Exploring the Relationship Between Weather Conditions and COVID-19 Rates

AUTHOR
Angel Martinez, Jordan Tasho, Matthew Saephan

PUBLISHED
December 14, 2023

Introduction

Covid-19 at its height was the most trending topic in the world and understanding its transmission rate is key to preventing another outbreak, as well as understanding the outbreak that has already occurred. This naturally led to our group pondering whether there were any external variables that might affect transmission rate that weren't discussed in the mainstream media. Our group chose to look at weather as a driving factor in Covid transmission rates, specifically temperature as well as geographical location as this has an effect on weather patterns. We gathered the necessary data from major cities across the US that were well geographically distributed throughout the countries; these being Los Angeles (CA), Jacksonville (FL). Detroit (MI), New York City (NY) and Houston (TX).

Why does any of this matter? Well, in order to prevent another outbreak we must understand what drove the first outbreak to become so deadly. What was the reason when it came to the rapid rate at which the virus spread? Although it will be impossible to completely stop the virus from being transmitted anymore, we can at least try to find a way to reduce the transmission rate and therefore potentially save lives with our research. In order to address this problem we decided to analyze weather and primarily temperature across the different cities as our driving factor for increasing transmission rates. While we can't change the external temperature we can suggest more stringent lockdowns during any periods of time (seasons) where we expect transmission rates to be exceptionally high. Another reason for the distribution of cities we chose, is that some cities are from extremely red states and others are from extremely blue states, so this allowed us to see how political policy impacted transmission as well. A hypothesis made from a states political color is that red states may experience higher infection rates due to there mask mandate and public events being less restrictive than blue states.

Some variables our data included were number of covid cases and death per date, and other weather variables such as precipitation, wind speed, dew, pressure, etc. Then given our team was able to add more variables such as `new_cases` which was calculated by (today's cases - yesterday cases). Then we also made `cases_growth` by doing (new_cases today / yesterday cases). `cases_growth` would be the main variable used to compare with weather variables as it accounts for different populations sizes since bigger populations experience more cases each day than smaller ones.

Upon a mid review section of our research, where we discussed our current research status to other groups of data scientists, A question that came up was how are we accounting for incubation periods. The data of infection doesn't correlate to the same date of weather because there is an incubation period where a person get infected but it isn't known until a couple days later. With this being said,according to the CDC, on average there is a 6.5 day period for "general" covid, 4.3 day period for the Delta variant, and a 3-4 day period for the Omicron variant. Therefore, our team decided that the best approach to account for incubation period with the data we have, is to get the average of the different averages and shift the covid data set dates back. The average ended up being 5 days and once we joined the covid data set to its weather data set, the day a person would have gotten infected more accurately matches the weather of that day than how we initially had it.

Since we are tying to find a correlation and will be trying to make prediction, we split our data into 60% training, 20% validation, and 20% testing set after we have added all the variables to our data. the training set will be used to for EDA and fitting parameters of a model, validation will be used for comparing models and tuning model parameters, then testing set will only be used a single time to test final model. This data splitting is important i norder to test the validation of our models since the testing set is unknown to us until the very end.

Therefore, the goal of this research is to find a correlation between covid rates and weather conditions. Out team have defined a successful outcome if we do find some significant evidence that given a weather variable, it can cause covid rates to either increase or decrease. The failure was defined if our analysis can't find significant evidence of a correlation existing

