**The Effects of Airport Traffic on the Transmission of COVID19**

Pelumi Ajayi, Heather Seo, Qiyu Wang

**Introduction**

The global pandemic of COVID-19 abruptly transformed our lives to an unprecedented degree and created complex challenges worldwide. It has become ever more important for governments to implement effective policies to curb the spread of the disease. As different countries adopt different border control restrictions, we wanted to examine how such policy implementation has impacted the spread of the disease. Thus, we want to study how the international connections through airlines correlated with the pandemic and therefore come up with policy recommendations for the executive branches of the governments, for them to choose policies based on facts and rational thinking rather than political ideologies that have divided the nation.

**Data Processing**

We found that the airport traffic and airline route dataset includes 15 U.S states (California, Colorado, Florida, Georgia, Hawaii, Illinois, Massachusetts, Michigan, North Carolina, New Jersey, Nevada, New York, Texas, Virginia, Washington) and six Canadian states (Alberta, British Columbia, Manitoba, Nova Scotia, Ontario, Quebec). In addition to investigating how the airport traffic may affect the percent change of covid19 cases, we wanted to consider other microlevel influencers.

To consider economic and medical factors that may influence the confirmed cases in each state, at state level, we found data on on the number of COVID-19 daily confirmed cases, population, population density (population per km^2), sex ratio, median age, GDP, unemployment rate, life expectancy, Gini Index, number of ICU beds, number of hospitcals and efficiency measures, lockdown policy implementation date, degree of lockdown and number of 2019 visitors, which comes from 14 different sources.

Considering the size of the air traffic data being huge, we utilized Spark and HDFS to speed up our calculations related to airline traffic, airline routes, centrality scores (Including indegree, ourdegree, closeness and pagerank). Also, we extracted data only pertaining to states of interest, calculated moving average, summed rows if data is at finer granularity, merged large datasets, and applied z-score standardization for numeric columns. In the process, Spark achieves a great improvement in computational efficiency and also offers us capacity to further expand our study into larger scales in the future to include more countries and periods of time.

**Challenges and Limitations**

One biggest challenge is the quality of available data. As we perform analysis on time-series measure, over half of our predictors are static because of the lack of availability. Also, the limitations we encountered with our dataset was due to the size of the sample. One of the key dataset only included 23 cities, acting as a bottleneck, so we could not scale up a larger dataset for machine learning. As a result, we did not expect our regression models to produce an extremely good prediction on the macro-level but more of some reasonable statistical insights on the individual counties at a given period to address our problem.

**Insights and Analytics**

1. Model free insights from data

To understand the overall trend of the data we created scatterplots and line plots to assess if there were any synchronized changes between percent of operating airplanes and the number of covid cases. In general, we observed a positive correlation between the patterns of line graphs. For example, in Florida, New York, and Ontario, we see as the percentage of air traffic operating goes high larger percentage increases in confirmed cases. This means that more traffic is related to more number of covid cases.

1. Model driven insights (results) from data

Based on the linear regressions and ANOVA reports, we clearly see that the air traffic is a statistically significant measure in almost all states in terms of predicting rate of change in confirmed cases. The R-squared scores and coefficients are significant enough to have more attention from the policy makers. Using California as an example, every percentage increase in the percentage of operating fights will cause an increase of 0.3%. Considering the population base of California of 39.5 millions, each percent increase of operating fights could possibly cause 118500 confirmed cases. And as the diffusion of infection is proportionally changing, it might snowball an even larger impact later on.

