Analyzing the Correlation Between COVID-19 Infections and Airport Traffic

Anna Tang, Clement Ip, Erica Ho

December 12, 2020

Contents

1 Introduction

1.1 Problem Statement

Team ACE* is commissioned by a Canadian Airline to determine how they should circulate flights during the trying times of the pandemic. They would like us to produce machine learning models that predicts how air traffic would fare in their main airport hubs here in Canada. In the event of a wave of infections, would people prefer to stay where they are or would they want to take the first flight home? What would people do if the curve of infections flatten?

It is difficult to guarantee predictions of exact percentages of airport traffic (compared to pre-pandemic rates) or give an estimate of how many flights they should send out. We compromised with the airline, and promised that we could give them general guidance on whether they should increase, decrease, or maintain the number of flights based on the extrapolated traffic percentage of pre-pandemic numbers.

For simplicity, we will be looking at the following airports (from West to East), as show in Figure 1:

Vancouver (YVR), Calgary (YEG), Toronto (YYZ), and Montreal (YUL)

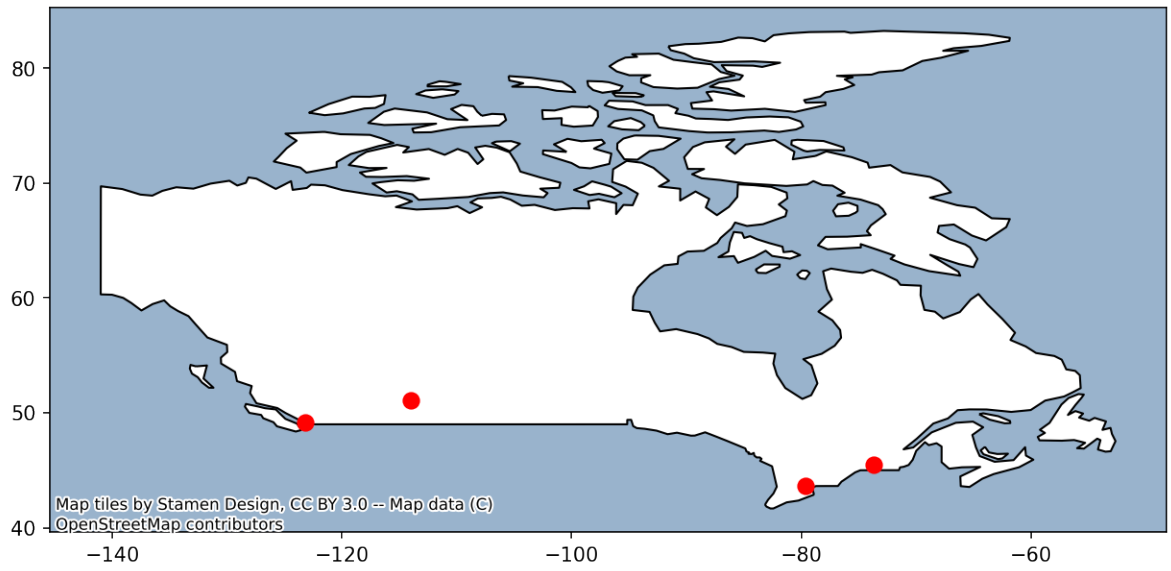

Figure 1: Locations of the airports from left to right: YVR, YEG, YYZ, YUL

*(A)nna, (C)lement, (E)rica = ACE

1.2 Assumptions

We will make the following assumptions when analyzing the data:

1. Every traveller goes to their own airport. All the cases in one province will only influence the traffic of their own airports.
    - (ex. Everyone living in British Columbia is associated to YVR, and the number of infections in BC will only have an effect on YVR airport traffic only.)
2. The values from the `PercentofBaseline` column uses the pre-pandemic baseline period from February 1st, 2020 - March 15th, 2020 to calculate the percentage of that day.
3. The values from the `PercentofBaseline` column aggregates all arrivals and departures on a certain date.

2 Methodology

2.1 Gathering and Cleaning the Data

Data for this project was mainly collected from two data sets:

- COVID-19's Impact on Airport Traffic from Geotab
- COVID-19 Data Repository from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University

2.1.1 Producing Usable Data Sets*

A script was used to merge all COVID-19 records dating from 2020-03-16 to 2020-10-16 from the John Hopkins University Coronavirus repository. After this process was completed, the data was filtered by country and only keeping records that were associated to Canada or the United States. Outliers such as the U.S. Virgin Islands, Diamond Princess and Grand Princess were omitted from the final data set. Similarly, another script processed the airport traffic data and only kept records that were of North American airports.

Once requirements were defined, the COVID-19 and airport traffic data sets produced were further filtered and separated to their respective province. A `Difference` column for the data sets containing the infections was added. This described the increase or decrease in cases from the day before.

*see *casesNa.csv, casesCanada.csv, casesUS.csv, airport_traffic_NA.csv* in *Archived-Data* and *Covid-Data*

2.1.2 Filtering Noise

After both the COVID-19 data and the airport data are graphed with their respective airports, certain patterns of noise started to appear. The number of new COVID-19 cases in Canada showed dips to zero on certain days and drastically jumps on the day afterwards. These patterns are not because of no infections, but human error and inconsistencies with reporting. For instance, reporting was not available early on in the pandemic, or cases were not being confirmed on holidays or weekends. The real number of infections on a pair of days is likely the average between zero and the reported number of the next day. Hence, to minimize the effect of data entry delay noise on our analysis, we smooth the case counts with LOESS smoothing.

At first glance, the airport traffic data looked very cyclic (over a week), which makes sense as people prefer travelling on certain days of the week over others. We do not want to look at the effects of weekdays versus weekends or holidays on airport traffic, but instead the long-run trend of airport traffic. This is why the airport traffic is LOESS smoothed to negate the short-run effects of calendar days on the data.

