# Final Exam

## DS-1000: Spring 2023

Alex Tomack

Due Date: 2023-04-28

# Overview

This is your final exam for DS-1000. It consists of seven questions plus an additional extra credit question. It is cumulative in the sense that you are expected to apply concepts and skills learned over the course of the semester.

# Grading

The final exam is due by 11:59PM on Friday, April 28th. Five points will be deducted for each day late it is received. Submissions received after midnight on Sunday, April 30th will not be graded.

Please upload **two** versions of this midterm. The first is a PDF of the **knitted** output, just like your problem sets which is used by the graders. The second is this .Rmd file in its raw form which is used by the professor to **apply a machine learning algorithm to check for violations of the honor code (see below)**. An additional 5 points will be deducted for failing to submit both files in the requested formats.

# Resources

You are permitted to rely on any course resources from the Spring 2023 semester. These include all lecture slides, recordings, problem sets, answer keys, homeworks, and lecture notes, as well as any and all posts to Campuswire.

Campuswire access will be restricted during the week of the final exam You are only permitted to post clarifying questions about the exam, and these should only be made visible to the instructor and TAs. The graders, TAs, and the Professor will remove questions that ask for help on the contents of the exam.

# Honor

Unlike the problem sets, you are **prohibited** from working on this midterm together. You must digitally sign your name below, confirming that you did not collaborate on this exam with any of your classmates, share work, or otherwise discuss its contents.

# Independent Work Statement

Please sign your name in the space provided by typing out your full name in place of the underline:

“I, Alex Tomack, am aware of the serious nature of plagiarism and affirm that I did not collaborate with other students while completing this final exam. I understand that violations of this agreement will result in a zero on the final exam, a failing grade for the semester, and a hearing with the Undergraduate Honor Council.”

# Question 0

Require `tidyverse` and load the `covid_prepped.Rds` (https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/9_Advanced_Topcis/data/covid_prepped.Rds?raw=true) data to an object called `covidData`.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0     ✓ purrr   1.0.0
## ✓ tibble  3.2.0     ✓ dplyr   1.1.0
## ✓ tidyr   1.2.1     ✓ stringr 1.5.0
## ✓ readr   2.1.3     ✓ forcats 0.5.2
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
require(scales)
```

```
## Loading required package: scales
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
covidData<-read_rds("https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/9_Advanced_Topi
cs/data/covid_prepped.Rds?raw=true")
```

# Codebook

The codebook for this dataset is produced below. Refer to this when interpreting regression coefficients!

| Name | Description |
|---|---|
| trump.votes | Total number of votes cast for Trump in 2020 |

| Name | Description |
|---|---|
| perc.trump.2020 | Proportion of all votes that were cast for Trump in 2020 |
| covid.deaths | Total number of Covid-19 related deaths in each county as of fall of 2020 |
| population | County population |
| perc.non.hisp.white | Percent of the county that is non-Hispanic white |
| perc.non.hisp.black | Percent of the county that is non-Hispanic black |
| perc.non.hisp.asian | Percent of the county that is non-Hispanic asian |
| perc.hispanic | Percent of the county that is Hispanic |
| perc.male | Percent of the county that is male |
| perc.65up | Percent of the county that is 65 years or older |
| unemp.rate | County unemployment rate (unemployed / in the labor force) |
| lfpr | County labor force participation rate |
| weekly.wages | Average weekly wages in the county |
| perc.rural | The percent of the county that is classified as rural |
| perc.manuf | Percent of the county that is employed in manufacturing |
| perc.trump.2016 | Proportion of all votes that were cast for Trump in 2016 |
| covid.death.rate | Number of Covid-19 related deaths per 1,000 people in each county as of the fall of 2020 |
| log.pop | County population (logged) |

# Question 1 [5 points]

Consider the following research question: “Were counties that had more deaths due to Covid-19 less likely to vote for Donald Trump in 2020?”

Please provide two arguments, one for both YES answer, and one for the NO answer to this research question, stating your theoretical assumptions [2 points] and your hypothesis [0.5 points].

> - YES: Write several sentences describing why the answer might be “yes”. Make sure to clearly state your theoretical assumptions! Yes– if there are more deaths, citizens are more likely to blame the president. It should be their job to control the spread and keep the US population safe, and increasing quantities of deaths would indicate they’re failing that responsibility.

> - NO: Write several sentences describing why the answer might be “NO”. Make sure to clearly state your theoretical assumptions! No– Trump’s supporters are more likely to downplay the severity of Covid. If people are dying, it has nothing to do with the president.

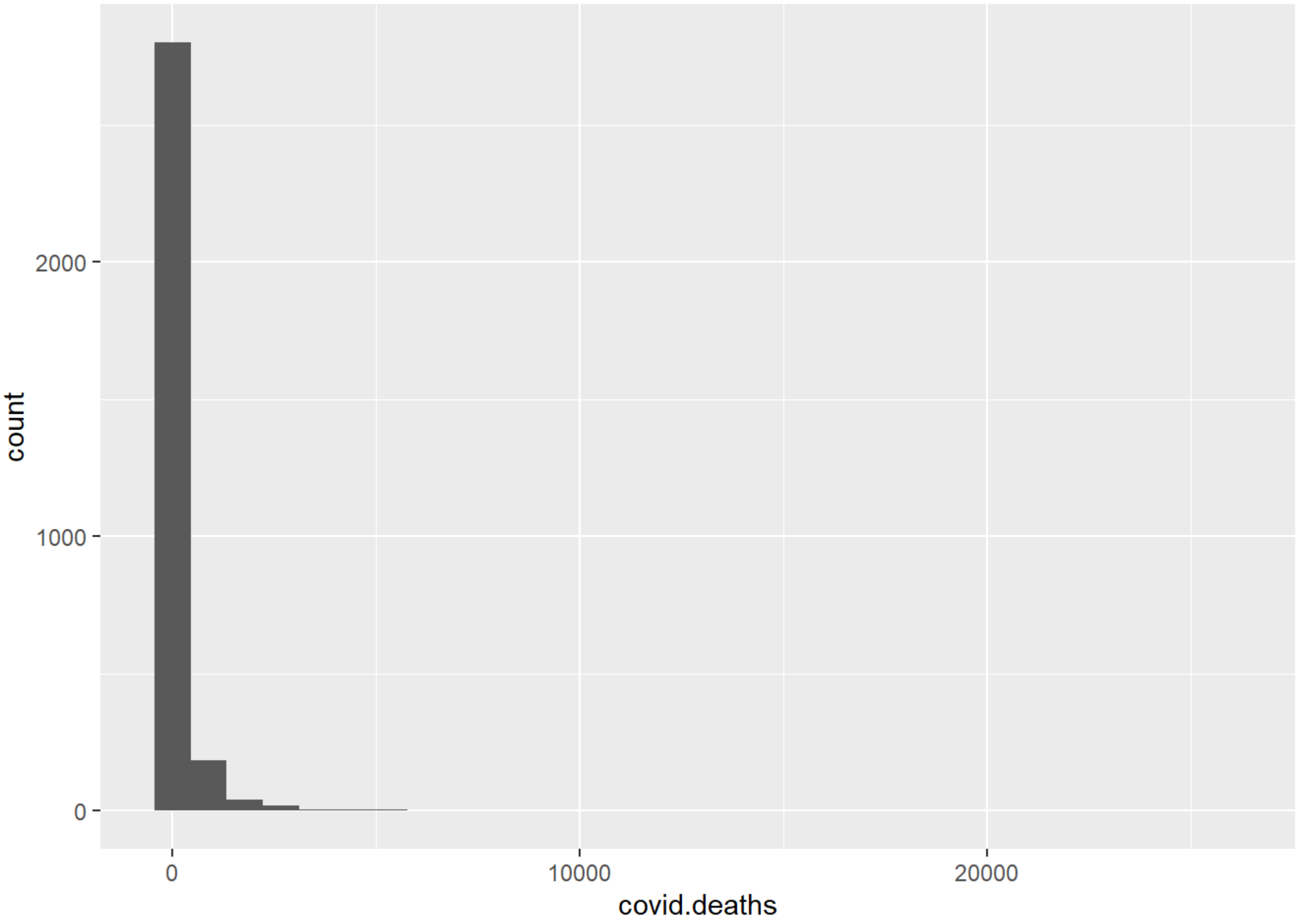
# Question 2 [5 points]

What are the main $X$ and $Y$ variables for this analysis [1 point]? Provide univariate [1 points] and multivariate [1 point] visualizations of them both. Do you need to transform either of these variables? If so, what is the optimal transformation [1 point]? Finally, interpret the multivariate visualization in light of the research question above. Which answer does it support [1 point]?

> - Write a sentence here defining which variables are the main $X$ and $Y$ variables. For our $X$ variable we’ll be looking at # of covid deaths and/or the covid.death.rate. For the $Y$ variable I’ll be looking at percent of trump support per county. For quantity of covid deaths, we’d need to log the variable, or you could normalize the distribution using the covid.death.rate.