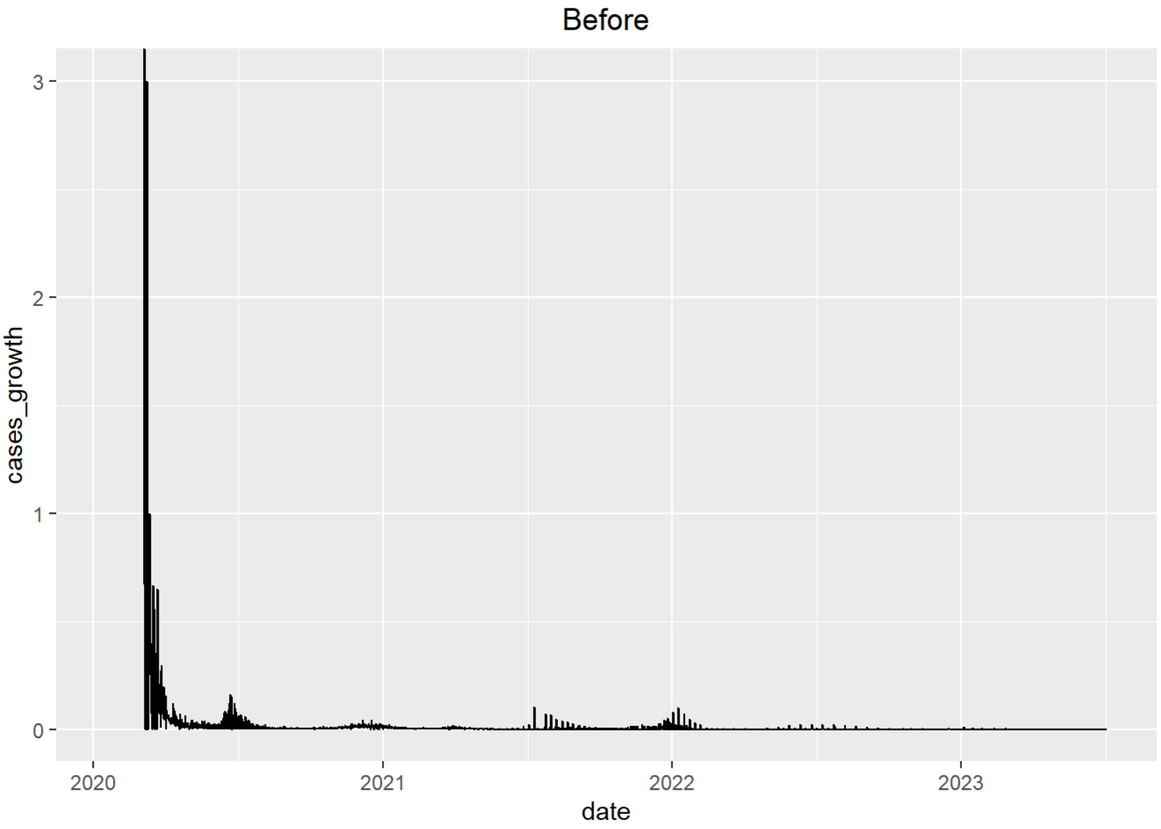
Related Works

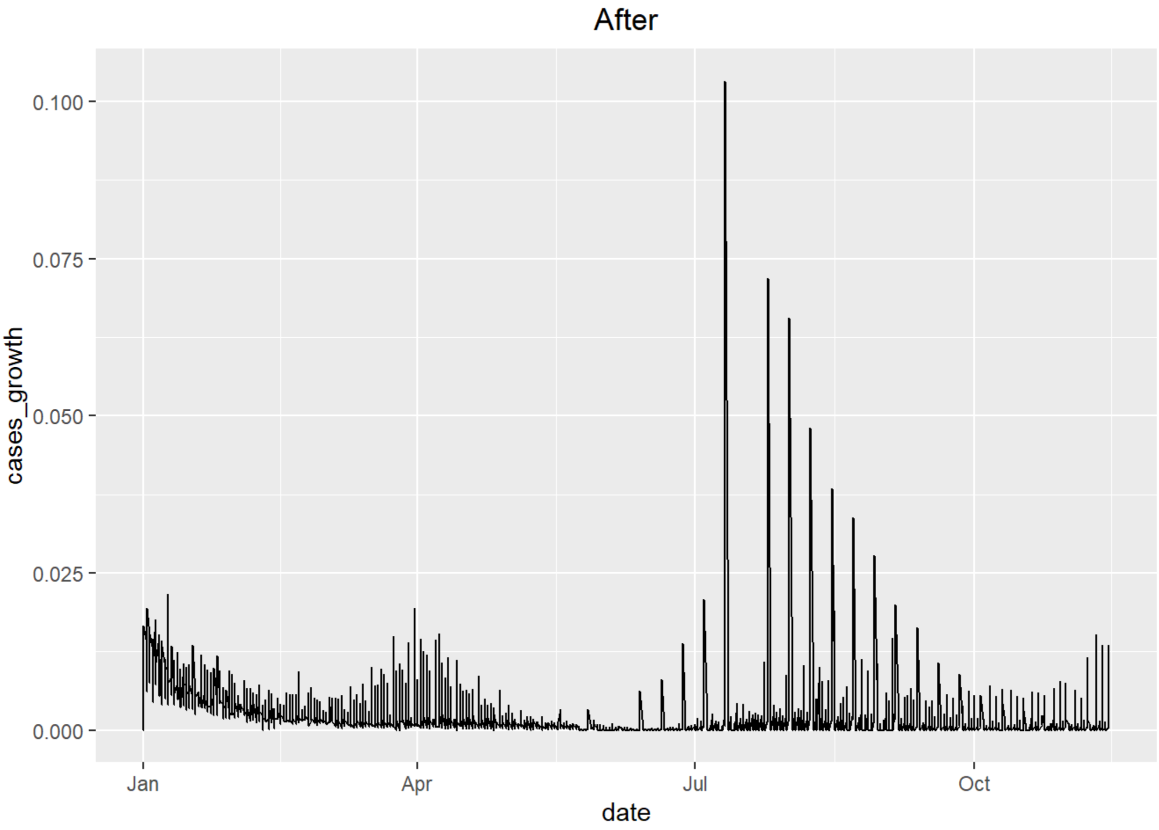
The National Library of Medicine had research article which had a simaler question; they studied the correlation between weather variables and Covid cases/deaths from Italy. They did end up finding a correlation and was able to use other models to find out which weather variable has more of a significant inpact.

Tomasz Zuk and cologues study focuses on assessing the impact of various factors, including climatic conditions (temperature and humidity), census features, and health center capacities, on the spread of COVID-19. The researchers developed predictive models using machine learning techniques, trained on datasets from Kaggle, and evaluated their performance using standard metrics. Results indicate that weather variables, particularly temperature and humidity, are more relevant in predicting the mortality rate compared to population, age, and urbanization.

