category of the storm. Generally there is a correlation between the category of the storm and the barometric pressure as the category is based on windspeeds. If there is no difference between the barometric pressures of the two distributions it is likely that the category is not going to be different, but we should still do tests.

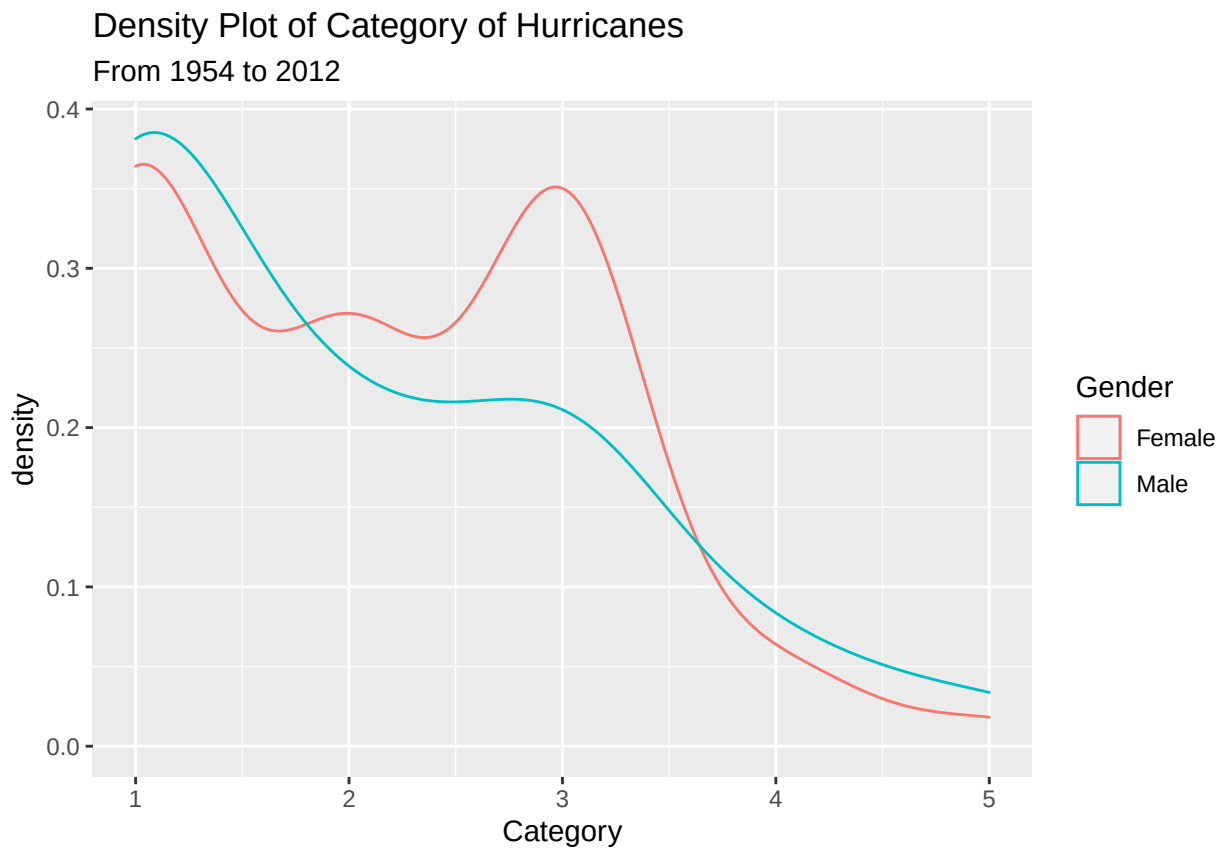
```
cor(hurr$Minpressure_Updated.2014, hurr$Category)
```

```
## [1] -0.8780495
```

As we can see there is a strong negative correlation as suggested, the lower the barometric pressure the higher the storms category will likely be.

Let's take a quick look at the density plot of the two distributions.

```
ggplot(hurr, aes(x = Category, color = Gender, group = Gender)) +
  geom_density() +
  xlab("Category") +
  labs(title = "Density Plot of Category of Hurricanes", subtitle = "From 1954 to 2012")
```



The shapes of the two distributions is slightly different. We see that there is a peak at category 3 for female named hurricanes where we don't really see as strong of a peak at category 3 for males. This could prove to be a slight difference between male and female named hurricanes.

Just to be safe, we can do a Wilcoxon rank-sum test on the category of the storm.

```
wilcox_test(hurr$Category ~ hurr$Gender, dist = "exact")
```

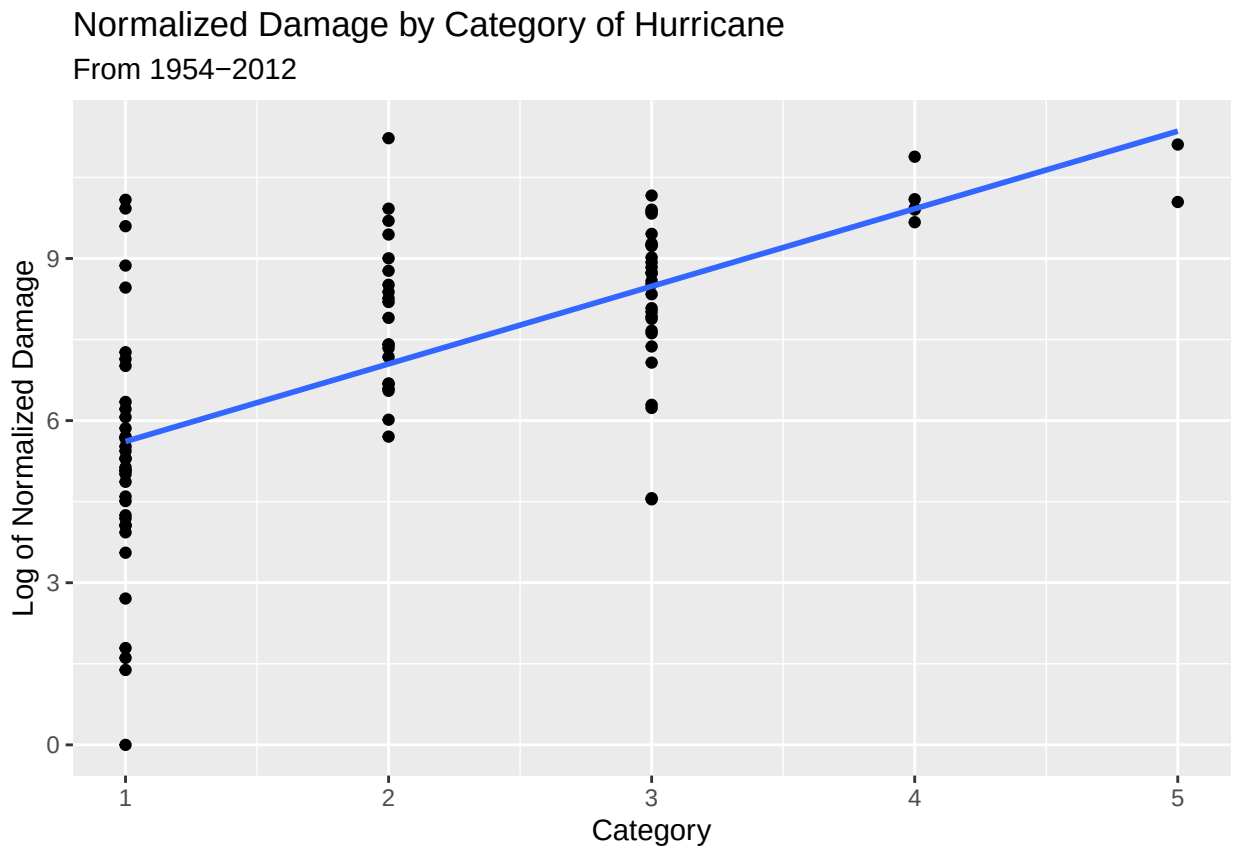
```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  hurr$Category by hurr$Gender (Female, Male)
## Z = 0.75507, p-value = 0.4525
## alternative hypothesis: true mu is not equal to 0
```

Again, we get a large P-Value which suggests that the distribution of categories is not statistically different between male and female named hurricanes. Although there does seem to be a higher density of category 3 hurricanes that have female names then male.

One thing we could do is try to fit a model that includes our predictor variables, basically we are left with the category of the storm and the barometric pressure. We can predict these models and see if the predictions are different for male and female named hurricanes.

```
ggplot(hurr, aes(x = Category, y = log(NDAM))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x) +
```

```
  ylab("Log of Normalized Damage") +
  labs(title = "Normalized Damage by Category of Hurricane", subtitle = "From 1954-2012")
```



As an initial test, we see that log damage doesn't really follow a linear pattern when using category. This suggests we should use something like a gam for a predictive model.

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following objects are masked from 'package:BSDA':
##
##     Gasoline, Wheat
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

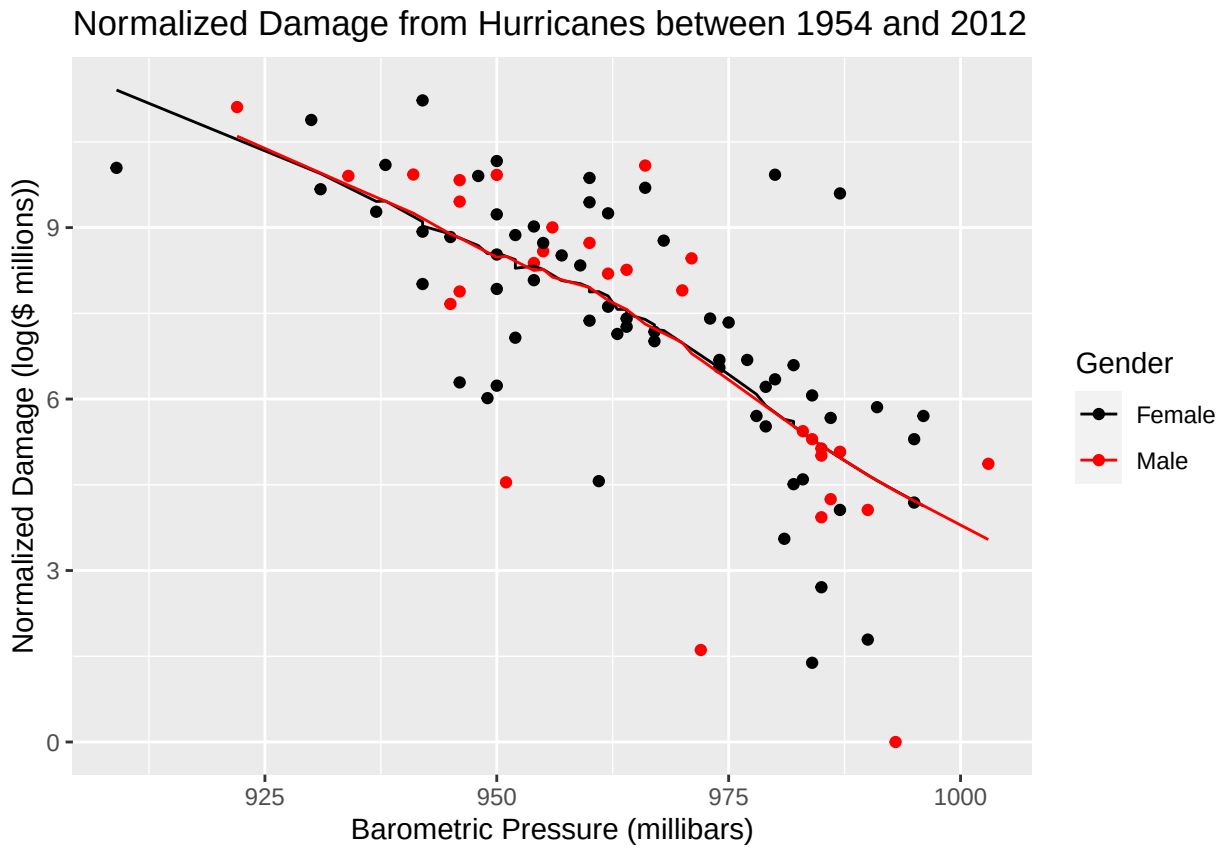
```
hurr.gam <- gam(log(NDAM) ~ s(Minpressure_Updated.2014, Category), data = hurr)
```

Now that we have a predictive model, lets see what the predictions are based on our original data.

```
hurr$preddam <- predict(hurr.gam, newdata = hurr)
```

Now we can plot our predictions in relation to our two predictive variables.