```
#covid deaths-- unlogged
covidData%>%
  ggplot(aes(x=covid.deaths))+
  geom_histogram()
```

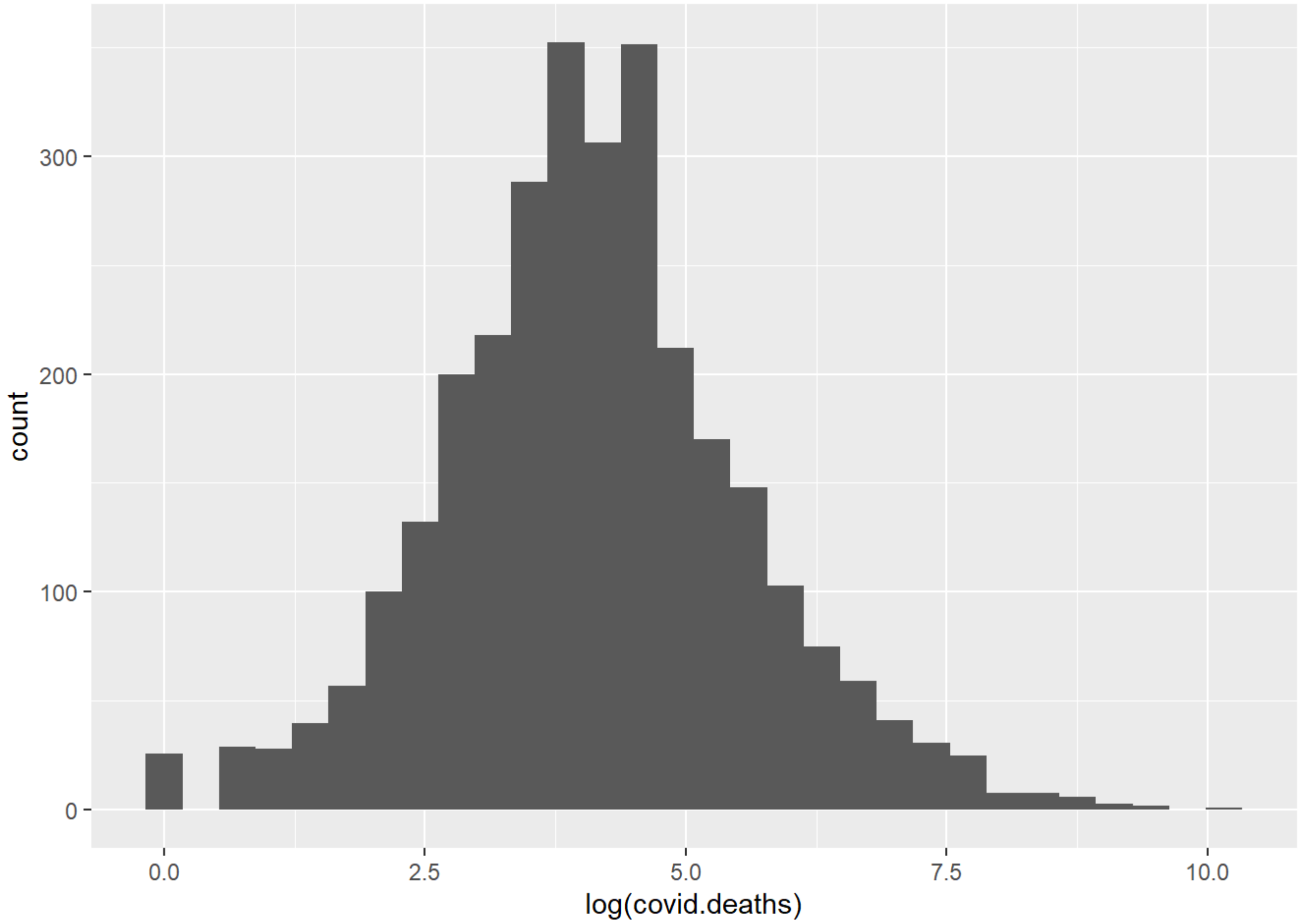
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# covid deaths--Logged
covidData%>%
  ggplot(aes(x=log(covid.deaths)))+
  geom_histogram()
```

