Covid-19

Albin Abraham

5/3/2020

### Introduction

This year a novel strain of coronavirus called SARS-CoV-2 has been spreading across the world at an alarming rate. It is a respiratory virus that causes a range of mild to severe symptoms. Since some of the mild cases don’t show any symptoms at all, asymptomatic transmitters for the virus have been a huge concern. Compared to it’s most recent counterpart, SARS-CoV (2003 pandemic), both viruses have 86% similar genetic structure, yet SARS-CoV-2 has nearly a 10 fold the strength when binding to the host cells. These differences in binding mechanisms are due to how effective the virus’ spike proteins are. In the case of SARS-CoV-2, the spike proteins are incredibly effective at targetting and spreading among human cells. Transmitting this virus can be done through close contact with an infected person, with recent studies claiming most transmissions of the virus occur within 6 feet. It is contracted through respiratory droplets. Some recent studies have shown significant evidence suggesting varying lengths of time the droplets may be suspended in air, also known as the aersol transmission of the virus. Simply put, when an infected person coughs, sneezes, or even talks, he or she will eject liquid particles containing the virus into the air, which then can be ingested by someone else, resulting in infection. Some studies suggest the virus can survive on surfaces from several hours to even days. However, it is important to note most of the data suggests contraction of this virus is most likely done through the air. Going into this experiment, one of the important variables we wanted to analyze was humidity. Based on studies for coronavirus, including SARS-CoV-2, coronaviruses cannot exist in dry air and require a liquid medium to travel to the host cells. In the research community, transimission that doesn’t require a liquid transport vesicle is referred to as airborne transmission. An airborne virus, such as measles, may spread in any form of air, including dust, resulting in a very high transmission rate. Based on this premise, we wanted to analyze humidity to see if transmission increased in counties that had a better environment to transmit the infection, hopefully confirming the claims made by modern research on SARS-CoV-2. Unfortunately, due to regulation of data sets, we couldn’t access humidity by day and were forced to rely solely on temperatures.

Many countries have implemented a travel ban from foreign countries. South Korea has implemented public “phone booths” for rapid testing. Some countries like Italy, China, France, and the UK have put restrictive mass quarantines in place where people will get fined and ticketed for leaving. Some people have even suggested potentially lethal solutions like injecting alcohol, bear bile, or cleaning disinfectant. As of now, there is not conclusive data on an effective treatment or vaccine. The CDC and WHO have made many suggestions, from wearing masks and gloves to social distancing. With rising infection rates and the world-wide death toll still climbing, the solution to repressing and eliminating the virus is as important as ever. Due to the huge variance in community impact in the US, we are seeing significantly varying opinions on what policies people believe should be implemented. Whether it is misinformation from a few clinicians in California or people preaching about their rights, the actual data for these communities needs to be analyzed PROPERLY to make effective and safe decisions. This report was designed to analyze a small scale of communities, United States counties, to see if there are any obvious trends in the data provided by the CDC. Out of the ten counties analyzed, we made sure to take comparatively larger counties, ranging from east to west coast and southern to northern US.

We will utilize R to test the correlation between different variables, focusing on population density. We define density of a region to be the population of the county divided by the area of land mass of the county. After seeing the significant impact on densely populated regions, like New York City, we proposed the following hypothesis:

As population density increases, positive tests/cases per population increases.

With this hypothesis in mind, we wanted to test a few other variables. Labeled “Escore,” this variable is the number of enplanements in millions taken from the largest international airport in a 50 mile radius. Additionally, MeanDistance is the average distance a civilian in that state must travel to reach an airport. Lastly, DaysClosed is the variable we used to help predict the effectiveness of our current quarantining policies in these communties. In the US, states has closed down non-essential labor, but the dates these policies were enforced varied. We recorded these variations in this variable by accounting for how many of the 28 days the county’s non-essential businesses were closed.

### Methodology and Data:

Disclaimer: To properly test the variables proposed above, we would need a much larger sample. Admittedly, a more optimal method would be to generate samples randomly from a large sized data frame with an earlier Day 0. A procedure this scale would be a huge investment of time and the methods in this study were designed to be a precursor for further large-scale investigation.

The data collected in this study was provided by the CDC and ranged from early January to late April. For this analysis we analyzed vector lengths of 28, representing 28 days following the first 100 infections that were present in the county. The reason for choosing 100 was to ensure community spread infection was occurring in the county before analyzing the data points. 28 days was the chosen vector length because our smallest vector length of days equal to or greater than 100 infections was Jefferson County, with 28 days. In the code below, we organized the data by subsetting the dataframes to counties and assigning our variables to each county. From there, we combined the measures and variables together and performed a pairs anaylsis. Using the pairs function, we could see density impacted both measures significantly, and as a result we chose to factor density into our graphic analysis of each variable. To confirm this we ran a linear regression model on our variables, resulting in the significant correlation between positive infections and the county’s population density. Finally, due to some suspicious outliers in the data frame, we created barplots and a unique expanded analysis to further investigate the communities that had suspicious results.

Although our sample was small, we attempted to diversify our sample by taking counties from fairly different regions. This was to ensure significant differences in temperature, population sizes/areas, etc. Most notably, we made sure to find a more rural community to analyze, Jefferson County, which was difficult due to our margin for 100 infections. This will become more and more easy to analyze as the data sets increase in length and infections in communities grow.

##### Code for subsetting each data frame, example being Jefferson County:

JFD <- subset(zdata, zdata$State == "AL" & zdata$County.Name == “Jefferson County”) JFD <- JFD[5:length(JFD)] JFD <- t(JFD) colnames(JFD) <- “Cases” JFD <- as.data.frame(JFD) JFD <- subset(JFD, JFD$Cases > 100 & JFD$Cases <= 700)

The first few lines were to subset our primary dataframe, zdata, and flip the rows with columns using the t() function. Also, the JFD[5:length(JFD)] command was used because the first few data points in each column were identification numbers we didn’t want included in our dataframe. Then we simply named the column and used another subset commands to specify exactly which cases were above 100 infections. The limiting command <=700 was necessary for some of the counties because we had to shrink down the vector length to 28.

Using these two commands we were able to organize the data based on day 0-27 and keep the cases in order by data: JFD$Day <- 0:27 JFD[,c(“Day”,“Cases”)] -> JFD

From there it was just a matter of adding on our variables and manually inputted informations to the preexisting county dataframe by using commands like this: JFD$TEMP <- c(75,83,84,85,76,74,60,67,72,78,80,82,84,73,80,77,62,73,73,67,63,62,70,77,67,59,73,77)

### Issues

The most glaring weaknesses in this report is the sample size and the counties were hand chosen. The best way to overcome this is by increasing the amount of days and lower the day 0 required number of infections. To propose a solution to this, we could take something like 25 infections as our day 0 threshold since 100 infections was a little bit excessive. This decrease would allow us to analyze communities that weren’t impacted on such a large scale. It would also allow us to analyze more days, and consequently more infection rates, ultimately leading to more accurate averages. Cutting out the 25-100 days could have dramatically impacted the infection rate in certain communities because of one particular variable, DaysClosed. In some cases, communities were closed all 28 days. Extending to 25 infections would bring in earlier dates, and ultimately, could give more accurate results for the quarantine correlations.

