1. Abstract and Introduction (2%)

For this final project, I chose COVID-19 as my project topic. Because COVID-19 has been one of the most influence events that not only for a country, but also for the whole world. It is not only affecting people’s health and lives, but also affecting the world’s social activities and economy.

Because COVID-19 can be considered as a new event, there haven’t exist formal and accurate datasets that record worldwide data for the COVID-19. Therefore, I chose the COVID-19 datasets that recorded data of the New York State. One of the most important datasets I chose is the New York State Statewide COVID-19 published by health.data.ny.gov, which records daily total tests performed, daily new positive, cumulative tests performed, and cumulative positive cases of each County from March to October. With these data, I think maybe there are some relations between them. For example, it seems obvious that more tests performed will lead to more positive cases. However, this is just assumption that need to be tested to confirm. Since I have data of each County, I think maybe there are some other Counties’ attributes that affect the cumulative positive cases. Therefore, after searching, I also found datasets of each County’s population, work from home population, age over 65 population, as well as total households with income between $75,000 and $99,999.

I selected attributes `population of age over 65` and `total households with income between $75,000 and $99,999` specificity because I assumed that older people maybe are more easily to be infected, and people with higher household with income are more willing to get test and not be infected. Based on these assumptions, I chose these two attributes to help prove my assumption.

1. Data Description and Exploratory Data Analytics (3%)

Since I want to explore attributes that affect the cumulative positive cases, I have chosen five datasetes. The first and main one is “New\_York\_Statewide\_COVID-19\_Testing.csv” (TESTING), which includes features date, county, new tests performed, cumulative tests performed, new positive cases, and cumulative positive cases.

The second dataset is “us-counties.csv” (CASE\_DEATH), which contains features date, county, cases, and deaths. I chose this dataset because it could be merged with the first dataset by the same county or same date, so that I can combine the ‘deaths’ features with features in the “New\_York\_Statewide\_COVID-19\_Testing.csv” dataset.

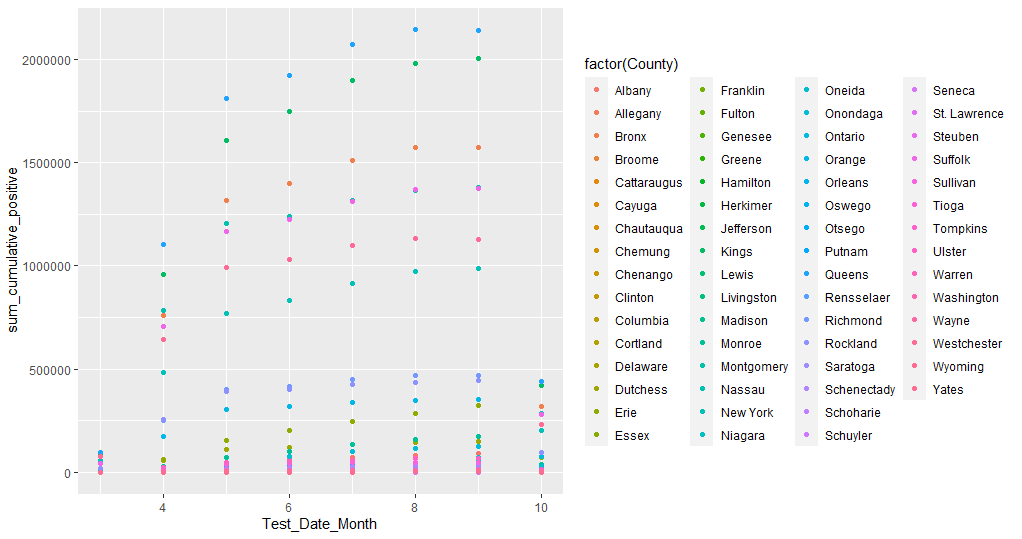
The third dataset I chose is “Population\_by\_Age\_and\_Sex\_-\_Counties.csv” (POPULATION\_BY\_AGE\_SEX), which contains features of populations of different ages and sexes of different counties. Since this dataset includes all counties of all States in the United State. I chose features total population, total population 65 years and over, total population male 65 and over, and total population female 65 and over. I specifically chose the age range 65 years and over because I wanted to test if county with higher old people’s population proportion will also have a high positive test proportion.

The fourth dataset I chose is “Income\_and\_Benefits\_-\_Counties.csv” (INCOME\_AND\_BENEFITS), which contains features of populations of different ranges of income with household of different counties. For this dataset, I chose features county name, total households, and total household with income from 750000 to 99999. I chose this range specifically because I wanted to see if counties with higher income range proportion will have a lower positive cases rate since maybe rich people will be more careful and have more protections for their health, which lead to a high tests performance rate and low positive cases rate.