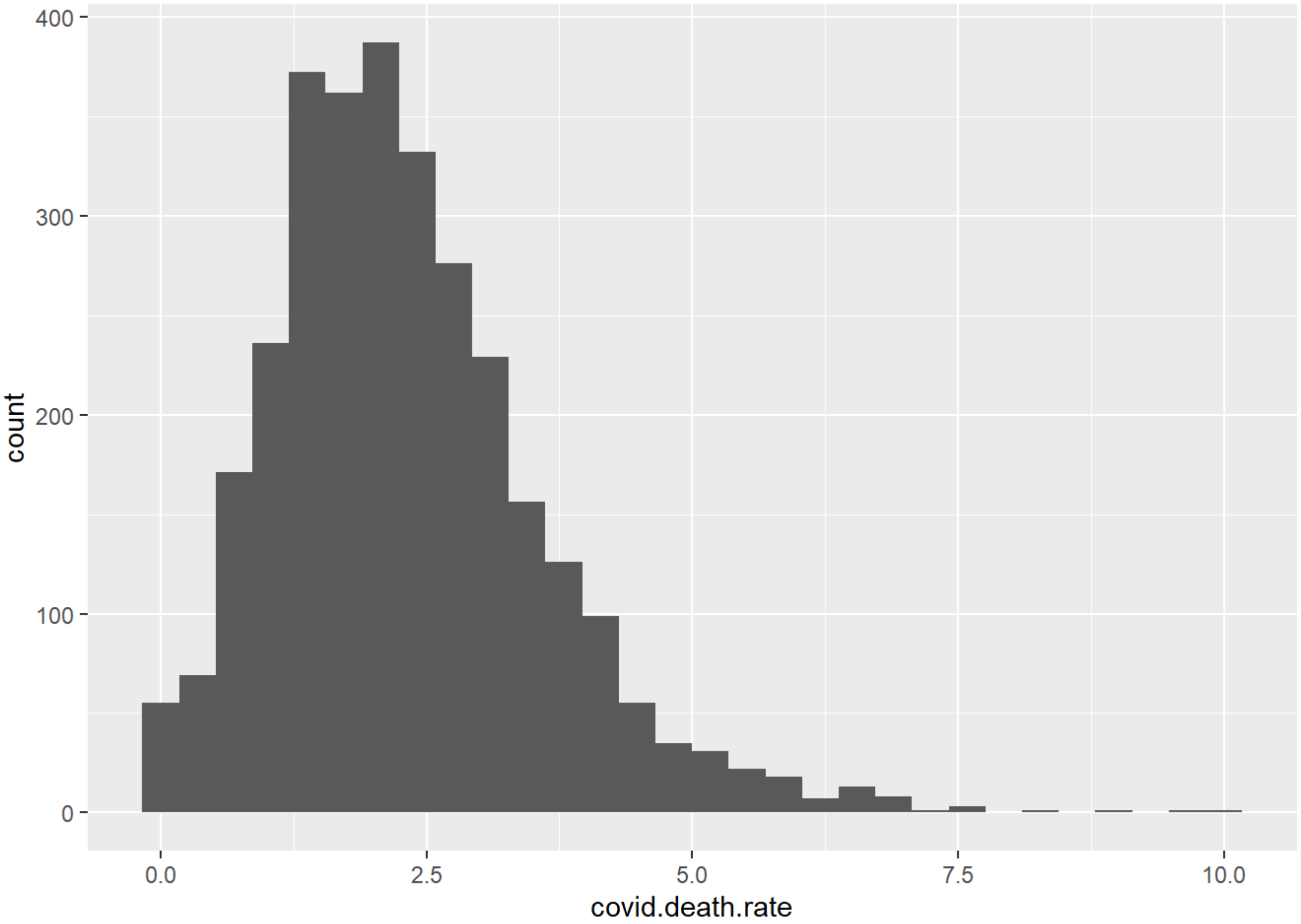
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 48 rows containing non-finite values (`stat_bin()`).
```