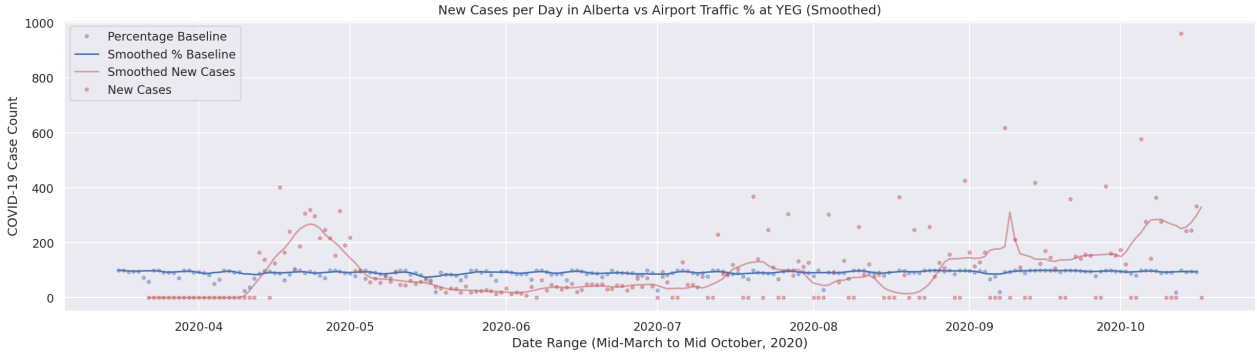

Figure 2: Number of New Cases in AB and YEG Traffic: Filtered

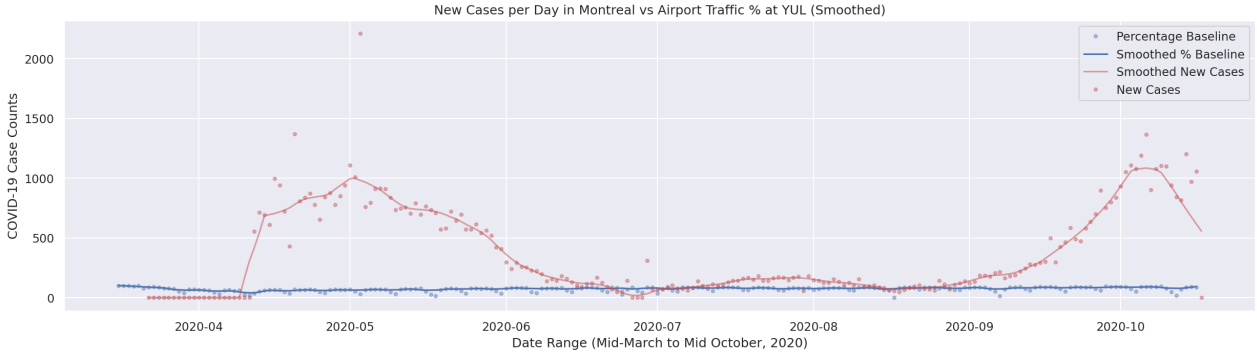

Figure 3: Number of New Cases in QC and YUL Traffic: Filtered

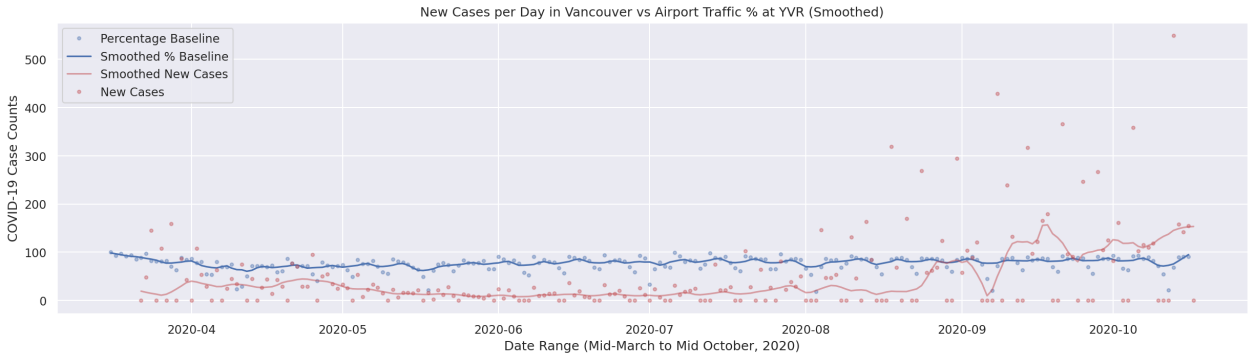

Figure 4: Number of New Cases in BC and YVR Traffic: Filtered

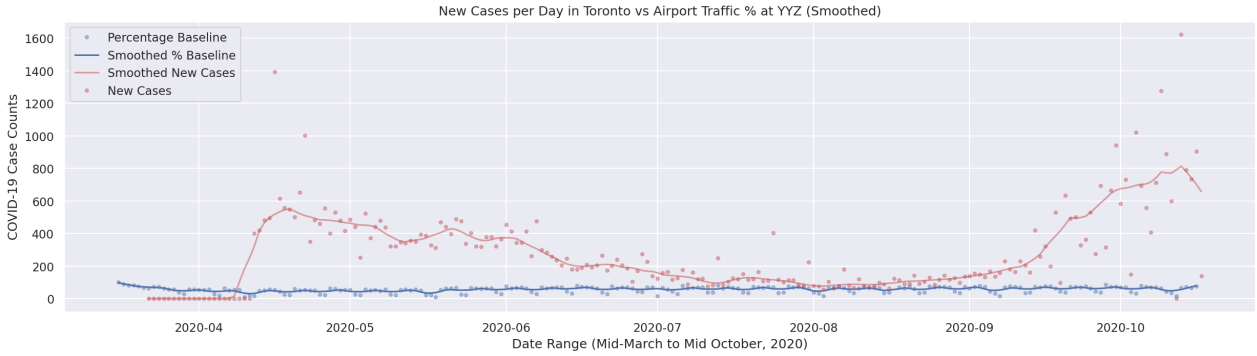

Figure 5: Number of New Cases in ON and YYZ Traffic: Filtered

2.2 Analyzing Statistical Correlation

After filtering the noise presented in the COVID-19 and airport traffic data, a correlation analysis was done to determine whether there exists a relationship between the `PercentofBaseline` airport traffic variable and the two `Confirmed` and `Difference` variables. The two COVID related variables represented total confirmed cases, and the daily change in confirmed COVID cases in the same province as the airport analyzed. The smoothed `Difference` and `PercentofBaseline` values were modelled using linear regression, and the strength of the correlation was later determined by the calculated correlation coefficient (r-value). This process was repeated with modelling the relationship with the `PercentofBaseline` and `Confirmed`.

