Andrew LaFortune
Project 3 – Spark COVID 19 Data Research
5/6/2022

Introduction and Business Scenario

As the COVID-19 pandemic seemingly draws to a close, it is worth reflecting on what measures for preventing the spread or "flattening the curve" worked and didn't work over the course of two years. The most popular model for COVID-19 analysis is the counts of infected, recovered, and deceased counts. These metrics show what has happened after the fact, and don't necessarily capture the nuances of what led to the results. This is great for reflection days or months after a plan has gone into effect, but policy makers need to have an idea of what might happen as a result of different policy actions like limiting restaurant capacity or re-opening schools.

The business case imagined for this project is one where the office of Governor Tim Walz is seeking to create a toolkit for future elected officials that will allow them to visualize the near future in terms of the public's response and compliance with preventative measures, and the probable effect on those key measurements of infections, recoveries, and deaths. In the COVID-19 pandemic most preventative measures had to do with limiting social movement and interaction. If people limit their human interaction to a small group, it is less likely that the virus can spread quickly through the population. There are a few ways verify that people are obeying or disobeying stay-at-home orders. They can be asked to self-report their activity, but there is no guarantee that everyone will answer and humans are likely to provide answers they think are acceptable. Their cellphones can be tracked with GPS data, but that data is often expensive because of its appeal to advertisers and the difficulty in collecting it. Since this is already a government project, it would make sense to look at government resources like traffic cameras. The Minnesota Department of Transportation (MNDOT) maintains over 100 Weigh in Motion (WIM) and Automatic Traffic Recorder (ATR) stations across the state which provide counts of the number of cars that drive past them on an hourly basis.

Using the traffic volume data from WIM stations from the years 2017-2022, the proposed goal of this project was to compare the traffic in pandemic years (2020-2022) to non-pandemic years and align those results with counts of COVID infections and deaths (made available by the New York Times) and the timeline of major policy changes in the pandemic like stay-at-home and mask mandates. Then, using a variety of time-series forecasting and regression models, see if a model could be developed that would predict the trajectory of traffic volume as a measure of social mobility for a defined set of policy choices. For example, if daily infection counts were decreasing and daily traffic volume was increasing, a decision to implement strict capacity requirements might have little effect on people's real behavior. Ultimately, this tool could prove useful and highly valuable in saving lives as the best policy decisions could be made to get out of a pandemic quickly, and the frustration and inconsistency of the COVID-19 pandemic might be avoided if the forecasted outcomes are accurate enough.

The Data

Data for this task came from two primary sources and one secondary source. The first primary source was the WIM Traffic Volume data. It came in the form of .csv files from the MNDOT Data Products page with raw counts of vehicles for each station in the state in hourly increments. Each station could survey traffic going in multiple directions across multiple lanes of traffic with separate rows of data for each. These raw counts were aggregated by day and by station so that the details of travel direction and lane number were ignored, then joined with station information tables which allowed each station to be assigned to a county. The original WIM traffic comparison (Appendix A) aggregated traffic volume by district. The state of Minnesota has 8 different districts (Appendix B), and the designation of each county is available on a different MNDOT site as an html table.

The main goal was to show how traffic volume differed in pandemic years compared to a typical non-pandemic year. WIM data was available from 2017 onward, so the data from 2017 to 2019 was averaged to generate a "typical" traffic volume for each district on a given day. To account for traffic variations by day of the week, the date strings from the raw data were converted to ISO week date format. This way, the first Monday of 2020 would be compared to the average first Monday of other years rather than comparing January 6[st], 2020 to an average January 6[th]. With this conversion done, the raw vehicle count data for average years and pandemic years was plotted with an equal scale for each district with an upper limit of two million vehicles per day. This first plot in Figure 1 shows that volume is much higher in the metro district as would be expected, and the Statewide trends are more or less the same as the trends within the Metro district.