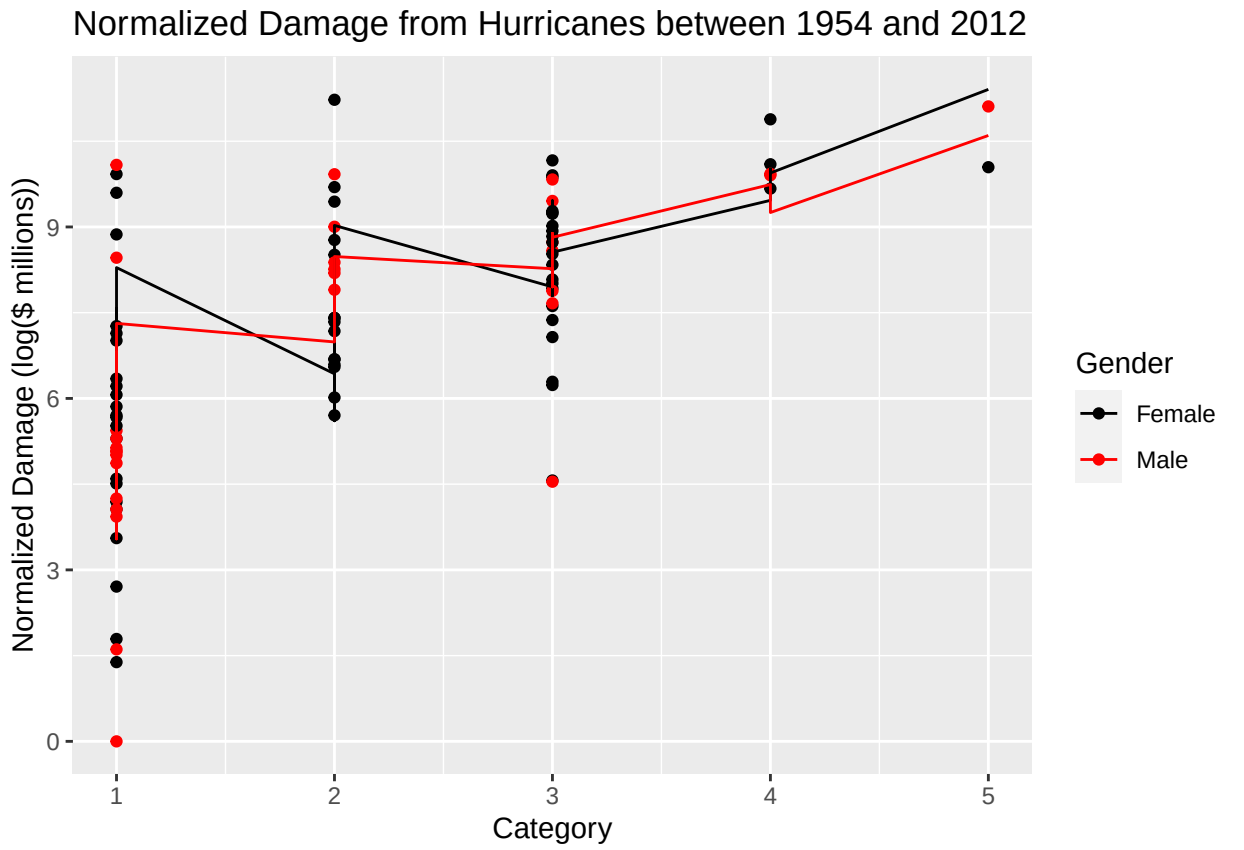
```
ggplot(hurr, aes(x = Minpressure_Updated.2014, y = log(NDAM), group = Gender, color = Gender)) +
  geom_point() +
  geom_line(aes(x = Minpressure_Updated.2014, y = preddam), data = hurr) +
  xlab("Barometric Pressure (millibars)") +
  ylab("Normalized Damage (log($ millions))") +
  labs(title = "Normalized Damage from Hurricanes between 1954 and 2012") +
  scale_color_manual(values = c("black", "red"))
```



We see that the splines are basically completely overlapping, as far as our predictive model is concerned, there really is no difference in male and female named hurricanes when it comes to barometric pressure predicting log normalized damage. Let's see if there is any difference when it comes to category.

```
ggplot(hurr, aes(x = Category, y = log(NDAM), group = Gender, color = Gender)) +
  geom_point() +
  geom_line(aes(x = Category, y = preddam), data = hurr) +
  xlab("Category") +
```

```
ylab("Normalized Damage (log($ millions))") +
labs(title = "Normalized Damage from Hurricanes between 1954 and 2012") +
scale_color_manual(values = c("black", "red"))
```



We do see some slight differences when it comes to category of storm. There really isn't a meaningful different. You could mabye say that for category 1 and 2 that female named hurricanes are predicted to cause more log damage. For categories 4 and 5, there really isn't much data so the right tail of the predictions are pretty dependent on those few datapoints and aren't as reliable.

Really we do not see much of a difference between male and female named hurricanes when looking at a model that predicts damage from barometric pressure and storm category.

The last difference we can identify is the total count of hurricanes with female names compared to hurricanes with male names over the same time period (1954-2012). There are twice as many female named hurricanes than male, this alone could effect different measures done to compare the two. Overall though it seems that for the most part there is no major differences between female named hurricanes and male when using the binary male or female marker.

Part B: Positive patience

Does putting people in a positive mood affect their patience? Ifcher and Zarghamee (2011)[1] performed an experiment in which subjects were asked a series of questions of the form

"What amount of money, $\$p$, if paid to you today would make you indifferent to $\$m$ paid to you in t days?"

Here, p is called the *present value* and m the *future value*. The idea is that if people are patient, the value of p will be equal to or very close to m. If people are impatient, p may be quite a lot less than m. For any answer to any question, the *discount* is

$$1 - \frac{\text{present value}}{\text{future value}}$$

So if $p = m$, there was no discounting. If p was small compared to m, the discount was close to 1. (There were no negative values of the discount in the original study.)

In the experiment, 34 subjects were randomly assigned to the treatment group and watched clips from *Robin Williams: Love on Broadway* before being asked the discounting questions. (Robin

[1] https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.101.7.3109

Williams was alive at the time of the experiment.) This treatment was intended to induce "positive affect." The other 35 subjects were assigned to the control group and watched images of landscapes and wildlife in Denali National Park, Alaska. This was considered a neutral video. After the videos, each subject was asked the same set of 30 discounting questions (in random order.)

Ifcher and Zarghamee concluded that "compared to neutral affect, mild positive affect significantly reduces time preference over money." That is, watching the Robin Williams clips reduced discounting compared to the control. How strong was the evidence for this? We note that Camerer et al.[2] recently attempted a replication and did not get significant results, but the death of Robin Williams makes the interpretation of the replication ambiguous. Instead, we will analyze data from Ifcher and Zarghamee's original study.

The STATA file `ifcher11.dta` contains data from the experiment. The variables include:

- `id`: the unique identifier for the individual in the experiment. There are 69 unique IDs.
- `fv`: future value (m in the notation above)
- `pv`: present value (p in the notation above)
- `delay`: the number of days from the present until the future value could be received (t in the notation above)
- `treatment`: "1" is the positive affect treatment (Robin Williams); "0" is the neutral control (Alaskan landscapes.) Note: The data header says "2" is control, which is not correct.
- `happiness`: Self-reported happiness on a scale from 1 to 7, where 7 means "completely happy."

Questions

1. (5 points.) We'll first take a look at discounting in the study as a whole (before comparing treatment and control groups.) We will simplify Ifcher and Zarghamee's original analysis (which uses cluster standard errors) by averaging the discount over all of each individual's answers to get 69 exchangeable data points (under the null.)

   (a) Calculate the mean value of "discount" for each individual. That is, for each individual, find the average of $1 - p/m$ over all 30 questions. (One method would be to use the `aggregate()` command — see the notes for examples — but there are lots of ways to do this.) For simplicity, ignore the small number of NA's when calculating means. You should end up with 69 values. Call this variable `meanDiscount` or something similar.

   (b) Draw a clearly-labeled graph of the distribution of `meanDiscount`.

   (c) Find a 95% confidence interval for the population *median* value of `meanDiscount`.

2. (5 points.) Perform a nonparametric test of the hypothesis that the distribution of `meanDiscount` is the same for both the treatment and the control. Justify your choice of test, and give a P-value and a substantive conclusion.

[2] `https://experimentaleconreplications.com/finalreports/Ifcher%20&%20Zarghamee%202011.pdf`

3. (5 points.) Ifcher and Zarghamee also re-analyzed the data after excluding all answers for which there was no discounting: that is, leaving out answers for which $p = m$. One justification for this might be that if people were never going to discount anyway, then including them in the analysis doesn't add useful information.

   (a) Recalculate the mean discount for each individual after the answers for which $p = m$ have been omitted.

   (b) Perform a nonparametric test of the hypothesis that the distribution of `meanDiscount` is the same for both the treatment and the control, *after* excluding answers with no discounting.

   (c) Is it appropriate to exclude the non-discounting answers? That is, does excluding the answers for which there was no discount improve the analysis (e.g. by increasing power)? Or does it make it worse (e.g. by introducing bias or additional variance)? Explain.

Part B: Positive Patience

We will start by importing the data.

```
library(haven)
pos <- read_dta(file = "ifcher11.dta")
head(pos)
```

```
## # A tibble: 6 x 13
##      id question treatment    pv delay    fv happiness college gender  race
##   <dbl>    <dbl>     <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>  <dbl> <dbl>
## 1  1000        1         1  20      28  24.3         6       3      2     1
## 2  1000        2         1  18       1  18.3         6       3      2     1
## 3  1000        3         1  18       3  18.3         6       3      2     1
## 4  1000        4         1  32.8     1  32.8         6       3      2     1
## 5  1000        5         1  15      56  24.3         6       3      2     1
## 6  1000        6         1  11.3    28  11.3         6       3      2     1
## # ... with 3 more variables: religion <dbl>, practice_religion <dbl>,
## #   family_income <dbl>
```

Question 1.

Part (a)

We will start by creating a new column that includes the discount for each question, so that we can aggregate the discount.

```
pos$discount <- 1 - (pos$pv/pos$fv)
head(pos)
```

```
## # A tibble: 6 x 14
##      id question treatment    pv delay    fv happiness college gender  race
##   <dbl>    <dbl>     <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>  <dbl> <dbl>
## 1  1000        1         1  20      28  24.3         6       3      2     1
## 2  1000        2         1  18       1  18.3         6       3      2     1
## 3  1000        3         1  18       3  18.3         6       3      2     1
## 4  1000        4         1  32.8     1  32.8         6       3      2     1
## 5  1000        5         1  15      56  24.3         6       3      2     1
## 6  1000        6         1  11.3    28  11.3         6       3      2     1
## # ... with 4 more variables: religion <dbl>, practice_religion <dbl>,
## #   family_income <dbl>, discount <dbl>
```

Now we will aggregate the data to find the mean discount for each individual. When we aggregate, we can also take the mean of various other columns to be able to keep some of the variables with the data. In many cases these extra columns are all the same for each question. We will then save just the two columns we care about to it's own variable: mean discount, and treatment.

```
pos.agg <- aggregate(pos, by = list(pos$id), FUN = mean, na.rm = TRUE)
meanDiscount <- pos.agg %>% select(treatment, discount)
head(meanDiscount)
```

```
##   treatment   discount
## 1         1 0.16081337
## 2         1 0.01191801
## 3         1 0.29115223
## 4         1 0.08250264
## 5         1 0.27121878
## 6         1 0.35030085
```

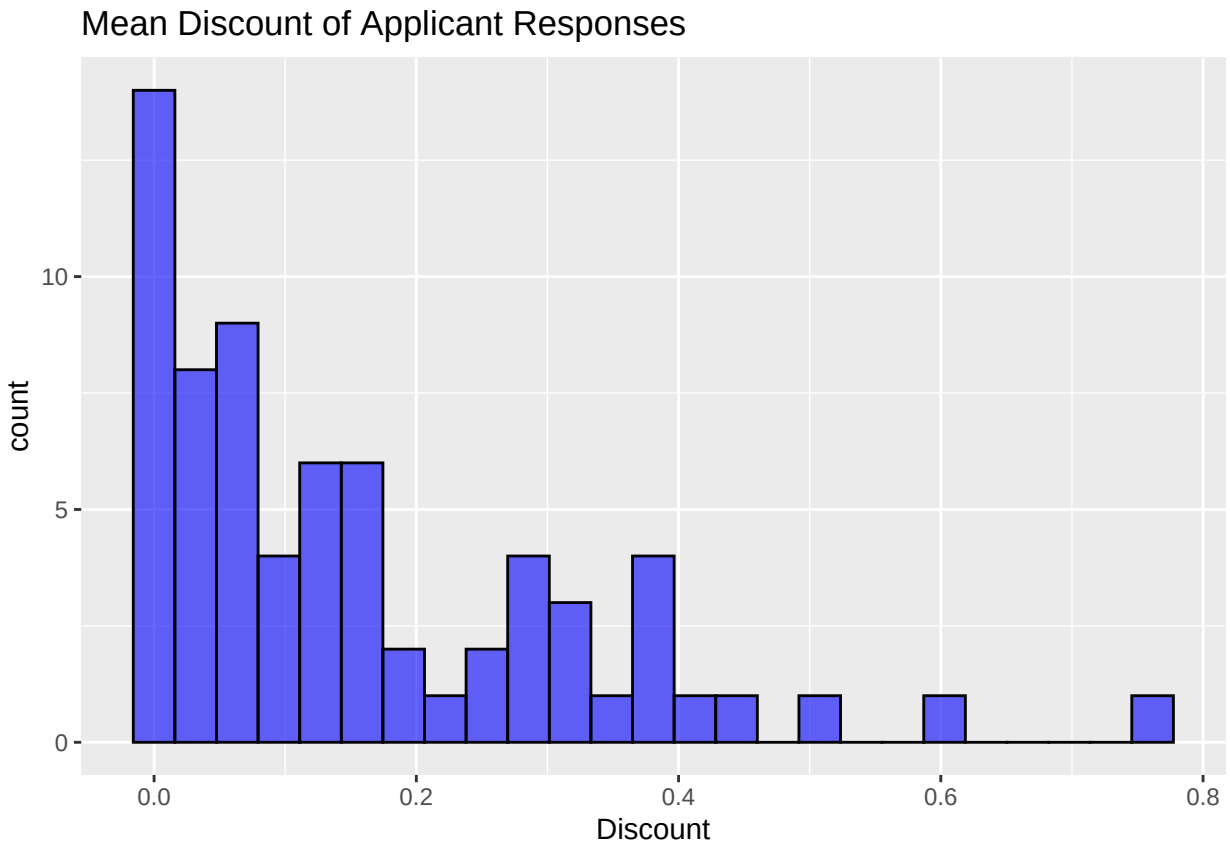
```
aggregate(pos, by = list(pos$id), FUN = mean, na.rm = TRUE) %>%
   select(treatment, discount) %>%
   head()
```

```
##   treatment   discount
## 1         1 0.16081337
## 2         1 0.01191801
## 3         1 0.29115223
## 4         1 0.08250264
## 5         1 0.27121878
## 6         1 0.35030085
```

Part (b)

We can use a density plot in ggplot to see the distribution of the mean discount of each participant in the study.

```
ggplot(meanDiscount, aes(x = discount)) +
  geom_histogram(bins = 25, fill = "blue", color = "black", alpha = 0.6) +
  xlab("Discount") +
  labs(title = "Mean Discount of Applicant Responses")
```



The histogram plot shows that the distribution of the mean discount is a right skewed distribution.

Part (c)

We can start by finding the median of our mean discount as a point of reference.

```
median(meanDiscount$discount)
```

```
## [1] 0.09267998
```

```
est <- numeric(9999)
```

To find a confidence interval we first must find a value of k such that the probability that the confidence interval does not contain the true population mean is no more than 5%. We can do this using a vector of different values of k and using pbinom.

```
k = c(25, 26, 27, 28, 29)
1 - 2 * pbinom(k - 1, 69, 0.5)
```

```
## [1] 0.9845677 0.9705065 0.9467106 0.9088135 0.8519678
```

We see that the value of k that gives us a value that is at least 0.95 is 26. We can sort the values in mean discount and then move 26 spaces from each endpoint, these will be our endpoints of our confidence interval.

```r
meanDiscount.sort <- meanDiscount[order(meanDiscount$discount), 2]
```

```r
meanDiscount.sort[26]
```

```
## [1] 0.05145358
```

```r
meanDiscount.sort[69 - 26 + 1]
```

```
## [1] 0.1608134
```

This gives us a 95% confidence interval of about [0.051, 0.161] for the population median of the mean discount from all respondents.