Although linear regression is simple and provided a good starting point in the analysis, it did not quite fit our expectations since the confirmed number of cases does not grow at a linear rate. Other methods, such as polynomial regression, provided a better fit for the shape of the data.

2.2.1 Interpreting Correlation During Infection Waves

During the onset of a new wave of COVID-19 cases, the rate of infection grew at a soaring rate and potentially influenced one's decision on whether or not they should be travelling. As number of confirmed cases grew, how did air travel do? What was the response after the initial slope of the wave?

To answer these questions, linear regression was used to analyze the statistical correlation during the rise and drop of infections in the first waves. The overall correlation between `PercentofBaseline` and `Confirmed` was used as a baseline for comparison, to estimate how long recovery took.

2.2.2 Comparing Airport Traffic with North American COVID Data

It is expected in our globalized world that airport traffic in one province is not only affected by the state of the pandemic within the provinces' borders, but also by the state of the pandemic outside of the province, country, and continent. In particular, the correlation between the two variables `PercentofBaseline` of the airport's departures and arrivals and North America's total `Confirmed` cases was analyzed and plotted. The graphs are colour-coded by date, and fitted with either a polynomial regression model or a K-Neighbours Regressor. Training and validation data were used to evaluate the performance of the models, and resulted in reasonable prediction scores.

2.3 Machine Learning Modelling

For each airport, there are four different factors which are put into the regression:

1. Total Confirmed Cases (North America)
2. New Cases per Day (North America)
3. Number of Total Confirmed Cases for the province that the airport serves
4. New Confirmed Cases for the province that the airport serves.

The goal of the model is to predict what the `PercentofBaseline` of airports should be, meaning at what percent capacity the airport should operate at (compared to pre-pandemic rates), given the current COVID-19 cases counts in both North America as a whole, and in the airport's province.

We chose a gradient booster regression model as the four different factors would each be weighed differently when it comes to how each factor may affect the airport traffic. A neural network regressor would also be ideal, if we had more data to work with. The presented data resulting from the pandemic is what we would consider to be as a "weak learner" and the loss function would adjust the greedily produced regression trees to minimize the residuals made from the predictions. Since gradient boosting makes greedy predictions sequentially, there is a small concern of overfitting. A Min-Max Scaler is also included in the model, as the factors need to be balanced out as total case counts are much larger than the difference between two day's case counts. The same data split from the previous step was used and resulted in high validation and training scores (0.8 - 0.9), with a minimal difference.

3 Results

3.1 Analysis

Following the steps outlined in our Methodology, we present the results of each of the respective airports below.

3.1.1 Vancouver (YVR)

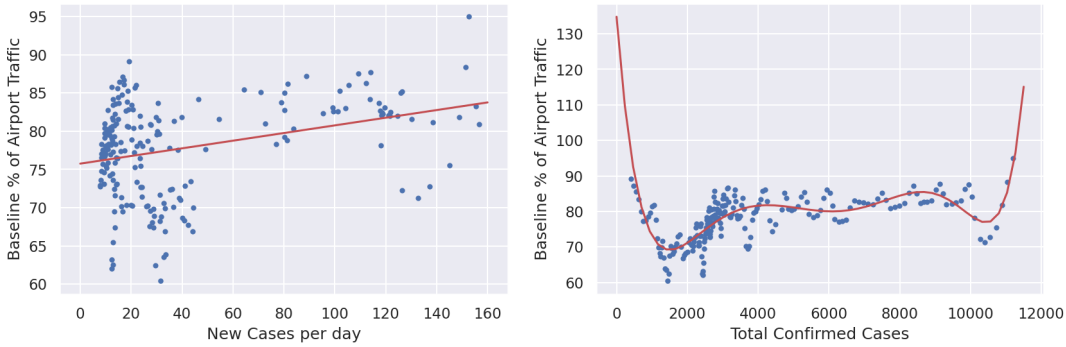

Figure 6: YVR Airport Baseline Versus Number of New Cases (Difference) and Total Confirmed Cases in BC

Looking at the number of new cases per day near YVR versus YVR `Baseline` airport traffic, it does not seem like there is much of a correlation. The data for the new cases numbers isn't very well balanced, which makes sense as BC hasn't previously been too hardly hit by the pandemic.

We turn to look at total `Confirmed` cases instead, which gives us much more information. While there is no clear linear trend, we do see that there are two dips in airport traffic around the 2000 and the 10000 mark of confirmed cases. A polynomial fitting on the graph as shown below reveals the two dips clearly. Upon further inspection, these case counts happen to be the number of confirmed case counts reported during the period of time corresponding to the onset of the first and second waves of the pandemic in BC. The new percentage baseline of airport traffic recovers back to a consistent $80 - 90\%$

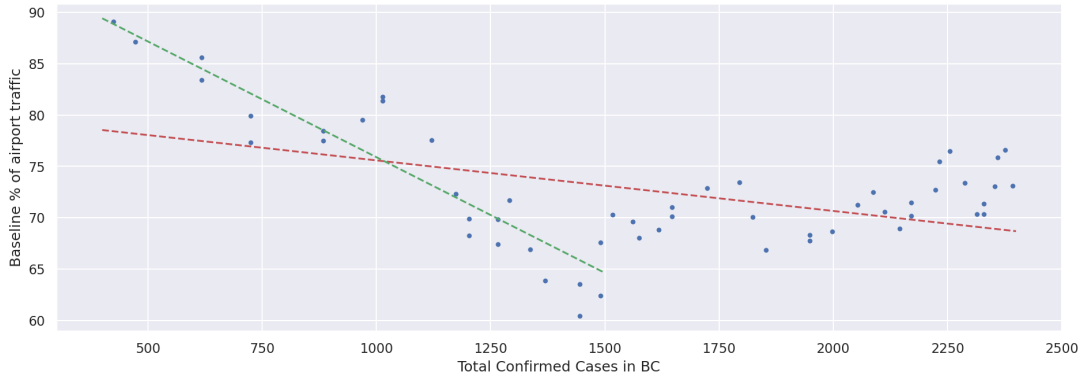

Figure 7: Linear Analysis of YVR First Wave

Isolating the first wave (data from Mid-March to Mid-May) indicates that while there was a clear linear decrease in airport traffic as confirmed cases increased. The green dotted line in Figure 7 below shows a clear linear correlation with a $r-$value of -0.92 between the number of `Confirmed` cases and airport traffic in the first 25 days of the pandemic. When case counts hit around 1500, airport traffic began going up again, albeit slower compared to the initial decrease.