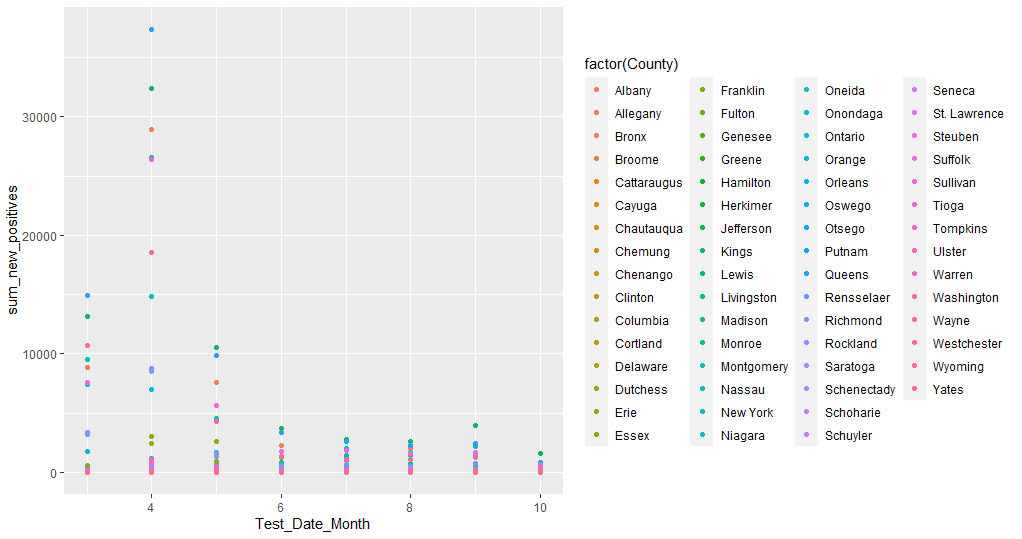
The last dataset I chose is “Worked\_at\_Home\_-\_Counties.csv” (WORK\_AT\_HOME), which contains work from home populations of different counties. I chose features county name and total population worked at home. I chose this dataset because I think there may be relationship between the number of positive cases and total population worked at home. For example, higher work at home rate may lead to lower positive cases rate.

1. Analysis (5%)

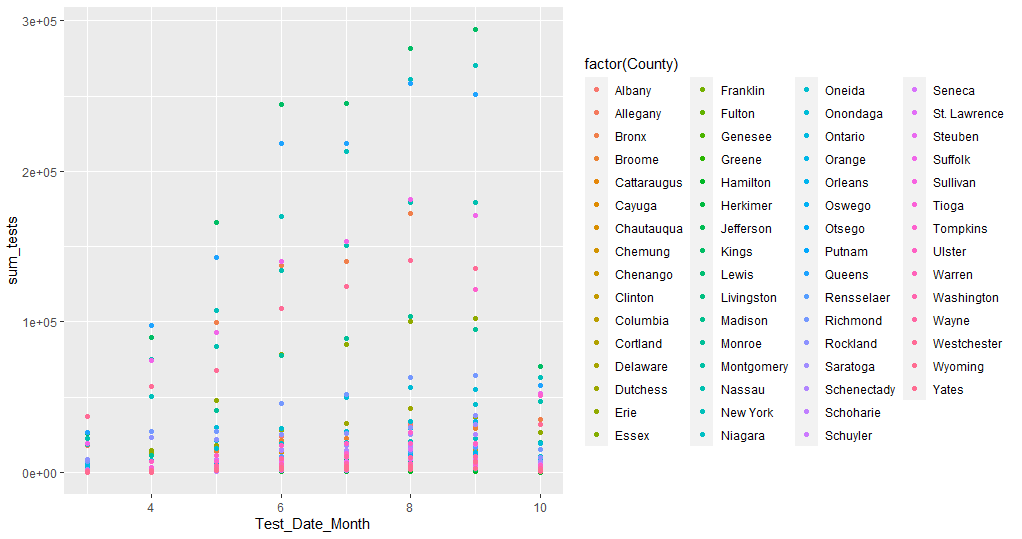
Because I have five datasets, so my first step is loading them into R and filtering out as well as extracting features that I am going to use to combine into a new dataset, which I am going to use for modeling. After renaming columns and extracting features that I want from the datasets, I changed the types of dates in both TESTING and CASE\_DEATH from char to date. I want to apply two models which are regression model and classification model. There are two hypothesis that I want to test with regression model. Although I first want to set the counties’ death number as the dependent variable and test if there exists regression relationship between it and other features, after using the summary() method and plotting, I found out that there are lots of zero death case among counties in different months, which will not allow it to be tested as a dependent variable. Therefore, instead of the death number, I decided to use the cumulative positive cases number as my regression models’ dependent variable. The first one is there exists relationship between the monthly cumulative positive cases and monthly tests performed among all counties in the New York State. For this hypothesis, I merged the TESTING and CASE\_DEATH datasets based on the same counties and dates, then I grouped the dataset by month which created a new dataset contains information of each county of each month. Below is a graph contains four point plots with months as x axis and four independent variables as y axis and different colors as different counties. From the graphs, they show that the cumulative positive cases seems to has the same pattern for each county as the daily sum tests performed and the cumulative tests performed.



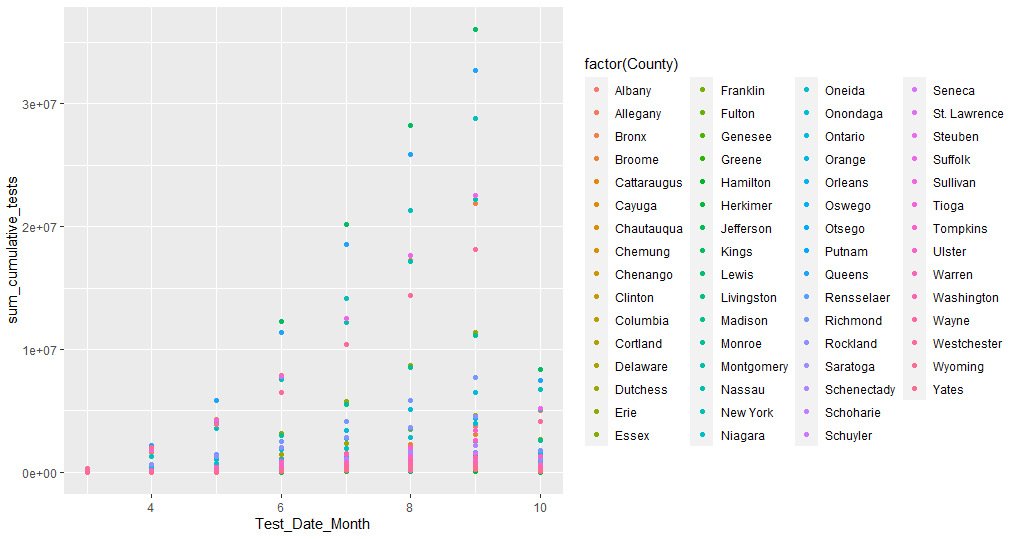
**Figure 3.1.1** *point plot of Test\_Date\_Month and sum\_cumulative\_positives based on Countie*s