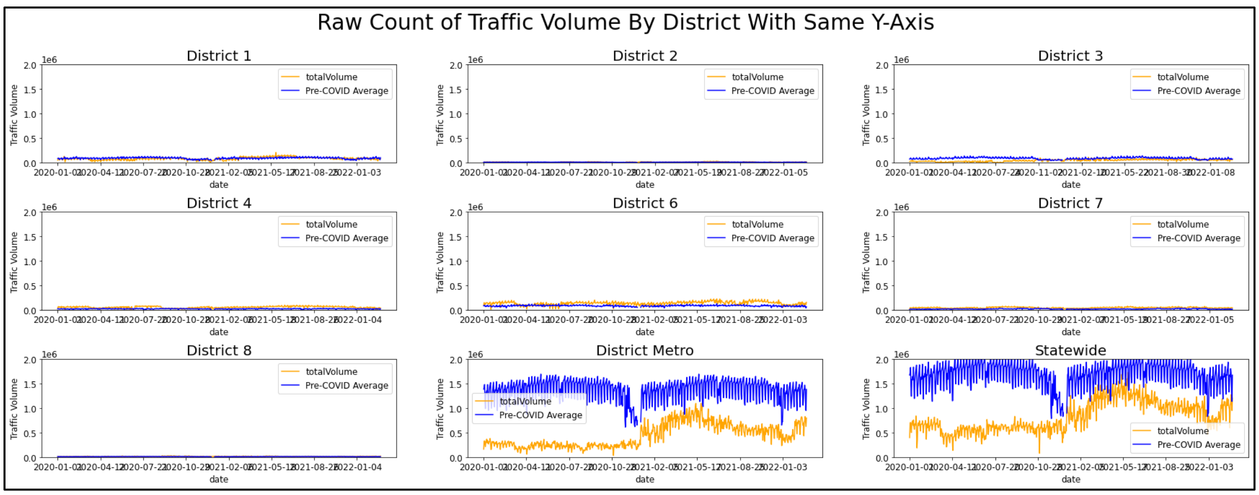


Figure 1: Traffic volume by district with equal volume axis scaling.

Next, to better show the patterns in traffic in the smaller districts, no scale limit was passed to the pyplot plotting function, so the data range was inferred automatically for each district. The results are captured in Figure 2, and show that traffic counts vary more day-to-day in smaller districts, and in many cases the COVID traffic volume was higher than the pre-COVID traffic. It is worth noting that no adjustments were made to account for new WIM stations being added, which might explain the higher COVID traffic volume in Districts 4, 6, and 7. In any case the raw volume graph was rather noisy and did not provide an easy way to examine the difference between COVID and pre-COVID traffic across the entire time period.