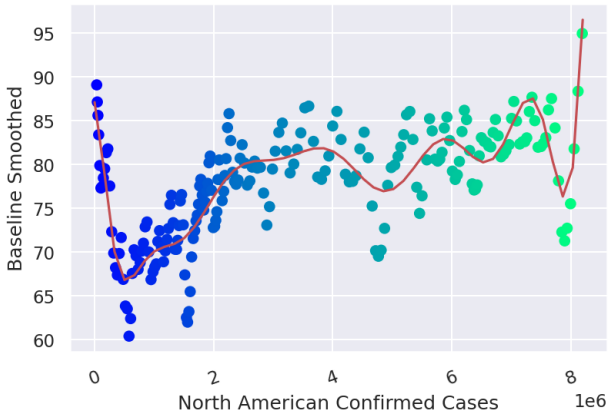

Figure 8: YVR Airport Traffic vs. Confirmed Cases in North America

When comparing North America COVID cases to the YVR traffic, the same initial dip is seen, along with a gradual recovery. The graph in Figure 8 is colour-mapped by date, allowing us to see that airport traffic may also be affected by the date, or more specifically, what stage of the pandemic they are currently in.

3.1.2 Calgary (YEG)

During initial observations, the wave (Mid April - June) and the steady increase of new cases after the summer stood out at first glance.

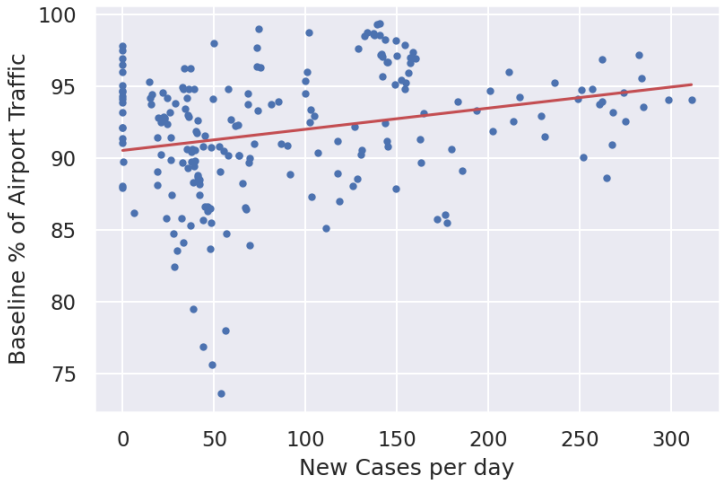

Figure 9: YEG Airport Baseline Versus Number of New Cases (Difference)

Results of the statistical correlation analysis showed that there was little to no correlation between

the traffic percentage and the number of new cases per day(`Difference`). This made sense, since the plotted airport traffic remained at a constant rate, and being only the 4^{th} and 37^{th} busiest airport in Canada and North America, respectively.

When analyzing airport traffic between the total number of `Confirmed` cases, results showed that there was a slight positive correlation between the two variables with using a linear regression. Upon further examination using a K-Neighbors Regressor, the resulting model produced a fit with many oscillations and even some dramatic dips when the number of confirmed cases reaches 7000 and 15000. Using the initially plotted graphs, the first dip corresponds to the end of the first wave, while the second one relates to the increase noted at the end of the summer.

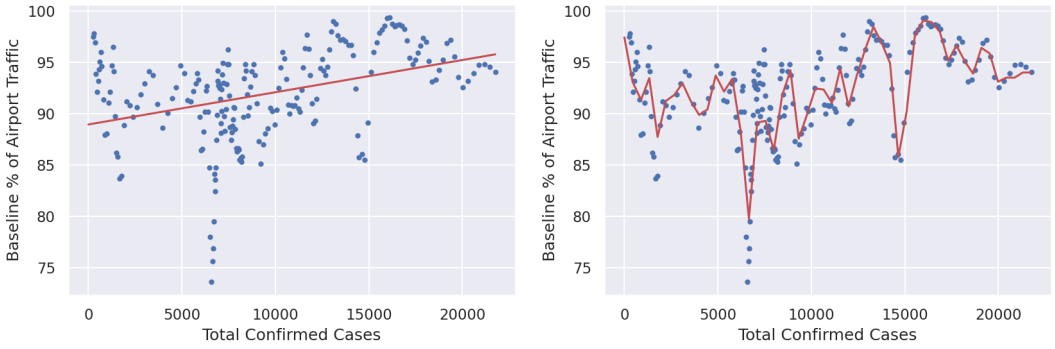

Figure 10: YEG Airport Baseline Versus Number of Confirmed Cases
Linear Regression (Left) & K-Neighbours Regressor (Right)

Similar trends were used to analyze the regression of the airport traffic percentage if the number of `Confirmed` cases grew at the rate presented in the first wave in April. The initial rise, compared the overall curve showed a steep 25% drop at the end, whereas the latter had a 10% decrease. In other words, if the pandemic continued on at the rate of the first wave's uprising, it would not take long to cause another severe drop in traffic. However, recovery from the initial wave showed a 20% increase which is nearly enough to account for the amount lost during the it's onset.

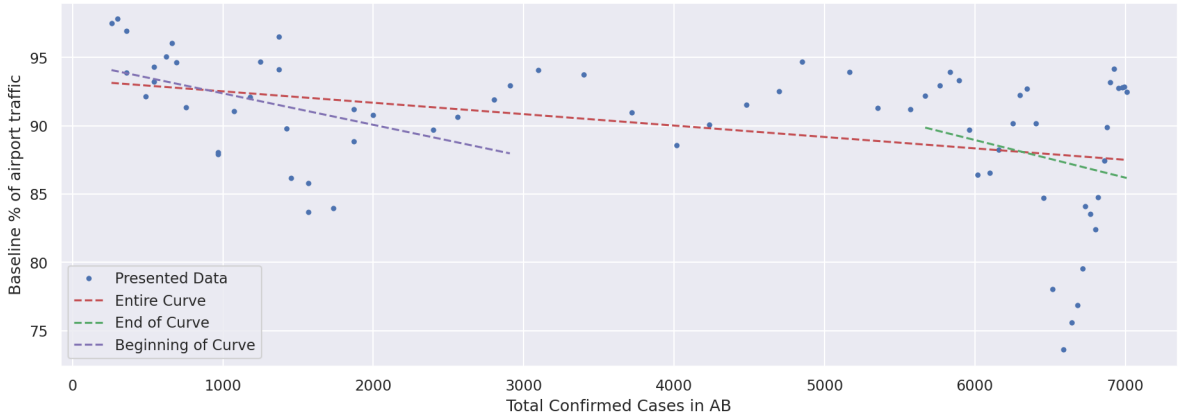

Figure 11: YEG Airport Baseline Versus Confirmed Cases in First Wave

A comparison with YEG airport traffic with the total number of `Confirmed` cases in North America shows a drastic dip at 1500000 at the early stages of the pandemic. However, this dip does recover as time goes by and with implementation of new safety measures.