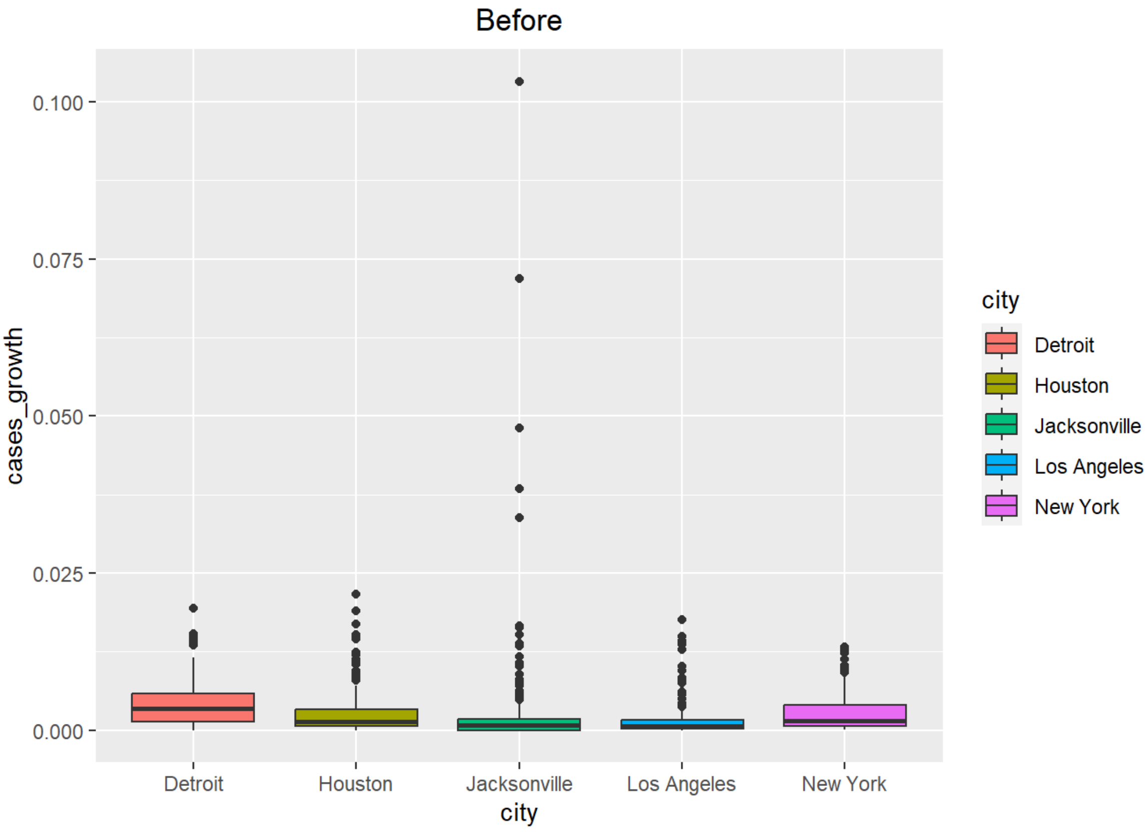
Exploratory Data Analysis

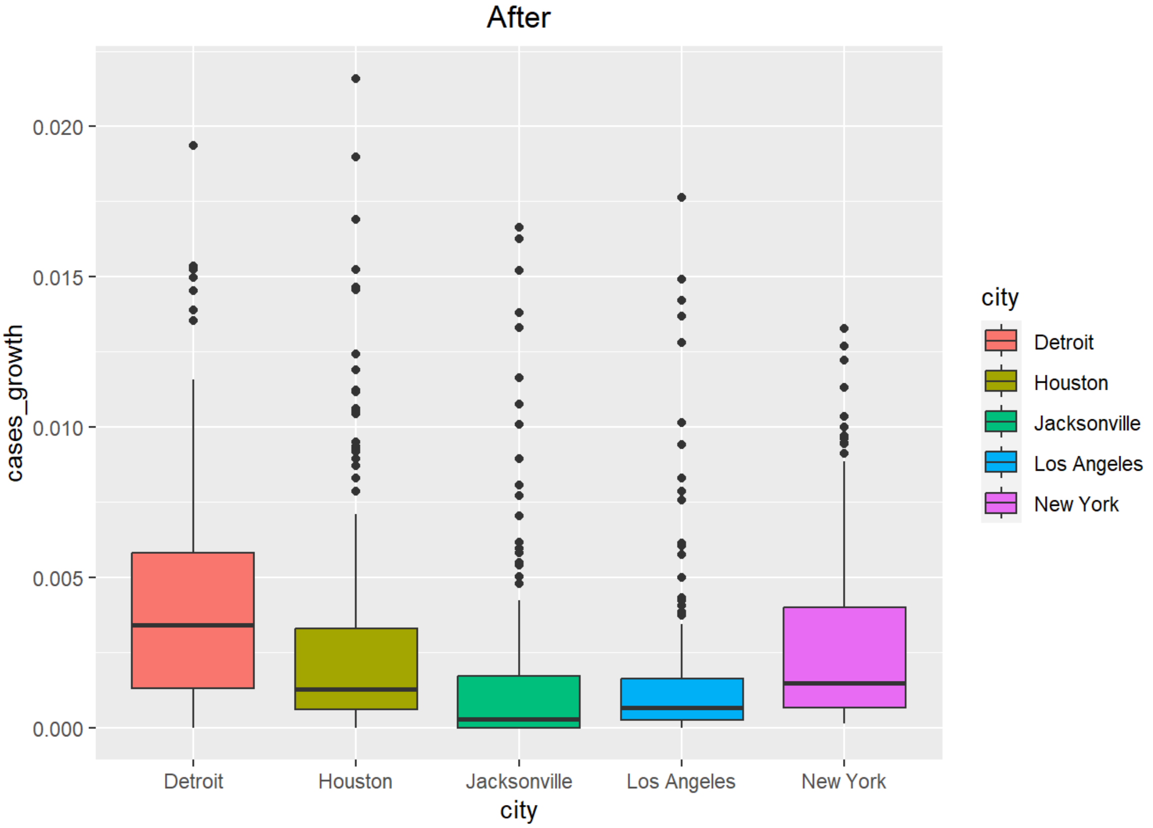
Before we begin to constructing EDA's and applying models to our data set, we noticed that some of our data had observations since the pandemic started. Although it is more data, it would essentially make our maid covid variable `cases_growth` be not so helpful as it will record the amount of growth from the beginning. If you think about it if yesterday there was 1 person infected and today 3 people infected, than that's a total growth of 200%. these number can through off our graphs and models and therefor, our team though it was best to just include 2021. As some data sets are missing some 2022 data.