Just as a final test we can use the sign test in the package BSDA to see if it gives us the same confidence interval as it uses a similar process.

```r
SIGN.test(meanDiscount$discount, md = 0.1)
```

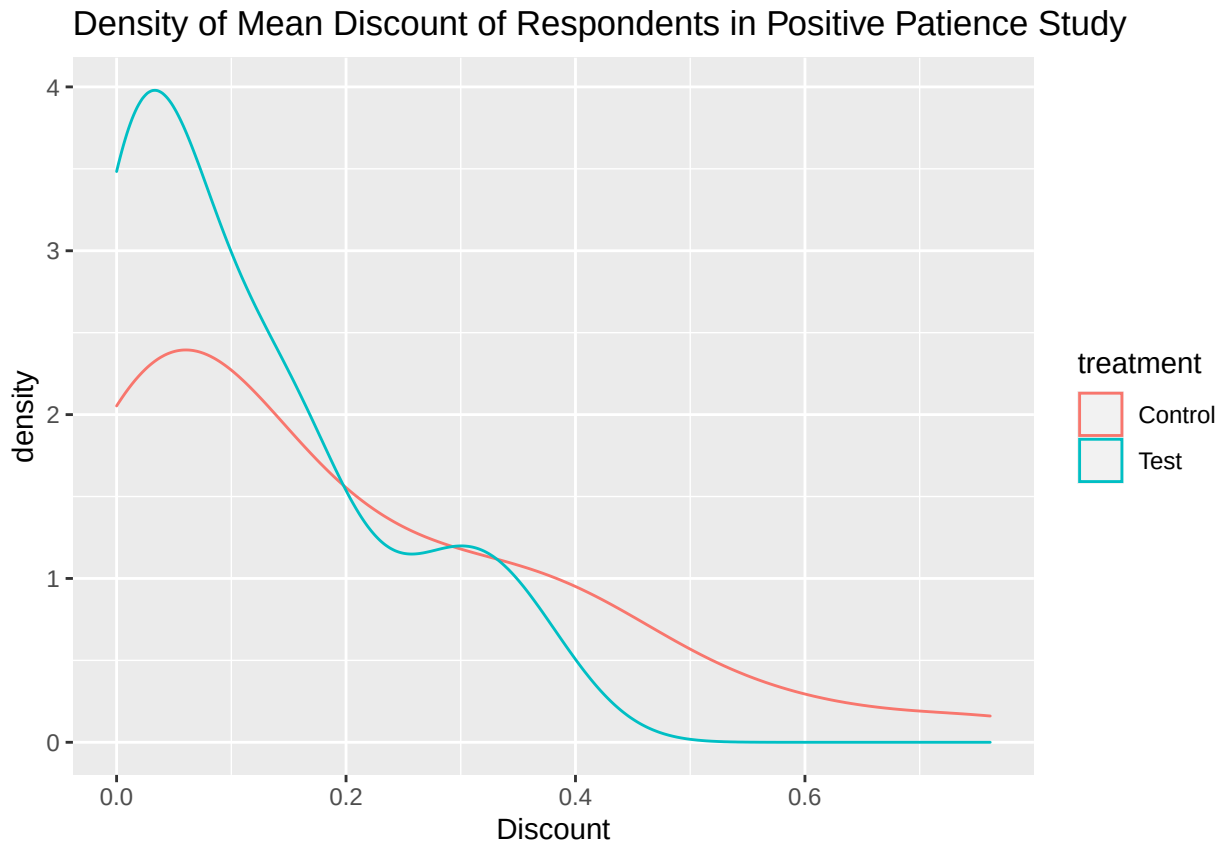
```
##
##  One-sample Sign-Test
##
## data:  meanDiscount$discount
## s = 34, p-value = 1
## alternative hypothesis: true median is not equal to 0.1
## 95 percent confidence interval:
##  0.05186125 0.15737033
## sample estimates:
## median of x
##  0.09267998
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9467 0.0519 0.1568
## Interpolated CI       0.9500 0.0519 0.1574
## Upper Achieved CI     0.9705 0.0515 0.1608
```

We see that the sign test library also gives us a confidence interval of about [0.051, 0.161].

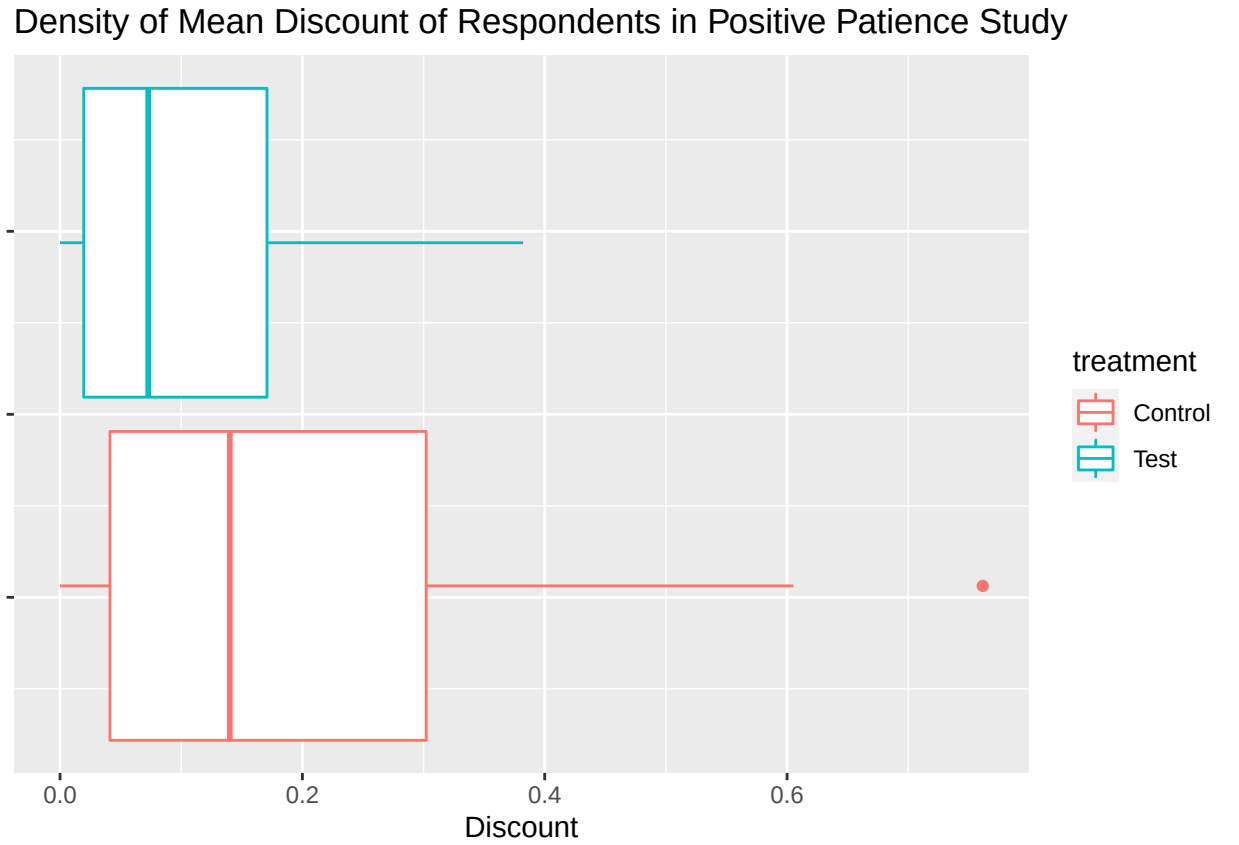
Question 2.

We want to choose a non-parametric test to see if the distributions are statistically similar between the control and test. Right away we can see that we want some sort of permutation test. Before choosing a test and doing it, let's first look at the distributions plotted on the same plot to see what that looks like.

```r
meanDiscount$treatment <- recode_factor(meanDiscount$treatment,
                                        '0' = "Control", '1' = "Test")
ggplot(meanDiscount, aes(x = discount, color = treatment, group = treatment)) +
  geom_density() +
  xlab("Discount") +
  labs(title = "Density of Mean Discount of Respondents in Positive Patience Study")
```




```
ggplot(meanDiscount, aes(x = discount, color = treatment, group = treatment)) +
  geom_boxplot() +
  xlab("Discount") +
  labs(title = "Density of Mean Discount of Respondents in Positive Patience Study") +
  theme(axis.text.y = element_blank())
```



We see from the density plot that the two groups seem to have somewhat differing shapes to their distributions. Although the shape is different, both still have the greatest density around a discount of 0.05.

The test that we will use to determine if the two distributions are the same is the Wilcoxon rank-sum test. This is a permutation test in which the response variable, discount, is shuffled and then a mean is computed for the two groups. This is done many times and we see if over all those scrambles of the response variable produced meaningful differences in means of the two groups. In this case we want to do a two tailed test as we don't really care which way the shift goes, we just care if there is a substantial difference in the distributions.

The Wilcoxon rank-sum test is chosen here because we do not have data that seems like it is pulled from a specific distribution, and measuring the difference in means does make sense as we are looking at the mean of each respondents discount. We have an outlier that could effect our results, but the outlier is not particularly egregious.

```
wilcox_test(discount ~ treatment, dist = "exact", conf.int = TRUE, data = meanDiscount)
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  discount by treatment (Control, Test)
## Z = 1.4432, p-value = 0.1515
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
##  -0.01159138  0.11504716
## sample estimates:
## difference in location
```

```
##              0.03778543
```

For this test we get a P-Value of about 0.1515, we also get a confidence interval that encapsulates 0. This confidence interval should be taken with a grain of salt though. We didn't really see any reason to believe that the treatment groups distribution could be explained by a shift of the control group in the density plot.

This P-value sits squarely in a position where it is hard to evaluate it's significance. While by some standards a P-Value of about 0.15 is high, it really isn't extremely high or low. Our sample sizes aren't very large at 34 and 35 for the test and control group respectively. With these sizes, we run the risk of not really having enough power in the test which makes it hard for us to reject the null even if the populations actually are different.

With all that said, we should fail to reject the null hypothesis that the distributions are the same, but with the caveat that our power is likely not too great so there is a decent chance of making a Type II error. As in many cases in statistics it would definitely be helpful to get a larger sample size so that we could increase our power and be able to reject the null if in fact the alternative is true.

Question 3.

Part (a)

We need to go back to our original data and subset it to only choose instances where the respondents in the study have a present value that does not equal the future value.

```
pos.neq <- subset(pos, pv != fv)
length(pos$id) - length(pos.neq$id)
```

```
## [1] 599
```

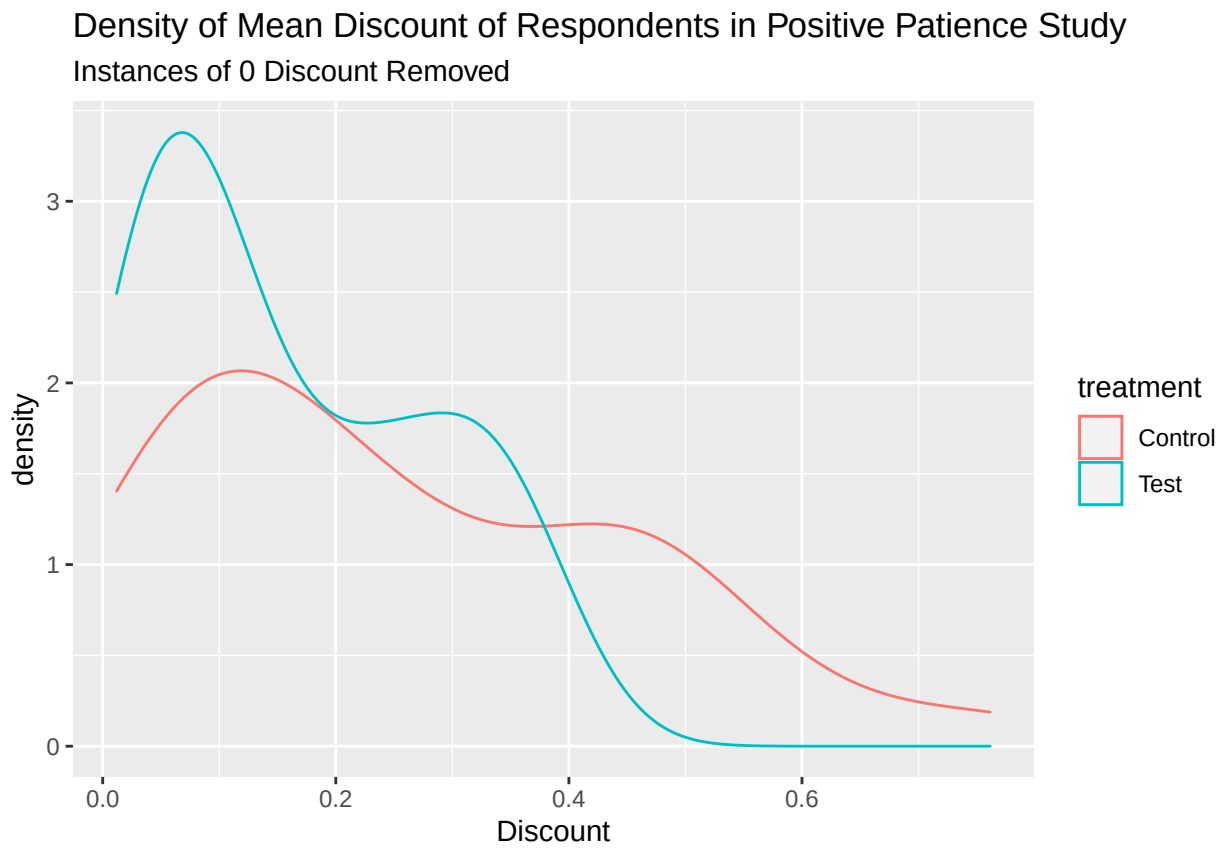
We can see that this removes about 600 answers to questions, and the total length of the original data was about 2000 so we removed around 25% of our data. Now we can reaggregate the data.

```
pos.agg.neq <- aggregate(pos.neq, by = list(pos.neq$id), FUN = mean, na.rm = TRUE)
meanDiscount.neq <- pos.agg.neq %>% select(treatment, discount)
head(meanDiscount.neq)
```