**Figure 3.1.2** *point plot of Test\_Date\_Month and sum\_new\_positives based on Countie*s

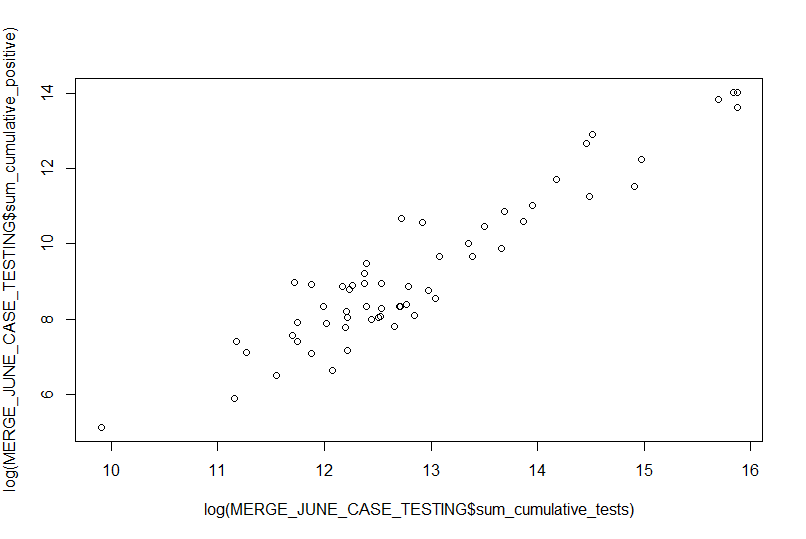


**Figure 3.1.3** *point plot of Test\_Date\_Month and sum\_tests based on Countie*s

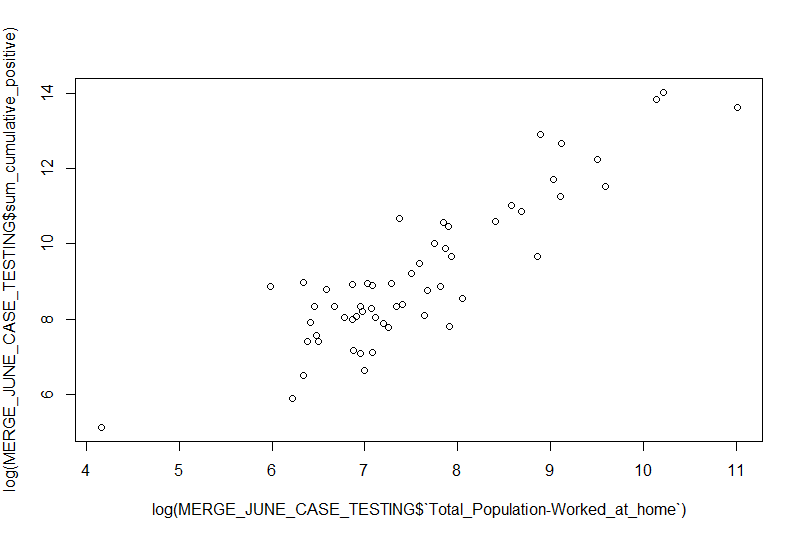


**Figure 3.1.4** *point plot of Test\_Date\_Month and sum\_cumulative\_tests based on Countie*s

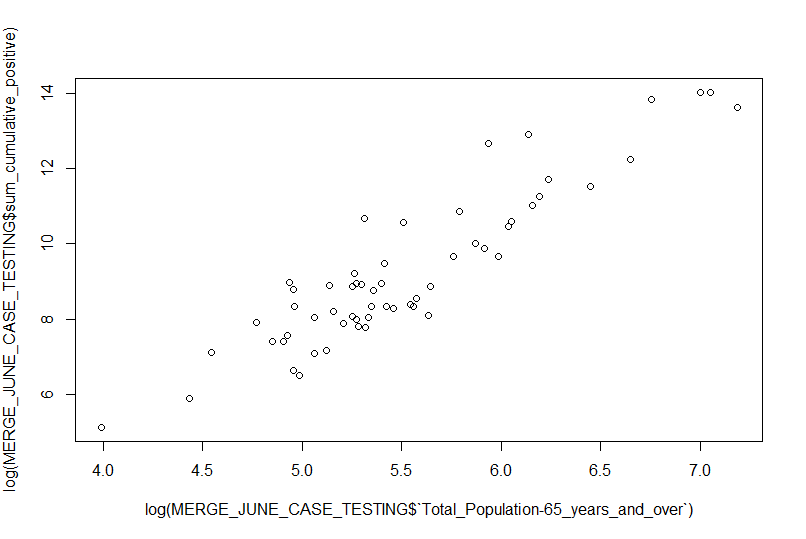
Because datasets POPULATION\_BY\_AGE\_SEX, INCOME\_AND\_BENEFITS, WORK\_AT\_HOME are updated in June 2020. Therefore, the second hypothesis for regression model is that there exists relationships between the cumulative positive cases rate in June for each county, work for home proportion, specific age range proportion, and specific income range proportion. For this hypothesis, I merged features that I want from the five datasets into a new dataset. Below I plotted three graphs between the dependent variable, cumulative positive cases number, and independent variables, cumulative tests number, total work at home population and total population of 65 years and over. For these graphs, I used logarithmic scales respond to skewness towards large values.