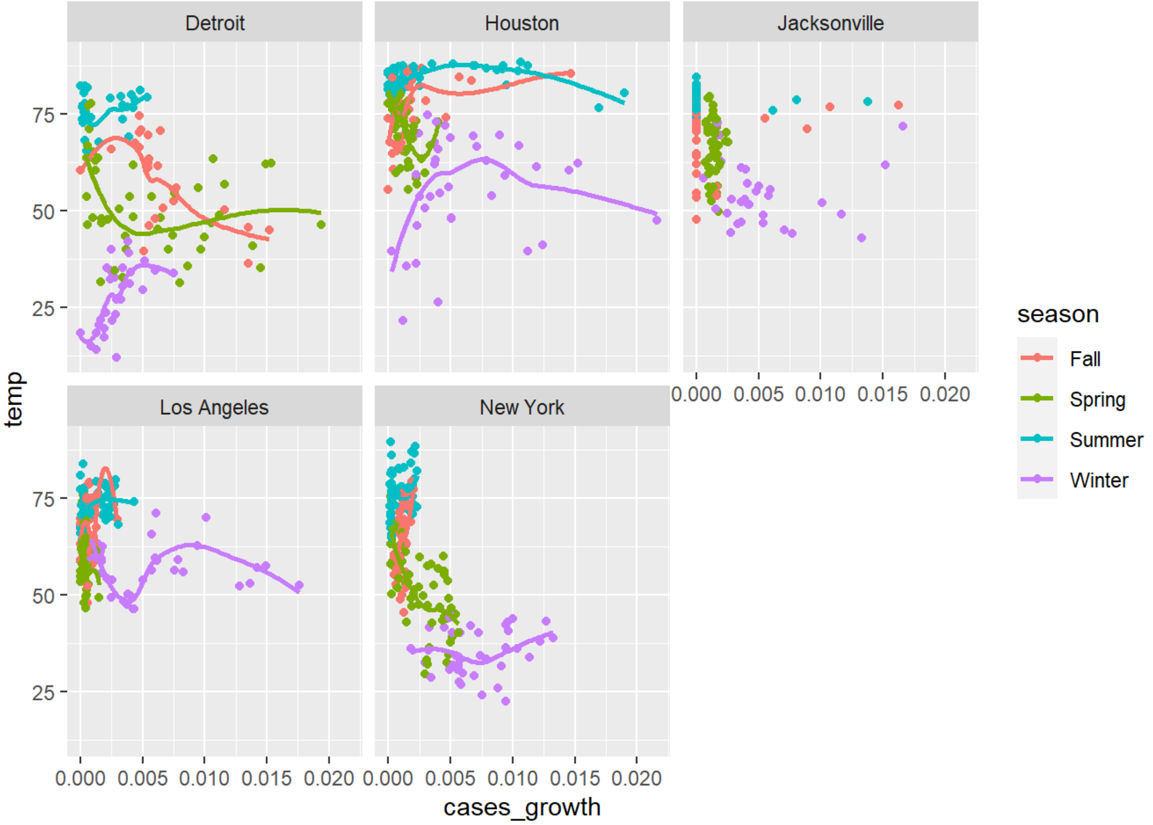


Another notable mention is that our training set has some very large outliers with Florida that could cause results to be misinterpreted. Since it was only 5 observations with such high `cases_growth`, the team found it appropriate to filter these observations out. This lead to our data being more consistent which will be beneficial to making better and more accurate predictions if a correlation is found.



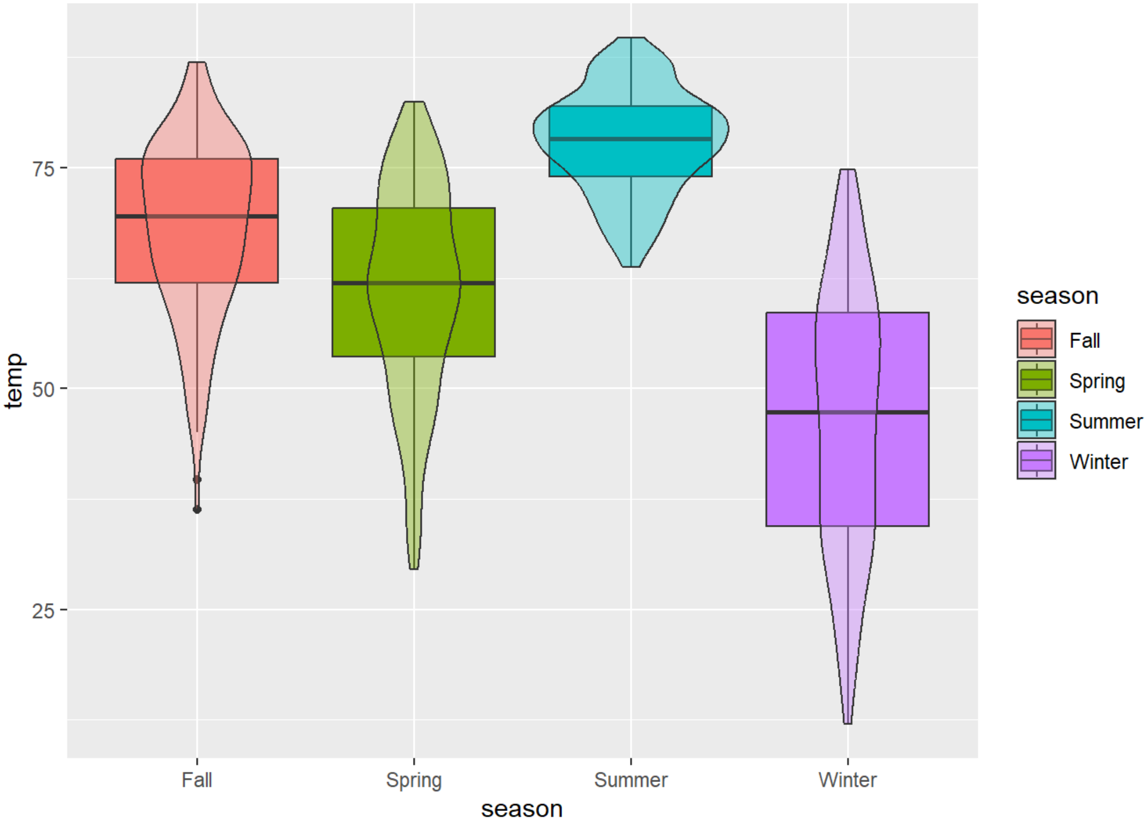


One of graphs we wanted to investigate is see how temperature could have an affect on covid cases. we graphed out the temperature on the y-axis and the cases_growth on the x-axis. A hypothesis that can be made is that the lower the temperature, the more plots we can begin seeing more towards the right side of the graph. This would be a start to investigating if colder temperatures causes higher covid cases.