An additional proposed solution is to create an alternate dataframe containing all possible counties that have fulfilled the above conditions and test the significance of other claims. This can be done by generating significantly large samples and performing a proportion test/t-test on these generated samples. In fact, using a concept like central limit theorem, we could generate an average of multiple samples from the population to predict the population’s average infection rate, assuming our testing in the US proficiently scales up to the magnitude required to generate an n large enough that CLT can be reasonably applied. An example of this application could be taking multiple counties in a state to propose the state population’s average infection. Then, policies could be made on a federal level based on certain state thresholds, and similarly, using the data from the unique counties we could plan a course of action on a more localized level.

Finally, the last issue is with our linear regression analysis. While we have optimal variable conditions for a simple linear regression, we did not have enough data points to make significant claims, relying on the regression lines generated. For example, when looking at something like a pairs function, we need many points, much more than 10, to reasonably predict some slightly correlated trends. Due to our restricted sample size, we could be missing correlations that may be present in larger samples. In fact, even our highly correlated variable, density, could be a fluke due to the chance we selected the outlier counties. One example of this is a slight negative correlation to temperature, but we can’t conclusively test this because there aren’t enough data points.

### Analysis

To properly analyze most of this data, we had to account for density, as explained previoulsy in this report. To do that, we had to factor in density to our infection rate measure and our total cases measure to see how the data sets and graphs might have been altered, to better account for the unique impact this variable might have had. To create a proper pairs function and to use the linear regression modeling in R, we also created two data frames : BigTen and BigTenAnalysis. BigTen was simply to aggregate the values calculated in the code below and BigTenAnalysis was a subset used to analyze our specific variables. For example we can see population, area, and density all included in BigTen, but only density was included in BigTenAnalysis. This was to limit or linear regression analysis and to keep our pairs function limited to our 2 measures and 5 variables.

In terms of process for the analysis, we wanted to do the following:

Step 1: Use boxplots to verify if there were any significant outliers that shouldn’t be present, essentially errors in the data

Step 2: Analyze the pairs correlations to predict how to organize our data and graphic representations

Step 3: Confirm any potential correlations seen using the pairs function by using a linear regression analysis

Step 4: Accomodate our graphs accounting for significant variables.

Step 5: Rank our BigTen counties based on these graphic results and distinguish the impact of significant variables (Example: the shift between Queens County and Miami-Dade County in our graphs once accounting for population density)

######################################################################################################  
#################################### Daily Infection / Population ####################################  
######################################################################################################  
  
x <- sequence(28)  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (JFD$`Cases`[i+1] - JFD$`Cases`[i])}  
JFD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (CCD$`Cases`[i+1] - CCD$`Cases`[i])}  
CCD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (LAD$`Cases`[i+1] - LAD$`Cases`[i])}  
LAD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (PHD$`Cases`[i+1] - PHD$`Cases`[i])}  
PHD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (QND$`Cases`[i+1] - QND$`Cases`[i])}  
QND$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (CKD$`Cases`[i+1] - CKD$`Cases`[i])}  
CKD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (MDD$`Cases`[i+1] - MDD$`Cases`[i])}  
MDD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (KGD$`Cases`[i+1] - KGD$`Cases`[i])}  
KGD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (HAD$`Cases`[i+1] - HAD$`Cases`[i])}  
HAD$InfectionsByDay <- tempv  
  
tempv <- sequence(28)  
for(i in x){  
 tempv[i] <- (WAD$`Cases`[i+1] - WAD$`Cases`[i])}  
WAD$InfectionsByDay <- tempv  
  
####################################################################################################  
################################ Average Infections Per Day Per Pop ################################  
####################################################################################################  
  
JFD$InfectionsByDay / JFD$Population -> JFD$PositiveTestPerPop  
CCD$InfectionsByDay / CCD$Population -> CCD$PositiveTestPerPop  
LAD$InfectionsByDay / LAD$Population -> LAD$PositiveTestPerPop  
PHD$InfectionsByDay / PHD$Population -> PHD$PositiveTestPerPop  
QND$InfectionsByDay / QND$Population -> QND$PositiveTestPerPop  
CKD$InfectionsByDay / CKD$Population -> CKD$PositiveTestPerPop  
MDD$InfectionsByDay / MDD$Population -> MDD$PositiveTestPerPop  
KGD$InfectionsByDay / KGD$Population -> KGD$PositiveTestPerPop  
HAD$InfectionsByDay / HAD$Population -> HAD$PositiveTestPerPop  
WAD$InfectionsByDay / WAD$Population -> WAD$PositiveTestPerPop  
  
  
rep(mean(JFD$PositiveTestPerPop[1:27]),28) -> JFD$AverageInfPerDay  
rep(mean(CCD$PositiveTestPerPop[1:27]),28) -> CCD$AverageInfPerDay  
rep(mean(LAD$PositiveTestPerPop[1:27]),28) -> LAD$AverageInfPerDay  
rep(mean(PHD$PositiveTestPerPop[1:27]),28) -> PHD$AverageInfPerDay  
rep(mean(QND$PositiveTestPerPop[1:27]),28) -> QND$AverageInfPerDay  
rep(mean(CKD$PositiveTestPerPop[1:27]),28) -> CKD$AverageInfPerDay  
rep(mean(MDD$PositiveTestPerPop[1:27]),28) -> MDD$AverageInfPerDay  
rep(mean(KGD$PositiveTestPerPop[1:27]),28) -> KGD$AverageInfPerDay  
rep(mean(HAD$PositiveTestPerPop[1:27]),28) -> HAD$AverageInfPerDay  
rep(mean(WAD$PositiveTestPerPop[1:27]),28) -> WAD$AverageInfPerDay  
  
  
JFD$AverageInfPerDay/JFD$Population -> JFD$AvInfPerPop  
CCD$AverageInfPerDay/CCD$Population -> CCD$AvInfPerPop  
LAD$AverageInfPerDay/LAD$Population -> LAD$AvInfPerPop  
PHD$AverageInfPerDay/PHD$Population -> PHD$AvInfPerPop  
QND$AverageInfPerDay/QND$Population -> QND$AvInfPerPop  
CKD$AverageInfPerDay/CKD$Population -> CKD$AvInfPerPop  
MDD$AverageInfPerDay/MDD$Population -> MDD$AvInfPerPop  
KGD$AverageInfPerDay/KGD$Population -> KGD$AvInfPerPop  
HAD$AverageInfPerDay/HAD$Population -> HAD$AvInfPerPop  
WAD$AverageInfPerDay/WAD$Population -> WAD$AvInfPerPop  
  
####################################################################################################  
############################### Average Infections Per Day Per Area ################################  
####################################################################################################  
JFD$Cases / JFD$Area -> JFD$CasesPerArea  
CCD$Cases / CCD$Area -> CCD$CasesPerArea  
LAD$Cases / LAD$Area -> LAD$CasesPerArea  
PHD$Cases / PHD$Area -> PHD$CasesPerArea  
QND$Cases / QND$Area -> QND$CasesPerArea  
CKD$Cases / CKD$Area -> CKD$CasesPerArea  
MDD$Cases / MDD$Area -> MDD$CasesPerArea  
KGD$Cases / KGD$Area -> KGD$CasesPerArea  
HAD$Cases / HAD$Area -> HAD$CasesPerArea  
WAD$Cases / WAD$Area -> WAD$CasesPerArea  
  