**Reasons for Selection of a Specific approach**

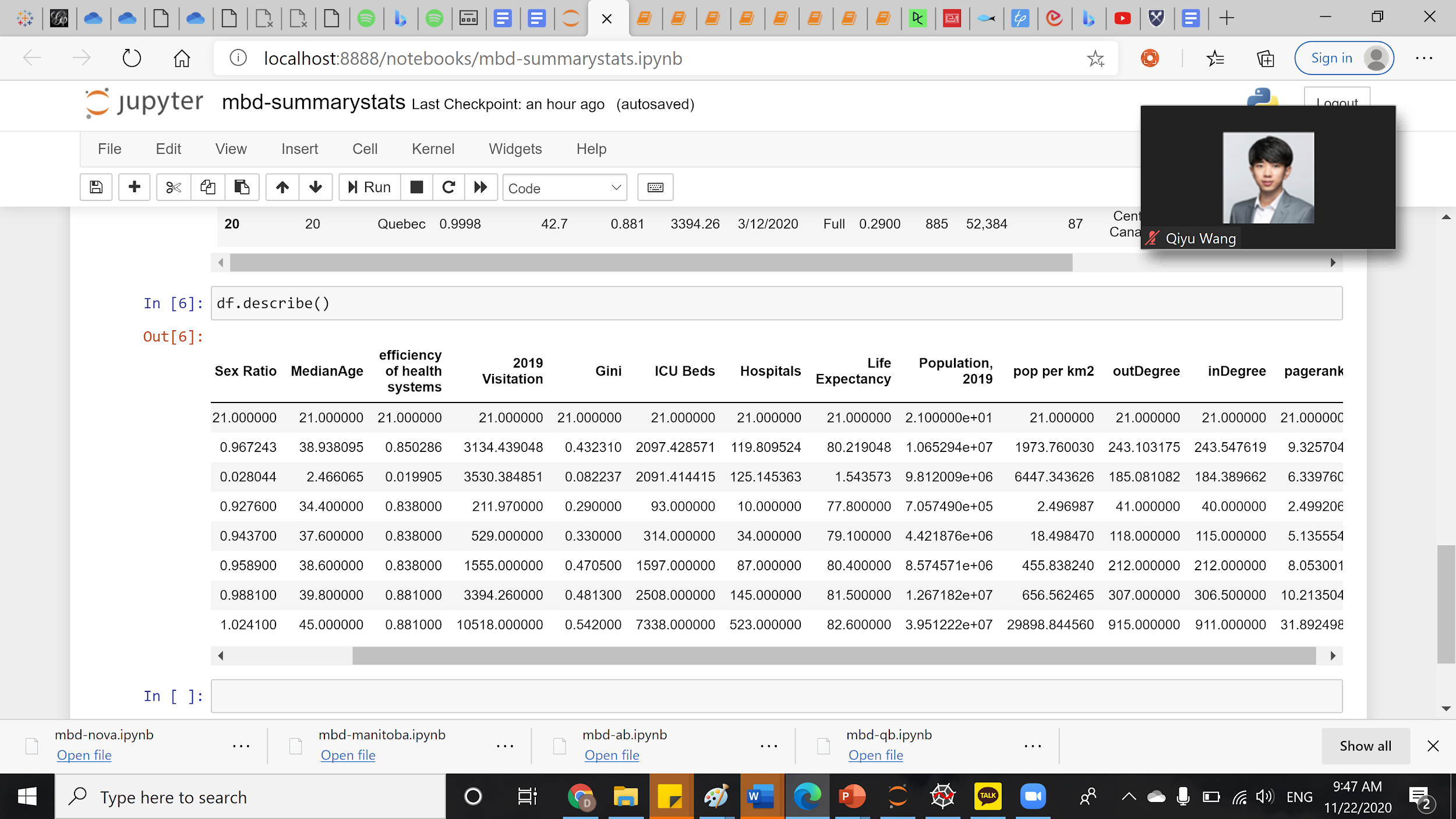
We generated line graphs and ANOVA analysis to examine the statistical significance of the predictors and the correlations between traffic and covid cases. We decided to study only the first period of pedantic (3/16 - 6/26) to simplify the problem because as we include more timeline the problem can be more complex to analyze and more sensitive to things that’s hard to be measured quantitatively such as local governing policies. Also, the problems are studied on two different levels. We want to study individual counties over the period to take a closer look at the impact of air traffics on pedemanic. And then we look at a larger picture which includes more states/provinces and more variables to see the feasibility to predict the vulnerability of a country to pandemic based on the data we obtained. Finally, for this study, we measure vulnerability of a country to pandemic based on the rate of change in infected cases.