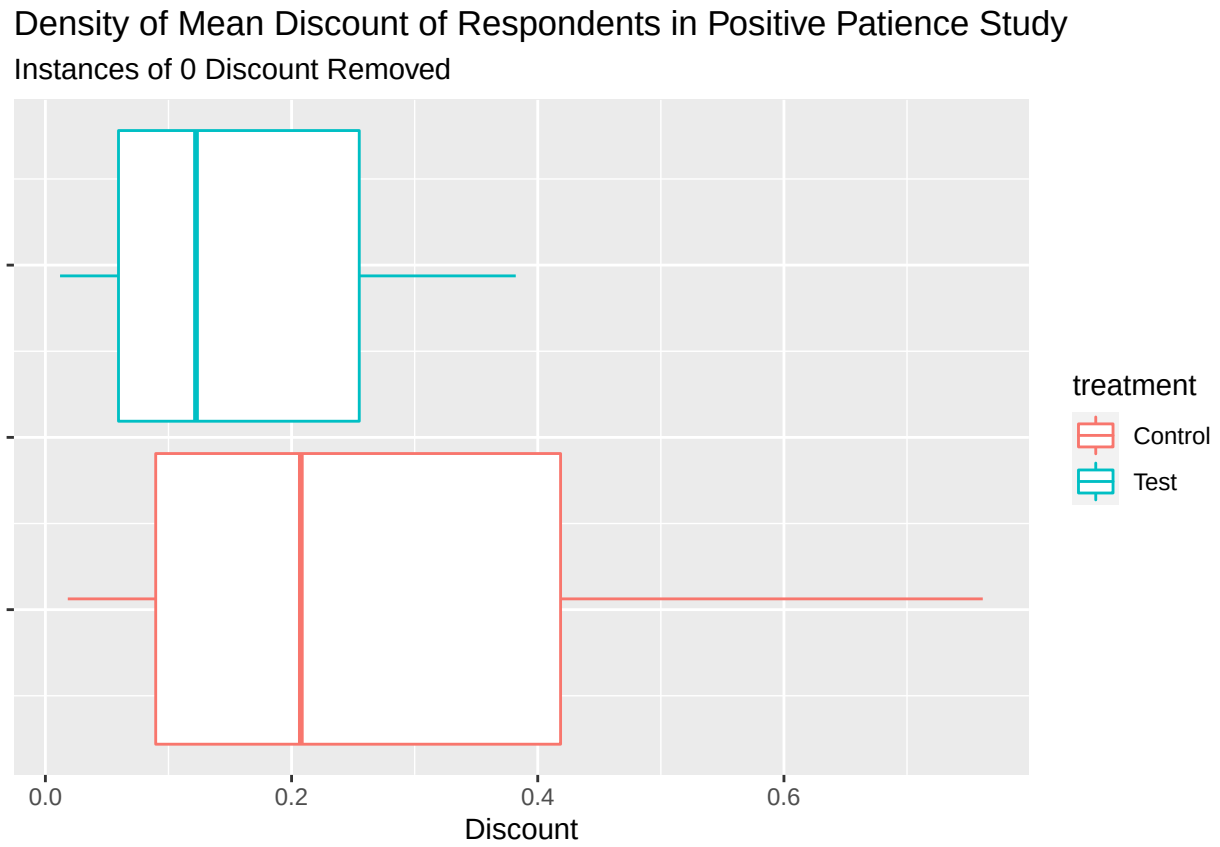
```
##   treatment   discount
## 1         1 0.21929096
## 2         1 0.01191801
## 3         1 0.30119196
## 4         1 0.09519536
## 5         1 0.32546253
## 6         1 0.35030085
```

Let's look at some plots to see how the distributions changed by removing instances where the discount was 0.

```
meanDiscount.neq$treatment <- recode_factor(meanDiscount.neq$treatment,
                                            '0' = "Control", '1' = "Test")
ggplot(meanDiscount.neq, aes(x = discount, color = treatment, group = treatment)) +
  geom_density() +
  xlab("Discount") +
  labs(title = "Density of Mean Discount of Respondents in Positive Patience Study",
       subtitle = "Instances of 0 Discount Removed")
```




```
ggplot(meanDiscount.neq, aes(x = discount, color = treatment, group = treatment)) +
  geom_boxplot() +
  xlab("Discount") +
  labs(title = "Density of Mean Discount of Respondents in Positive Patience Study",
       subtitle = "Instances of 0 Discount Removed") +
  theme(axis.text.y = element_blank())
```



Removing the 0 discounts from the data did make a rather substantial change in the distributions in comparison of leaving them in. We also see that our outlier that we had in the control group is not an outlier.

Part (b)

For consistency, and for the same reasons, we will be using the Wilcoxon rank-sum test on the subset of our data.

```
wilcox_test(discount ~ treatment, dist = "exact", conf.int = TRUE, data = meanDiscount.neq)
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  discount by treatment (Control, Test)
## Z = 1.9606, p-value = 0.05033
## alternative hypothesis: true mu is not equal to 0
## 95 percent confidence interval:
##  -0.0003966364  0.1653611800
## sample estimates:
## difference in location
##             0.07014378
```

With our subset of our data that doesn't include discounts of 0, we get a P-value that is significant. With a P-Value of 0.05 we can reject the null in favor of the alternative hypothesis that the distributions are different. While in this case we did find that the distributions are different between control and treatment

are different, see the discussion in the next part to see if this really means that this method provides principled and significant evidence or not.

Part (c)

To decide if it is appropriate to exclude the non-discounting answers, we need to understand if it is done for a principled reason. I would lean on the side of saying that it is not appropriate in terms of a study. The idea of the study is to determine if mood influences patience, I believe that in cases where the discount is 0 still demonstrates whether the persons mood effects patience. If a person is willing to not discount the payment even after waiting that says something about their patience and then their mood could be evaluated to see if that had an effect on their patience. If there is no significance between mood and patience, then maybe patience when it comes to monetary compensation is based on some other variable.

We can look at how many participants were removed entirely because they had a 0 discount.

```
length(pos.agg$id)
```

```
## [1] 69
```

```
length(pos.agg.neq$id)
```

```
## [1] 58
```

We see that 11 participants were not included in the second analysis. This leads us to wonder why they didn't want a discount to begin with. Maybe the amount of money was so insignificant that any small change to it was not really going to impact them. For example someone who is financially secure might not see much difference between being given $20 instead of $25, whereas someone living paycheck to paycheck would see the extra $5 as something meaningful.

```
paste("Control Variance with 0s:", var(pos.agg$discount[which(pos.agg$treatment == 0)]))
```

```
## [1] "Control Variance with 0s: 0.037859816233869"
```

```
paste("Treatment Variance with 0s:", var(pos.agg$discount[which(pos.agg$treatment == 1)]))
```

```
## [1] "Treatment Variance with 0s: 0.0133903600931754"
```

```
paste("Control Variance without 0s:", var(pos.agg.neq$discount[which(pos.agg.neq$treatment == 0)]))
```

```
## [1] "Control Variance without 0s: 0.0383950331474936"
```

```
paste("Treatment Variance without 0s:", var(pos.agg.neq$discount[which(pos.agg.neq$treatment == 1)]))
```

```
## [1] "Treatment Variance without 0s: 0.0145680246605553"
```

By reducing the sample size we are likely to decrease our power, meaning that it would take a larger difference in distributions to find significant findings. We see that in this case we did reject the null and our concern would be Type I error, but we should have a good reason to reduce the power and I do not believe that

this is a principled reason to do so. By removing the 11 participants that all had a discount of 0 we would increase our variance in both groups.

In the end I do not believe that excluding non-discounting answers is appropriate in terms of the overall study. It seems as though we would be removing answers from the study that just didn't meet our expectations that somehow patience could be measured monetarily for all respondents in this way. If Ifcher and Zarghamee saw people responding with a 0 discount as an issue prior to running the experiment, they should have constructed the study in such a way that this is avoided. As they did not construct the study in such a way, they opened them up to having people who would not care enough about the money to want a discount, and therefore should include all results.

Part C: Coronavirus and partisanship

The file `covid-s681-042520.txt` contains the following variables, measured on almost all U.S. counties.

- `date` gives the date of the counts (April 25th for all observations.)
- `fips` gives the unique FIPS code for the county.
- `county` gives the name of the county. (Exception: New York City is counted as one observation even though it's five counties.)
- `state` gives the state the county is in.
- `census_region` gives the region the county is in: Northeast, Midwest, South, or West.
- `cases` gives the total number of confirmed COVID-19 cases in that county up to that date.
- `deaths` gives the total number of COVID-19 deaths in that county up to that date.
- `pop` gives the population of the county.
- `land_area` gives the land area of the county in square miles.
- `white` gives the percentage of the population who are white.
- `black` gives the percentage of the population who are black.
- `DemShare` gives the proportion of the two-party vote (Democratic and Republican) that went to the Democratic candidate for President in 2016.

Starting at a map of COVID-19 cases per capita, one might be struck by its resemblance to a map of the results of the 2016 election: counties that voted more Democrat look like they have higher rates of infection, even after accounting for population.

Questions

1. (5 points.) Describe the relationship between `DemShare` and COVID-19 cases per 10,000 people at the county level in a graph or graphs and in words.
2. (10 points.) It could be that, for whatever reason, COVID-19 spreads faster in Democratic counties, all other things being equal. But is there an alternative explanation? Build a model that attempts to see if the relationship between `DemShare` and COVID-19 cases per 10,000 people persists after sensibly accounting for other important features provided in the data set. If there's an alternative explanation what is it, and is it convincing?

NOTE: If for personal reasons, you don't want to complete this question, let me know and I'll send you a different question.

Part C: Coronavirus and partisanship

We will start by importing the coronavirus data and saving it to a variable.

```
corona <- read.table("covid-s681-042520.txt", header = TRUE)
head(corona)
```

```
##         date fips  county   state census_region cases deaths    pop land_area
## 1 2020-04-25 1001 Autauga Alabama         South    37      2  55395    594.44
## 2 2020-04-25 1003 Baldwin Alabama         South   154      3 200111   1589.78
## 3 2020-04-25 1005 Barbour Alabama         South    33      0  26887    884.88
## 4 2020-04-25 1007    Bibb Alabama         South    35      0  22506    622.58
## 5 2020-04-25 1009  Blount Alabama         South    31      0  57719    644.78
## 6 2020-04-25 1011 Bullock Alabama         South    12      0  10764    622.81
##   white black   DemShare
## 1  78.1  18.4 0.24598218
## 2  87.3   9.5 0.20187742
## 3  50.2  47.6 0.47164121
## 4  76.3  22.1 0.21772975
## 5  96.0   1.8 0.08614472
## 6  27.2  69.9 0.75605055
```

Question 1.

Before any graphing or analysis, we should create a column that looks at the COVID-19 cases per 10,000 people.

```
corona$case.bypop <- corona$cases / (corona$pop / 10000)
head(corona)
```

```
##         date fips  county   state census_region cases deaths    pop land_area
## 1 2020-04-25 1001 Autauga Alabama         South    37      2  55395    594.44
## 2 2020-04-25 1003 Baldwin Alabama         South   154      3 200111   1589.78
## 3 2020-04-25 1005 Barbour Alabama         South    33      0  26887    884.88
## 4 2020-04-25 1007    Bibb Alabama         South    35      0  22506    622.58
## 5 2020-04-25 1009  Blount Alabama         South    31      0  57719    644.78
## 6 2020-04-25 1011 Bullock Alabama         South    12      0  10764    622.81
##   white black   DemShare case.bypop
## 1  78.1  18.4 0.24598218   6.679303
## 2  87.3   9.5 0.20187742   7.695729
## 3  50.2  47.6 0.47164121  12.273589
## 4  76.3  22.1 0.21772975  15.551409
## 5  96.0   1.8 0.08614472   5.370848
## 6  27.2  69.9 0.75605055  11.148272
```

There are some rows where we do not have some data so that causes us to not have data for the coronavirus cases by population. For now we will ignore those datapoints.

```
corona <- subset(corona, !is.na(corona$case.bypop))
```

We will start by creating a map of the US by county that shows coronavirus cases per 10,000 residents.

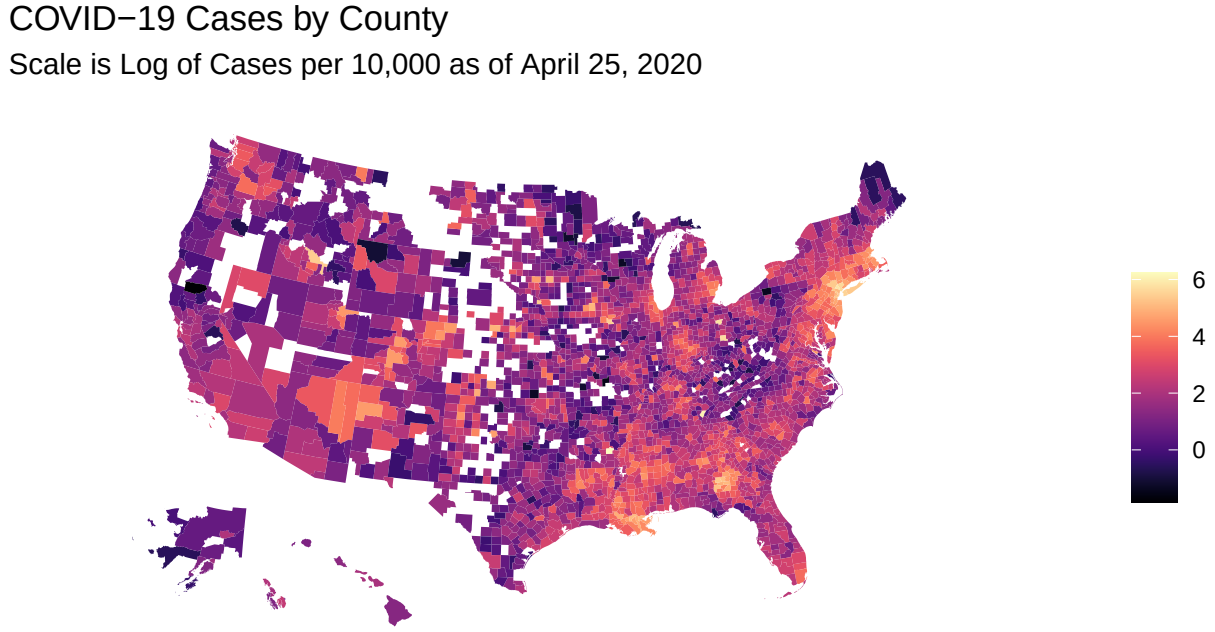
```
## devtools::install_github("UrbanInstitute/urbnmapr")

library(tidyverse)
library(urbnmapr)
library(viridis)
```

```
## Loading required package: viridisLite
```