**Figure 3.2.1** *plot of sum\_cumulative\_tests and sum\_cumulative\_positives*



**Figure 3.2.2** *plot of Total\_Population-Worked\_at\_home and sum\_cumulative\_positives*

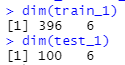


**Figure 3.2.3** *plot of Total\_Population-65\_years\_and\_over and sum\_cumulative\_positives*

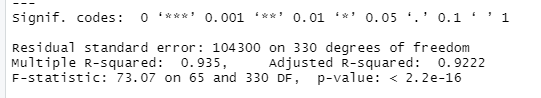
From the graphs above, they show that there exist trending pattern because the dependent variable, sum\_cumulative\_positive, and all the other independent variables.

1. Model Development and Application of model(s) (12%)

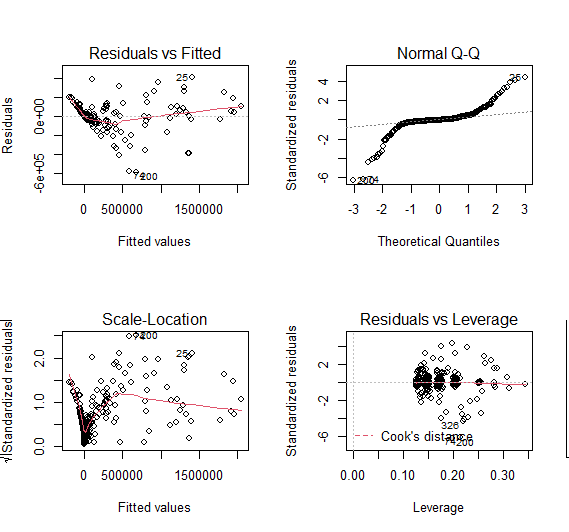
From part 3 above, I assumed that there exists relationship between the monthly cumulative positive cases and monthly tests performed among all counties in the New York State. For this hypothesis, I decided to use the Linear Regression model. Therefore, after I merged the TESTING and CASE\_DEATH datasets based on the same counties and dates, and grouped the dataset by month, I changed the type of County feature from char to factor, which then it can be used in the regression model. To train and test for the Linear Regression model, I used sample() method to separate the new ALL\_MONTH\_TESTING\_GROUP dataset into two datasets, with 80 percent rows in the train dataset, and 20 percent rows in the test dataset. Below is the screenshot of the dimensions of the train dataset and test dataset.



I then used the train dataset to build the first linear regression model with function lm() and with sum\_cumulative\_positive as dependent variable, and all the other features as independent variables.



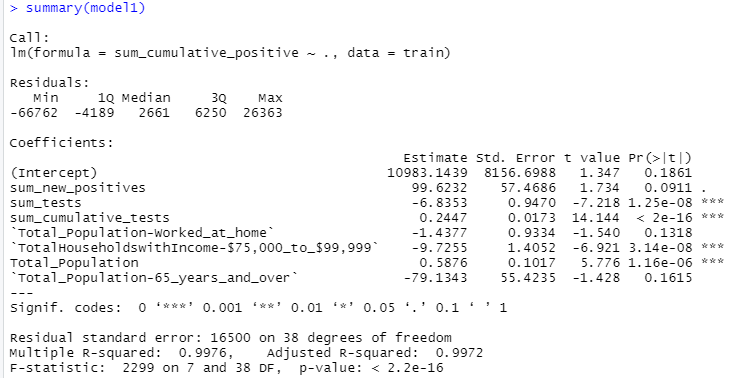
From the output of the summary of this regression model, it shows that it has F-stat value equals to 73.07 which is far greater than 1. Therefore, there is a relationship between the dependent variable and independent variables. The Adjusted R-squared is equal to 0.9222 which shows that this model fit.



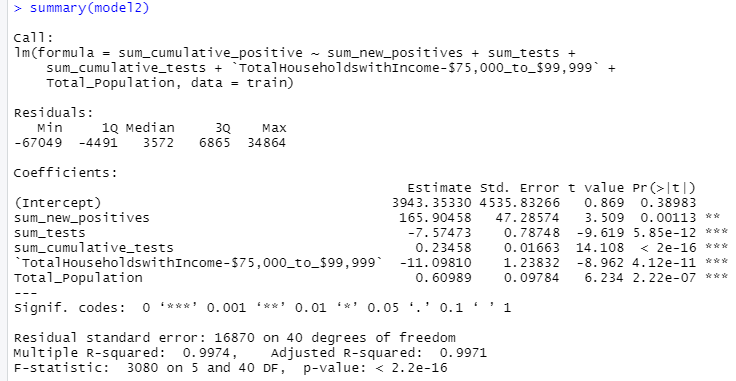
**Figure 4.1.1** *plot of the first Linear Regression model*

The above graph is the plot of the linear regression model. The top left Residuals vs. Fitted plot shows that there is no distinctive pattern of the model which means there isn’t exist any non-linear relationships between the variables. The right top Normal Q-Q plot shows that most of the residuals follow a straight line well but not all of them. The bottom left Scale-Location plot shows that most of the residuals are spread equally along the ranges of predictors. The bottom right Residuals vs. Leverage plot shows that there is no influential case or cases since I can barely see Cook’s distance line.