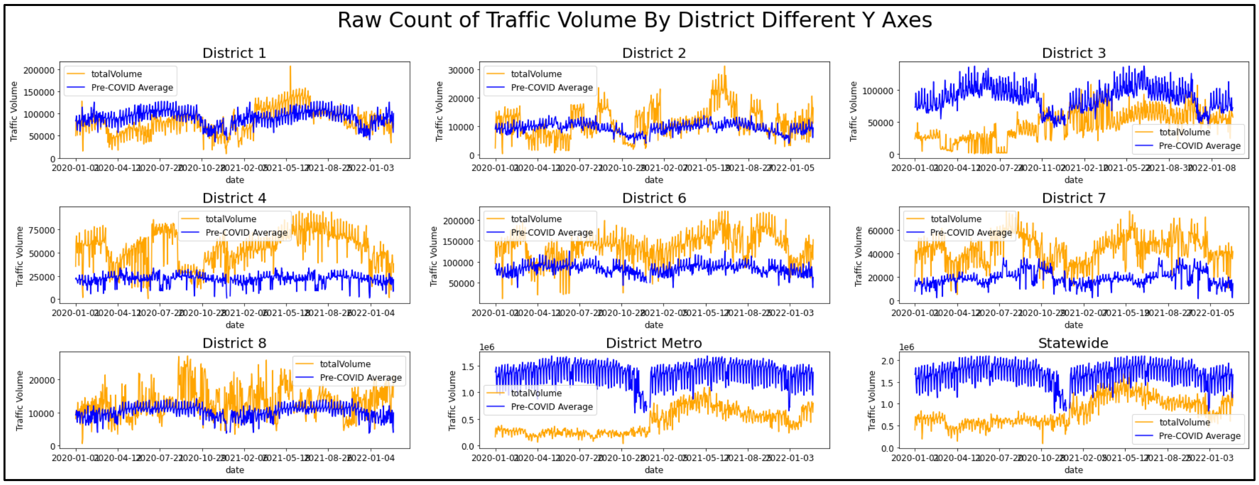


Figure 2: Traffic volume by district with variable volume scaling

To reduce the noise in the Figure 2 plot, the two data series were combined by dividing the COVID traffic volume by the average non-COVID traffic volume. This led to the results in Figure 3, which shows a time series for each district over the two-year COVID period representing the proportion of average traffic on each date. Again, some of the variability in the smaller districts led to noisy graphs like in District 8, and some outliers distorted the scale of graphs like in District 7 and District 4. The only graphs that didn’t consistently go far above the pre-COVID volumes with high variability were the statewide and Metro District graphs, which closely mimic each other. At this point the idea of creating a per-district COVID policy response model was abandoned in favor of focusing on a better model for the state response on aggregate.

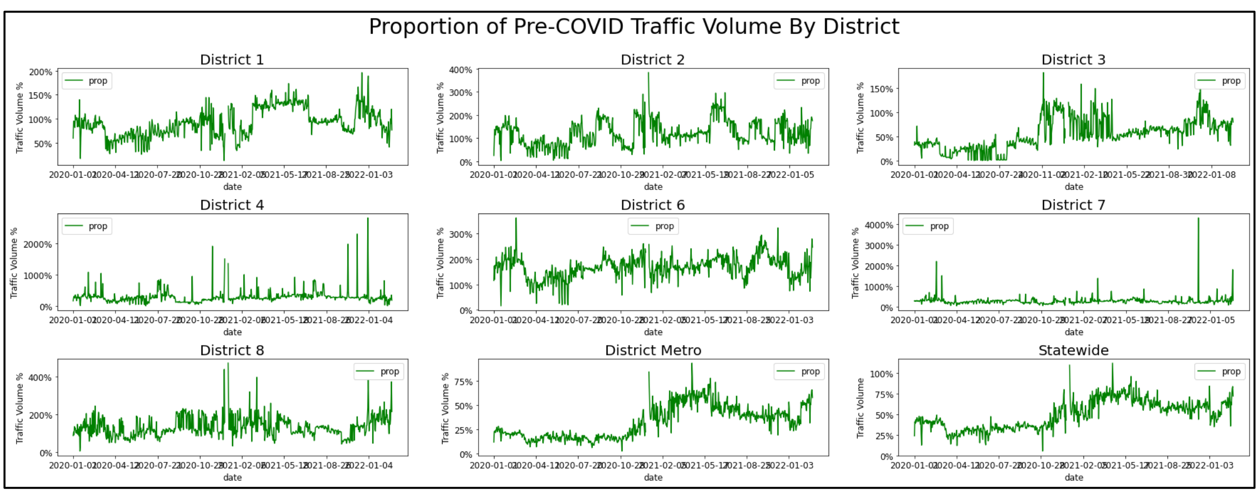


Figure 3: Proportion of average traffic volume by district during COVID

Figure 4 shows the enlarged view of traffic volume for the entire state as a proportion of the typical traffic volume in all dates of the years 2020 and 2021. It is unclear why the traffic volume was so low to begin with in 2020 before any pandemic measures had come out. Some possibilities might include particularly harsh winter weather leading to less travel, but data was not readily available to verify any single cause.