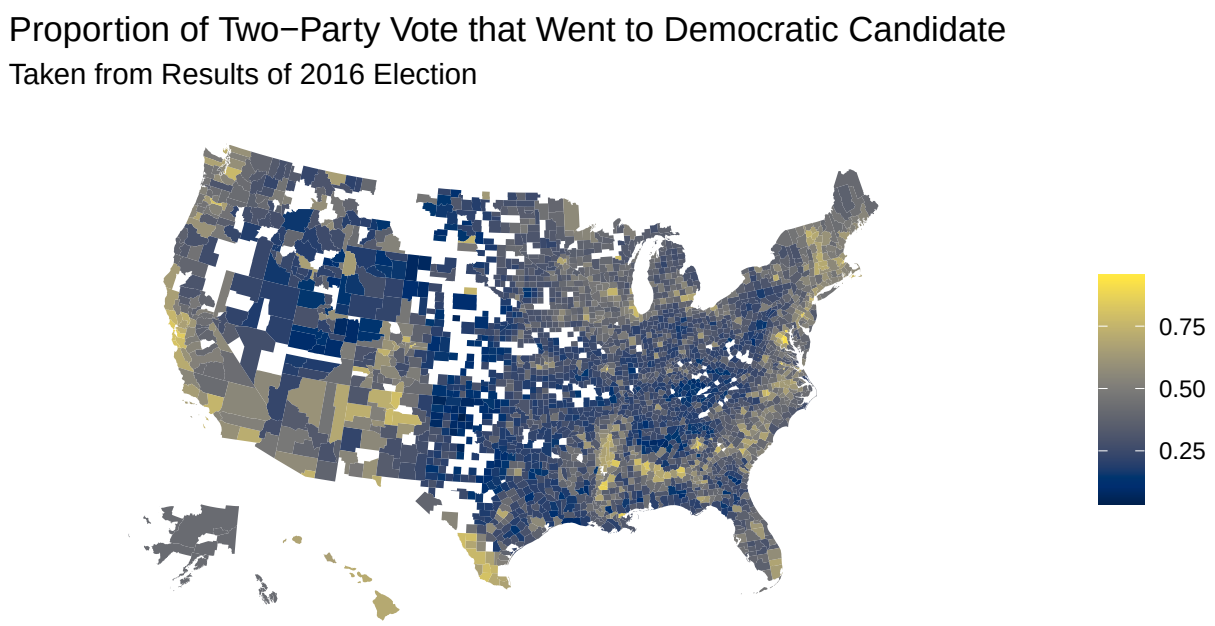
```
data(counties)
counties$county_fips <- as.numeric(counties$county_fips)
corona_data <- left_join(corona, counties, by = c("fips" = "county_fips"))

p1 <- corona_data %>%
  ggplot(aes(long, lat, group = group, fill = log(case.bypop))) +
  geom_polygon(color = NA) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "COVID-19 Cases by County",
       subtitle = "Scale is Log of Cases per 10,000 as of April 25, 2020") +
  scale_fill_viridis(option="magma") +
  theme(legend.title = element_blank(),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.background = element_blank())
p1
```




```
p2 <- corona_data %>%
  ggplot(aes(long, lat, group = group, fill = DemShare)) +
  geom_polygon(color = NA) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "Proportion of Two-Party Vote that Went to Democratic Candidate",
```

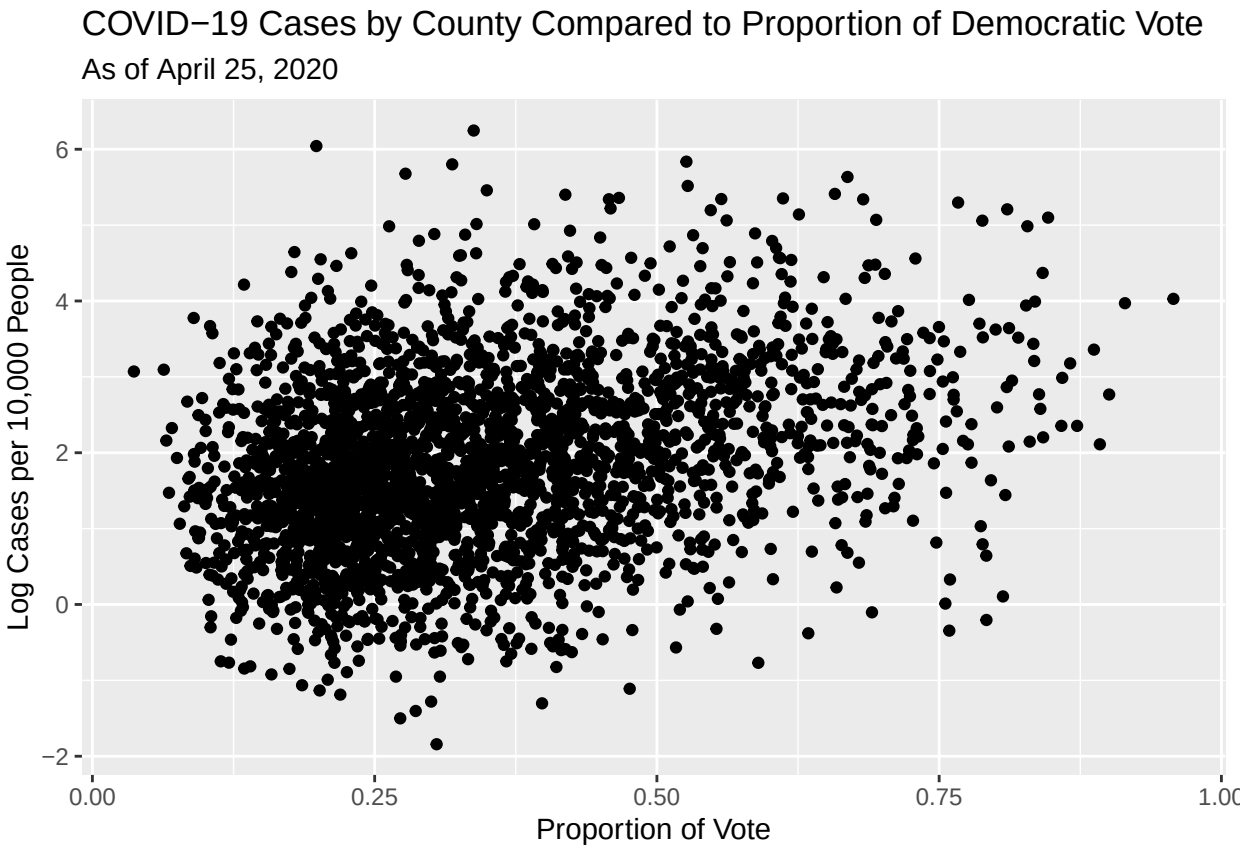
```
       subtitle = "Taken from Results of 2016 Election") +
  scale_fill_viridis(option="cividis") +
  theme(legend.title = element_blank(),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.background = element_blank())
p2
```



When looking at these two plots of the US map, there definitely are some striking similarities between infection rates and counties political leanings. This is especially true when focusing on the coasts.

We can plot each county by it's democratic leanings compared to the COVID-19 cases per 10,000. As we know that there are places in the United States that there are vastly more infections than other places we should do a log transformation to make our plot show the distribution best.

```
ggplot(corona, aes(x = DemShare, y = log(case.bypop))) +
  geom_point() +
  xlab("Proportion of Vote") +
  ylab("Log Cases per 10,000 People") +
  labs(title = "COVID-19 Cases by County Compared to Proportion of Democratic Vote",
       subtitle = "As of April 25, 2020")
```



There does seem to be some level of positive correlation between the log of COVID-19 cases and Democratic vote. This means that it seems like as the proportion of democratic votes increases, so too does the log of cases of COVID-19 per 10,000 people. The correlation doesn't seem to be terribly strong.

```
cor(corona$case.bypop, corona$DemShare)
```

```
## [1] 0.220031
```

As believed there is a rather weak correlation between the two variables. With that said, we have a great number of counties (~2800), so a dispersion such as this is probable even if there is some correlation between the two variables.

Question 2.

As we look at building a model, let's look at the variables we have available to us to predict the cases of COVID-19 per 10,000 county residents.

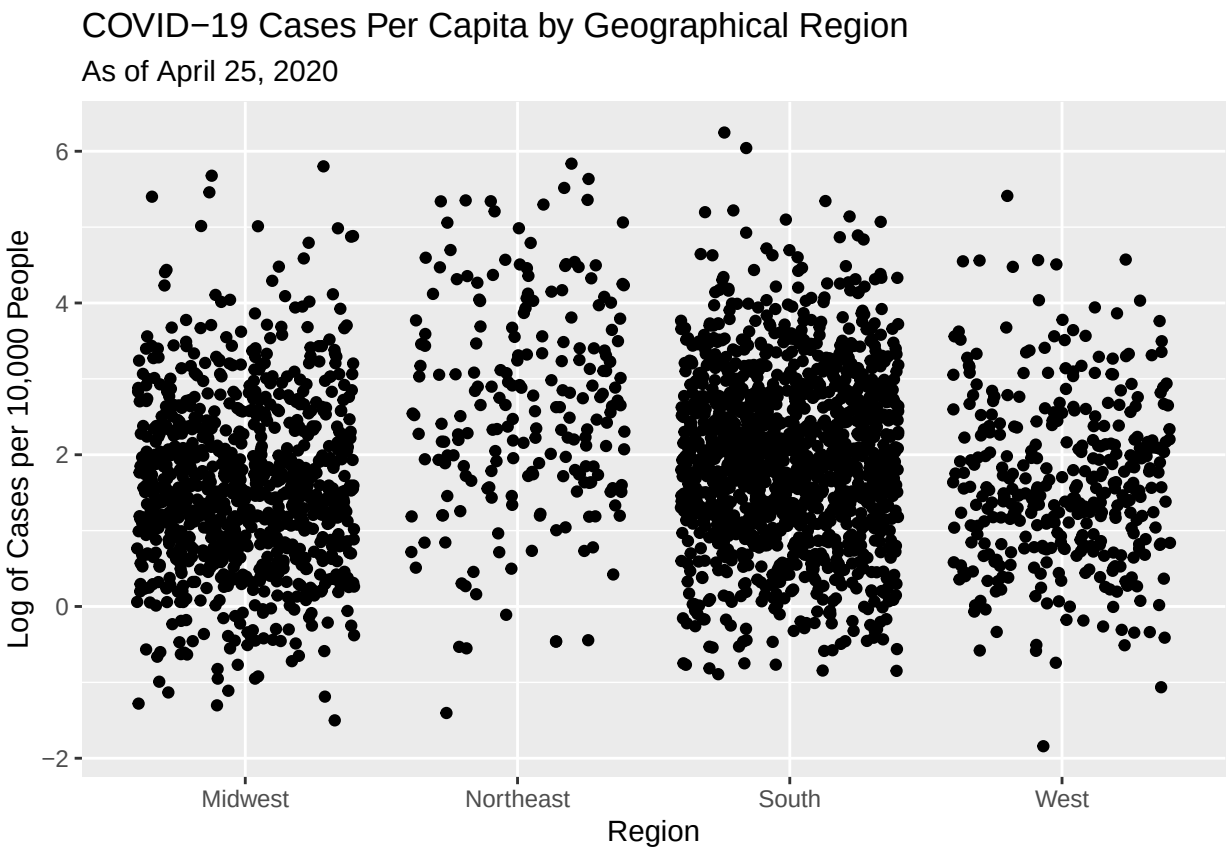
```
summary(corona)
```

```
##      date                fips          county             state
##  Length:2793        Min.   : 1001   Length:2793        Length:2793
##  Class :character   1st Qu.:18105   Class :character   Class :character
##  Mode  :character   Median :29070   Mode  :character   Mode  :character
##                     Mean   :30185
```

```
##                    3rd Qu.:45052
##                    Max.   :56043
##                    NA's   :1
##  census_region         cases              deaths              pop
##  Length:2793       Min.   :     1.0   Min.   :    0.00   Min.   :     453
##  Class :character  1st Qu.:     5.0   1st Qu.:    0.00   1st Qu.:   14743
##  Mode  :character  Median :    20.0   Median :    0.00   Median :   31269
##                    Mean   :   333.1   Mean   :   17.18   Mean   :  113391
##                    3rd Qu.:    83.0   3rd Qu.:    3.00   3rd Qu.:   78564
##                    Max.   :155124.0   Max.   :11419.00   Max.   :10116705
##
##    land_area           white           black          DemShare
##  Min.   :     2.0   Min.   :10.80   Min.   : 0.00   Min.   :0.03676
##  1st Qu.:   421.4   1st Qu.:79.10   1st Qu.: 0.80   1st Qu.:0.22634
##  Median :   592.6   Median :91.30   Median : 2.90   Median :0.31186
##  Mean   :  1018.4   Mean   :84.71   Mean   :10.06   Mean   :0.34595
##  3rd Qu.:   892.5   3rd Qu.:95.90   3rd Qu.:12.50   3rd Qu.:0.43241
##  Max.   :145504.8   Max.   :98.80   Max.   :85.30   Max.   :0.95749
##
##    case.bypop
##  Min.   :  0.1586
##  1st Qu.:  2.9744
##  Median :  5.9574
##  Mean   : 13.7996
##  3rd Qu.: 13.8320
##  Max.   :516.1059
##
```

The first variable that makes sense as a possible predictor is census_region. We could maybe use state, but there are so many that the geographical region as a factor is likely better. It seems plausible that climate and culture could effect transmission rate, and we know the various regions of the Us have varying climates and cultures. We can start by plotting these variables together.