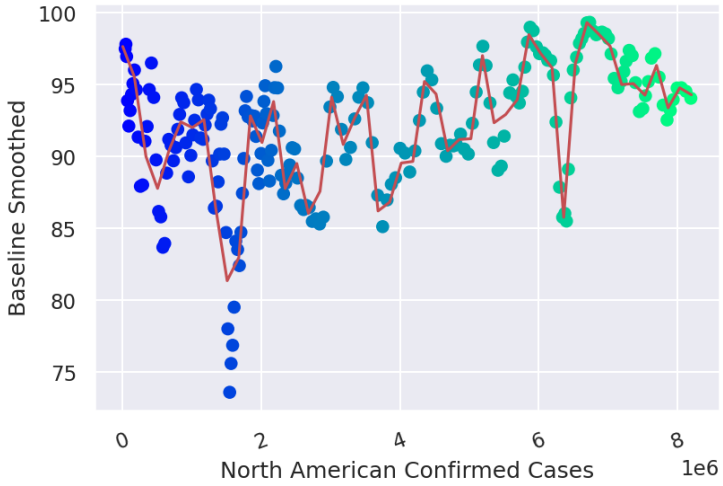

Figure 12: YEG Airport Baseline Versus Total Confirmed Cases in North America Using K-Neighbours Regressor

3.1.3 Toronto (YYZ)

Moving east, it appears that Ontario has been severely impacted by COVID-19, as seen by the number of confirmed cases.

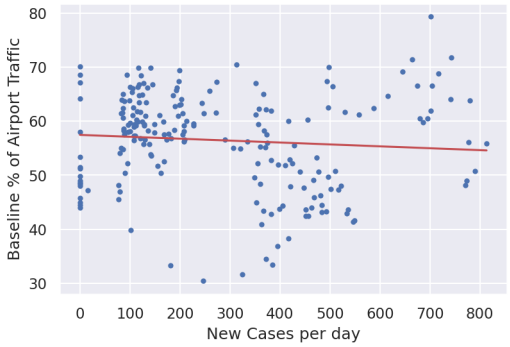

Figure 13: YYZ Airport Baseline Versus New Cases in Toronto

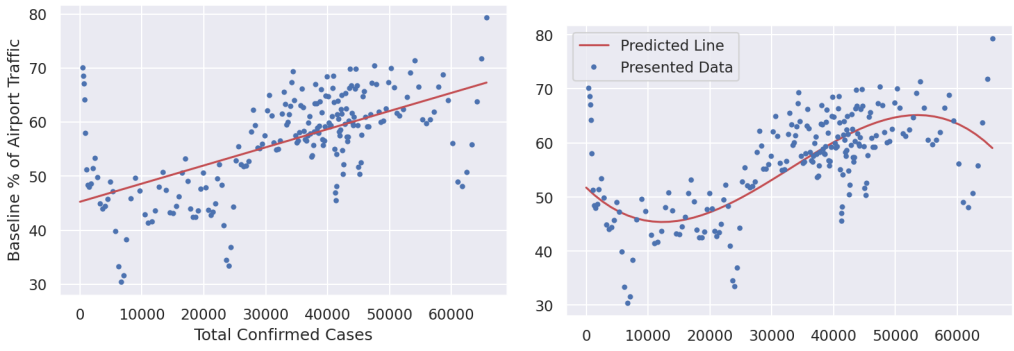

Figure 14: YYZ Airport Baseline Versus Confirmed Cases in 1st Wave

Unsurprisingly enough, there is no correlation between the number of new cases and the airport

traffic, but there is a strong relationship between airport traffic and the total number of `Confirmed` cases.

An attempt to fit the correlation with a polynomial regression was made, and the dip at 10000 cases can be related to the initial fall seen in April. As the pandemic progressed with more safety measures added in, the percentage of airport traffic went back up again, but faces another potential fall.

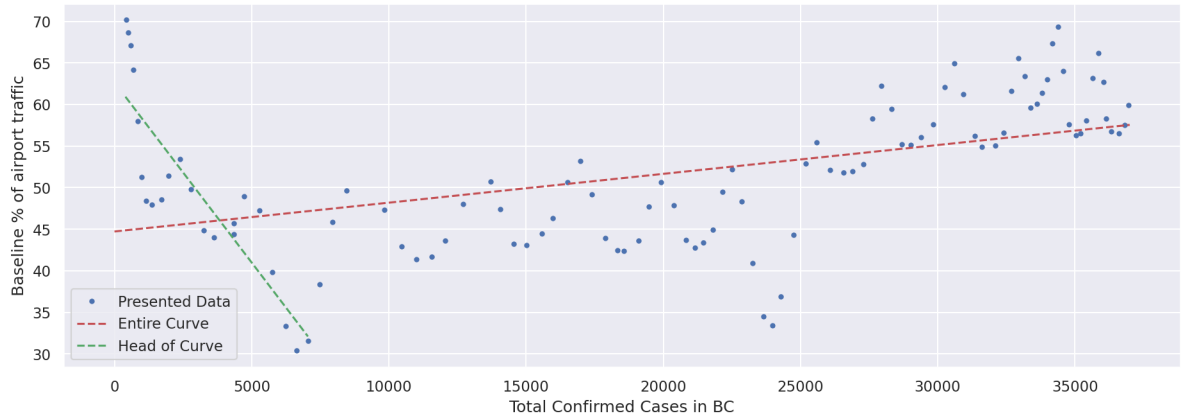

Figure 15: YYZ Airport Baseline Versus New and Confirmed Cases in 1st Wave

Analyzing the trends in `PercentOfBaseline` airport traffic during the waves, it appears that airport traffic was heavily impacted during the first month. However, it appear to eventually make a recovery as the first influx subsided.