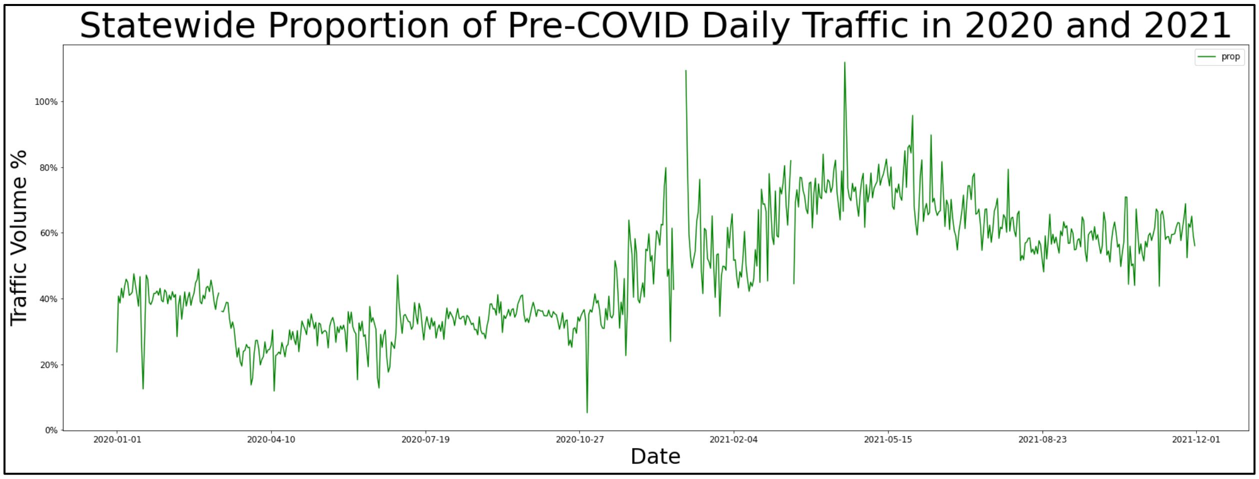


Figure 4: Enlarged view of statewide traffic volume during COVID years as a proportion of pre-pandemic volumes

At this point some patterns emerged like a steep drop-off in volume around March of 2020 when the pandemic began, and a gradual rise in traffic volume from then onwards as essential workers went back to work followed by more workers and others as they made choices about what mobility was acceptable and safe during the pandemic. Then a steep rise in traffic volume occurs around the end of 2020 as re-opening began in phases.

To verify those very broad trends in the pandemic, a secondary data source was identified. A series of specific policy events was gathered from an article by BallotPedia and then categorized manually into Restriction, Relaxation, Report, and Policy as shown in the sample of data in Figure 4. Report and Policy events included things like re-opening guidance documents from the Department of Health and policy votes to extend the COVID state of emergency which didn't have significant effects on social mobility. The main focus of the data was the Restriction and Relaxation events which were categorized by Announcement, Effective, and Extension dates. To limit the amount of events to process only Effective dates were examined which turned out to be 7 Restriction events and 9 Relaxation events.

Date	Category	Order Type	Event Type	Description
2020-03-15	Education	Restriction	Announcement	All schools K-12 closing 3/18-3/27
2020-03-18	Education	Restriction	Effective	All K-12 schools close
2020-03-25	Education	Restriction	Extension	Extended school closure to May 1
2020-04-23	Education	Restriction	Extension	Schools closed remainder of year
2020-07-30	Education	Relaxation	Announcement	Safe Learning Plan for 2020-2021
2020-09-10	Education	Report	Announcement	Burbio: Most schools using online learning
2021-02-17	Education	Relaxation	Announcement	In-person instruction to resume 2/22

Figure 5: Sample of policy event categorization

The last primary data source for this project was the raw counts of daily COVID infections and deaths published on Github by the New York Times. This data is available as counts by county, state, and country. Other levels of detail like COVID cases in prisons and colleges are also available, but not used in this scenario. Since the traffic data had already been aggregated to the state level, that was the only file needed. All that needed to be done to prepare the data was to filter it down to only date entries for Minnesota and separate the infection and death counts into different data series. Figure 6 consists of the plot from Figure 4 with vertical lines added for major Restriction and Relaxation events along with plots of the cumulative counts of COVID cases and COVID deaths for each day in 2020 and 2021.