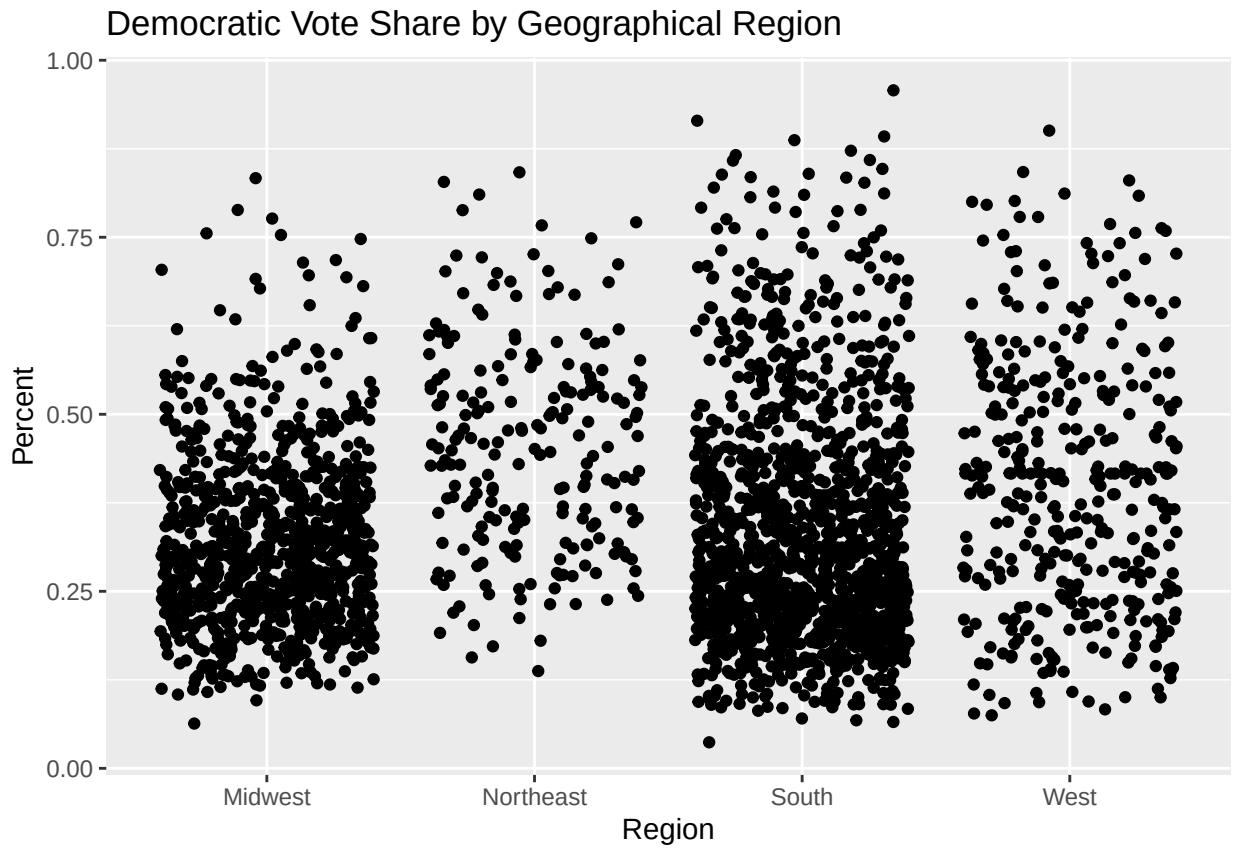
```
corona$census_region = as.factor(corona$census_region)
ggplot(corona, aes(x = census_region, y = log(case.bypop))) +
  geom_jitter() +
  xlab("Region") +
  ylab("Log of Cases per 10,000 People") +
  labs(title = "COVID-19 Cases Per Capita by Geographical Region",
       subtitle = "As of April 25, 2020")
```



There does seem to be some level of difference between geographical regions, we see that the Northeast as a whole seems to have higher rates per capita than other regions. This seems like it could be helpful in our model and we will include it for now.

Let's look at some plots to see how DemShare relates to the region.

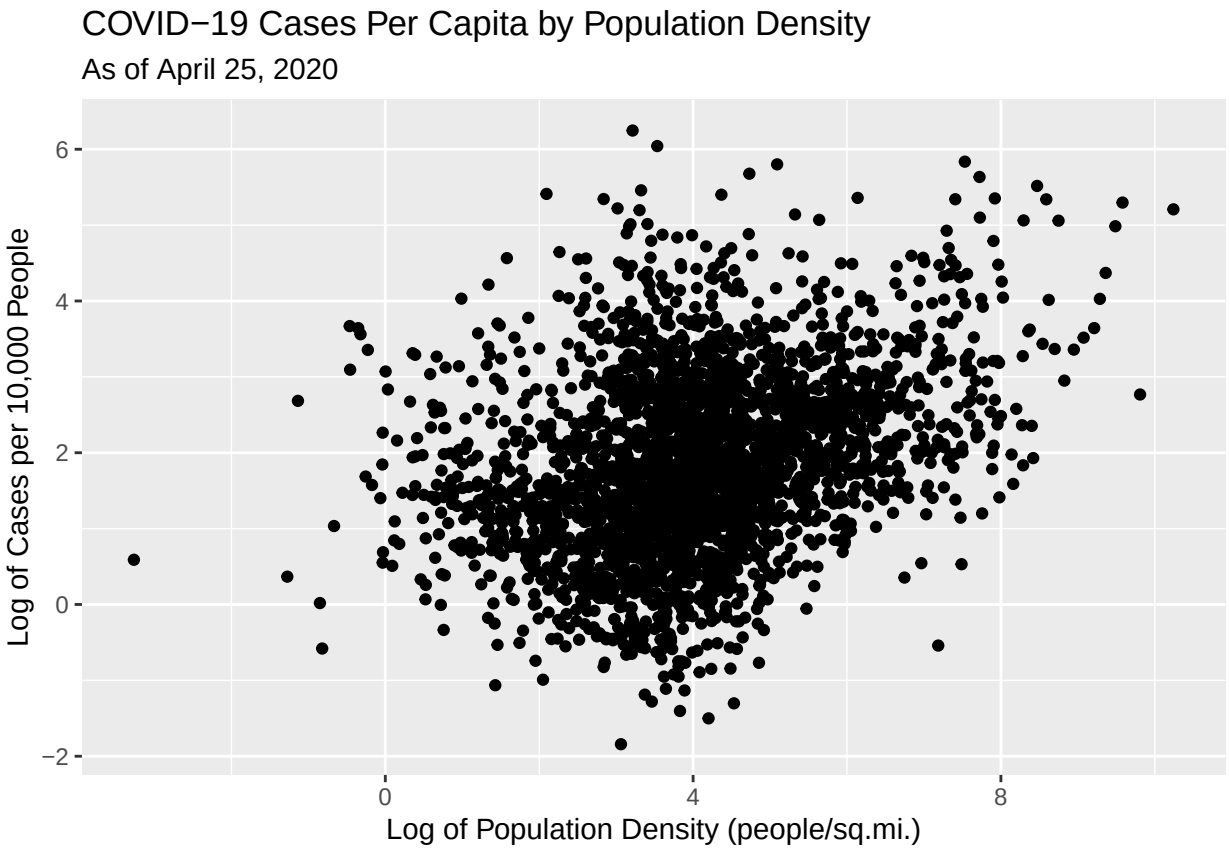
```
ggplot(corona, aes(x = census_region, y = DemShare)) +
  geom_jitter() +
  xlab("Region") +
  ylab("Percent") +
  labs(title = "Democratic Vote Share by Geographical Region")
```



When compared to the jitter plot of COVID-19 cases per 10,000 people this plot looks extremely similar. The bulk of the Northeast counties are higher on average than the South and Midwest, and the West is pretty evenly dispersed.

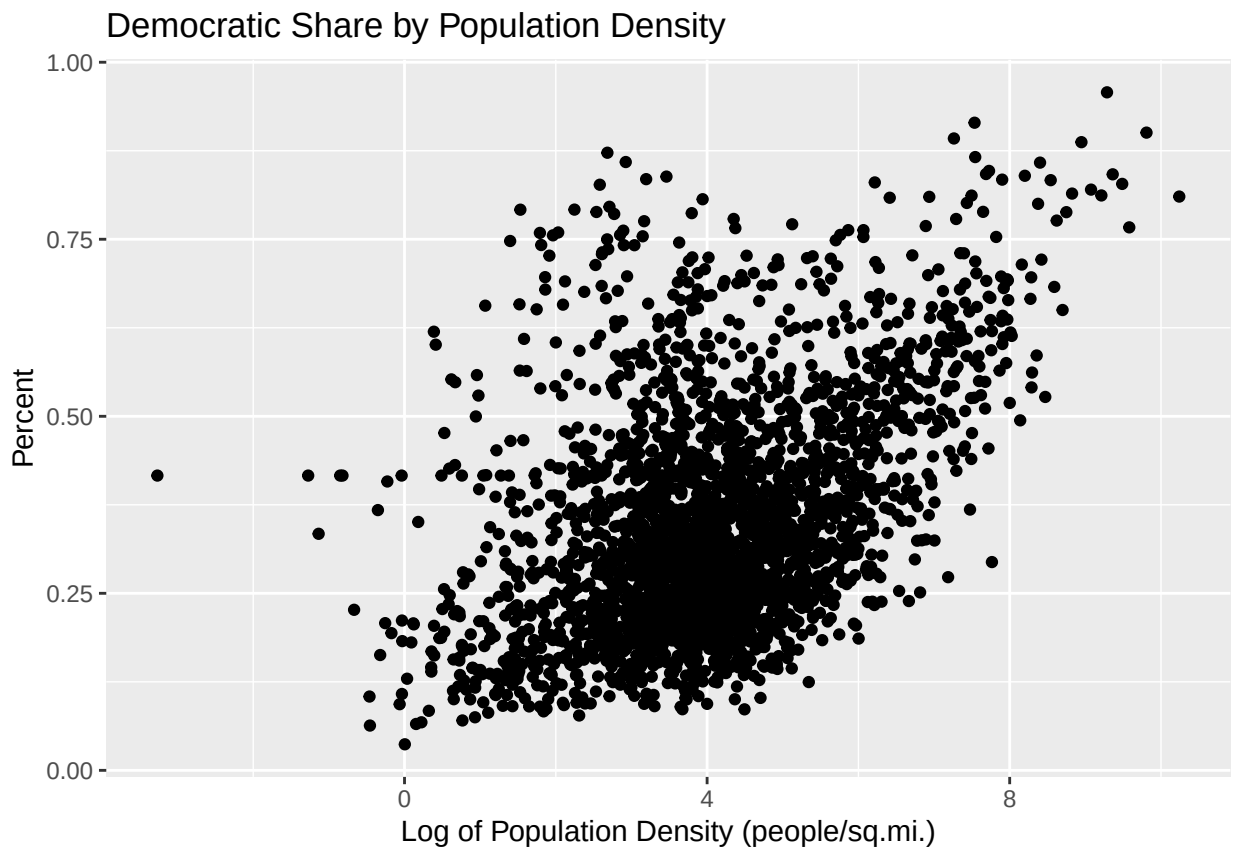
The next variable we can look at is population and area of the county. It probably doesn't make sense to use these variables on their own, but instead make a variable that is the population density. We know that viruses tend to spread more readily when people live closer together so population density seems like a plausible predictor for per capita cases of COVID-19. Population density is another variable that we will need to do some type of transformation on as there are some counties that are largely different than others.

```
corona$popdensity <- corona$pop / corona$land_area
ggplot(corona, aes(x = log(popdensity), y = log(case.bypop))) +
  geom_point() +
  xlab("Log of Population Density (people/sq.mi.)") +
  ylab("Log of Cases per 10,000 People") +
  labs(title = "COVID-19 Cases Per Capita by Population Density",
       subtitle = "As of April 25, 2020")
```



We see that there does seem to be some level of positive correlation between log of population density and log of per capita cases of COVID-19.

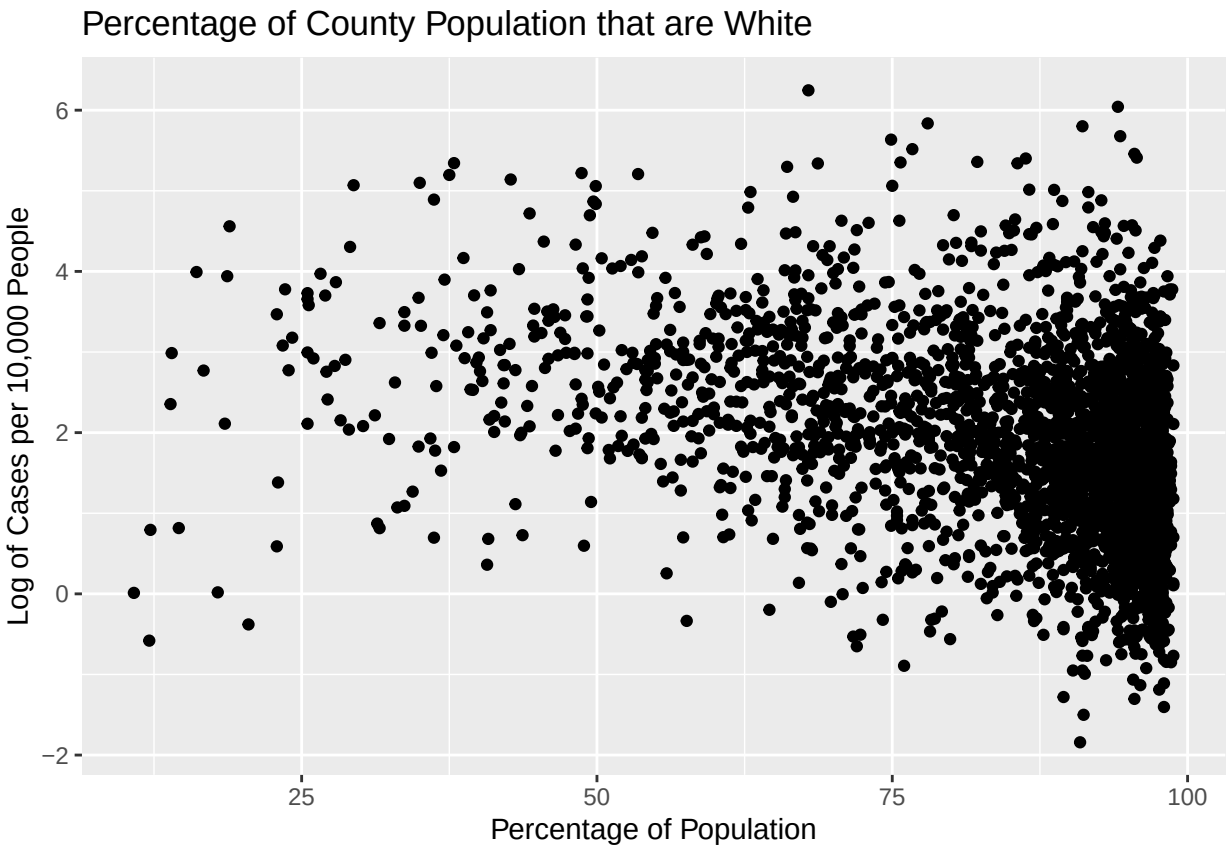
```
ggplot(corona, aes(x = log(popdensity), y = DemShare)) +
  geom_point() +
  xlab("Log of Population Density (people/sq.mi.)") +
  ylab("Percent") +
  labs(title = "Democratic Share by Population Density")
```



Again we see an extremely similar plot when looking at DemShare and population density. We would expect something like this as generally cities tend to be more democratic leaning.