```
#covid deaths
covidData%>%
  ggplot(aes(x=covid.death.rate))+
  geom_histogram()
```

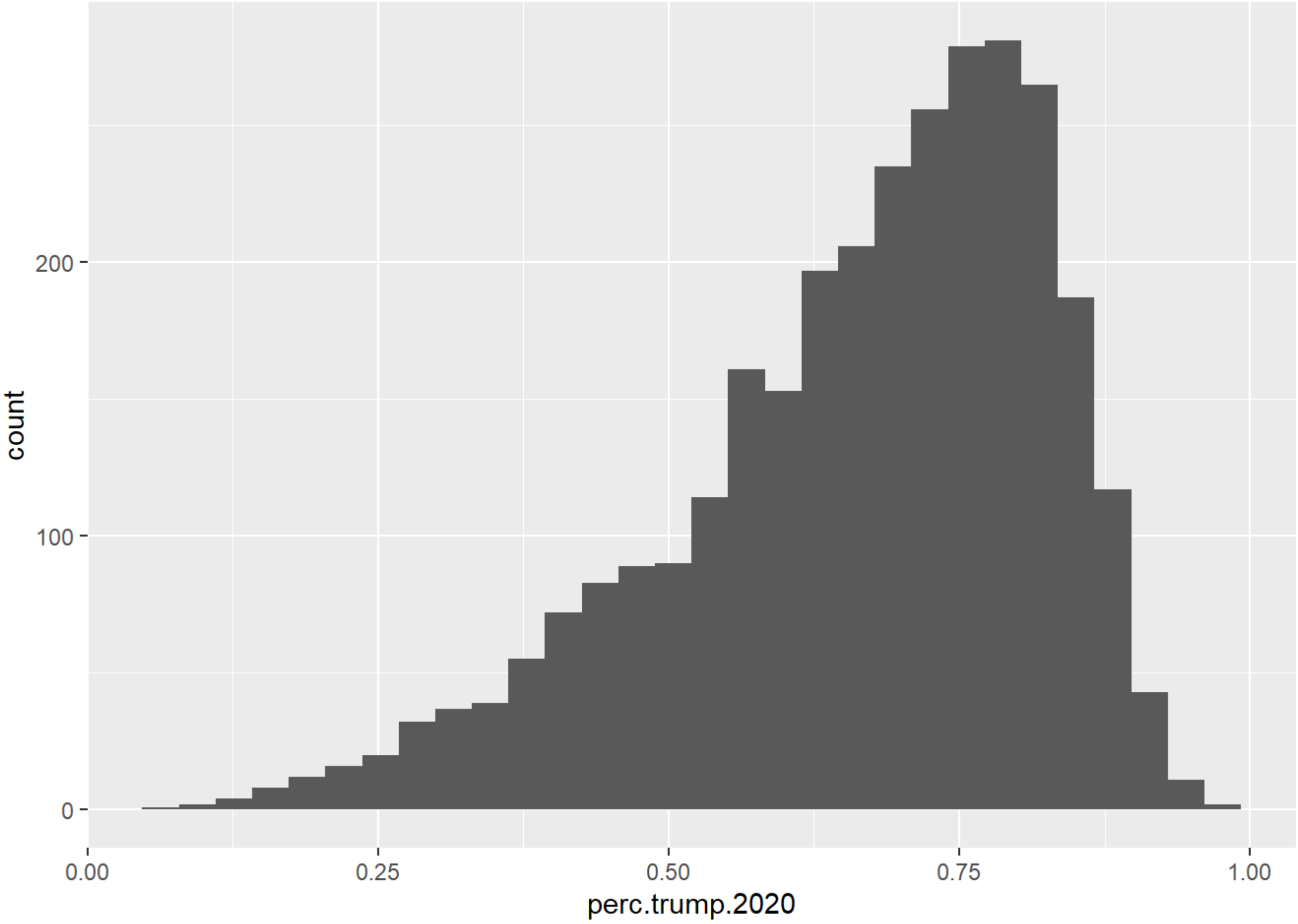
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



> - Write a few sentences here discussing the univariate visualization of $X$. When we look at the raw quantity of covid deaths, we see that there’s a highly right-skewed distribution. I went ahead and logged the variable, which normalizes the distribution of deaths. I also wanted to see what it looks like when we standardize the # of deaths as a proportion of deaths by county, which doesn’t normalize the data, but it makes it significantly less right-skewed.

```
#percent of trump votes in 2020
covidData%>%
  ggplot(aes(x=perc.trump.2020))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