JFD$Cases / (JFD$Density\*1000000) -> JFD$CasesPerDensity  
CCD$Cases / (CCD$Density\*1000000) -> CCD$CasesPerDensity  
LAD$Cases / (LAD$Density\*1000000) -> LAD$CasesPerDensity  
PHD$Cases / (PHD$Density\*1000000) -> PHD$CasesPerDensity  
QND$Cases / (QND$Density\*1000000) -> QND$CasesPerDensity  
CKD$Cases / (CKD$Density\*1000000) -> CKD$CasesPerDensity  
MDD$Cases / (MDD$Density\*1000000) -> MDD$CasesPerDensity  
KGD$Cases / (KGD$Density\*1000000) -> KGD$CasesPerDensity  
HAD$Cases / (HAD$Density\*1000000) -> HAD$CasesPerDensity  
WAD$Cases / (WAD$Density\*1000000) -> WAD$CasesPerDensity  
  
JFD$InfectionsByDay / JFD$Area -> JFD$PositiveTestPerArea  
CCD$InfectionsByDay / CCD$Area -> CCD$PositiveTestPerArea  
LAD$InfectionsByDay / LAD$Area -> LAD$PositiveTestPerArea  
PHD$InfectionsByDay / PHD$Area -> PHD$PositiveTestPerArea  
QND$InfectionsByDay / QND$Area -> QND$PositiveTestPerArea  
CKD$InfectionsByDay / CKD$Area -> CKD$PositiveTestPerArea  
MDD$InfectionsByDay / MDD$Area -> MDD$PositiveTestPerArea  
KGD$InfectionsByDay / KGD$Area -> KGD$PositiveTestPerArea  
HAD$InfectionsByDay / HAD$Area -> HAD$PositiveTestPerArea  
WAD$InfectionsByDay / WAD$Area -> WAD$PositiveTestPerArea  
  
JFD$InfectionsByDay / (JFD$Density\*1000000) -> JFD$PositiveTestPerDensity  
CCD$InfectionsByDay / (CCD$Density\*1000000) -> CCD$PositiveTestPerDensity  
LAD$InfectionsByDay / (LAD$Density\*1000000) -> LAD$PositiveTestPerDensity  
PHD$InfectionsByDay / (PHD$Density\*1000000) -> PHD$PositiveTestPerDensity  
QND$InfectionsByDay / (QND$Density\*1000000) -> QND$PositiveTestPerDensity  
CKD$InfectionsByDay / (CKD$Density\*1000000) -> CKD$PositiveTestPerDensity  
MDD$InfectionsByDay / (MDD$Density\*1000000) -> MDD$PositiveTestPerDensity  
KGD$InfectionsByDay / (KGD$Density\*1000000) -> KGD$PositiveTestPerDensity  
HAD$InfectionsByDay / (HAD$Density\*1000000) -> HAD$PositiveTestPerDensity  
WAD$InfectionsByDay / (WAD$Density\*1000000) -> WAD$PositiveTestPerDensity  
  
######################################################################################################  
#################################### Day 27 Infections Per Population ################################  
######################################################################################################  
  
100\*JFD$Day27Infections/(JFD$Population\*10000000) -> JFD$Day27PercentPositives  
100\*CCD$Day27Infections/(CCD$Population\*10000000) -> CCD$Day27PercentPositives  
100\*LAD$Day27Infections/(LAD$Population\*10000000) -> LAD$Day27PercentPositives  
100\*PHD$Day27Infections/(PHD$Population\*10000000) -> PHD$Day27PercentPositives  
100\*QND$Day27Infections/(QND$Population\*10000000) -> QND$Day27PercentPositives  
100\*CKD$Day27Infections/(CKD$Population\*10000000) -> CKD$Day27PercentPositives  
100\*MDD$Day27Infections/(MDD$Population\*10000000) -> MDD$Day27PercentPositives  
100\*KGD$Day27Infections/(KGD$Population\*10000000) -> KGD$Day27PercentPositives  
100\*HAD$Day27Infections/(HAD$Population\*10000000) -> HAD$Day27PercentPositives  
100\*WAD$Day27Infections/(WAD$Population\*10000000) -> WAD$Day27PercentPositives  
  
######################################################################################################  
######################################## 10 County Data Frames #######################################  
######################################################################################################  
  
as.data.frame(matrix(nrow=10,ncol=10,0)) -> BigTen  
rownames(BigTen)<-c("Jefferson", "Contra Costa", "Los Angeles","Philadelphia","Queens","Cook","Miami-Dade","King","Harris","Wayne")  
colnames(BigTen)<-c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")  
BigTen[1,]<-JFD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[2,]<-CCD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[3,]<-LAD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[4,]<-PHD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[5,]<-QND[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[6,]<-CKD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[7,]<-MDD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[8,]<-KGD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[9,]<-HAD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen[10,]<-WAD[1,c("Day27PercentPositives","AvInfPerPop", "AveTemp","Escore","MeanDistance","Area","Population","Density", "DaysClosed","Day27Infections")]  
BigTen

## Day27PercentPositives AvInfPerPop AveTemp Escore MeanDistance  
## Jefferson 0.010622155 48.696922 73.32143 1.3 28.76  
## Contra Costa 0.006490468 17.827180 65.39286 25.0 12.77  
## Los Angeles 0.009382470 3.408235 66.78571 39.6 12.77  
## Philadelphia 0.055328283 127.478961 57.53571 15.0 20.44  
## Queens 0.144456665 234.536715 55.35714 29.5 16.30  
## Cook 0.030046602 21.459069 54.07143 76.8 15.87  
## Miami-Dade 0.029676113 39.946537 82.46429 20.8 12.16  
## King 0.014016866 22.195915 52.25000 21.9 18.85  
## Harris 0.010560153 8.100258 79.35714 20.0 16.15  
## Wayne 0.071720983 150.436428 52.32143 16.8 21.35  
## Area Population Density DaysClosed Day27Infections  
## Jefferson 1124 0.659 0.0005862989 24 700  
## Contra Costa 716 1.154 0.0016117318 28 749  
## Los Angeles 4058 10.040 0.0024741252 27 9420  
## Philadelphia 134 1.584 0.0118208955 28 8764  
## Queens 109 2.273 0.0208532110 22 32835  
## Cook 945 5.150 0.0054497354 24 15474  
## Miami-Dade 1899 2.717 0.0014307530 14 8063  
## King 2116 2.253 0.0010647448 21 3158  
## Harris 1703 4.713 0.0027674692 20 4977  
## Wayne 612 1.749 0.0028578431 23 12544

BigTenAnalysis <- BigTen[,c("Day27PercentPositives","AvInfPerPop","AveTemp","Escore","MeanDistance","Density","DaysClosed")]  
  
BigTenAnalysis

## Day27PercentPositives AvInfPerPop AveTemp Escore MeanDistance  
## Jefferson 0.010622155 48.696922 73.32143 1.3 28.76  
## Contra Costa 0.006490468 17.827180 65.39286 25.0 12.77  
## Los Angeles 0.009382470 3.408235 66.78571 39.6 12.77  
## Philadelphia 0.055328283 127.478961 57.53571 15.0 20.44  
## Queens 0.144456665 234.536715 55.35714 29.5 16.30  
## Cook 0.030046602 21.459069 54.07143 76.8 15.87  
## Miami-Dade 0.029676113 39.946537 82.46429 20.8 12.16  
## King 0.014016866 22.195915 52.25000 21.9 18.85  
## Harris 0.010560153 8.100258 79.35714 20.0 16.15  
## Wayne 0.071720983 150.436428 52.32143 16.8 21.35  
## Density DaysClosed  
## Jefferson 0.0005862989 24  
## Contra Costa 0.0016117318 28  
## Los Angeles 0.0024741252 27  
## Philadelphia 0.0118208955 28  
## Queens 0.0208532110 22  
## Cook 0.0054497354 24  
## Miami-Dade 0.0014307530 14  
## King 0.0010647448 21  
## Harris 0.0027674692 20  
## Wayne 0.0028578431 23

summary(lm(formula = BigTen$Day27Infections~BigTen$AveTemp+BigTen$Escore + BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed))