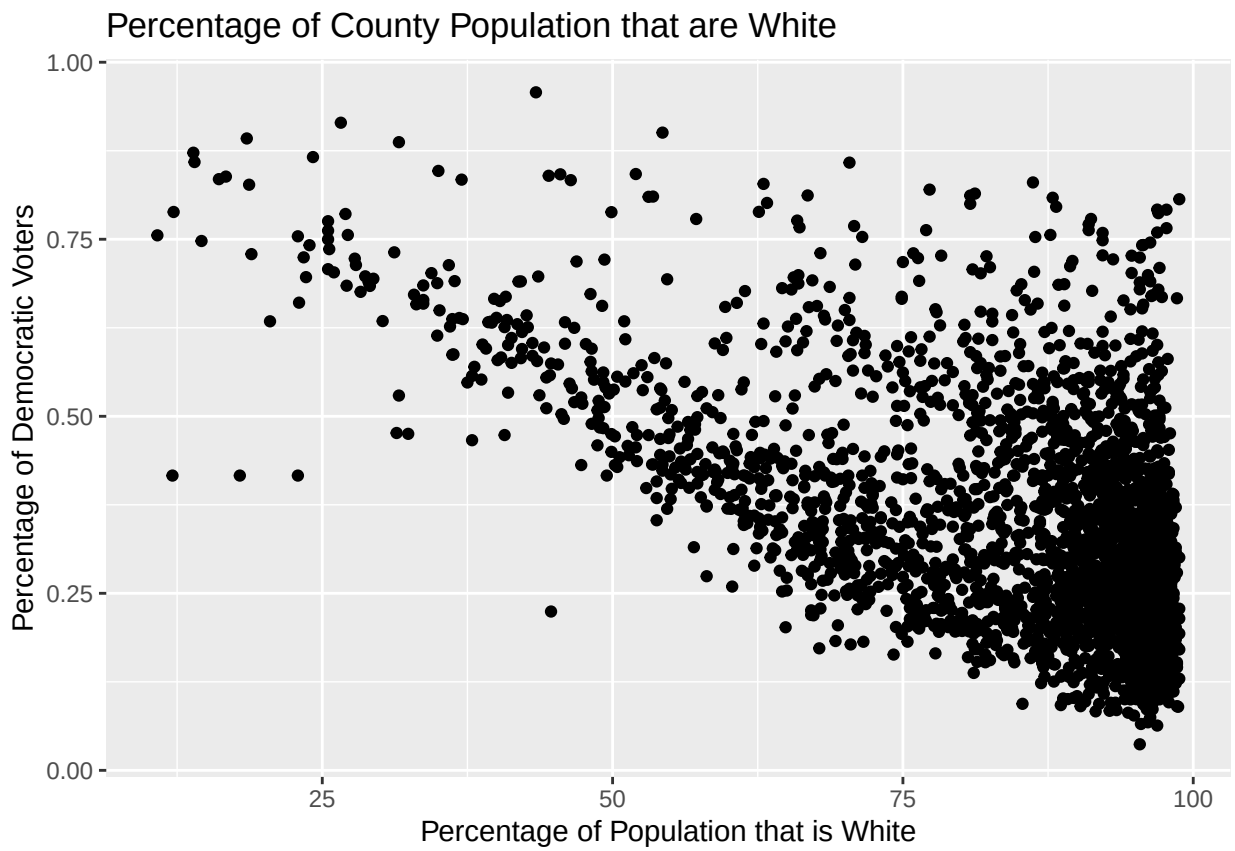
The last two variables we could use as predictors are percentage of the counties population that are white and black. While no obvious reasons for why these variables might be good predictors for COVID-19 transmission are readily apparent to me, it's best to see what kind of relationship the two have.

```
ggplot(corona, aes(x = white, y = log(case.bypop))) +
  geom_point() +
  xlab("Percentage of Population") +
  ylab("Log of Cases per 10,000 People") +
  labs(title = "Percentage of County Population that are White")
```



Interestingly enough there does seem like there could be some level of correlation between the percentage of the population which is white and log of per capita cases of COVID-19. The relationship almost looks quadratic, but it is hard to be certain because there are so few points less than 50% compared to points greater than 50%.

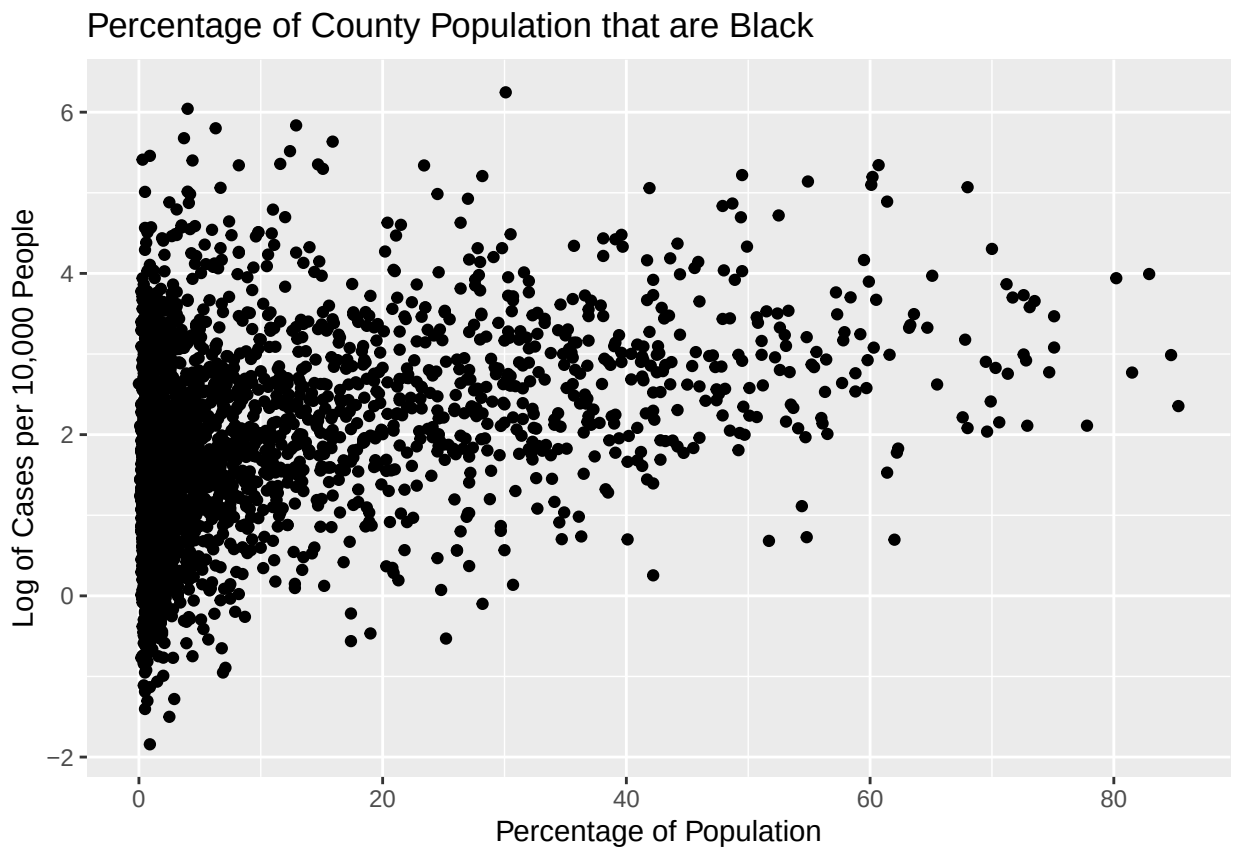
```
ggplot(corona, aes(x = white, y = DemShare)) +
  geom_point() +
  xlab("Percentage of Population that is White") +
  ylab("Percentage of Democratic Voters") +
  labs(title = "Percentage of County Population that are White")
```



When looking at DemShare compared to percentage of white voters, the plot isn't perfectly the same as COVID-19 cases per 10,000 people but there definitely are similarities.

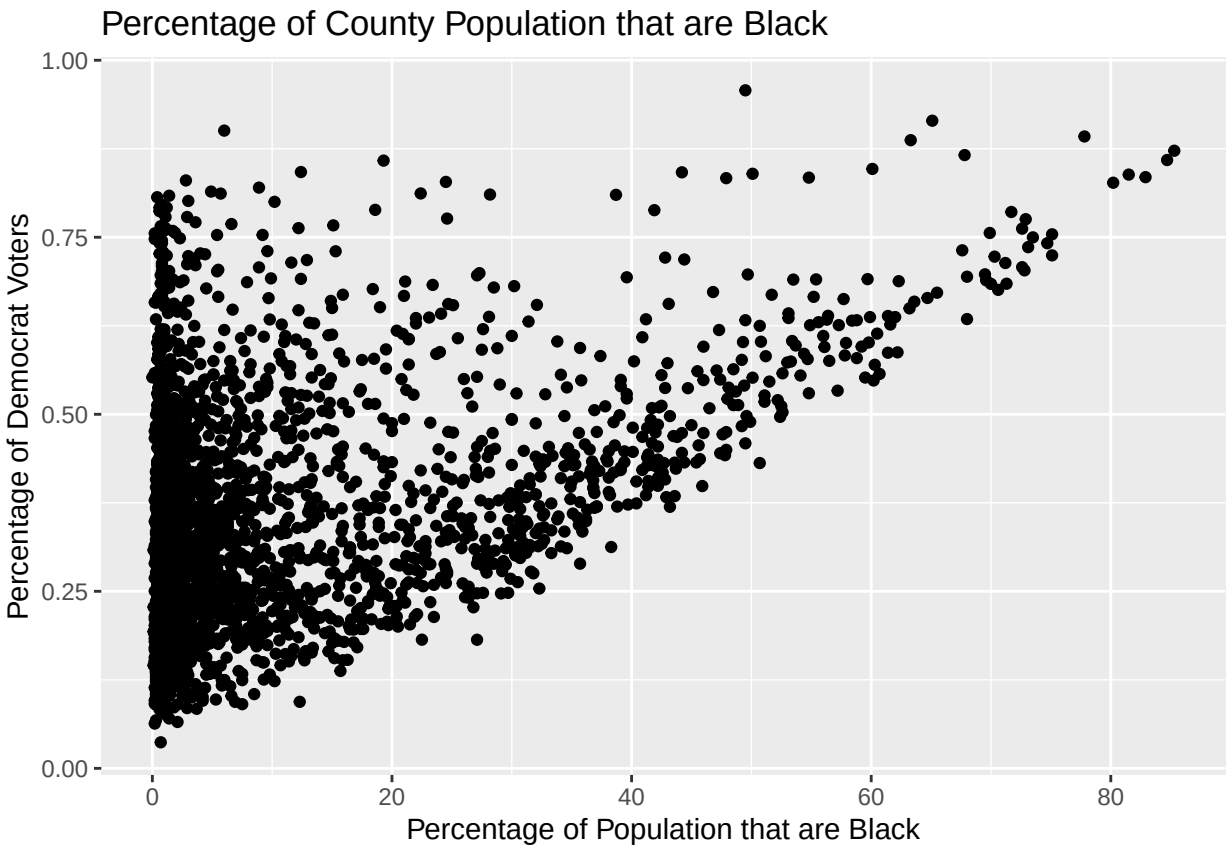
Lastly we can look at percentage of the population of counties that are black.

```
ggplot(corona, aes(x = black, y = log(case.bypop))) +
  geom_point() +
  xlab("Percentage of Population") +
  ylab("Log of Cases per 10,000 People") +
  labs(title = "Percentage of County Population that are Black")
```



It looks as though there could be a trend in the percentage of population of a county that is black as well, although this relationship looks more linear than the previous variable.

```
ggplot(corona, aes(x = black, y = DemShare)) +
  geom_point() +
  xlab("Percentage of Population that are Black") +
  ylab("Percentage of Democrat Voters") +
  labs(title = "Percentage of County Population that are Black")
```



Again, the black variable doesn't have the exact same interaction with DemShare, but it is similar.

The only other data points we have wouldn't make sense as predictors, such as number of deaths, that just seems like cheating. Now that we have looked at every possible predictor in our data, we can create a model that includes DemShare.

I think that in this case that a parametric model, such as linear regression, is our best choice. From linear regression we have more interpretability of the model than we would with a non-parametric method like a spline or loess.

In order to evaluate our models better we will split our data randomly into a training data set and a test data set. The training data will be 80% of the total data set.

```
set.seed(8691)
index <- sort(sample(nrow(corona), nrow(corona) * 0.8))
corona.train <- corona[index,]
corona.test <- corona[-index,]
```

```
corona.lm <- lm(log(case.bypop) ~ DemShare + census_region + log(popdensity) +
                  white + I(white^2) + black, data = corona.train)
summary(corona.lm)
```

```
##
## Call:
## lm(formula = log(case.bypop) ~ DemShare + census_region + log(popdensity) +
##     white + I(white^2) + black, data = corona.train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5453 -0.6786 -0.0541  0.5869  4.7441
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.0046536  0.4318484  -0.011    0.991
## DemShare                0.0136735  0.2148042   0.064    0.949
## census_regionNortheast  0.7193399  0.0935834   7.687 2.25e-14 ***
## census_regionSouth     -0.2657934  0.0604700  -4.395 1.16e-05 ***
## census_regionWest       0.1200866  0.0802476   1.496    0.135
## log(popdensity)         0.1513539  0.0179959   8.410  < 2e-16 ***
## white                   0.0421342  0.0102838   4.097 4.33e-05 ***
## I(white^2)             -0.0003435  0.0000702  -4.893 1.06e-06 ***
## black                   0.0268684  0.0034092   7.881 5.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 2225 degrees of freedom
## Multiple R-squared:  0.2515, Adjusted R-squared:  0.2488
## F-statistic: 93.43 on 8 and 2225 DF,  p-value: < 2.2e-16
```

Just by our initial model we see that the only coefficient with a non-significant P-Value is DemShare. We haven't checked our assumptions fully so we won't take our P-Values too literally at this point, but we already are seeing that DemShare probably doesn't improve this model.