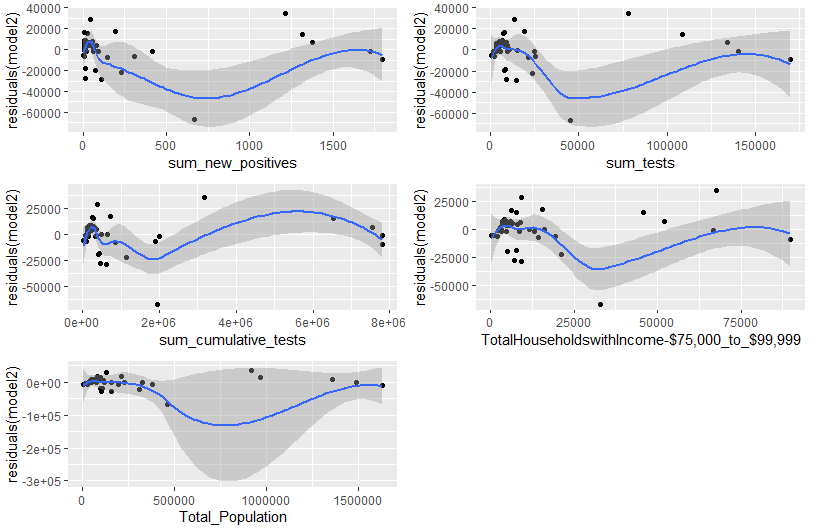
The second hypothesis for regression model is that there exists relationships between the cumulative positive cases rate in June for each county, work for home proportion, specific age range proportion, and specific income range proportion. After I split the dataset into two datasets with 80 percent data in the train dataset, and 20 percent data in the test dataset, I used the lm() method to make default model with sum\_cumulative\_positive as the dependent variable, and all the other features as independent variables.



From the summary above it shows that the F-statistic is 2299 for this default model which means the relationship exists between these variables. The Adjusted R-squared equals to 0.9972 which means the model fit. However, by looking at the p-value for each independent variable I found out that not all of them have significant value less than 0.05. Therefore, I decided to update the first model with removing variables Total\_Population-Worked\_at\_home and Total\_Population-65\_years\_and\_over, and see if the new model2 will be better than the model1.

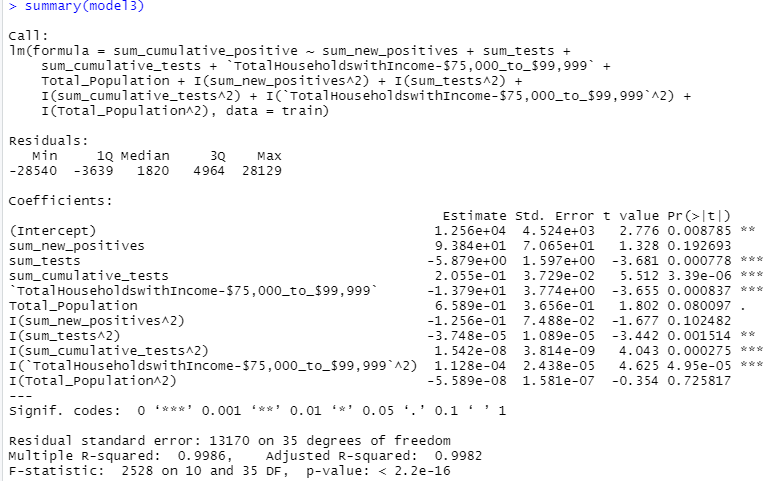


From the summary above of model2, it shows that even now all the independent variables have significant p-values, the model did not have a great improve. To try to improve the model, I plotted the residual plot with all significant features from model2 and wanted to see there exist any nonlinear pattern.



**Figure 4.1.2** *plot of residuals of five independent variables*

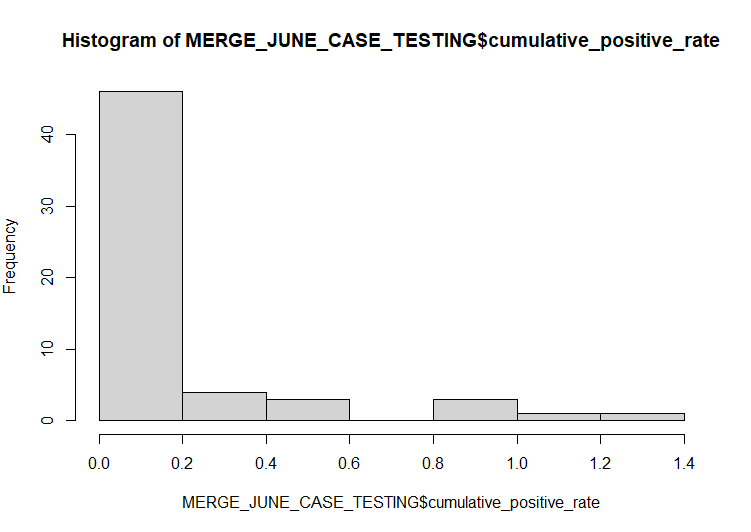
From the residual plots above, they shows that there exist some non-linear pattern for all these features. Therefore, to try to enhance the model, I added a square term to check for non-linearity and build the model3.



By comparing the model3’s Adjusted R-squared and the other two models, which model3’s Adjusted R-squared is higher than the other two, therefore, it seems like the model got improved a little bit. To further optimize it, I removed the insignificant variables from model3 and created the model4.