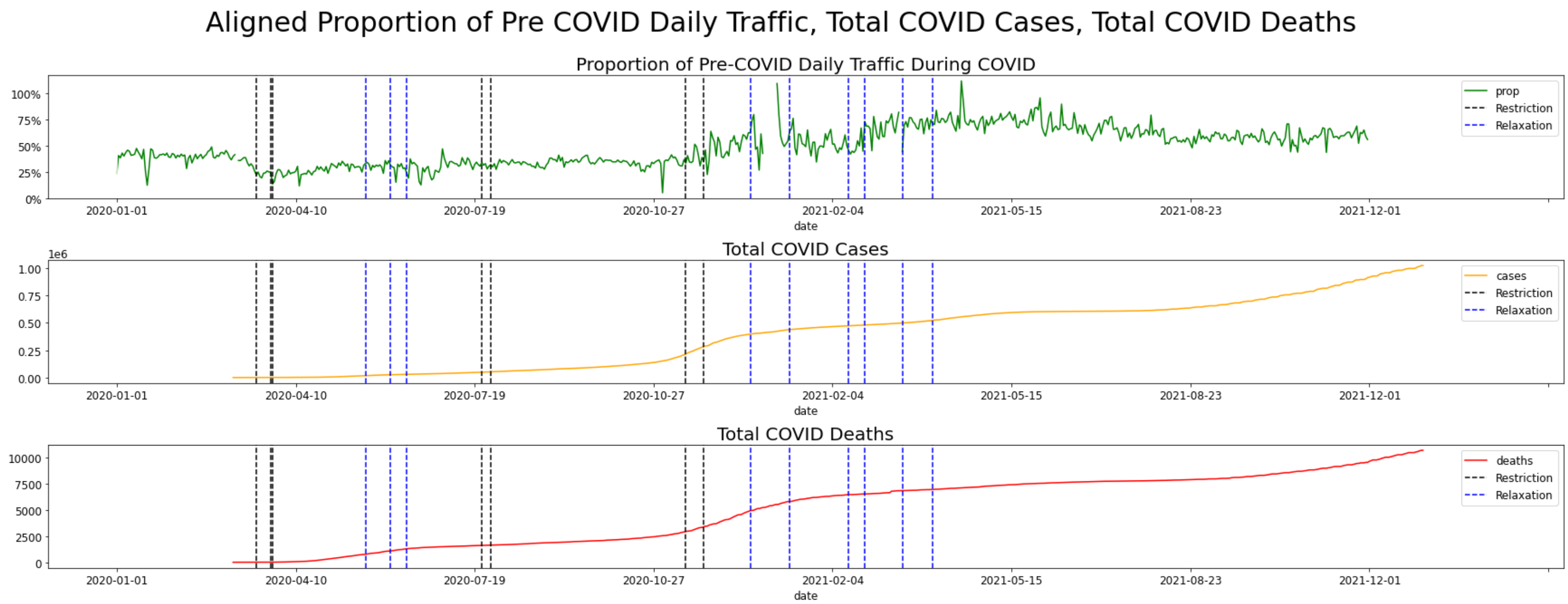


Figure 6: Aligned timelines of traffic volumes, COVID cases, and COVID deaths.