**Discussion and Recommendation**

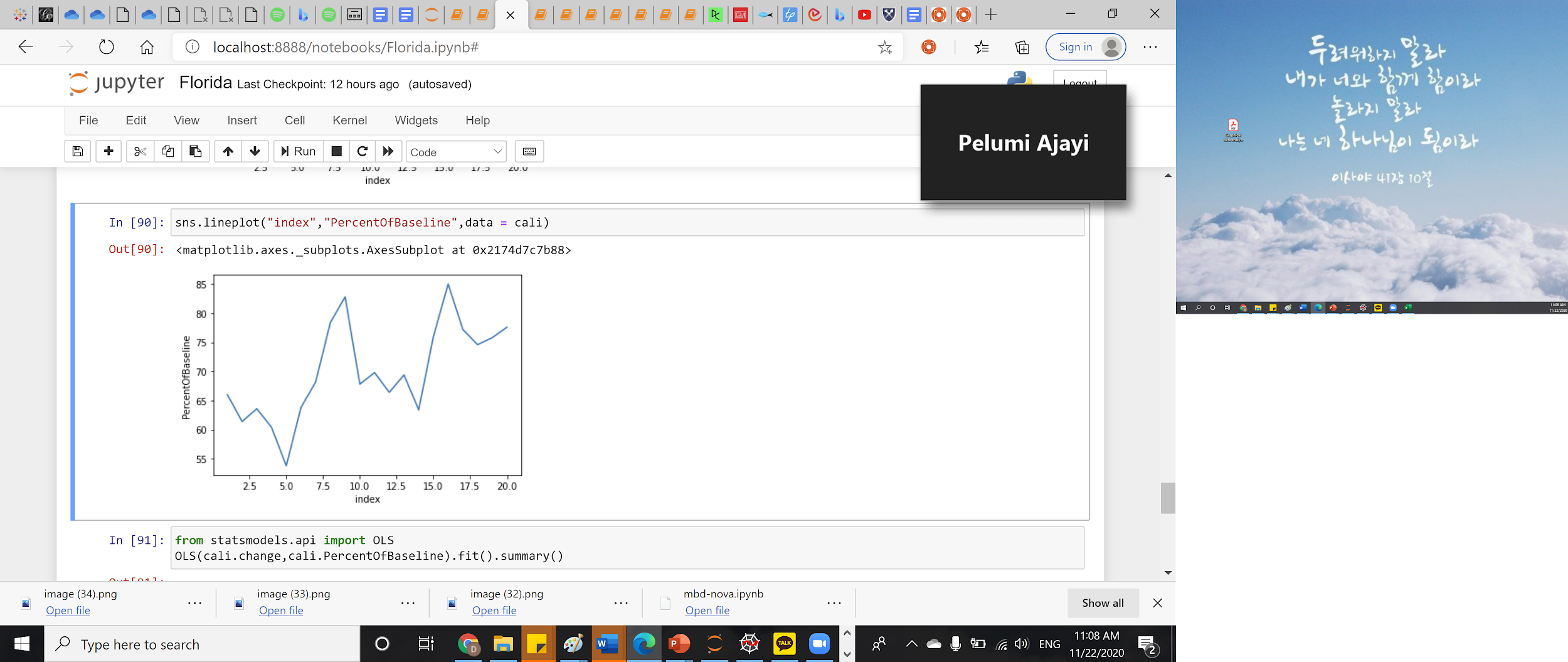
The results we gained are generalizable to the United States as a whole because 15 states represent different regions -- North, South, East, and West -- of the U.S, showing us the general trend in different regions. The results we found could also be used to determine the vulnerability of each state. Depending on the results, we clearly see a correlation between confirmed cases and international airline connections including specific statistical measures to certain areas, and based on this information government officials should be able to determine in areas where applicable to impose lockdown/quarantine to address the increasing rate of covid cases, especially for places that has relatively high indegree/pagerank centrality.