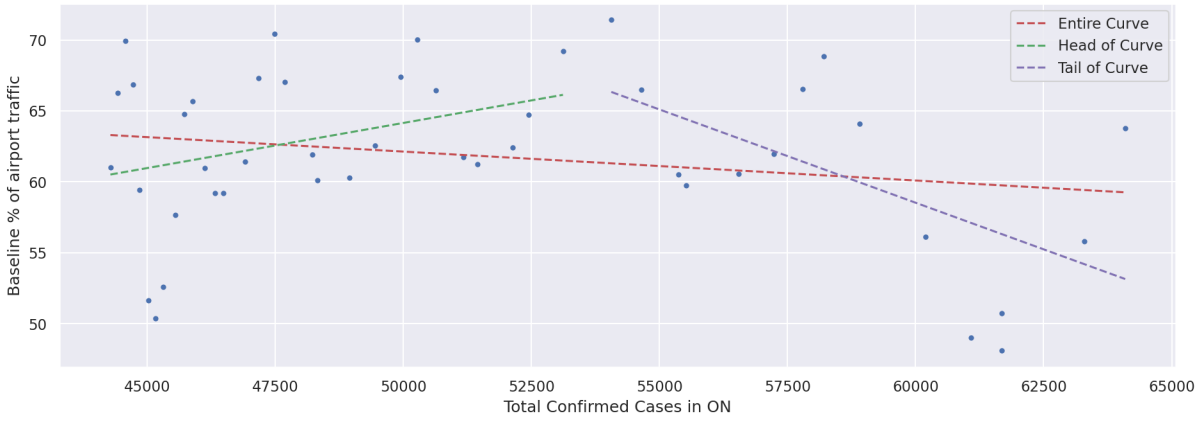

Figure 16: YYZ Airport Baseline Versus Confirmed Cases in 2nd Wave

This was not the case for the second wave. As seen in Figure 16, the airport traffic started to fall as well. The tail of the curve did indicate to have somewhat of a moderate correlation with the dropping airport traffic.

The initial dip in airport traffic is also reflected when comparing with the total number of `Confirmed` cases in North America, but also shows a steady increase of 20% overtime. This increase is seen to be stable and even indicates a continuation of an upwards trend.

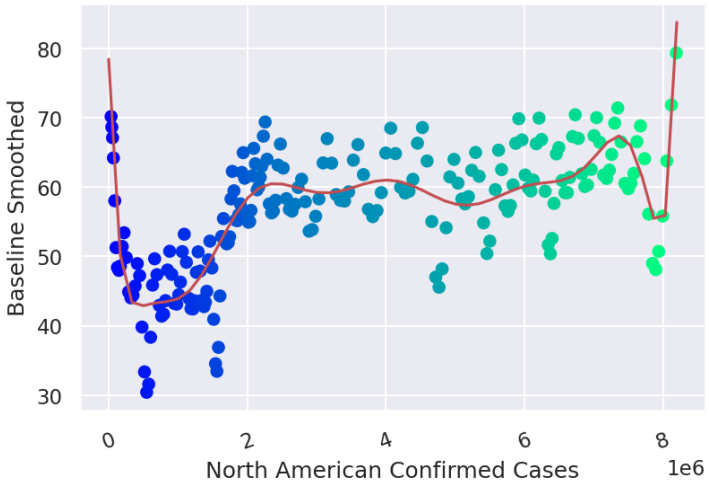

Figure 17: YYZ Airport Baseline Versus Total Confirmed Cases in North America

3.1.4 Montreal (YUL)

A similar case can be made in Quebec, as they too have been heavily impacted by COVID-19.

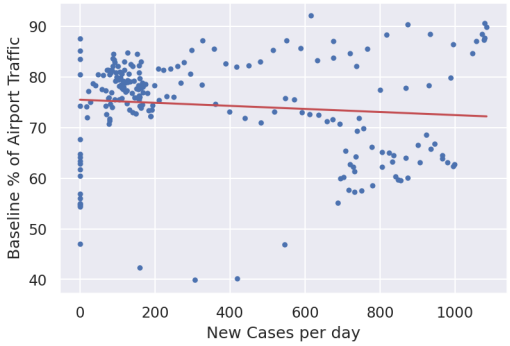

Figure 18: YUL Airport Baseline Versus Montreal's New Cases (Difference)

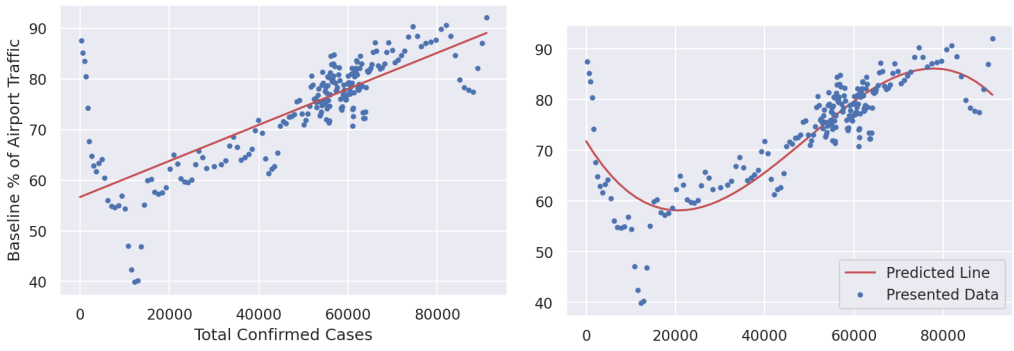

Figure 19: YUL Airport Baseline Versus Total Confirmed Cases in Montreal - Linear and Polynomial Fitting

While there is no clear statistical correlation between airport traffic and the number of new cases per day (`Difference`), the statistical analysis indicates a somewhat strong positive correlation

between airport traffic and total `Confirmed` cases with an r-value of 0.77. Similar to the other airports, the relation between airport traffic and Total `Confirmed` cases is not linear, but rather has a positive "curvilinear" relationship (first dips down rapidly, but slowly recovers). As seen in the graph below, we can fit a polynomial curve to the data, with 70% accuracy which highlights the large dip from the start of the first curve and the subsequent recovery.