##   
## Call:  
## lm(formula = BigTen$Day27Infections ~ BigTen$AveTemp + BigTen$Escore +   
## BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed)  
##   
## Residuals:  
## 1 2 3 4 5 6 7 8 9 10   
## 997.9 -992.6 4277.5 -4925.7 2390.7 -1458.8 -385.1 -3850.9 -1388.5 5335.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22381.16 23575.17 0.949 0.3962   
## BigTen$AveTemp -111.74 198.82 -0.562 0.6041   
## BigTen$Escore 127.41 108.93 1.170 0.3071   
## BigTen$MeanDistance 25.41 419.93 0.061 0.9546   
## BigTen$Density 1190041.13 284083.89 4.189 0.0138 \*  
## BigTen$DaysClosed -670.01 435.99 -1.537 0.1992   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4944 on 4 degrees of freedom  
## Multiple R-squared: 0.8788, Adjusted R-squared: 0.7273   
## F-statistic: 5.8 on 5 and 4 DF, p-value: 0.05667

summary(lm(formula = BigTen$Day27Infections~BigTen$AveTemp+BigTen$Escore + BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed-1))

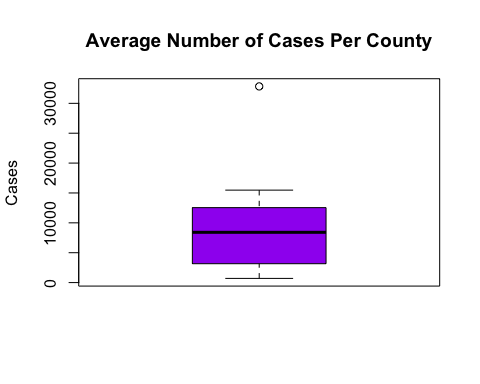
##   
## Call:  
## lm(formula = BigTen$Day27Infections ~ BigTen$AveTemp + BigTen$Escore +   
## BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed -   
## 1)  
##   
## Residuals:  
## 1 2 3 4 5 6 7 8 9 10   
## -1329.7 -851.0 3516.7 -5592.8 2783.7 -2447.4 710.9 -919.4 -2212.3 7325.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## BigTen$AveTemp 55.95 90.36 0.619 0.56292   
## BigTen$Escore 184.28 90.07 2.046 0.09614 .   
## BigTen$MeanDistance 248.36 344.67 0.721 0.50345   
## BigTen$Density 1286517.91 262653.11 4.898 0.00448 \*\*  
## BigTen$DaysClosed -425.38 348.20 -1.222 0.27630   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4895 on 5 degrees of freedom  
## Multiple R-squared: 0.9312, Adjusted R-squared: 0.8624   
## F-statistic: 13.54 on 5 and 5 DF, p-value: 0.006252

# Residuals checks for errors, essentially,  
 # Note there are still possibilities for errors.

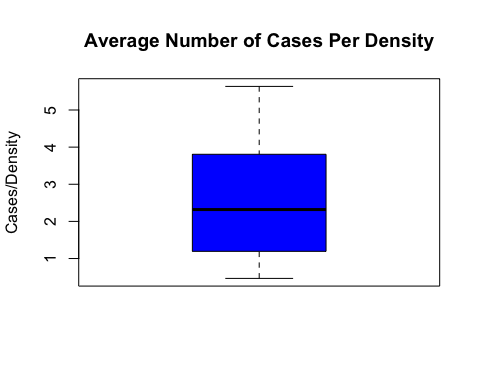
### Results

Once accounting for density, the boxplots indicate there were no major outliers. In the case of Queens County, they were the only outlier if not accounting for density. The only correlated value confirmed by linear regression modeling was density, and even then we can’t confirm this without a larger sample. Once graphing the scatterplots and accounting for densities we chose to put all 10 counties total cases and rates of positive infections by day on two graphs. Using these graphs to analyze potential trends, we concluded the low correlations for temperature, daysclosed, escore, and meandistance were fairly just. There was one important thing to note that requires a deeper analysis and more comprehensive data, and that is the rate analysis. We can see in the infection rate graph the rate of MOST counties had a significant decline in rates, and if not a decline, most definitely a plateau was present. These slopes could be indicative of social distancing or quarantining effectively reducing the rate of infection. There are two counties that weren’t significantly impacted and they were Cook County and Queen’s County. Although Queens county showed a plateau, Cook County appeared to have a linear increase in rate. This was concerning since Cook County has been ranked as one of the most successful counties during quarantine by recent GPS studies on travel. To further investigate this, we subsetted the two counties to have longer days than the 28 days to see if we saw the trend occur later on in the graphs.

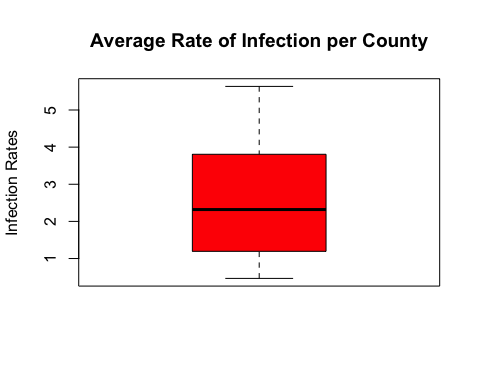
##########################################################################################  
######################################### Graphs #########################################  
##########################################################################################  
  
#Boxplots for all cases in all ten counties   
# - explain that it looks the way it does due to a lack of data sets(counties)  
# - if we use more samples then there would be a more normalized distribution  
# - if this is done effectively, then perhaps a better conclusion could have been drawn.  
  
#Boxplots for Average number of infections per Density for all counties  
# - explain why this does not account for all possible variables related to this study  
# - explain how/why density would make up for the errs in this graph  
# - accounting for density accounts for regions that are in need of intervention  
  
#Boxplot for Average number of infection rates for all counties  
# - The boxplot helps identify outliers for Average number of infection rates   
# - An analysis of the individual counties would help us distinguish the impact of different variables.   
# - Example :  
# - We would want to check outlier communities with low average infection rates to see if a specific variable was causing/correlated to this decrease(outlier)  
# - if enough outliers were to be analyzed and support a particular claim/policy then that data would suggest that it is an effective policy  
  
# Calculate Average number of cases per county and store that as one vector  
AverageCases <- sequence(10)  
AverageCases[1] <- JFD$Day27Infections[1]  
AverageCases[2] <- HAD$Day27Infections[1]  
AverageCases[3] <- KGD$Day27Infections[1]  
AverageCases[4] <- LAD$Day27Infections[1]  
AverageCases[5] <- CKD$Day27Infections[1]  
AverageCases[6] <- QND$Day27Infections[1]  
AverageCases[7] <- MDD$Day27Infections[1]  
AverageCases[8] <- WAD$Day27Infections[1]  
AverageCases[9] <- CCD$Day27Infections[1]  
AverageCases[10] <- PHD$Day27Infections[1]  
boxplot(AverageCases, col = "purple", main = "Average Number of Cases Per County", ylab = "Cases")