This aggregation of three key metrics in one figure along with some secondary markers allows for a clearer understanding of the entire pandemic timeline. The first set of policy restrictions indicated by dashed vertical black lines are the closures of schools on March 18th, closure of public businesses on March 23rd, and stay-at-home order on March 27th. A sharp drop in traffic volume appears immediately following those measures. It is interesting that the traffic volume was lower in 2020 before COVID cases reached Minnesota and restrictions were in place. Following the first drop-off and initial restrictions, traffic volume makes a steady climb until a series of relaxations with many businesses allowed to operate at 50% capacity. Surprisingly, the traffic volume decreases a bit following these changes. This may be due to the higher than usual indulgence in outdoor activities by Minnesotans following a long winter of being locked in. It is also interesting that the number of total cases and deaths was steadily rising during this first period of relaxation.

Next, in July of 2020 public schools closed again and the state mask mandate went into effect. No changes were made to business capacity, and only a slight decrease in traffic volume below the pre-COVID average was seen. Fall of 2020 saw an acceleration of COVID case and death counts, which prompted new restrictions. Private gatherings were limited by household and indoor entertainment venues like bars, restaurants, and gyms were closed. Following these restrictions, the rate of increase in total COVID cases slowed, but the traffic volume across the state increased rapidly with a peak near pre-pandemic levels. It isn't clear from the data why this happened, but it may indicate that this second round of major restrictions was not as effective as the first in getting Minnesotans to change their behavior to stop the spread of the virus. Once the infection rate plateaued near the end of 2020, businesses began to reopen at partial capacity and traffic volume increased again. On February 13th, 2021, the maximum capacity of restaurants and other indoor venues was increased to 250, and on

February 22nd, 2021, students were allowed back in-person. Following these changes, the traffic volume accelerated to a plateau around 80% where it remained throughout the remaining relaxations.

Data Analysis Conclusions

Qualitatively, the changes in COVID cases and deaths before restrictions went into effect seem to make sense. The changes at the end of 2020 seem like they did not have as much effect on people's behavior as traffic volume continued to increase after those changes. The late 2020 restrictions seem to reduce the rate of transmission as the COVID cases graph has an inflection point near the new restrictions on private gatherings and indoor dining. This might mean that people were still gathering in small groups and driving to do so, but not interacting with as many people in public. More data about where people were spending time might be able to fill in the gaps here.

Relaxations in general don't seem to be tied as closely to changes in COVID case and death counts. In fact, the first relaxation in June of 2020 is close to an inflection point in the death counts graph which means the policy was in place before the rate of increase in deaths had begun to slow. The second series of relaxations came after the infection rate plateaued in December. Following the relaxations, the infection rate did not increase significantly. It makes sense that there would not be any major decreases in infection and death rate following relaxations, and the fact that there are not significant increases in infection and death rate signal that the relaxations were safe to implement.

Machine Learning Model

Having completed a retrospective analysis of the relation between policy changes, traffic volume, and COVID cases and deaths, the primary goal of developing a tool that might be passed on to future elected officials for modeling the potential affects of policy changes was approached. To do this, a partial time series of expected traffic volumes percentages for 30 days prior to an event would be used to train a Facebook Prophet model which can predict the traffic volume proportion for the 10 days after an event. The effectiveness of the model was measured by the Root Mean Squared Error (RMSE) of the predicted traffic volume proportion for the 10 days after an event compared to the ground truth value.

Since the effects of Relaxations were less prominent in the qualitative analysis, only Restriction events will be tested. Four different dates in 2020 were chosen for testing: March 18th (stay-at-home order), July 25th (state mask mandate), November 13th (limitations on private gatherings and indoor dining), and November 23rd (halt on indoor dining and private gatherings). A baseline test was created by averaging the traffic volume over the years 2017 and 2018, then using the traffic volume data from 2019 with four randomly selected dates as events to see how well the Prophet model could predict in a typical year. The graph of traffic volume proportion in 2019 is available in Appendix C.