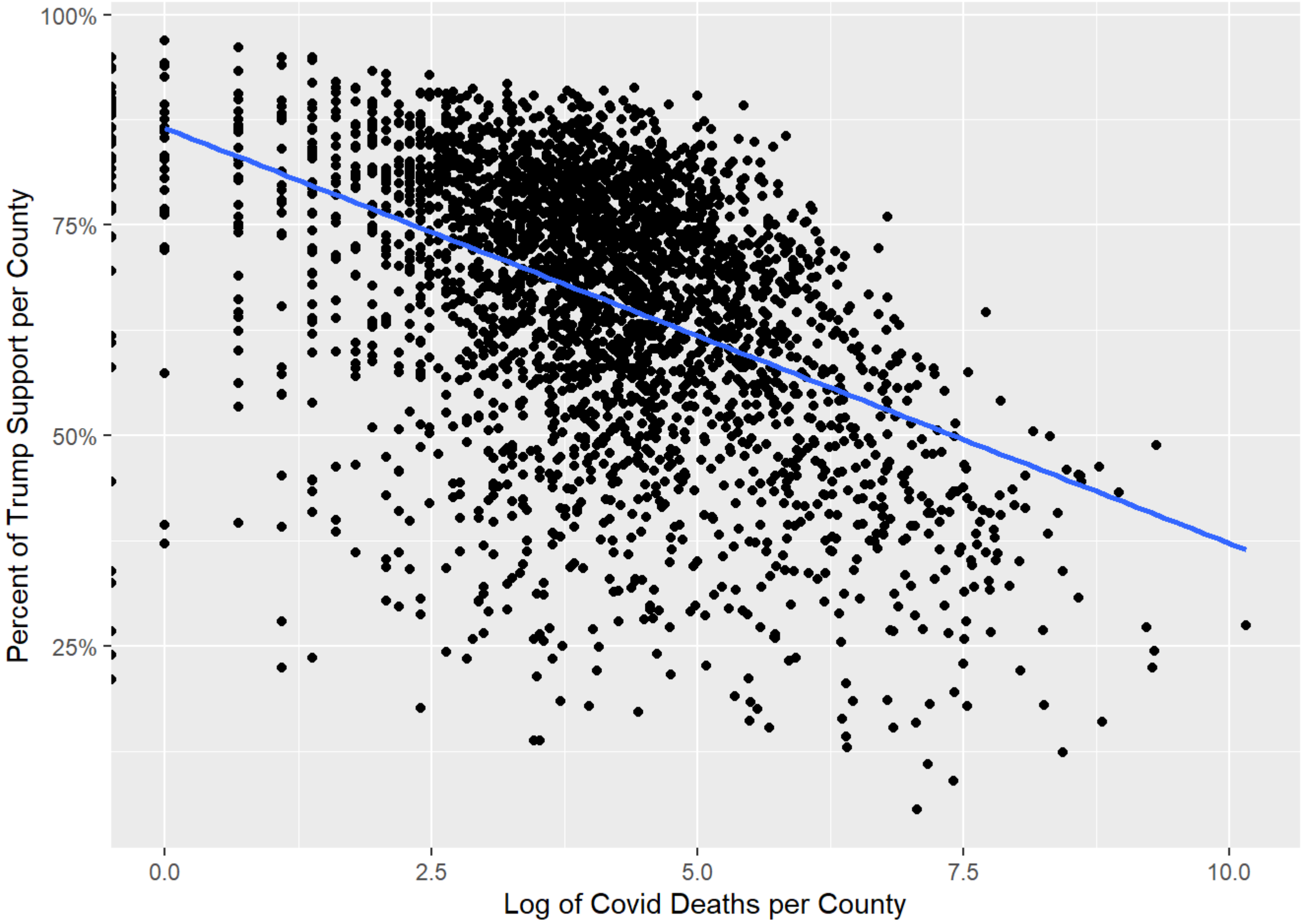


> - Write a sentence or two here discussing the univariate visualization of $Y$. The distribution of trump support by county is already normalized as a percentage, so it doesn’t have to undergo any transformation.

```
# mutlivariate using log(covid.deaths)
covidData%>%
  ggplot(aes(x=log(covid.deaths), y=perc.trump.2020))+
  geom_point()+
  scale_y_continuous(labels = scales::percent)+
  labs(x="Log of Covid Deaths per County",
      y="Percent of Trump Support per County")+
  geom_smooth(method="lm", se=F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 48 rows containing non-finite values (`stat_smooth()`).
```

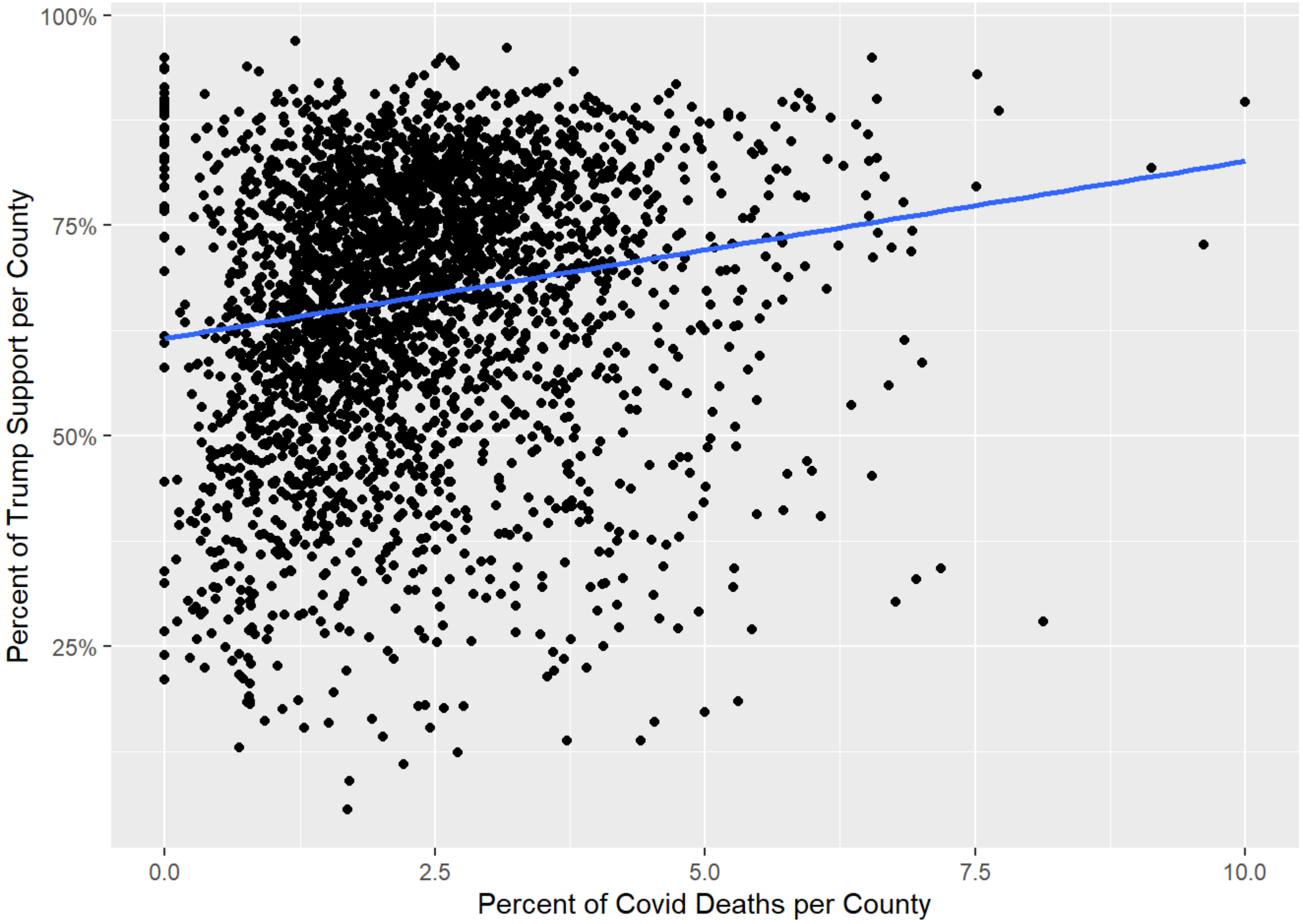


```
m1<-lm(formula=perc.trump.2020~log(covid.deaths+1), covidData)
summary(m1)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ log(covid.deaths + 1), data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65358 -0.07569  0.03000  0.10415  0.29338
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.863307   0.007516  114.86   <2e-16 ***
## log(covid.deaths + 1) -0.048724   0.001727  -28.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1433 on 3065 degrees of freedom
## Multiple R-squared:  0.2061, Adjusted R-squared:  0.2059
## F-statistic: 795.8 on 1 and 3065 DF,  p-value: < 2.2e-16
```

```
# multivariate analysis using percent of covid deaths
covidData%>%
  ggplot(aes(x=covid.death.rate, y=perc.trump.2020))+
  geom_point()+
  scale_y_continuous(labels = scales::percent)+
  labs(x="Percent of Covid Deaths per County",
      y="Percent of Trump Support per County")+
  geom_smooth(method="lm", se=F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
m2<-lm(formula=perc.trump.2020~covid.death.rate, covidData)
summary(m2)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59597 -0.08936  0.02941  0.11858  0.33402
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.615661   0.005994 102.707   <2e-16 ***
## covid.death.rate 0.021057   0.002283   9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1586 on 3065 degrees of freedom
## Multiple R-squared:  0.02702,    Adjusted R-squared:  0.0267
## F-statistic:  85.1 on 1 and 3065 DF,  p-value: < 2.2e-16
```