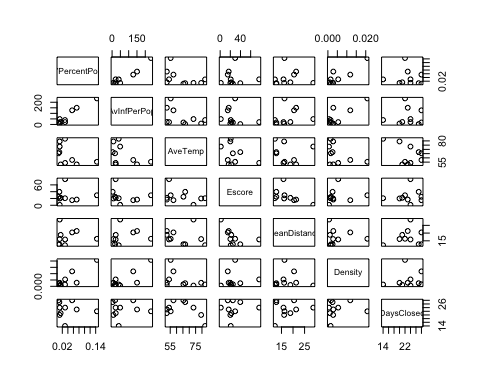
AvCasesPerDensity <- c(1:10)  
AvCasesPerDensity[1] <- AverageCases[1]/(JFD$Density[1] \* 1000000)  
AvCasesPerDensity[2] <- AverageCases[2]/(HAD$Density[1] \* 1000000)  
AvCasesPerDensity[3] <- AverageCases[3]/(KGD$Density[1] \* 1000000)  
AvCasesPerDensity[4] <- AverageCases[4]/(LAD$Density[1] \* 1000000)  
AvCasesPerDensity[5] <- AverageCases[5]/(CKD$Density[1] \* 1000000)  
AvCasesPerDensity[6] <- AverageCases[6]/(QND$Density[1] \* 1000000)  
AvCasesPerDensity[7] <- AverageCases[7]/(MDD$Density[1] \* 1000000)  
AvCasesPerDensity[8] <- AverageCases[8]/(WAD$Density[1] \* 1000000)  
AvCasesPerDensity[9] <- AverageCases[9]/(CCD$Density[1] \* 1000000)  
AvCasesPerDensity[10] <- AverageCases[10]/(PHD$Density[1] \* 1000000)  
boxplot(AvCasesPerDensity, col = "blue", main = "Average Number of Cases Per Density", ylab = "Cases/Density")



AverageInfectionRates <- c(1:10)  
JFD$AverageInfPerDay[1] -> AverageInfectionRates[1]  
CCD$AverageInfPerDay[1] -> AverageInfectionRates[2]  
LAD$AverageInfPerDay[1] -> AverageInfectionRates[3]  
PHD$AverageInfPerDay[1] -> AverageInfectionRates[4]  
QND$AverageInfPerDay[1] -> AverageInfectionRates[5]  
CKD$AverageInfPerDay[1] -> AverageInfectionRates[6]  
MDD$AverageInfPerDay[1] -> AverageInfectionRates[7]  
KGD$AverageInfPerDay[1] -> AverageInfectionRates[8]  
HAD$AverageInfPerDay[1] -> AverageInfectionRates[9]  
WAD$AverageInfPerDay[1] -> AverageInfectionRates[10]  
boxplot(AvCasesPerDensity, col = "red", main = "Average Rate of Infection per County", ylab = "Infection Rates")



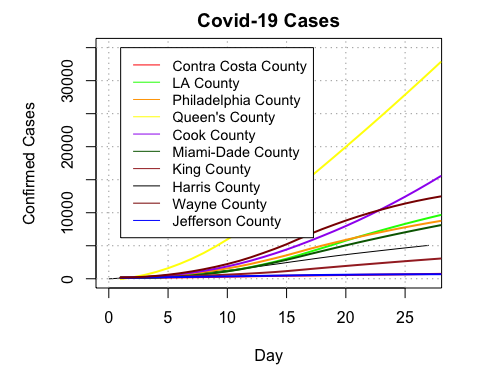
# T tests  
  
#t.test(BigTen$AvInfPerPop) # Average infection Rates / population size for each respective county  
#t.test(AverageCases) # Average number of cases per county  
#t.test(AverageInfectionRates) # Average Infection Rates per county of the 28 day timespan  
#t.test(AvCasesPerDensity) # Average Cases per density of each respective county  
  
#Graph Order   
  
# Analysis of Results  
#1.) Pairs analysis   
# - show possible correlations and state why density is used to adjust graphs  
# - why are we using density accounted measurement to rank(compare) variables.   
  
# Linear Regression  
  
#2.) Graph of Cases Per Day  
# - explain reasoning why New York is on top and suggest comparing variables while accounting for density will draw more accurate conclusions.  
# - based on pairs analysis  
  
#3.) Graph of Cases per Density # Miami will be on top  
# - Present Density graph and explain reasoning  
  
# use these variables to show a corelation to exposer  
#4.) Graph of Cases per Density with Temperature Ranking  
# - Since Data set is small there may not be a valid conclusion to be drawn from this case  
  
#5.) Graph of Cases per Density with Escore  
# - Since Data set is small there may not be a valid conclusion to be drawn from this case  
  
#6.) Graph of Cases per Density with mean Distance  
# - Since Data set is small there may not be a valid conclusion to be drawn from this case  
  
#7.) Graph of Cases per Density with days closed  
# - Since Data set is small there may not be a valid conclusion to be drawn from this case  
  
#8.) Graph of Individual Counties and explain why chicago and New York need further analysis.  
  
#9.) Graph of Counties Positive tests per Density overlapped  
# - propose new study/ another way to ana  
  
  
pairs(BigTenAnalysis)



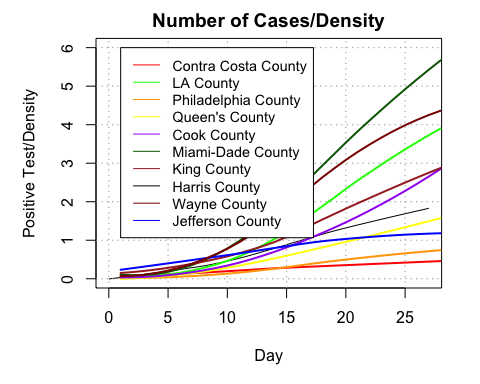
######################################################################################################  
################################## Linear Regression & Analysis ######################################  
######################################################################################################  
# Linear regression on Increase of cases/population size per day.  
summary(lm(formula = BigTen$Day27Infections~BigTen$AveTemp+BigTen$Escore + BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed))

##   
## Call:  
## lm(formula = BigTen$Day27Infections ~ BigTen$AveTemp + BigTen$Escore +   
## BigTen$MeanDistance + BigTen$Density + BigTen$DaysClosed)  
##   
## Residuals:  
## 1 2 3 4 5 6 7 8 9 10   
## 997.9 -992.6 4277.5 -4925.7 2390.7 -1458.8 -385.1 -3850.9 -1388.5 5335.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22381.16 23575.17 0.949 0.3962   
## BigTen$AveTemp -111.74 198.82 -0.562 0.6041   
## BigTen$Escore 127.41 108.93 1.170 0.3071   
## BigTen$MeanDistance 25.41 419.93 0.061 0.9546   
## BigTen$Density 1190041.13 284083.89 4.189 0.0138 \*  
## BigTen$DaysClosed -670.01 435.99 -1.537 0.1992   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4944 on 4 degrees of freedom  
## Multiple R-squared: 0.8788, Adjusted R-squared: 0.7273   
## F-statistic: 5.8 on 5 and 4 DF, p-value: 0.05667