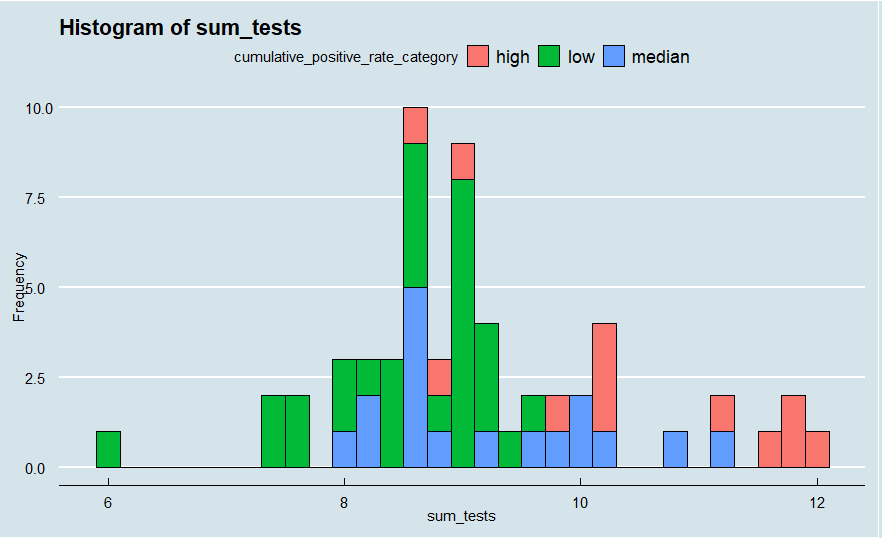
The second model I tried to build is Decision Tree Classifier model. I created a new feature cumulative\_positive\_rate which is the cumulative positive cases rate for each County in June.

1606627810(1)

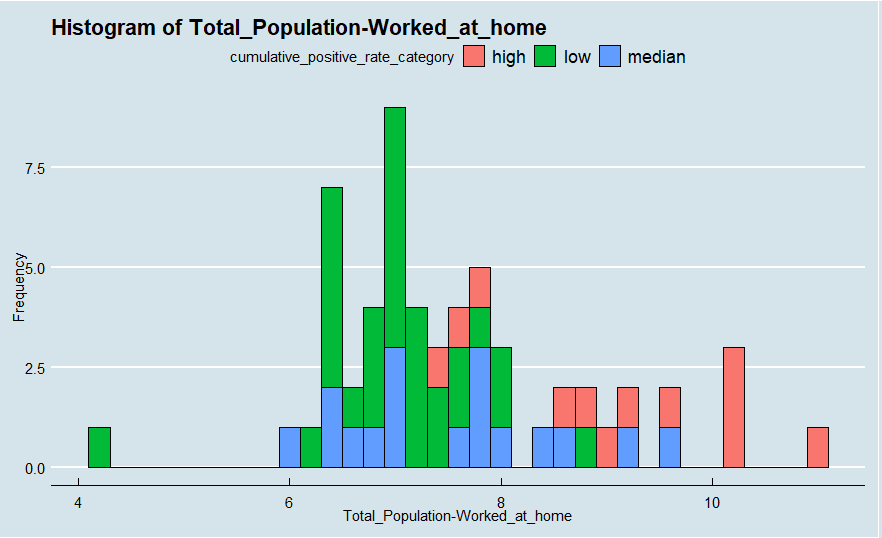


**Figure 4.2.1** *histogram plot of cumulative\_positive\_rate*

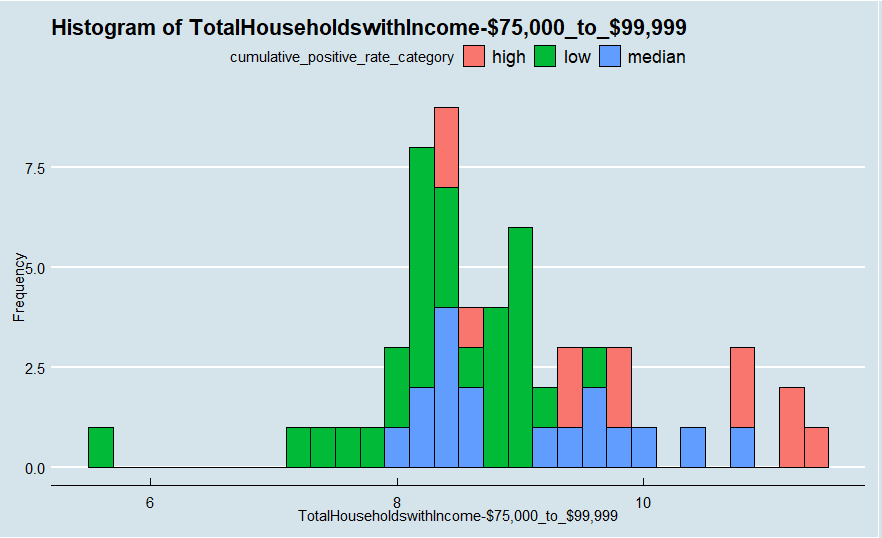
By checking with the summary and hist plot of this feature, I created another new feature cumulative\_positive\_rate\_category with cumulative\_positive\_rate less than 0.0820 as low, cumulative\_positive\_rate higher than 0.2000 as high, and the between as median. I then created a new classification dataset with 9 independent features, and this new factor type cumulative\_positive\_rate\_category feature. I separated this dataset into two datasets with 80 percent data as train dataset and 20 percent data as test dataset. Below is three histogram graphs between distributions of three important features and the cumulative\_positive\_rate\_category.