Given the results of the graph, there does seem to be some type of relationship with winter temp plots extending out the furthest which means more cases growth. But places like Detroit is showing opposite signs and Houston seems to have no relationship. The other weather variables was also applied to this graph however there wasn't any patterns that pop out that could be relevant. To further investigate the how much the different seasons temperature variables, we can make a box plot and violin plot ans see how there different averages place.

```
ggplot(US_training_filtered, aes(x = season, y = temp, fill = season)) +
  geom_boxplot() + geom_violin(alpha = 0.4)
```



Models & Methods

Anova & TukeyHSD

For the beginning stages of our model building, we wanted to first conduct our analysis on more general variables within our data before we dive deeper to the actually weather variables per day. A model used was the ANOVA (analysis of variance) test. This models estimates how quantitative dependent variable changed according to the levels of one or more categorical independent variable. In order to use the ANOVA test, it must stand true to 3 conditions before proceeding:

1. the observations are independent within and across groups
2. the data within each group are nearly normal
3. the variability across the groups are nearly equal

If the conditions are met, then the ANOVA test can be used. If its p-value is less than the significant value, than further testing can be done by using the TukeyHSD (honest significant difference) test. The important of the TukeyHSD test is that it can find which categorical variable are different by comparing the means od each possible pair of groups.

Using the box plot and violin graph from the temperature of our seasons above, we can see that the different seasons are independent within and across groups, data within each season are normal, and variabilty across froups are nearly equal.

Thefore ANOVA test can be used on the diffrent Seasons to get a better idea on much mush does temperture cahnge in the different seasons

```r
anova_result <- aov(temp ~ season, data = US_training_filtered)
summary(anova_result)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
season        3 103942   34647   293.7 <2e-16 ***
Residuals   874 103114     118                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

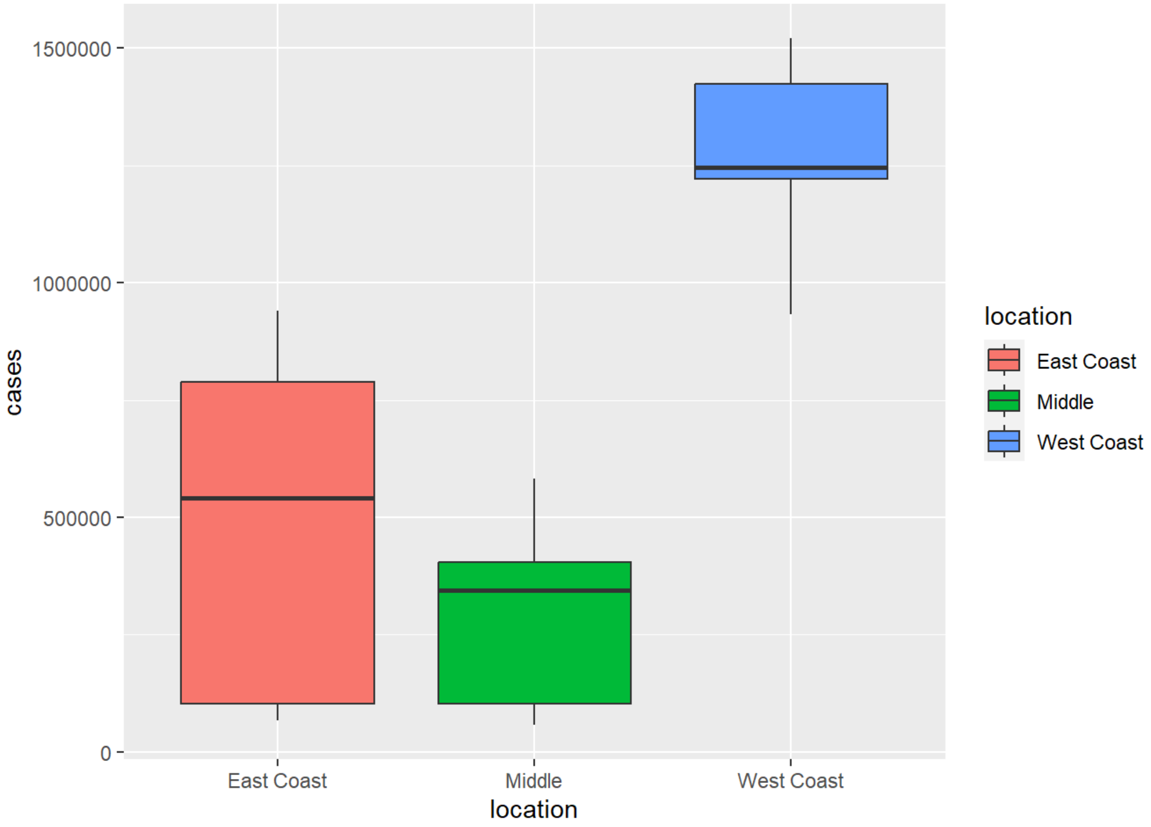
```r
tukey_result <- TukeyHSD(anova_result)
tukey_result
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = temp ~ season, data = US_training_filtered)

$season
                    diff        lwr        upr p adj
Spring-Fall    -7.118363  -9.714449  -4.522278     0
Summer-Fall     9.781535   7.077512  12.485559     0
Winter-Fall   -21.853586 -24.773506 -18.933666     0
Summer-Spring  16.899899  14.410816  19.388981     0
Winter-Spring -14.735222 -17.457302 -12.013142     0
Winter-Summer -31.635121 -34.460330 -28.809912     0
```

Given the p-value from the ANOVA analysis, we can reject the null hypothesis and use the TukeyHSD to see what seasons are causing the difference. The TukeyHSD gives us an adjusted p-value for each compassion and based on these results, each seasons has a significant difference than the others in temperature. The reason why this testing is relevant to our research is so that we can compare the differences between Summer and Winter since these seasons are the extreme differences in temperature which will be used to see if summer weather has a difference in covid transmission compared to winter weather. Given the results from `Winter-Summer`, we can see that there is a difference between Winter and Summer temperature is -31.67 degrees. the negative sign mean that winter is -31.64 lower than summer. This can be a good step when we begin further testing

Another test that we can use the ANOVA and TukeyHSD again is see how different is covid cases compared to the general location of the cities we chose. As mentioned before, we added new variables to the data set to know if the observation is from the east coast, central, or west coast. This can be useful to give us more information on how covid transmission differ across the US so that we can refer back to given the results when we finalize a conclusion.

```r
ggplot(US_training_filtered, aes(x = location, y = cases, fill = location)) +
  geom_boxplot()
```



Given the box plots, it matches all the conditions of the ANOVA tests in order to proceed. The 3 condition test is kinda questionable in some of the points so the results of the ANOVA tests shouldn't be taken too accurately but its good enough to find a pattern.