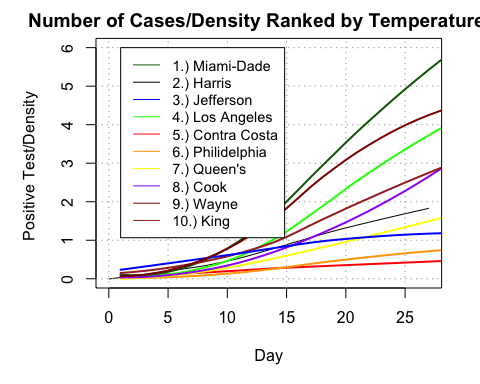
# helps show correlation with density  
# other variables aren't as conclusive since our data does not have enough data entries.  
  
par(mfrow = c(1,1))  
par(mar = c(5,5,2,2)) # alters the margins and prevent possible margin errors.  
  
  
scatter.smooth(HAD$Day,HAD$Cases, type="o", col="white",ylim = c(0,35000),xlab = "Day", ylab = "Confirmed Cases", main = "Covid-19 Cases", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$Cases~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$Cases~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$Cases~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$Cases~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$Cases~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$Cases~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$Cases~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$Cases~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$Cases~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
  
legend(1, 35000, legend=c("Contra Costa County", "LA County", "Philadelphia County", "Queen's County", "Cook County", "Miami-Dade County", "King County", "Harris County", "Wayne County", "Jefferson County"), col=c("red", "green","orange", "yellow", "purple", "dark green", "brown", "black", "dark red", "blue"), lty=1, cex=0.9)



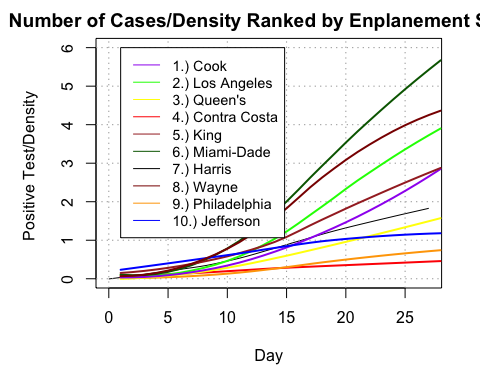
#\*\*\*\*\*\*\*\*\* use this (the graph below) to rank all of the ranking variables  
# explain how we processed this variable.   
# explain all of the factors related to each county  
  
# MDD had highest point  
  
scatter.smooth(HAD$Day,HAD$CasesPerDensity, type="o", col="white",ylim = c(0,6),xlab = "Day", ylab = "Positive Test/Density", main = "Number of Cases/Density", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$CasesPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$CasesPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$CasesPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$CasesPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$CasesPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$CasesPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$CasesPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$CasesPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$CasesPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
legend(1, 6, legend=c("Contra Costa County", "LA County", "Philadelphia County", "Queen's County", "Cook County", "Miami-Dade County", "King County", "Harris County", "Wayne County", "Jefferson County"), col=c("red", "green","orange", "yellow", "purple", "dark green", "brown", "black", "dark red", "blue"), lty=1, cex=0.9)



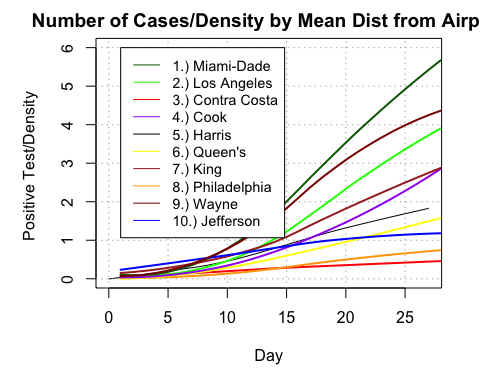
# Ranked by Temperature  
scatter.smooth(HAD$Day,HAD$CasesPerDensity, type="o", col="white",ylim = c(0,6),xlab = "Day", ylab = "Positive Test/Density ", main = "Number of Cases/Density Ranked by Temperature(F)", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$CasesPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$CasesPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$CasesPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$CasesPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$CasesPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$CasesPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$CasesPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$CasesPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$CasesPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
  
legend(1,6, legend=c("1.) Miami-Dade", "2.) Harris", "3.) Jefferson", "4.) Los Angeles", "5.) Contra Costa", "6.) Philidelphia", "7.) Queen's", "8.) Cook", "9.) Wayne", "10.) King"), col = c("dark green", "black", "blue", "green", "red", "orange", "yellow", "purple", "dark red", "brown"), lty = 1, cex = 0.9)



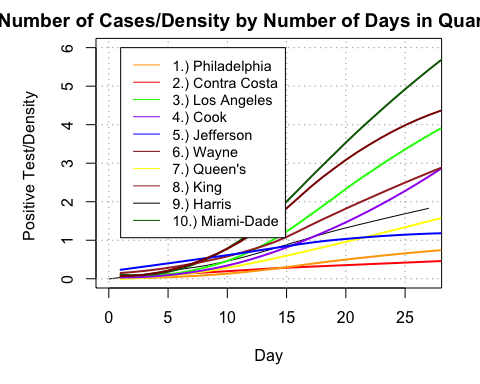
# Ranked by Escore  
scatter.smooth(HAD$Day,HAD$CasesPerDensity, type="o", col="white",ylim = c(0,6),xlab = "Day", ylab = "Positive Test/Density ", main = "Number of Cases/Density Ranked by Enplanement Score", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$CasesPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$CasesPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$CasesPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$CasesPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$CasesPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$CasesPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$CasesPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$CasesPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$CasesPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
legend(1, 6, legend=c("1.) Cook", "2.) Los Angeles", "3.) Queen's", "4.) Contra Costa", "5.) King", "6.) Miami-Dade", "7.) Harris", "8.) Wayne", "9.) Philadelphia", "10.) Jefferson"), col=c("purple", "green", "yellow", "red", "brown", "dark green", "black", "dark red", "orange", "blue"), lty=1, cex=0.9)



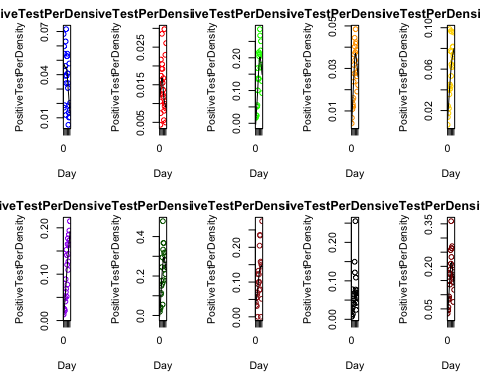
# Ranked by Mean Distance from an airport  
scatter.smooth(HAD$Day,HAD$CasesPerDensity, type="o", col="white",ylim = c(0,6),xlab = "Day", ylab = "Positive Test/Density ", main = "Number of Cases/Density by Mean Dist from Airport", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$CasesPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$CasesPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$CasesPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$CasesPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$CasesPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$CasesPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$CasesPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$CasesPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$CasesPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
legend(1, 6, legend=c("1.) Miami-Dade", "2.) Los Angeles", "3.) Contra Costa", "4.) Cook", "5.) Harris", "6.) Queen's", "7.) King", "8.) Philadelphia", "9.) Wayne", "10.) Jefferson"), col=c("dark green", "green", "red", "purple", "black", "yellow", "brown", "orange", "dark red", "blue"), lty=1, cex=0.9)