**Figure 4.2.2** *histogram plot of sum\_tests and cumulative\_positive\_rate\_category*

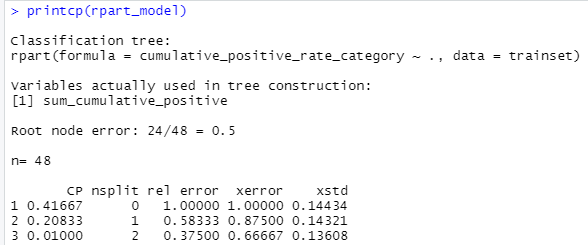


**Figure 4.2.3** *histogram plot of Total\_Population-Worked\_at\_home and cumulative\_positive\_rate\_category*

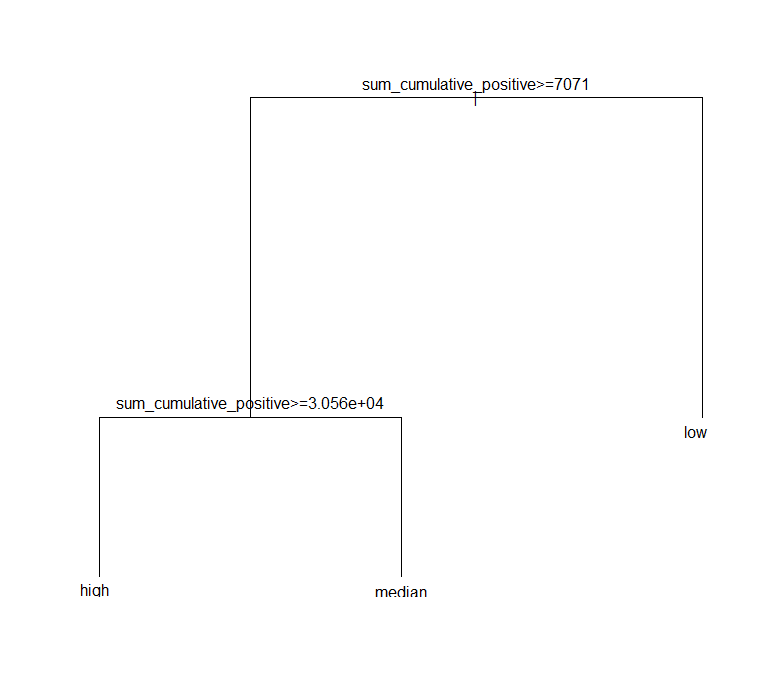


**Figure 4.2.4** *histogram plot of TotalHouseholdswithIncom-$75,000\_to\_$99,999 and cumulative\_positive\_rate\_category*

I used the rpart() function to build the decision tree model with cumulative\_positive\_rate\_category as the dependent variable and the other features as independent variables. I then used the printcp() method to check more details of this model.



The cp is the complexity parameter where the larger the cp is, the smaller the tree’s nsplit is. From the above output, it shows that when cp=0.01000 and nsplit=2, the rel error, xerror, and xstd are the smallest. Therefore, the tree looks like the graph below:



**Figure 4.2.5** graph of the Decision Tree Classifier model

The third model I built is the KNN classifier dataset. Because the features have different scale, and if the data is not normalized, it will lead to biased outcome, therefore, I used a normalize function to normalize all independent features before applying the model. After normalizing, I split the dataset into two datasets again with 80 percent data as the train dataset and 20 percent data as the test dataset. Because there are 48 rows in the train dataset, and the square root of 48 is around 6.928. Therefore, I created two models, knn\_6 with ‘k’ value equals to 6 and knn\_7 with ‘k’ value equals to 7. I will evaluate them using the confusionMatrix() method.

1. Conclusions and Discussion (3%) Describe your conclusions

For the first Linear Regression model with hypothesis that there exists relationship between the monthly cumulative positive cases and monthly tests performed among all counties in the New York State, I used the model to predict on the test dataset and evaluate the result using RMSE and R2.

1606614639(1)

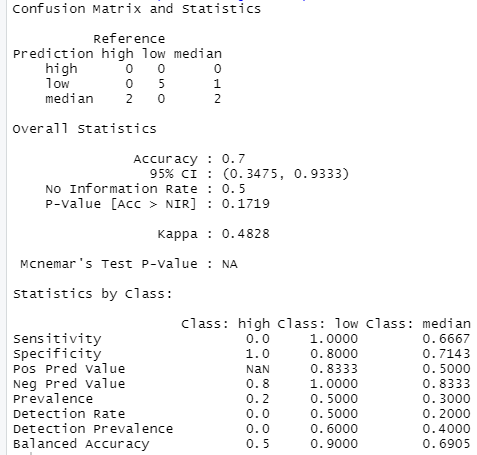
From the output above it shows that the RMSE is a very large number which indicates that the predicted value will be off a lot from the actual value. This means the model may not work too well.

For the second Linear Regression model with hypothesis there exists relationships between the cumulative positive cases rate in June for each county, work for home proportion, specific age range proportion, and specific income range proportion, I used the final model4 to predict on the test dataset and used RMSE and R2 to evaluate the result.

1606625628(1)

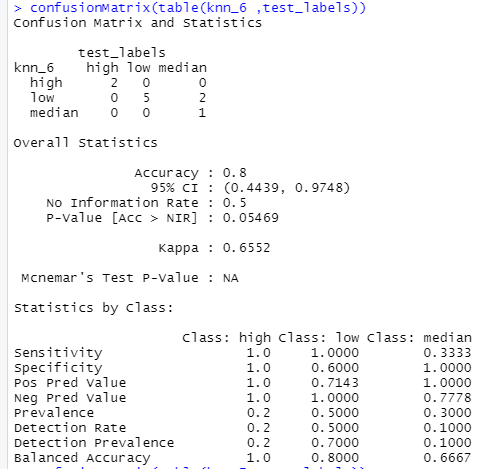
Even though the model has a high R2, the RMSE is still really high that means the predict values will off a lot from the actual values.