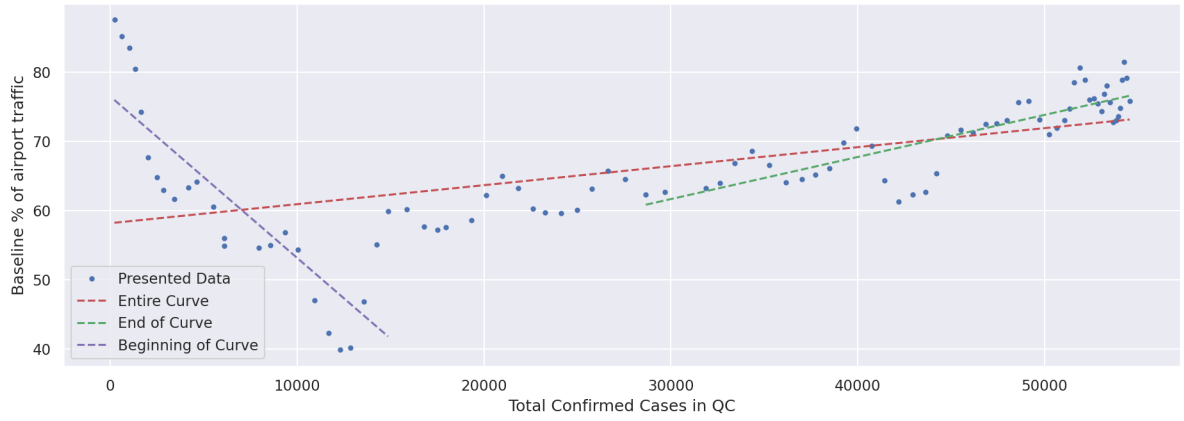

Figure 20: YUL Airport Baseline Versus Total Confirmed Cases in 1st Wave

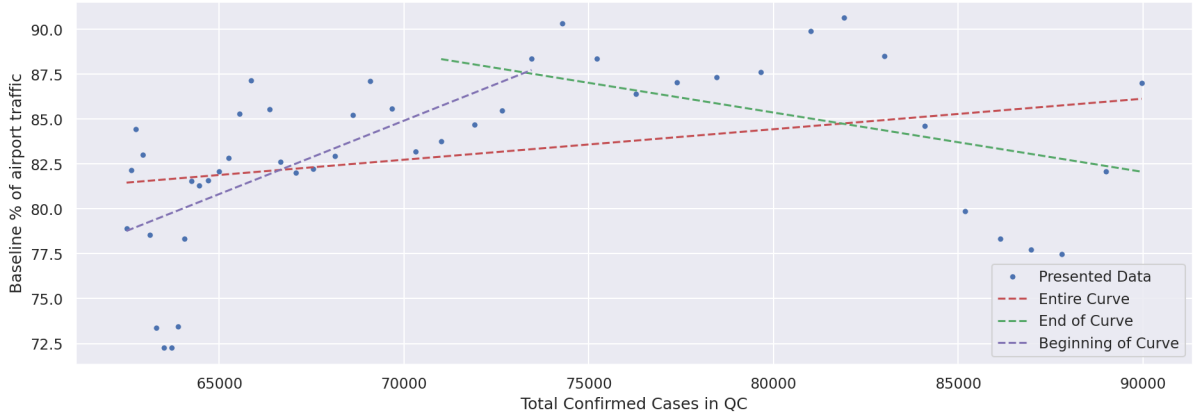

Figure 21: YUL Airport Baseline Versus Total Confirmed Cases in 2nd Wave

A linear analysis on the first and second half of first wave highlights two linear correlations, one positive and one negative. A moderate negative correlation can be identified in the first part of the first wave, with an $r-$value of -0.66. A strong positive correlation can be seen in the latter part of the first wave with a $r-$value of 0.84. A statistical analysis on the second wave was not able to find any significant linear correlation between sequential halves of the second wave.

Comparing the North American `Confirmed` cases to YUL `PercentOfBaseline` airport traffic shows similar results to Vancouver and Toronto with a large dip and subsequent recovery of around 20% over time. Despite all those small side dips, the upward trend shows stability.

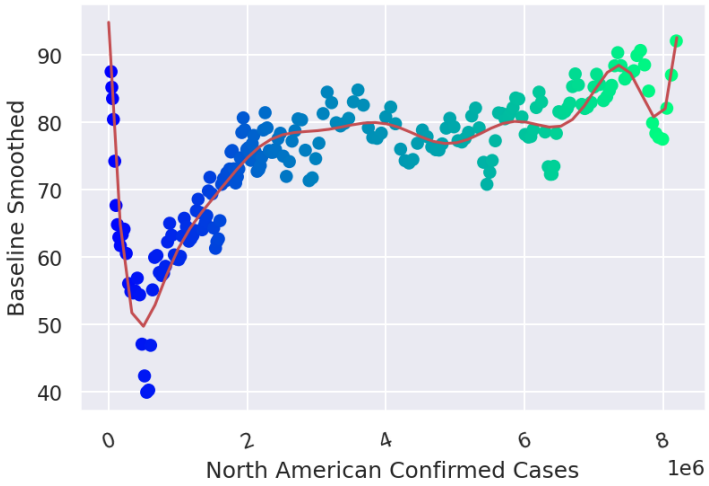

Figure 22: YUL Airport Baseline Versus Total Confirmed Cases in North America

4 Discussion and Limitations

Through our analysis of the four Canadian airports, the main takeaways are as follows:

1. At the start of each wave, baseline goes down. The graph between Total `Confirmed` Cases and `PercentOfBaseline` Airport Traffic for each Canadian Airport all show a similar trend as the number of confirmed cases increases.

   Airport traffic in all airports decreased rapidly and hit the lowest point when the pandemic first hit. This can be due to the initial "shock" of the pandemic and travel restrictions being imposed.

2. Even when Total `Confirmed` Cases and Daily `Difference` between Confirmed Cases go up after the "shock", `Baseline` airport traffic is able to recover to a consistent sub-100% rate and shows a stabilization in `Baseline` rates in all four airports.

3. The position of the airports has had a profound effect on the relation between the number of `Confirmed` Cases and `Baseline` Airport Traffic.

   While YVR and YUL (which are the western-most and eastern-most major Canadian airports, respectively) have recovered to 80% capacity even as the number of `Confirmed` Cases has increased, YYZ has only recovered to 60% capacity. However, all three of these airports have been able to recover to a consistent percentage throughout the pandemic.