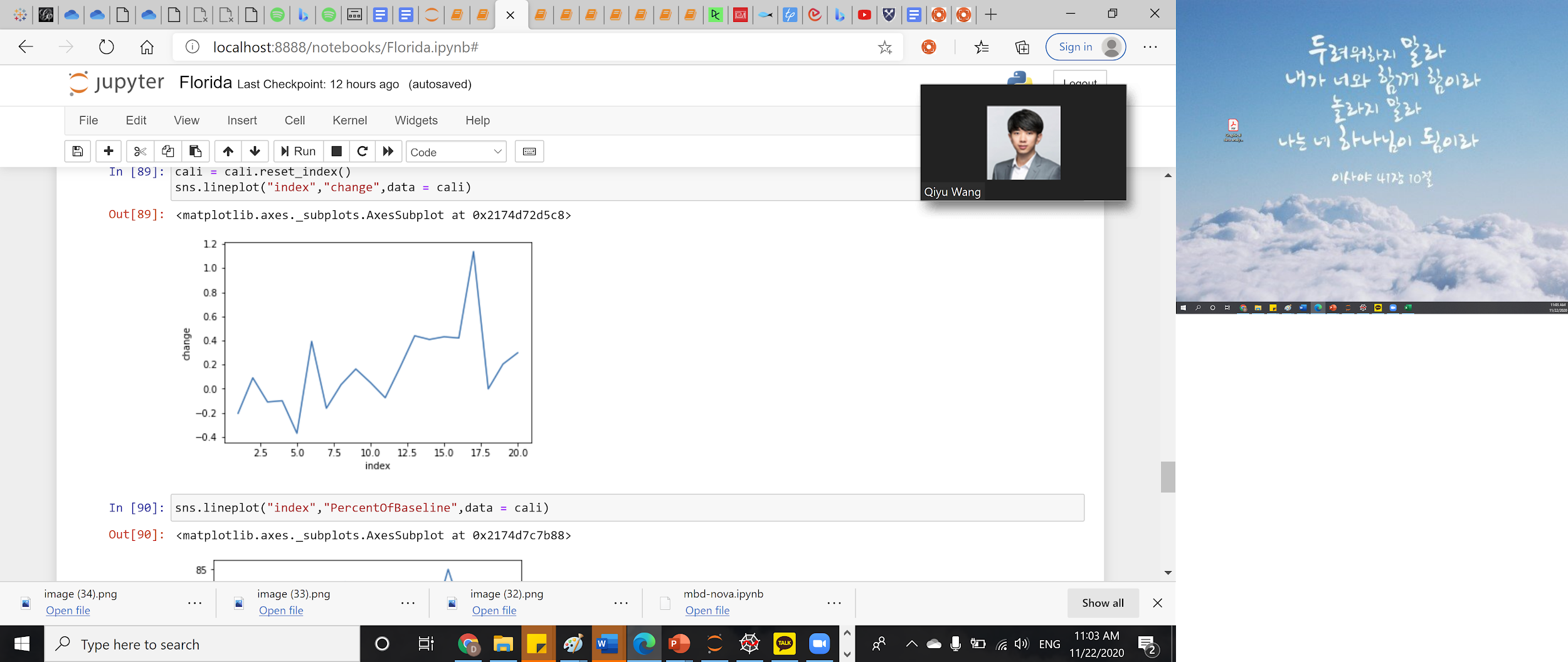
Some of the things that could be refined for this project are improving the quality of the dataset, which will both bring us better predictions and also a larger volume of data available as the bottleneck is resolved. Moreover, As we having more data, we could apply some different approach that could do better predictions in a larger scale, such as Spark ML for more computational efficiency and Random forest to capture more complex patterns.

<figure 1: Summary Statistics>



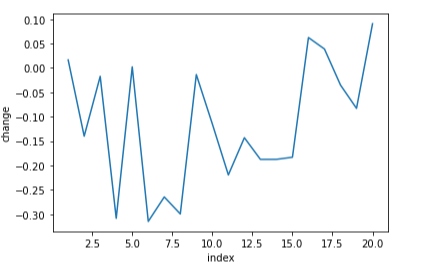
<figure 2, 3, 4: Florida>

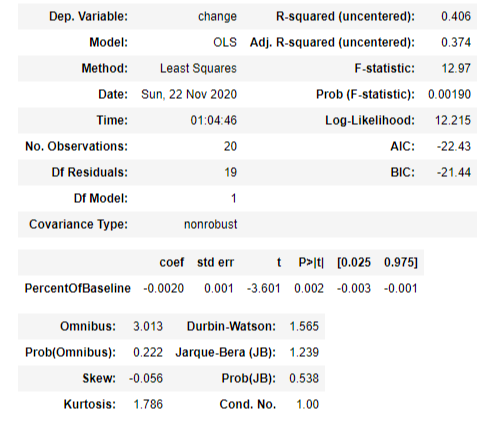
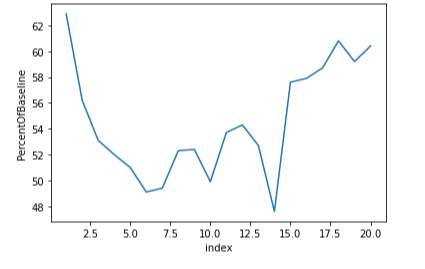






<figure 5, 6, 7: New York>





<figure 8, 9, 10: Ontario>