> - Write a few sentences interpreting the multivariate visualization. When we look at the mulitvariate visualization and regress the different transformations of our x variable on our y, we see that transforming the quantity of deaths by logging it is better. The graph which shows percentage of deaths and trump support indicates a positive relationship.. The other which shows quantity of deaths and trump support indicates a strong negative relationship. We can also notice in the regression summary that a greater amount of variation in trump support is explained by the log of covid deaths (20% per the R-squared/adj-R-Squared), while only 2% of the variation in trump support is explained by the percentage of deaths per county.

# Question 3 [5 points]

Now run a simple linear regression predicting `perc.trump.2020` as a function of `covid.deaths` [1 point] (DO NOT MODIFY EITHER VARIABLE FOR THIS QUESTION). What do you conclude? Make sure to interpret (1) the regression coefficient in substantive terms and (2) the confidence in the conclusion. [2 points] Why might we doubt that this regression captures the real relationship between Covid-19 deaths and Trump support? [2 points]

```
# regression of covid deaths on trump support in 2020
m3<-lm(formula=perc.trump.2020~covid.deaths, covidData)
summary(m3)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.deaths, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54828 -0.09222  0.02912  0.11542  1.22735
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.774e-01  2.865e-03  236.44   <2e-16 ***
## covid.deaths -6.325e-05  3.579e-06  -17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1532 on 3065 degrees of freedom
## Multiple R-squared:  0.09247,    Adjusted R-squared:  0.09218
## F-statistic: 312.3 on 1 and 3065 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here interpreting the regression output. We see that for every additional death, there is a -6.325e-05*100% change in trump support. Overall, we see a negative correlation between the raw quantity of deaths and support for Trump. However, while our covid.deaths variable is statistically significant, it only accounts for ~9% of the variation in trump support. It is not able to tell the entire story.

# Question 4 [5 points]

Now run a similar model that predicts `perc.trump.2020` as a function of the `covid.death.rate` (AGAIN, DO NOT MODIFY EITHER VARIABLE FOR THIS QUESTION). What do you conclude from this regression? As in Q3, make sure to describe the regression coefficient and interpret the p-value [2 points]. Why are the results so different? [3 points]

```
m4<-lm(formula=perc.trump.2020~covid.death.rate, covidData)
summary(m4)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59597 -0.08936  0.02941  0.11858  0.33402
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.615661   0.005994 102.707   <2e-16 ***
## covid.death.rate 0.021057   0.002283   9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1586 on 3065 degrees of freedom
## Multiple R-squared:  0.02702,    Adjusted R-squared:  0.0267
## F-statistic:  85.1 on 1 and 3065 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here interpreting the regression output and comparing it to the preceding result. For a 1% increase in death rate, we see a 0.02% increase in Trump support. Our variable is statistically significant, with a p-val < 0.0001, however only accounts for 2% of the variable in Trump support.

# Question 5 [5 points]

Evaluate model fit for the second specification ( `perc.trump.2020 ~ covid.death.rate` ) using 100-fold cross validation with an 80-20 split and save the RMSE to an object called `lm.simple.RMSE` [3 points]. What is the cross-validated average RMSE [1 point]? Make sure to interpret it in substantive terms [1 point]!

```
set.seed(123)
lm.simple.RMSE<-NULL
for (i in 1:100){
  sample<-sample(1:nrow(covidData),
                size=round(nrow(covidData)*0.8),
                replace=F)

  train<-covidData%>%slice(sample)
  test<-covidData%>%slice(-sample)

  rmse.model<-lm(formula=perc.trump.2020~covid.death.rate, data=covidData)

  test$pred<-predict(rmse.model, newdata=test)

  rmse <- sqrt(mean((test$perc.trump.2020 - test$pred)^2,na.rm=T))
  lm.simple.RMSE<-c(lm.simple.RMSE, rmse)
}
mean(lm.simple.RMSE)
```

```
## [1] 0.1582372
```

```
summary(rmse.model)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59597 -0.08936  0.02941  0.11858  0.33402
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.615661   0.005994 102.707   <2e-16 ***
## covid.death.rate 0.021057   0.002283   9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1586 on 3065 degrees of freedom
## Multiple R-squared:  0.02702,    Adjusted R-squared:  0.0267
## F-statistic:  85.1 on 1 and 3065 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here interpreting the RMSE result. Our CV average RMSE is 0.15, meaning our model which predicts trump support based on covid death rate averages errors of 15%.

# Question 6 [10 points]

Now re-run the second specification (i.e., the one estimating `perc.trump.2020 ~ covid.death.rate` ) but include the following additional $X$ variables: - Population: `population` (you need to log it first) - Share of the county that is non-Hispanic white: `perc.non.hisp.white` - Share of the county that is non-Hispanic black: `perc.non.hisp.black` - Share of the county that is non-Hispanic asian: `perc.non.hisp.asian` - Share of the county that is Hispanic: `perc.hispanic` - Share of the county that is male: `perc.male` - Share of the county that is 65 years or older: `perc.65up` - Unemployment rate: `unemp.rate` - Labor force participation rate: `lfpr` - Average weekly wages: `weekly.wages` - Share of the county that is classified as rural: `perc.rural` - Share of the county employed in manufacturing: `perc.manuf`

Save this regression model to an object called `lm.controls` . Does the answer to the research question change after including these controls? [5 points]. **Only interpret the coefficient on `covid.death.rate` for full credit! No need to interpret every coefficient!**

What about the RMSE? (Again, you need to run 100-fold cross validation with an 80-20 split, and save it to an object called `lm.controls.RMSE` .) Which specification is better, based on these results? [5 points]