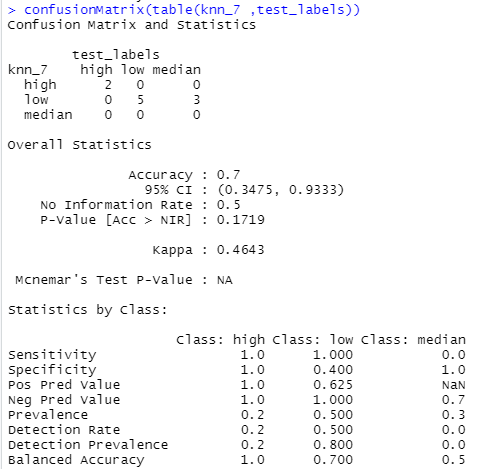
I used the Decision Tree Classifier model to predict on the test dataset with type=”class”. I then used the confusionMatrix() method to first evaluate the result.



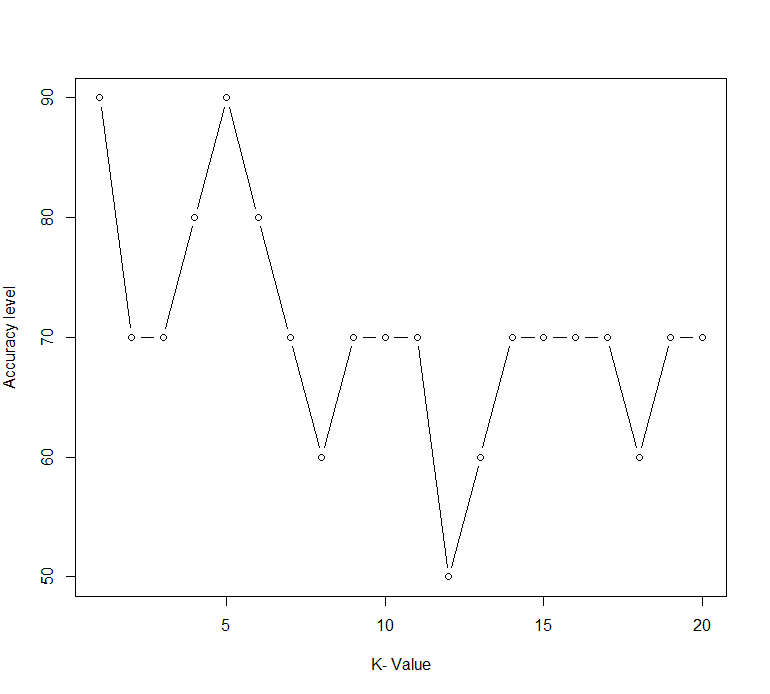
From the report above it shows that the prediction has 0.7 accuracy. However, since there are not too much training and testing data, the model may have a much lower accuracy if it applies to a larger dataset.

For the two KNN Classifier models knn\_6 and knn\_7, I also used confusionMatrix to evaluate the result after I used them to predict on the test dataset.



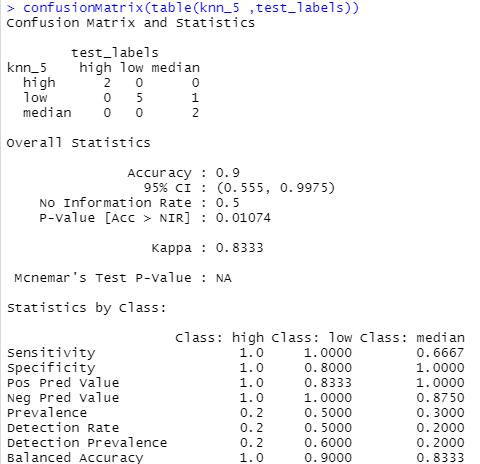


From the two reports above, they show that model with ‘k’ equals to 6 has a higher accuracy than the model with ‘k’ value equals to 7. To check which ‘k’ value is the best one, I created a loop that calculated the accuracy of the KNN model for ‘k’ values ranging from 1 to 20. From the plot below it shows that when ‘k’ value equals to 5, the model has the highest accuracy.



**Figure 5.1.1** *plot of accuracy with different k-value*

Therefore, I built another model with ‘k’ value equals to 5, which has a accuracy equals to 0.9, and it’s also higher than the Decision Tree model.



From all these work, it can be concluded that there does exist some relationship between the monthly cumulative positive cases and monthly tests performed among all counties in the New York State, and also relationship between the cumulative positive cases rate in June for each county, work for home proportion, specific age range proportion, and specific income range proportion. Moreover, the cumulative\_positive\_rate can be grouped into different levels based on these aspects.

Reference

New York State Statewide COVID-19 Testing：

<https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e>

The New York Times Coronavirus (Covid-19) Cases and Deaths in the United States：

<https://data.humdata.org/dataset/nyt-covid-19-data?force_layout=desktop>

Population by Age and Sex - Counties：In June

<https://covid19.census.gov/datasets/population-by-age-and-sex-counties>

#### Income and Benefits - Counties：In June

<https://covid19.census.gov/datasets/income-and-benefits-counties>

#### Worked at Home - Counties：In June

<https://covid19.census.gov/datasets/worked-at-home-counties/data>

https://datascienceplus.com/how-to-apply-linear-regression-in-r/

<https://data.library.virginia.edu/diagnostic-plots/>

GitHub URL:

https://github.com/HelenWu11/DataAnalytics2020\_RuoyanWu\_FinalProject