```
anova_result <- aov(cases ~ location, data = US_training_filtered)
summary(anova_result)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
location      2 1.274e+14 6.368e+13    1003 <2e-16 ***
Residuals   875 5.557e+13 6.351e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tukey_result <- TukeyHSD(anova_result)
tukey_result
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = cases ~ location, data = US_training_filtered)

$location
                          diff       lwr       upr p adj
Middle-East Coast     -175895.1 -221322.8 -130467.4     0
West Coast-East Coast  823505.2  771097.5  875912.9     0
West Coast-Middle      999400.3  944965.3 1053835.3     0
```

Based on these results, we can see that all different location are significantly different from one and other, however if you where to look at the west Cast, there is a more covid transmission cases going on. This model was more used to reference to make sure our cities are experimenting covid differently so that if we do find a correlation between weather and covid transmission, then we know that it was based on cities with wide variability. We originally wanted to test the general location with cases_growth, however it did not satisfy he 3rd condition of the ANOVA analysis at all. Therefor any results from the test would not be very effective.

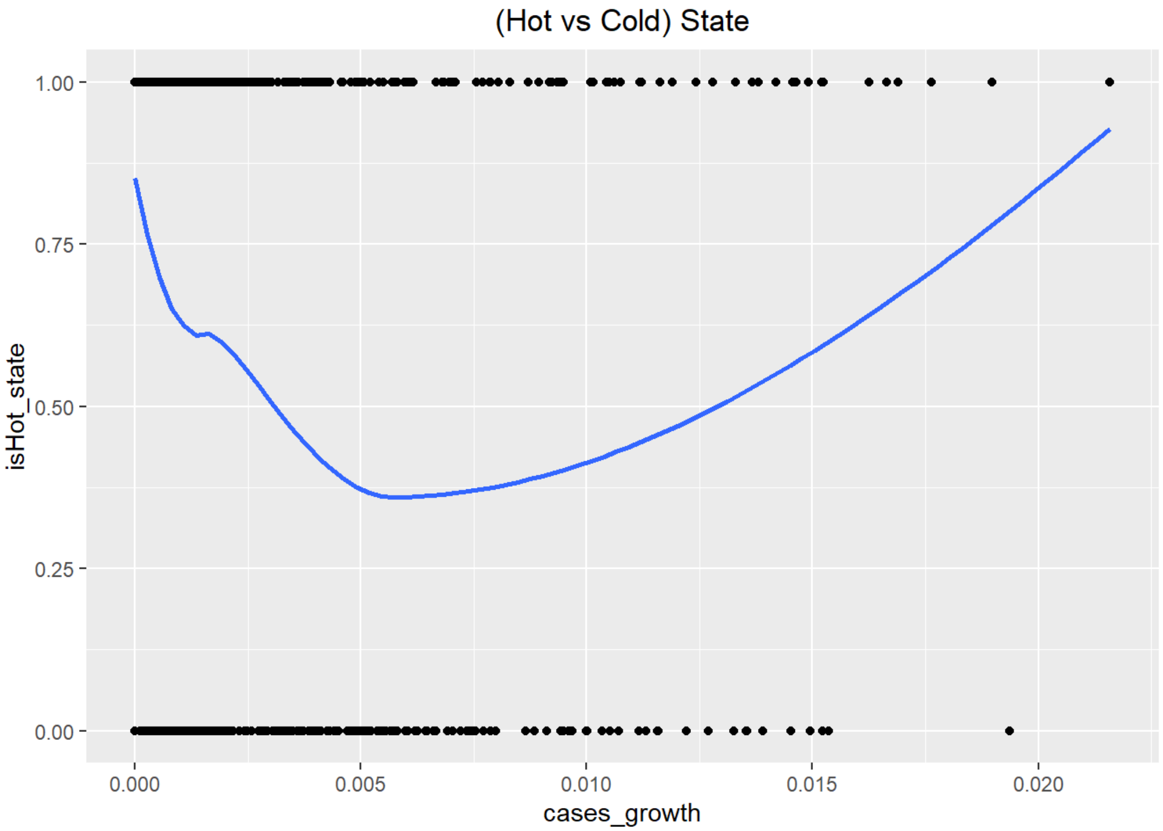
Logistic Regression

Logistic regression is a regression model for response variables in a binary format (0 or 1). In our data set, we generated some instances of binary data with our categorical values with only 2 diffrent values. we also made a rain variable that was gather based on the precipation data. overall, our binaray variable include:

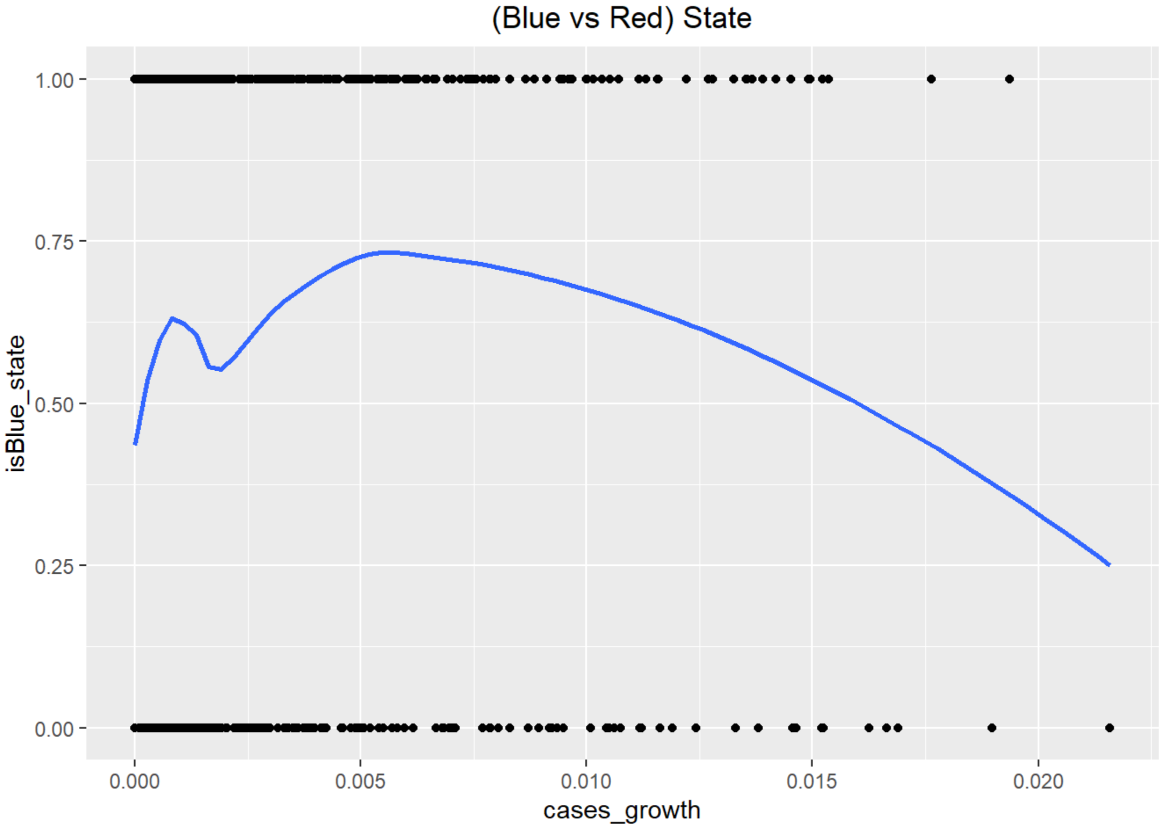
isBlue_state (1 = blue, 0 = red) isHot_state (1 = Hot, 0 = Cold) isRain (1 = it Rained, 0 = it didn't)

Just like most models, there are conditions that must be met before you apply the model to your data. for logistic regression you just have to graph the data and see if you get an inverse cubic type curve (like a stretched out integral symbol). If true, then you can proceed to apply the model and check if you have a p-value less than the significant value. This model can be useful to see if there is significant evidence if rain has more covid growth rates than no rain. If there is significant evidence, than we can make predictions such as computing the odds of a cases growth rate of 2% occurs if it rained.

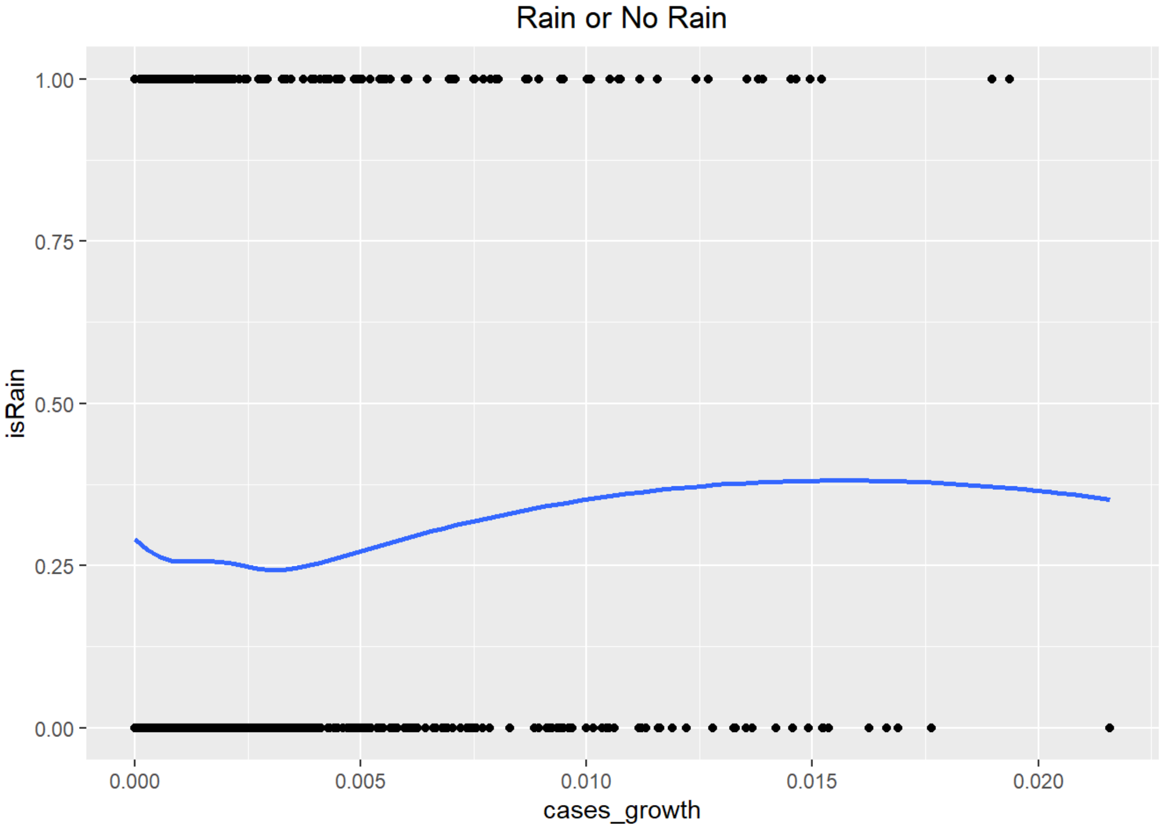
```r
# testing hot and cold states
ggplot(US_training_filtered, aes(x = cases_growth, y = isHot_state)) +
    geom_point() + geom_smooth(se = FALSE) +
    ggtitle("(Hot vs Cold) State") + theme(plot.title = element_text(hjust = 0.5))
```