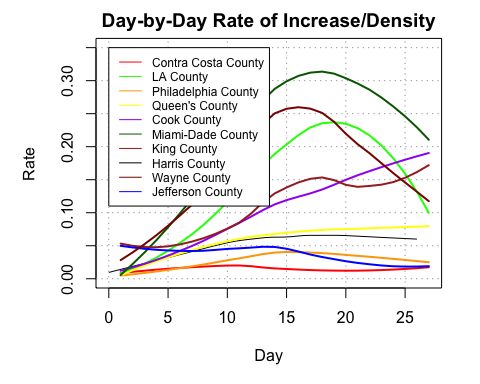
# Ranked by days in quarantine  
scatter.smooth(HAD$Day,HAD$CasesPerDensity, type="o", col="white",ylim = c(0,6),xlab = "Day", ylab = "Positive Test/Density ", main = "Number of Cases/Density by Number of Days in Quarantine", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$CasesPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$CasesPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$CasesPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$CasesPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$CasesPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$CasesPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$CasesPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$CasesPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$CasesPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
legend(1, 6, legend=c("1.) Philadelphia", "2.) Contra Costa", "3.) Los Angeles", "4.) Cook", "5.) Jefferson", "6.) Wayne", "7.) Queen's", "8.) King", "9.) Harris", "10.) Miami-Dade"), col=c("orange", "red", "green", "purple", "blue", "dark red", "yellow", "brown", "black", "dark green"), lty=1, cex=0.9)



# the following would be InfectionsbyDay/Density or slope per square mile  
# use this graph to propose further studies for analysis  
# declined after a given amount of time.  
# Explain further analysis of New York and Chicago  
  
par(mfrow = c(2,5))  
scatter.smooth(y = JFD$PositiveTestPerDensity, x = JFD$Day, main = "PositiveTestPerDensity/Day JFD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "blue")  
scatter.smooth(y = CCD$PositiveTestPerDensity, x = CCD$Day, main = "PositiveTestPerDensity/Day CCD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "red")  
scatter.smooth(y = LAD$PositiveTestPerDensity, x = LAD$Day, main = "PositiveTestPerDensity/Day LAD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "green")  
scatter.smooth(y = PHD$PositiveTestPerDensity, x = PHD$Day, main = "PositiveTestPerDensity/Day PHD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "orange")  
scatter.smooth(y = QND$PositiveTestPerDensity, x = QND$Day, main = "PositiveTestPerDensity/Day QND", ylab = "PositiveTestPerDensity", xlab = "Day", col = "gold")  
scatter.smooth(y = CKD$PositiveTestPerDensity, x = CKD$Day, main = "PositiveTestPerDensity/Day CKD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "purple")  
scatter.smooth(y = MDD$PositiveTestPerDensity, x = MDD$Day, main = "PositiveTestPerDensity/Day MDD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "dark green")  
scatter.smooth(y = KGD$PositiveTestPerDensity, x = KGD$Day, main = "PositiveTestPerDensity/Day KGD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "brown")  
scatter.smooth(y = HAD$PositiveTestPerDensity, x = HAD$Day, main = "PositiveTestPerDensity/Day HAD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "black")  
scatter.smooth(y = WAD$PositiveTestPerDensity, x = WAD$Day, main = "PositiveTestPerDensity/Day WAD", ylab = "PositiveTestPerDensity", xlab = "Day", col = "dark red")



# final Proposition graph  
# use this graph to propose further studies for analysis  
# compare and mark each counties applied quarantine and observe if the infection rates  
# declined after a given amount of time.  
par(mfrow = c(1,1))  
scatter.smooth(HAD$Day,HAD$PositiveTestPerDensity, type="o", col="white",ylim = c(0,0.35),xlab = "Day", ylab = "Rate", main = "Day-by-Day Rate of Increase/Density", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(CCD$PositiveTestPerDensity~CCD$Day)  
lines(predict(lo), col="red", lwd=2)  
lo <- loess(LAD$PositiveTestPerDensity~LAD$Day)  
lines(predict(lo), col="green", lwd=2)  
lo <- loess(PHD$PositiveTestPerDensity~PHD$Day)  
lines(predict(lo), col="orange", lwd=2)  
lo <- loess(QND$PositiveTestPerDensity~QND$Day)  
lines(predict(lo), col="yellow", lwd=2)  
lo <- loess(CKD$PositiveTestPerDensity~CKD$Day)  
lines(predict(lo), col="purple", lwd=2)  
lo <- loess(MDD$PositiveTestPerDensity~MDD$Day)  
lines(predict(lo), col="dark green", lwd=2)  
lo <- loess(KGD$PositiveTestPerDensity~KGD$Day)  
lines(predict(lo), col="brown", lwd=2)  
lo <- loess(JFD$PositiveTestPerDensity~JFD$Day)  
lines(predict(lo), col="blue", lwd=2)  
lo <- loess(WAD$PositiveTestPerDensity~WAD$Day)  
lines(predict(lo), col="dark red", lwd=2)  
legend(0, 0.35, legend=c("Contra Costa County", "LA County", "Philadelphia County", "Queen's County", "Cook County", "Miami-Dade County", "King County", "Harris County", "Wayne County", "Jefferson County"), col=c("red", "green","orange", "yellow", "purple", "dark green", "brown", "black", "dark red", "blue"), lty=1, cex=0.75)

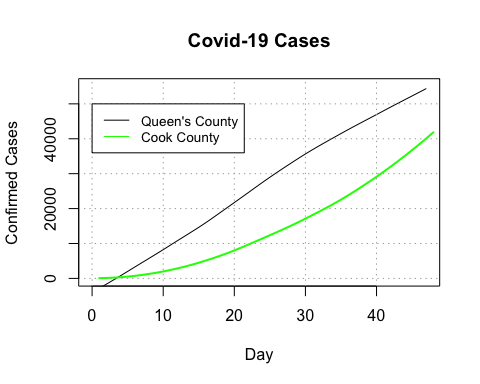


#\*\*\*\*\*

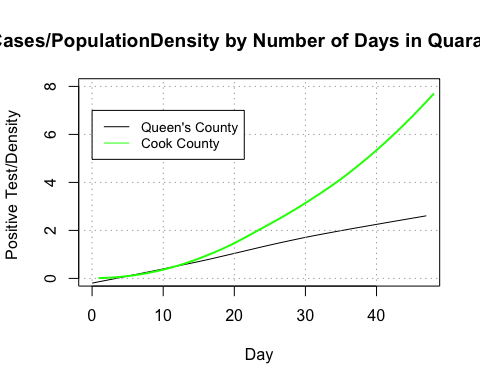
### Analysis Extended

Based on the extended graphs from this section, we can verify the prediction from above for Queens county and we can see a small impact in Cook County. However, we notice the sudden uptick after the hump, which implies there is something concerning to analyze in Cook County. If looking at the first two graphs, one can verify the exponential growth for Cook County cases, and the linear increase for growth rate. Again, these are both concerning since Cook County was suggested to be implementing social distancing so successfully and that we don’t necessarily have the most dense population. Even Queens is seeing a down turn in infection rates.