```
lm.controls<-lm(formula=perc.trump.2020~covid.death.rate+log.pop+perc.non.hisp.white+perc.non.hisp.black+perc.non.hisp.asian+perc.hispanic+perc.male+perc.65up+unemp.rate+lfpr+weekly.wages+perc.rural+perc.manuf, data=covidData)
summary(lm.controls)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate + log.pop + perc.non.hisp.white +
##     perc.non.hisp.black + perc.non.hisp.asian + perc.hispanic +
##     perc.male + perc.65up + unemp.rate + lfpr + weekly.wages +
##     perc.rural + perc.manuf, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42779 -0.05140  0.00772  0.06272  0.37207
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.175e+00  6.991e-02  16.805  < 2e-16 ***
## covid.death.rate     2.252e-02  1.593e-03  14.141  < 2e-16 ***
## log.pop             -3.513e-02  2.115e-03 -16.613  < 2e-16 ***
## perc.non.hisp.white  6.489e-03  2.887e-04  22.474  < 2e-16 ***
## perc.non.hisp.black -4.120e-04  2.884e-04  -1.428 0.153304
## perc.non.hisp.asian -3.383e-03  8.678e-04  -3.898 9.89e-05 ***
## perc.hispanic        4.122e-03  3.039e-04  13.563  < 2e-16 ***
## perc.male           -3.326e-01  9.025e-02  -3.685 0.000232 ***
## perc.65up           -7.040e-01  4.896e-02 -14.379  < 2e-16 ***
## unemp.rate          -1.952e+00  1.531e-01 -12.750  < 2e-16 ***
## lfpr                -7.981e-01  3.671e-02 -21.739  < 2e-16 ***
## weekly.wages         5.991e-06  1.112e-05   0.539 0.589975
## perc.rural           4.967e-04  9.785e-05   5.077 4.07e-07 ***
## perc.manuf           1.770e-02  2.024e-02   0.875 0.381897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09598 on 3053 degrees of freedom
## Multiple R-squared:  0.6451, Adjusted R-squared:  0.6436
## F-statistic: 426.9 on 13 and 3053 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here interpreting the regression output and answering the first question. For every percent change in covid death rates, there is a 2.252e-02% change in trump support. The only variables which are not statistically significant perc.non.ship.black, weekly.wages, and perc.manuf. In addition, 64% of the variation in trump support is explained by our regression variables.

```
set.seed(123)
lm.controls.RMSE<-NULL
for (i in 1:100){
  sample<-sample(1:nrow(covidData),
                size=round(nrow(covidData)*0.8),
                replace=F)

  train<-covidData%>%slice(sample)
  test<-covidData%>%slice(-sample)

  lm.controls<-lm(formula=perc.trump.2020~covid.death.rate+log.pop+perc.non.hisp.white+perc.non.hisp.black+perc.non.hisp.asian+perc.hispanic+perc.male+perc.65up+unemp.rate+lfpr+weekly.wages+perc.rural+perc.manuf, data=covidData)

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

  rmse <- sqrt(mean((test$perc.trump.2020 - test$pred)^2,na.rm=T))
  lm.controls.RMSE<-c(lm.controls.RMSE, rmse)
}
mean(lm.controls.RMSE)
```

```
## [1] 0.09578247
```

```
summary(lm.controls)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate + log.pop + perc.non.hisp.white +
##     perc.non.hisp.black + perc.non.hisp.asian + perc.hispanic +
##     perc.male + perc.65up + unemp.rate + lfpr + weekly.wages +
##     perc.rural + perc.manuf, data = covidData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42779 -0.05140  0.00772  0.06272  0.37207
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.175e+00  6.991e-02  16.805  < 2e-16 ***
## covid.death.rate     2.252e-02  1.593e-03  14.141  < 2e-16 ***
## log.pop             -3.513e-02  2.115e-03 -16.613  < 2e-16 ***
## perc.non.hisp.white  6.489e-03  2.887e-04  22.474  < 2e-16 ***
## perc.non.hisp.black -4.120e-04  2.884e-04  -1.428 0.153304
## perc.non.hisp.asian -3.383e-03  8.678e-04  -3.898 9.89e-05 ***
## perc.hispanic        4.122e-03  3.039e-04  13.563  < 2e-16 ***
## perc.male           -3.326e-01  9.025e-02  -3.685 0.000232 ***
## perc.65up           -7.040e-01  4.896e-02 -14.379  < 2e-16 ***
## unemp.rate          -1.952e+00  1.531e-01 -12.750  < 2e-16 ***
## lfpr                -7.981e-01  3.671e-02 -21.739  < 2e-16 ***
## weekly.wages         5.991e-06  1.112e-05   0.539 0.589975
## perc.rural           4.967e-04  9.785e-05   5.077 4.07e-07 ***
## perc.manuf           1.770e-02  2.024e-02   0.875 0.381897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09598 on 3053 degrees of freedom
## Multiple R-squared:  0.6451, Adjusted R-squared:  0.6436
## F-statistic: 426.9 on 13 and 3053 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here interpreting the RMSE output and comparing it to the preceding result. Our RMSE is much better, with an average of 9% variation in predicted 2020 trump support vs true 2020 trump support. Our CV based on the model which includes more variables is significantly better.

# Question 7 [5 points]

Finally, re-run the same specification once again, except add Trump's 2016 support ( `perc.trump.2016` ) as a control $X$ predictor, and save it to `lm.controls2` object. Does your conclusion change? Why? [4 points]

What about the model fit? Again using 100-fold cross validation with an 80-20 split, does the RMSE improve? [1 point]