```r
# testing blue and red states
ggplot(US_training_filtered, aes(x = cases_growth, y = isBlue_state)) +
    geom_point() + geom_smooth(se = FALSE) +
    ggtitle("(Blue vs Red) State") + theme(plot.title = element_text(hjust = 0.5))
```



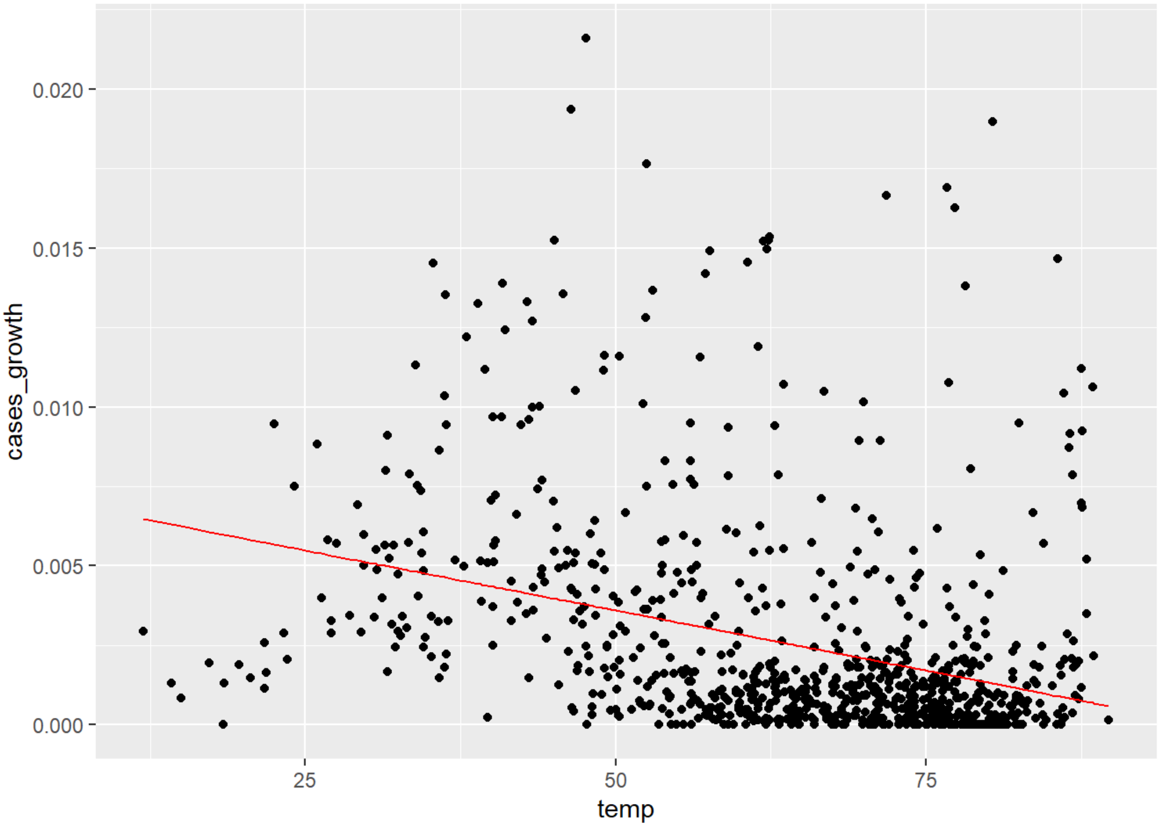
```
#testing if it rained or didn't
ggplot(US_training_filtered, aes(x = cases_growth, y = isRain)) +
    geom_point() + geom_smooth(se = FALSE)  +
    ggtitle("Rain or No Rain") + theme(plot.title = element_text(hjust = 0.5))
```

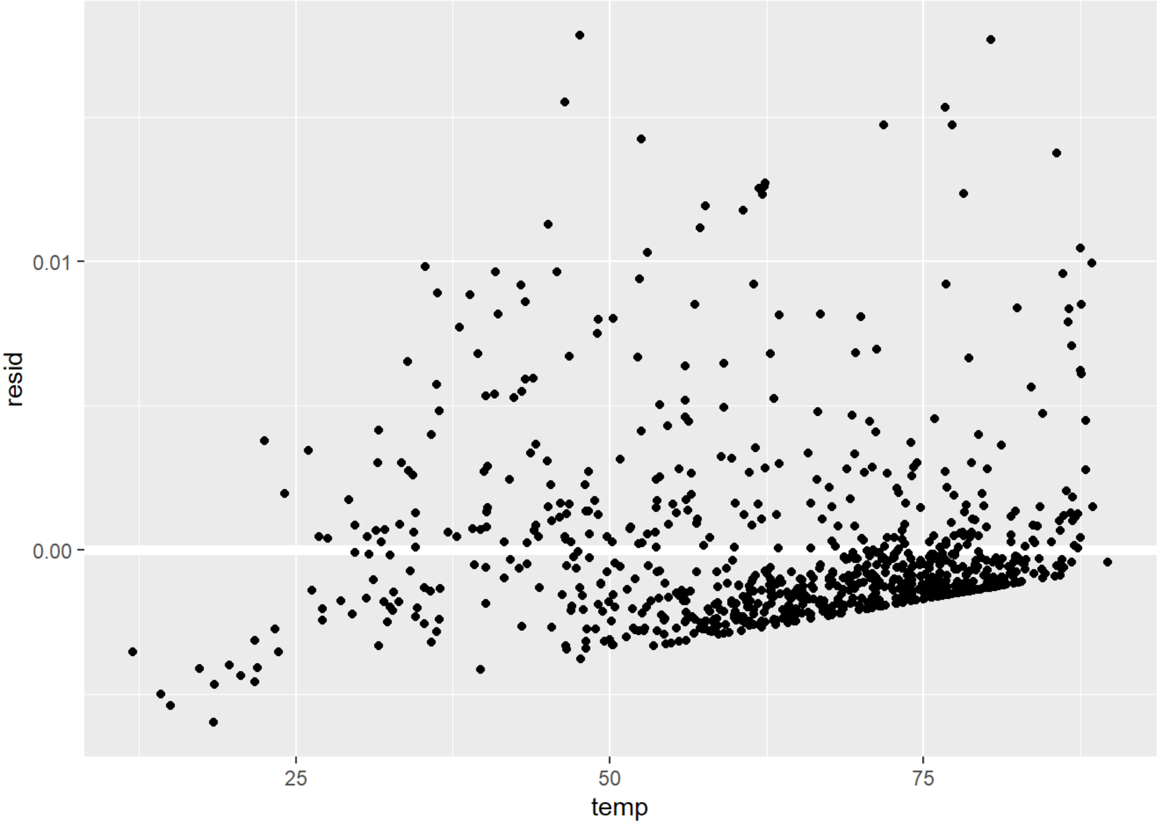


based on the results from the following graphs, the logistic model can not be applied here due to the line shape not be corect. If the model where to be applied, there are instances where it shows that there is a significant correlation between the variables. However, that is incorrect because based on the graphs, it does not have a good fit and in logistic regression, you cant change parameters to adjust anything due to the line being a representation of the plots.

Linear Regresion

Linear regression is where a lot of predictions can be made if a correlation exists. linear regression is a regression model that works if the plots form a linear relationship between 2 variables. It can also hand multiple variables such that we can compare `covid_cases` with all kinds of weather variables. Given the plots, it will do its best to make a prediction line where it could essentially be useful to us by being about to predict the number of cases_growth given a tempture.

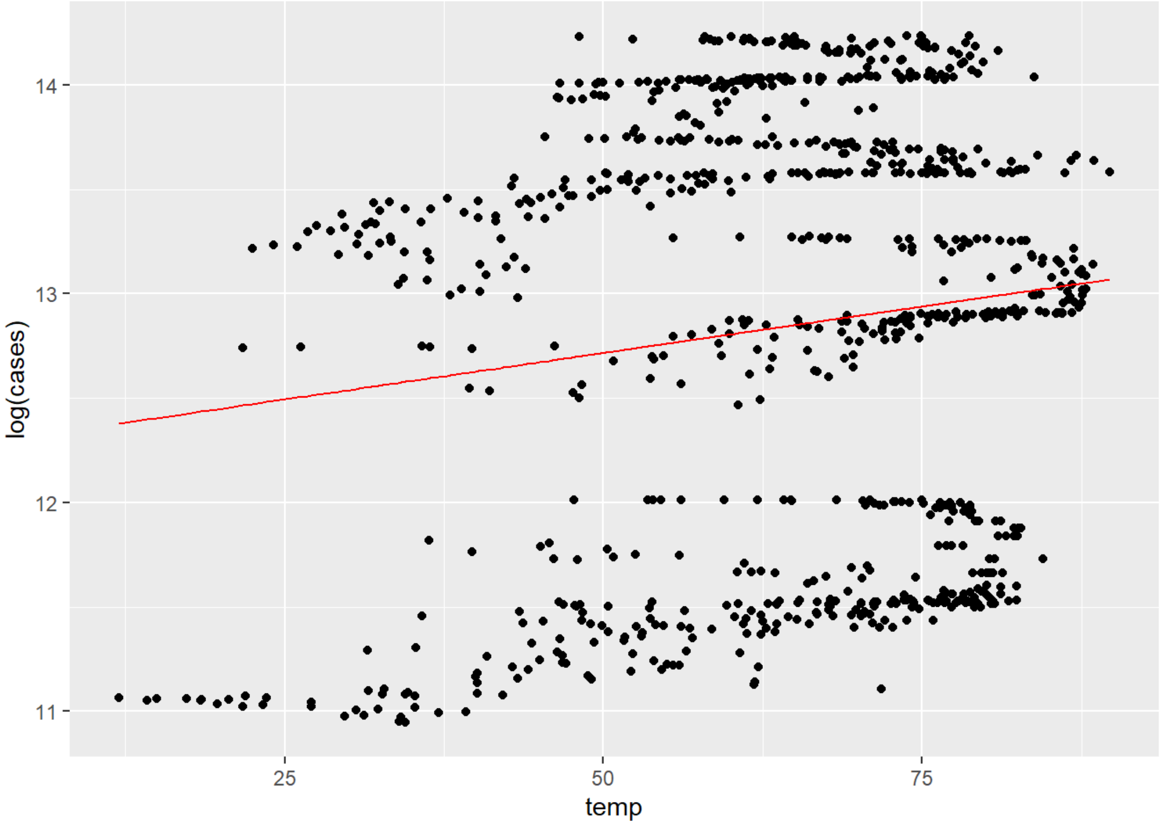




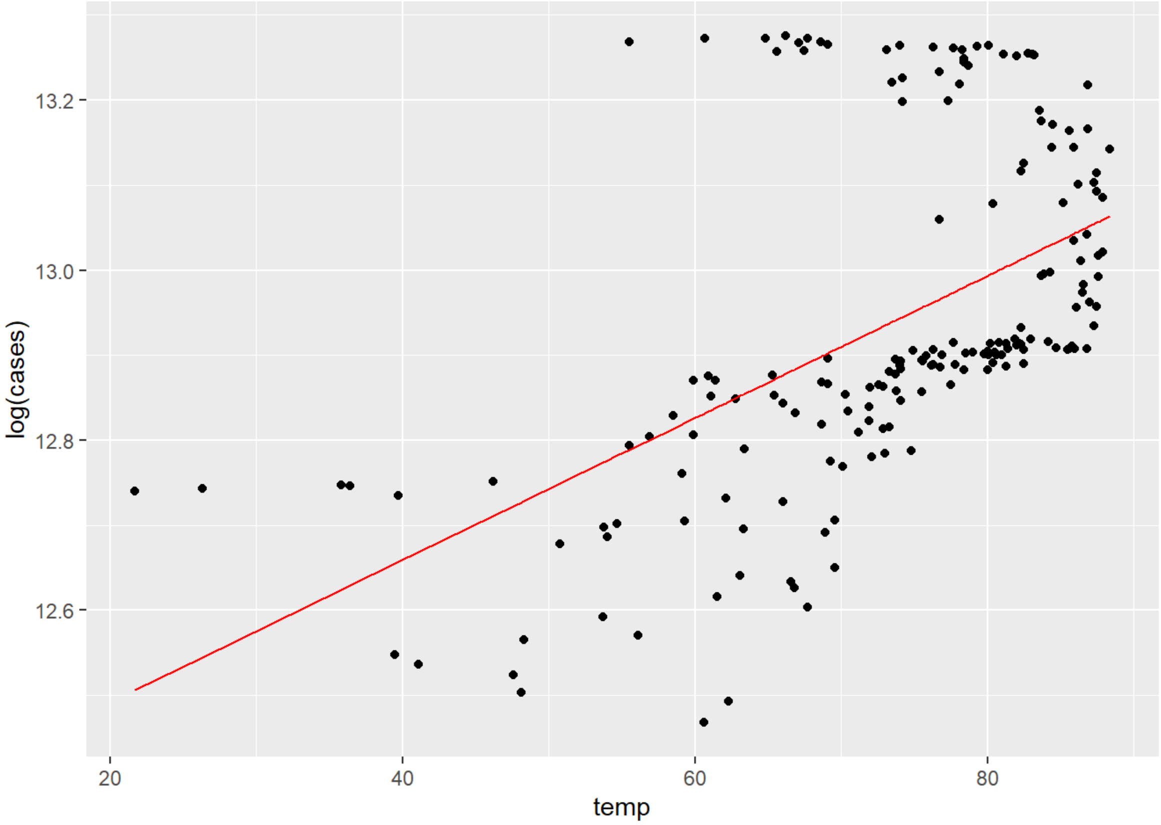
Based on the outcome, there is not a linear relationship with `cases_growth` and `temp`. Even though a prediction line is showing something, It is incorrect because if we plot the residuals, it shows that it is not random. Since the residuals are the difference between each actual and predicted points, it should exhibit some randomness because there should not be any more patterns within the data that the model did not capture. This same method of testing was applied with `cases_growth` to precip, dew, wind, pressure, etc and it all had the overall same outcome with no linear relationship

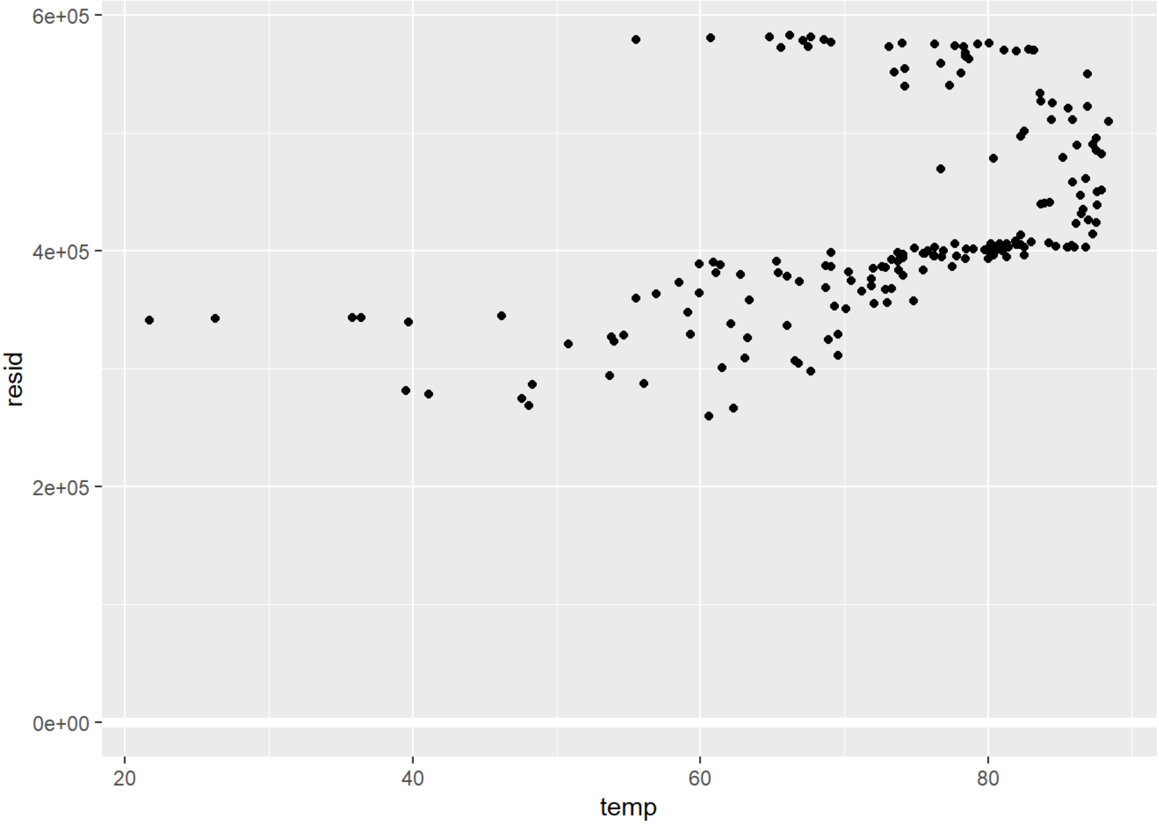
Logarthmic Regresion

Part of the linear regression family, logarithmic regresion is used for data that has a logarithmic relationship. logs are useful as it can help with finding a relationship with skewed data. As mentioned before, there was a relating from PubMed Central where they conducted a simaler analysis on the correlation between covid rates and weather conditions at Italy. They where able to find a correlation by taking the log of the number of cases with weather. This inspired our team to try it out on our data swell as we do have a number of cases column.