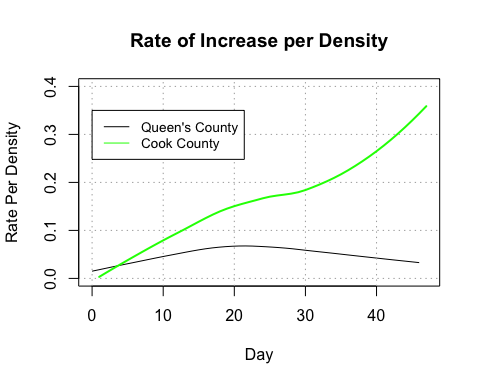
#5# Queens County starts on March 16th the length will be adjusted for to match Chicago #######################################################################  
ExQND <- subset(zdata, zdata$State == "NY" & zdata$County.Name == "Queens County")  
ExQND <- ExQND[5:length(ExQND)]  
ExQND <- t(ExQND)  
colnames(ExQND) <- "Cases"  
ExQND <- as.data.frame(ExQND)  
ExQND <- subset(ExQND, ExQND$Cases > 100 & ExQND$Cases < 53100)  
  
ExQND$Day <- 0:47  
ExQND[,c("Day","Cases")] -> ExQND  
ExQND$Escore <- rep(29.5,48)  
ExQND$MeanDistance<-rep(16.30,48)  
ExQND$Area<-rep(109,48)  
ExQND$Population <- rep(2.273,48)  
ExQND$Population/ExQND$Area -> ExQND$Density  
ExQND$DaysClosed <- rep(42,48)  
  
tempv <- sequence(48)  
x <- sequence(48)  
for(i in x){  
 tempv[i] <- (ExQND$`Cases`[i+1] - ExQND$`Cases`[i])}  
  
ExQND$InfectionsByDay <- tempv  
ExQND$InfectionsByDay / ExQND$Population -> ExQND$PositiveTestPerPop  
rep(mean(ExQND$PositiveTestPerPop[1:47]),48) -> ExQND$AverageInfPerDay  
ExQND$AverageInfPerDay/ExQND$Population -> ExQND$AvInfPerPop  
ExQND$Cases / ExQND$Area -> ExQND$CasesPerArea  
ExQND$Cases / (ExQND$Density\*1000000) -> ExQND$CasesPerDensity  
ExQND$InfectionsByDay / ExQND$Area -> ExQND$PositiveTestPerArea  
ExQND$InfectionsByDay / (ExQND$Density\*1000000) -> ExQND$PositiveTestPerDensity  
  
#6# Cook County starts on March 17th #########################################################################  
  
ExCKD <- subset(zdata, zdata$State == "IL" & zdata$County.Name == "Cook County")  
ExCKD <- ExCKD[5:length(ExCKD)]  
ExCKD <- t(ExCKD)  
colnames(ExCKD) <- "Cases"  
ExCKD <- as.data.frame(ExCKD)  
ExCKD <- subset(ExCKD, ExCKD$Cases > 100)  
  
ExCKD$Day <- 0:47  
ExCKD[,c("Day","Cases")] -> ExCKD  
ExCKD$Escore <- rep(76.8,48)  
ExCKD$MeanDistance<-rep(15.87,48)  
ExCKD$Area<-rep(945,48)  
ExCKD$Population <- rep(5.15,48)  
ExCKD$Population/ExCKD$Area -> ExCKD$Density  
ExCKD$DaysClosed <- rep(44,48)  
  
tempv <- sequence(48)  
x <- sequence(48)  
for(i in x){  
 tempv[i] <- (ExCKD$`Cases`[i+1] - ExCKD$`Cases`[i])}  
  
ExCKD$InfectionsByDay <- tempv  
ExCKD$InfectionsByDay / ExCKD$Population -> ExCKD$PositiveTestPerPop  
rep(mean(ExCKD$PositiveTestPerPop[1:47]),48) -> ExCKD$AverageInfPerDay  
ExCKD$AverageInfPerDay/ExCKD$Population -> ExCKD$AvInfPerPop  
ExCKD$Cases / ExCKD$Area -> ExCKD$CasesPerArea  
ExCKD$Cases / (ExCKD$Density\*1000000) -> ExCKD$CasesPerDensity  
ExCKD$InfectionsByDay / ExCKD$Area -> ExCKD$PositiveTestPerArea  
ExCKD$InfectionsByDay / (ExCKD$Density\*1000000) -> ExCKD$PositiveTestPerDensity  
  
scatter.smooth(ExQND$Day,ExQND$Cases, type="o", col="white",ylim = c(0,55000),xlab = "Day", ylab = "Confirmed Cases", main = "Covid-19 Cases", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(ExCKD$Cases~ExCKD$Day)  
lines(predict(lo), col="green", lwd=2)  
legend(0, 50000, legend=c("Queen's County", "Cook County"), col=c("black", "green"), lty=1, cex=0.85)



scatter.smooth(ExQND$Day,ExQND$CasesPerDensity, type="o", col="white",ylim = c(0,8),xlab = "Day", ylab = "Positive Test/Density ", main = "Cases/PopulationDensity by Number of Days in Quarantine", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(ExCKD$CasesPerDensity~ExCKD$Day)  
lines(predict(lo), col="green", lwd=2)  
legend(0, 7, legend=c("Queen's County", "Cook County"), col=c("black", "green"), lty=1, cex=0.85)



scatter.smooth(ExQND$Day,ExQND$PositiveTestPerDensity, type="o", col="white",ylim = c(0,0.4),xlab = "Day", ylab = "Rate Per Density", main = "Rate of Increase per Density", lwd = 2)  
grid(nx = NULL, ny = NULL, col = "darkgray", lty = "dotted")  
lo <- loess(ExCKD$PositiveTestPerDensity~ExCKD$Day)  
lines(predict(lo), col="green", lwd=2)  
legend(0, 0.35, legend=c("Queen's County", "Cook County"), col=c("black", "green"), lty=1, cex=0.85)



### Summary/Results & Discussion:

Overall, our results implied statistical significance for density, but no other variable. To really conclude this, we would need to do a much larger scale analysis, using randomized samples. There was a potential negative correlation between temperature and both of our measures as well.

The purpose of this report was to inform and predict potential candidates to look at for a more in depth linear regression analysis. With very little doubt, it seems density is a significant factor and should be analyzed on a much larger scale. Additionally, there is a significant amount of potential in days closed and temperature. The correlation between Escore and infections could not be due to the airport factor, but rather because larger airports are made because of larger demands from larger populations, implying the correlation may have just come from the density numerator, population. There was also potential correlation in distance from an airport as well, but this could also be an overlapping factor derived from area. Smaller states generally have lower areas per county, and as a result have lower distances to travel to an airport, making both values difficult to assess quantitatively.

This report accomplished it’s goal to suggest potential variables that could be correlated to differences in COVID-19 total infections and infection rates by county. This data analysis isn’t intended to be direct evidence that there is correlation, but an indicator that our variables density, temperature, and daysclosed are prime candidates for a much larger analysis using counties, whereas Escore and average distance may not be as prospective.

In conclusion, these results do validate our initial hypothesis and demand further analysis to properly validate this claim.

### Sources

CDC Data pulled from: usafacts.org/visualizations/coronavirus-covid-19-spread-map

All DaysClosed data was collected from the Coronavirus United States Wikipedia Page: <https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_United_State>

Escore pulled from Wiki <https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States>

MeanDistance data pulled from a Pearson study at: <https://www.mark-pearson.com/airport-distances/>.

Density variable was derived from area and population variables accumulated from county Wikipedia pages: <https://en.wikipedia.org/wiki/Jefferson_County,_Alabama>

All temperature data points were pulled from monthly collections at: <https://weather.com/>