```
lm.controls2<-lm(formula=perc.trump.2020~covid.death.rate+log.pop+perc.non.hisp.white+perc.non.h
isp.black+perc.non.hisp.asian+perc.hispanic+perc.male+perc.65up+unemp.rate+lfpr+weekly.wages+per
c.rural+perc.manuf+perc.trump.2016, data=covidData)
summary(lm.controls2)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate + log.pop + perc.non.hisp.white +
##     perc.non.hisp.black + perc.non.hisp.asian + perc.hispanic +
##     perc.male + perc.65up + unemp.rate + lfpr + weekly.wages +
##     perc.rural + perc.manuf + perc.trump.2016, data = covidData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109082 -0.013537 -0.000802  0.011000  0.222602
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.439e-01  1.843e-02   7.809 7.89e-15 ***
## covid.death.rate     6.581e-04  4.180e-04   1.574 0.115509
## log.pop             -8.204e-03  5.526e-04 -14.846  < 2e-16 ***
## perc.non.hisp.white  6.767e-04  7.842e-05   8.629  < 2e-16 ***
## perc.non.hisp.black -2.582e-04  7.332e-05  -3.521 0.000436 ***
## perc.non.hisp.asian -3.191e-05  2.211e-04  -0.144 0.885292
## perc.hispanic        1.038e-03  7.863e-05  13.196  < 2e-16 ***
## perc.male            1.320e-02  2.300e-02   0.574 0.566133
## perc.65up           -1.217e-01  1.275e-02  -9.549  < 2e-16 ***
## unemp.rate          -1.202e-02  4.000e-02  -0.300 0.763906
## lfpr                -8.099e-02  9.935e-03  -8.152 5.19e-16 ***
## weekly.wages        -4.123e-06  2.826e-06  -1.459 0.144673
## perc.rural           4.630e-05  2.496e-05   1.855 0.063760 .
## perc.manuf           2.611e-02  5.144e-03   5.075 4.10e-07 ***
## perc.trump.2016      9.421e-01  4.481e-03 210.247  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0244 on 3052 degrees of freedom
## Multiple R-squared:  0.9771, Adjusted R-squared:  0.977
## F-statistic:  9294 on 14 and 3052 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here answering the first questions. It drastically improves our regression, now with 97.7% of the variation in 2020 trump support explained by our regression variables (given by the adjusted R-Squared value). We can think about it logically– someone who voted for Trump in 2016 is more aligned with his way of thinking. They'd be more inclined to continue voting for something who advertises their values than to vote for someone else.

```
set.seed(123)
lm.controls2.RMSE<-NULL
for (i in 1:100){
  sample<-sample(1:nrow(covidData),
                size=round(nrow(covidData)*0.8),
                replace=F)

  train<-covidData%>%slice(sample)
  test<-covidData%>%slice(-sample)

  lm.controls2<-lm(formula=perc.trump.2020~covid.death.rate+log.pop+perc.non.hisp.white+perc.non.hisp.black+perc.non.hisp.asian+perc.hispanic+perc.male+perc.65up+unemp.rate+lfpr+weekly.wages+perc.rural+perc.manuf+perc.trump.2016, data=covidData)

  test$pred<-predict(lm.controls2, newdata=test)

  rmse <- sqrt(mean((test$perc.trump.2020 - test$pred)^2,na.rm=T))
  lm.controls2.RMSE<-c(lm.controls2.RMSE, rmse)
}
mean(lm.controls2.RMSE)
```

```
## [1] 0.02436621
```

```
summary(lm.controls2)
```

```
##
## Call:
## lm(formula = perc.trump.2020 ~ covid.death.rate + log.pop + perc.non.hisp.white +
##     perc.non.hisp.black + perc.non.hisp.asian + perc.hispanic +
##     perc.male + perc.65up + unemp.rate + lfpr + weekly.wages +
##     perc.rural + perc.manuf + perc.trump.2016, data = covidData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109082 -0.013537 -0.000802  0.011000  0.222602
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.439e-01  1.843e-02   7.809 7.89e-15 ***
## covid.death.rate     6.581e-04  4.180e-04   1.574 0.115509
## log.pop             -8.204e-03  5.526e-04 -14.846  < 2e-16 ***
## perc.non.hisp.white  6.767e-04  7.842e-05   8.629  < 2e-16 ***
## perc.non.hisp.black -2.582e-04  7.332e-05  -3.521 0.000436 ***
## perc.non.hisp.asian -3.191e-05  2.211e-04  -0.144 0.885292
## perc.hispanic        1.038e-03  7.863e-05  13.196  < 2e-16 ***
## perc.male            1.320e-02  2.300e-02   0.574 0.566133
## perc.65up           -1.217e-01  1.275e-02  -9.549  < 2e-16 ***
## unemp.rate          -1.202e-02  4.000e-02  -0.300 0.763906
## lfpr                -8.099e-02  9.935e-03  -8.152 5.19e-16 ***
## weekly.wages        -4.123e-06  2.826e-06  -1.459 0.144673
## perc.rural           4.630e-05  2.496e-05   1.855 0.063760 .
## perc.manuf           2.611e-02  5.144e-03   5.075 4.10e-07 ***
## perc.trump.2016      9.421e-01  4.481e-03 210.247  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0244 on 3052 degrees of freedom
## Multiple R-squared:  0.9771, Adjusted R-squared:  0.977
## F-statistic:  9294 on 14 and 3052 DF,  p-value: < 2.2e-16
```

> - Write a few sentences here. Now there’s only a 2.4% average error between the predicted 2020 trump support and true 2020 trump support. Our model is drastically improved! Ah ha! So fun!

# Extra Credit [5 points]

Using a random forest with a permutation test for variable importance, determine which $X$ variables are most important to predicting county-level support for Trump in 2020. Plot the result. What is the most important predictor? How important is the Covid-19 death rate?

```
# INSERT CODE HERE
```

- Write a few sentences here.