Based on the graph generated output, there seems to be a wierd pattern going on. after reviewing our data, I relized that the different cases from the different cities is causing this pattern to occur and thus the data must be from just one location since we need the cases variable to be growing in 1 path. This is true too in the research paper published my PubMed as they took the cases of Italy as a whole and not in different parts of Italy. Therefor, we filtered the data set to just 1 city and applied the logarithmic regression model again.





Based on these results, we where not able to find a correlation between cases and temperature as the article did. The Residuals follow the same pattern as the actual plotted points so there is no randomness going on and thus the model couldn't find anything.

Conclusion

In our extensive exploration, involving the application of ANOVA, TukeyHSD, Logistic Regression, Linear Regression, and Logarithmic Regression, we have discerned that there appears to be a lack of significant correlation between weather conditions and COVID transmission based on the tools that we used. If a connection was found then we could have gone to using the validating set to keep on applying our models for adjustments, then lastly use the testing set to ultimately give a better conclusion weather our models can me used for prediction or not. However since that is not the case, we cant proceed with our current methods. If more time was given, there is many more models and methods that can be further looked into to see if it can be applied to our data in order to find a correlation. Also instead of being very specific with daily observation of weather and covid, we can tone it down by getting the weekly average. Maybe this clears the space and a pattern becomes more visable. Although our testing ended up with a failure, it doesnt mean that there is no correlation between covid and weather. This only means that there wasn't any correlation with our methods and thus there is still more to investigate.

Sources

(N.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9387066/

US Department of Commerce, N. (2022, March 3). Climate. https://www.weather.gov/wrh/Climate?wfo=lox

Literature review of the effect of temperature and humidity on viruses. (n.d.-a). https://orf.od.nih.gov/TechnicalResources/Bioenvironmental/Documents/FINALPUBLISHEDPaperonHUMIDITYandViruses509.pdf

University of Manchester. (2021, December 10). New study shows link between weather and spread of covid-19. https://www.manchester.ac.uk/discover/news/new-study-shows-link-between-weather-and-spread-of-covid-19/

Malki, Z., Atlam, E.-S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020, September). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, solitons, and fractals. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7367008/

Find open datasets and Machine Learning Projects. Kaggle. (n.d.). https://www.kaggle.com/datasets

National Weather Service https://www.weather.gov/mfl/