   The exception is YEG, which is situated in the Prairies, and deals mainly with domestic flights instead of international flights. This is true as YEG's traffic fluctuates around 90% capacity, which is much higher than the other three airports. Since it mostly serves domestic flights, it makes sense that it is not as affected by international travel restrictions and hence was not as affected by the initial shock of the pandemic compared to the other three airports.

Overall, the three main observations are expected, as this shows that while less people are flying overall, most people have adapted to the new reality. We are also able to highlight the differences and similarities between how the four airports have been affected by the pandemic.

While the airport baseline data shows a recovery in flight traffic, the `PercentOfBaseline` percentage data we found only concerns the amount of flights going to and from an airport, and not the amount of people flying. We were not able to find any data on the amount of people on each flight. With COVID-19 restrictions and other factors to consider, we can probably assume that most flights now are carrying less people. So, just because there is a recovery in the `Baseline` Airport Traffic, it does not mean that the Canadian Airline is hemorrhaging any less money running these flights. Nevertheless, the `Baseline` percentage statistic is generally a good indication of how many flights should be sent.

The statistical analysis and machine learning models that we created only takes into account COVID-19 cases on the current day, this is not entirely accurate. The reason? People do not react that fast to new information and news stories as we would like them to. However, the number of cases on any day is dependent on the number of cases before that day since only an infected person can infect another person, so it shouldn't affect results by too much. In the long run, the difference between cases between each day does not do much to change people's minds when it comes to whether or not they should fly on an airplane (most flights are booked long ahead of time anyways).

A future extension of this project could include repeating the analysis for airports in the United States. In particular: Atlanta (ATL), Los Angeles (LAX), New York (JFK) and Seattle (SEA). These airports and their respective states have gone through multiple waves of infections and a higher percentage of travellers are still traveling compared to those in Canada.

5 Conclusion

5.1 Recommendations By Airport

5.1.1 Vancouver (YVR)

YVR is Canada's 2^{nd} busiest airport and 26^{th} in North America, hence traffic should remain consistent at around $80 - 90\%$. There may be some more dips, but it will recover over time as people get used to the new reality of the pandemic. British Columbia is fortunate enough to not be as heavily impacted as the other provinces, and YVR is the preferable gateway when for making international travels. The situation may not be as optimal as it was pre-pandemic, but it certainly will not drop to 60% like it did when the pandemic first hit in March and April. We recommend keeping the amount of flights at around $80 - 90\%$ baseline, and cautiously increasing flights in the next quarter.

5.1.2 Calgary (YEG)

Given how Alberta is seeing an upwards trend in the number of new cases, airport traffic will likely see a steady decrease as well. Stronger mandatory measures at the critical level have been put in place in December of 2020 to help curb the number of cases and any form of non-essential travel in or out of the province is heavily discouraged at this time. Using data from the first wave, a notable

drop in airport traffic was recorded and resulted in grave financial consequences in which are still being settled with today. Additionally, YEG attracts the least amount of traffic compared to our other hubs. Our recommendation is to lower the number of flights to $60 - 70\%$ at baseline for the next quarter, given the uncertainties of the holidays and the vaccine.

5.1.3 Toronto (YYZ)

When analyzing trends for the ongoing second wave, airport traffic increased by 6% and dropped yet again once the curve was starting to smooth out. This shows just how sensitive the travelling population is when making their decision to travel through YYZ. New data shows that the influx of confirmed cases will continue to increase and using our airport data as a baseline estimate, we can infer that there will be wanting people wanting to fly. YYZ is Canada's busiest airport and ranks 11^{th} overall in North America, so it is a major hub by these definitions. Going forward, we recommend keeping flights consistent at around $70 - 80\%$ as a start and later, a cautious increase of flights for the next quarter.

5.1.4 Montreal (YUL)

The situation in Quebec is very similar to the situation in Ontario, so we would present similar recommendations as for YUL. However, YUL ranks 3^{rd} and 37^{th} respectively in the busiest airports of Canada and North America, hence less busier than YYZ. We recommend starting with keeping capacity at the lower 70% range instead of around $70 - 80\%$ with YYZ for the start of next quarter.

5.2 Epilogue

As the pandemic continues, we may see a different trend between case counts and airport traffic as the holidays approach, and as a vaccine comes out. Airlines are monitoring the situation as it progresses and are cautiously implementing new safety measures to protect their travellers. While it is unlikely for airplane traffic to fully recover in the near future, there is an overall positive or neutral outlook for airplane traffic for all airports, and as people become used to the new reality. With vaccinations coming around the corner, we hope that this can restore the faith in travellers and make flying a memorable experience for everyone.

6 Project Accomplishment Statements

6.1 Anna Tang

- Generated scripts to organize the collected airport and COVID-19 Data (over 1GB) from the CSSE of John Hopkins University, and produced usable DataFrames for project analysis.
- Utilized different methods of regression to evaluate correlation between trends in travel and COVID-19 cases, and modeled the extrapolation of percentages of airport traffic.
- Performed linear regression on airport traffic during infection waves in order to examine behaviour on isolated timelines.
- Generated machine learning models for Calgary and Montreal that predicted airport traffic using K-Neighbours and Gradient Boosting Regressor.
- Wrote a program that calculated a county's closest airport given a set of coordinates, with an implementation of the Haversine function.
- Documented process and findings, and gave recommendations on increasing or decreasing the number of flights for each airport in the written report.

6.2 Clement Ip

- Obtained and analyzed data related to COVID-19 and sought to determine if it had any impact on airport traffic.
- Analyzed data specifically for airports located in Toronto and Montreal, and created machine learning models to predict the future impact of air travel.
- Generated plots and diagrams of our machine learning models for specific time frames in order to determine any significant changed in the data.
- Prepared summaries and descriptions on the methodology of the findings in the form of a written report.
- Provided recommendations to each of the airports on whether to increase/decrease both the number and capacity of flights.

6.3 Erica Ho

- Performed data cleaning, transformation and smoothing on collected airport and COVID-19 data to refine data frames for analysis.
- Coded framework and script for the regression and correlation statistical analysis of airport traffic for all four airports.
- Built Gradient Booster Regression Machine Learning model on four factors which consistently scored above 80% when analyzing airports.
- Concisely summarized Methods, Variables, Analysis and Conclusions of findings in a technical written report.