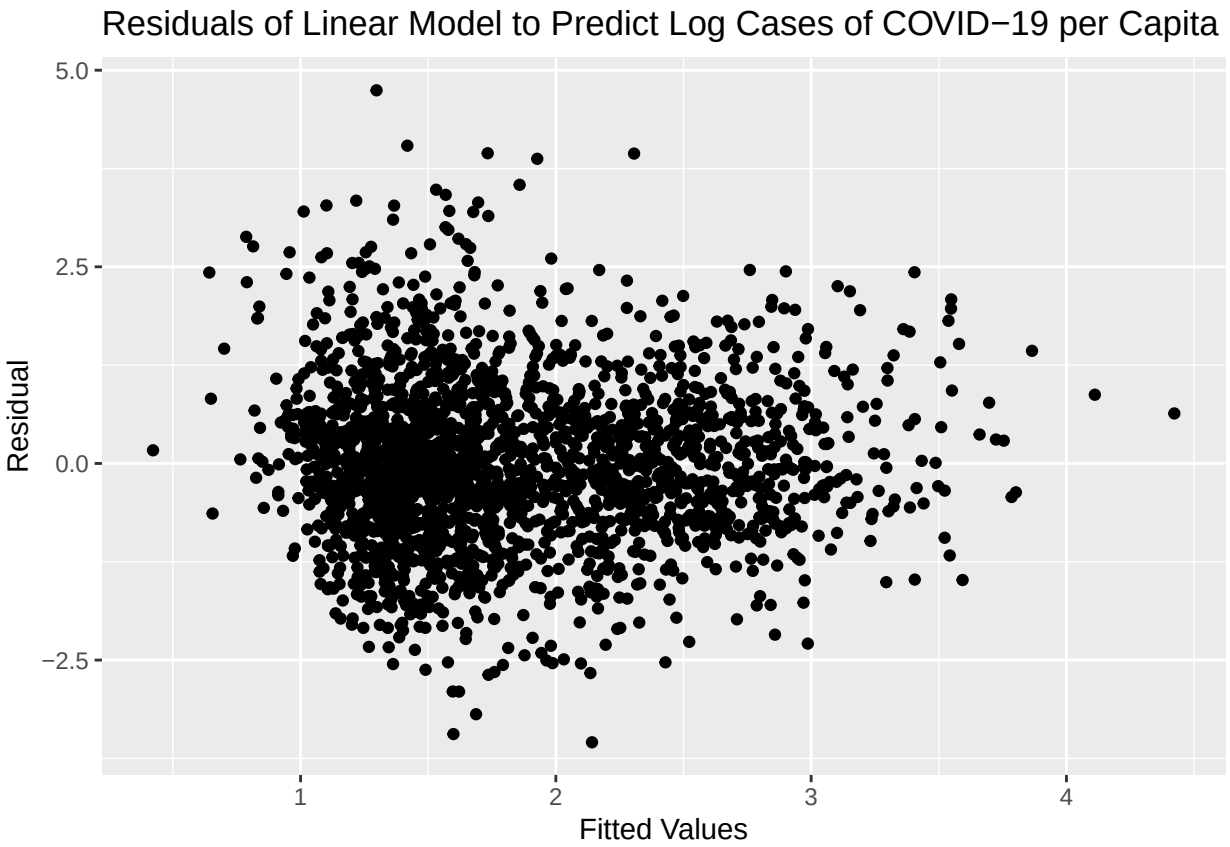
We can interpret the DemShare coefficient in the following way. With all else equal, a 1% increase in DemShare increases the log cases per 10,000 people by about 1.01%.

```
exp(0.0136735)
```

```
## [1] 1.013767
```

Let's take a look at the residuals from this model.

```
library(broom)
corona.df <- augment(corona.lm)
ggplot(corona.df, aes(x = .fitted, y = .resid)) +
  geom_point() +
  xlab("Fitted Values") +
  ylab("Residual") +
  labs(title = "Residuals of Linear Model to Predict Log Cases of COVID-19 per Capita")
```



The residual plot is not perfectly homoskedastistic, but it isn't terribly heteroskedastistic. We are mostly using this plot so that we can compare it to our model without DemShare in it. As another measure of the model, we can look at the AIC.

```
AIC(corona.lm)
```

```
## [1] 6484.724
```

Next we can look at the RMSE of the residuals of the training data.

```
corona.lm.rmse <- sqrt(mean((corona.df$log.case.bypop. - corona.df$.fitted)^2))
corona.lm.rmse
```

```
## [1] 1.02835
```

Lastly, let's see how this model predicts our test data set.

```
corona.lm.pred <- predict(corona.lm, newdata = corona.test)
corona.test$pred <- corona.lm.pred
```

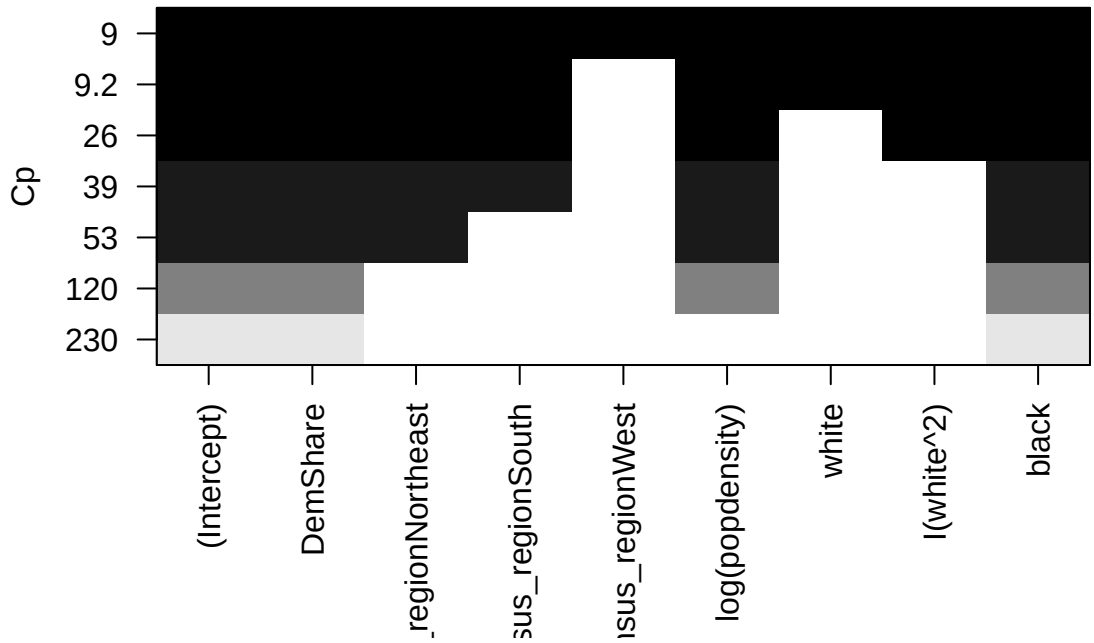
We will use RMSE again as a measurement of performance on the test dataset.

```
corona.test.rmse <- sqrt(mean((corona.test$pred - log(corona.test$case.bypop))^2))
corona.test.rmse
```

```
## [1] 0.9429298
```

Let's look at trying to do some feature selection to see if we can improve the model.

```
library(leaps)
corona.subset <- regsubsets(log(case.bypop) ~ DemShare + census_region + log(popdensity) +
                              white + I(white^2) + black, force.in = "DemShare", data = corona.train)
plot(corona.subset, scale = "Cp")
```



Actually the best model we could get by forcing in DemShare, is the model we had previously. So let's look at what happens when we remove DemShare.

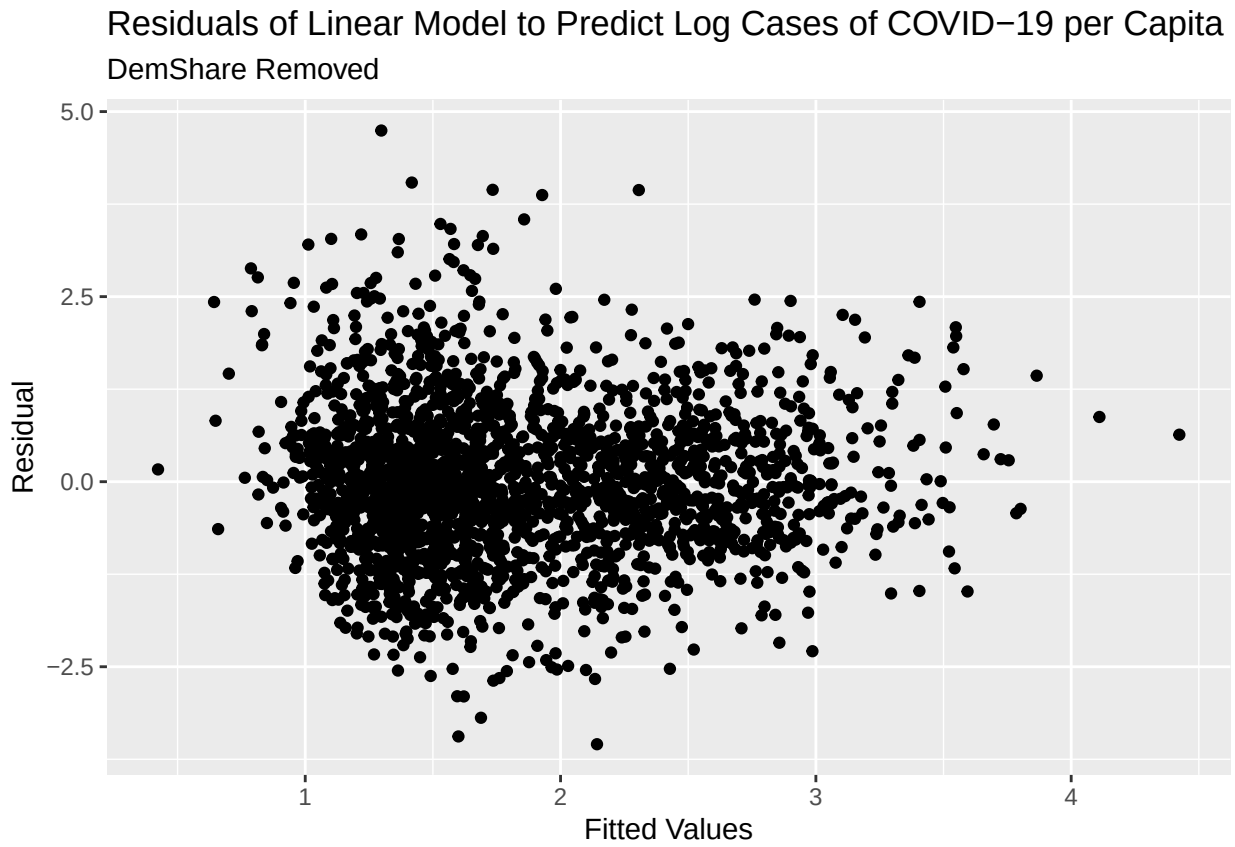
```
corona.lm2 <- lm(log(case.bypop) ~ census_region + log(popdensity) +
                   white + I(white^2) + black, data = corona.train)
summary(corona.lm2)
```

```
##
## Call:
## lm(formula = log(case.bypop) ~ census_region + log(popdensity) +
##     white + I(white^2) + black, data = corona.train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5465 -0.6776 -0.0540  0.5871  4.7439
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.002e-03  3.980e-01   0.015    0.988
## census_regionNortheast   7.203e-01  9.233e-02   7.801 9.35e-15 ***
## census_regionSouth      -2.670e-01  5.721e-02  -4.668 3.22e-06 ***
## census_regionWest        1.214e-01  7.733e-02   1.570    0.116
## log(popdensity)          1.519e-01  1.544e-02   9.840  < 2e-16 ***
## white                    4.198e-02  9.991e-03   4.202 2.75e-05 ***
## I(white^2)              -3.429e-04  6.942e-05  -4.939 8.43e-07 ***
## black                    2.689e-02  3.384e-03   7.947 3.01e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 2226 degrees of freedom
## Multiple R-squared:  0.2515, Adjusted R-squared:  0.2491
## F-statistic: 106.8 on 7 and 2226 DF,  p-value: < 2.2e-16
```

We can see how the residuals plot has changed by removing DemShare.

```
corona.df2 <- augment(corona.lm2)
ggplot(corona.df2, aes(x = .fitted, y = .resid)) +
  geom_point() +
  xlab("Fitted Values") +
  ylab("Residual") +
  labs(title = "Residuals of Linear Model to Predict Log Cases of COVID-19 per Capita",
       subtitle = "DemShare Removed")
```



The residuals plot looks nearly identical to the plot that included DemShare.

Now we can evaluate this new model based on the same measures we did on the previous model that included DemShare.

```
AIC(corona.lm2)
```

```
## [1] 6482.728
```

We get a slightly reduced AIC. This reduction does mean the model is better, but it is such a slight reduction that it likely isn't that much of an improvement.

```
corona.lm2.rmse <- sqrt(mean((corona.df2$log.case.bypop. - corona.df2$.fitted)^2))
corona.lm2.rmse
```

```
## [1] 1.028351
```

As far as the RMSE on the training data, we actually get the slightest increase in RMSE, an increase of 0.000001. This is such a negligible increase that I would say that the RMSEs are the same.

We will use RMSE again as a measurement of performance on the test dataset.

```
corona.lm2.pred <- predict(corona.lm2, newdata = corona.test)
corona.test$pred2 <- corona.lm2.pred
corona.test.rmse <- sqrt(mean((corona.test$pred2 - log(corona.test$case.bypop))^2))
corona.test.rmse
```

```
## [1] 0.9428813
```

The RMSE on the test data decreases slightly in the model that does not include DemShare, again the decrease is so small that it really isn't significantly worse or better. So this model likely isn't better or worse in terms of prediction than the model that included DemShare.

Lastly we can do an ANOVA test to further determine if these models are different.

```
anova(corona.lm, corona.lm2)
```

```
## Analysis of Variance Table
##
## Model 1: log(case.bypop) ~ DemShare + census_region + log(popdensity) +
##     white + I(white^2) + black
## Model 2: log(case.bypop) ~ census_region + log(popdensity) + white + I(white^2) +
##     black
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1   2225 2362.5
## 2   2226 2362.5 -1 -0.0043024 0.0041 0.9493
```

While we shouldn't take the P-Value literally, we see that we get an exceptionally high P-Value for the anova test suggesting that there is no statistical difference between the models. If we were selecting models and both are the same, then going with the simpler model would be desired, meaning we would use the model that doesn't include DemShare.

After these tests, we see that there is basically no difference in the models that include DemShare or don't include it. As there really is no difference, I would say that there really is no real credence to the fact that COVID-19 spreads faster based on how people vote.

I believe the true reason that there seems to be this relationship between DemShare and COVID-19 cases per capita is that there are variables that predict the spread of infection that also relate to DemShare. For example, the population density makes sense as a predictor of per capita infections because the closer people are to each other the more readily a virus can spread.

```
cor(corona$DemShare, corona$popdensity)
```

```
## [1] 0.3638465
```

We see that there is a correlation between population density and DemShare at the county level. Generally more heavily populated cities tend to have a higher percentage of Democrat voters than more rural area.

```
cor(corona$DemShare, corona$white)
```

```
## [1] -0.5472289
```

Another variable that we found that was significant in predicting COVID-19 spread is the percentage of the population that is white. This also has a decent correlation with DemShare, and a similar relationship. While it isn't known why the percentage of the population that is white would effect transmission, it does seem to be a good predictor in our model.

We have found convincingly that DemShare isn't a necessary predictor for log of COVID-19 cases per 10,000 people, but there are predictors that are related to DemShare. It would seem as though the counties that have a higher DemShare also share other variables that we can account for in our model and does not require the DemShare to be included.