The results of the test were not promising. The plots in Figures 7 and 8 show that for 7 out of 8 tests in the baseline and COVID test cases only one did not sharply diverge towards negative traffic volumes. This is most likely due to the short training period for each model, which was intentional to model the frequency of changes in pandemic situations and the need to make policy decisions based on limited input data. A tool that can be trained on multiple series would be more helpful in the future since many training series of 30 days could be compiled for each county, district, or state in the country. The quantitative results captured in Figures 9 and 10 show that the baseline model performed much better than the COVID model thanks to its one successful prediction with an RMSE of 6% of traffic volume for an average test RMSE of 54% compared to the COVID test RME of 109%. Based on these results, a different time series modeling tool would need to be used for predicting effects of restriction policies at different times in a pandemic.

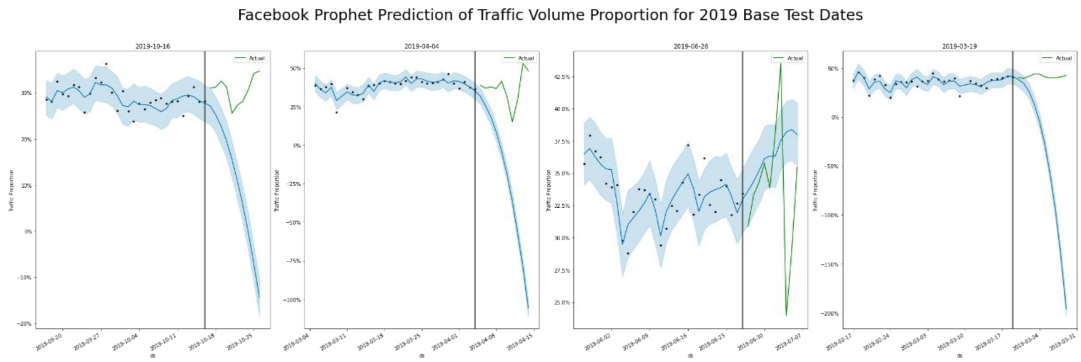


Figure 7: Prophet prediction for baseline tests

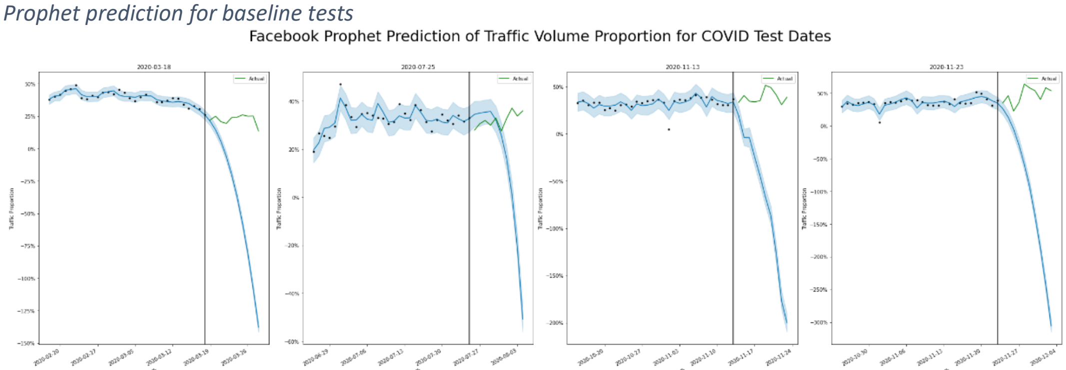


Figure 8: Prophet prediction for COVID test events

	train_rmse	test_rmse
2019-10-16	0.019980	0.249948
2019-04-04	0.026919	0.756229
2019-06-26	0.012664	0.059368
2019-03-19	0.039418	1.127075

Average Train RMSE: 0.024745229394833243
Average Test RMSE: 0.5481550936573452

Figure 9: Baseline test results

	train_rmse	test_rmse
2020-03-18	0.018858	0.808842
2020-07-25	0.027070	0.349142
2020-11-13	0.049890	1.321108
2020-11-23	0.050118	1.877342

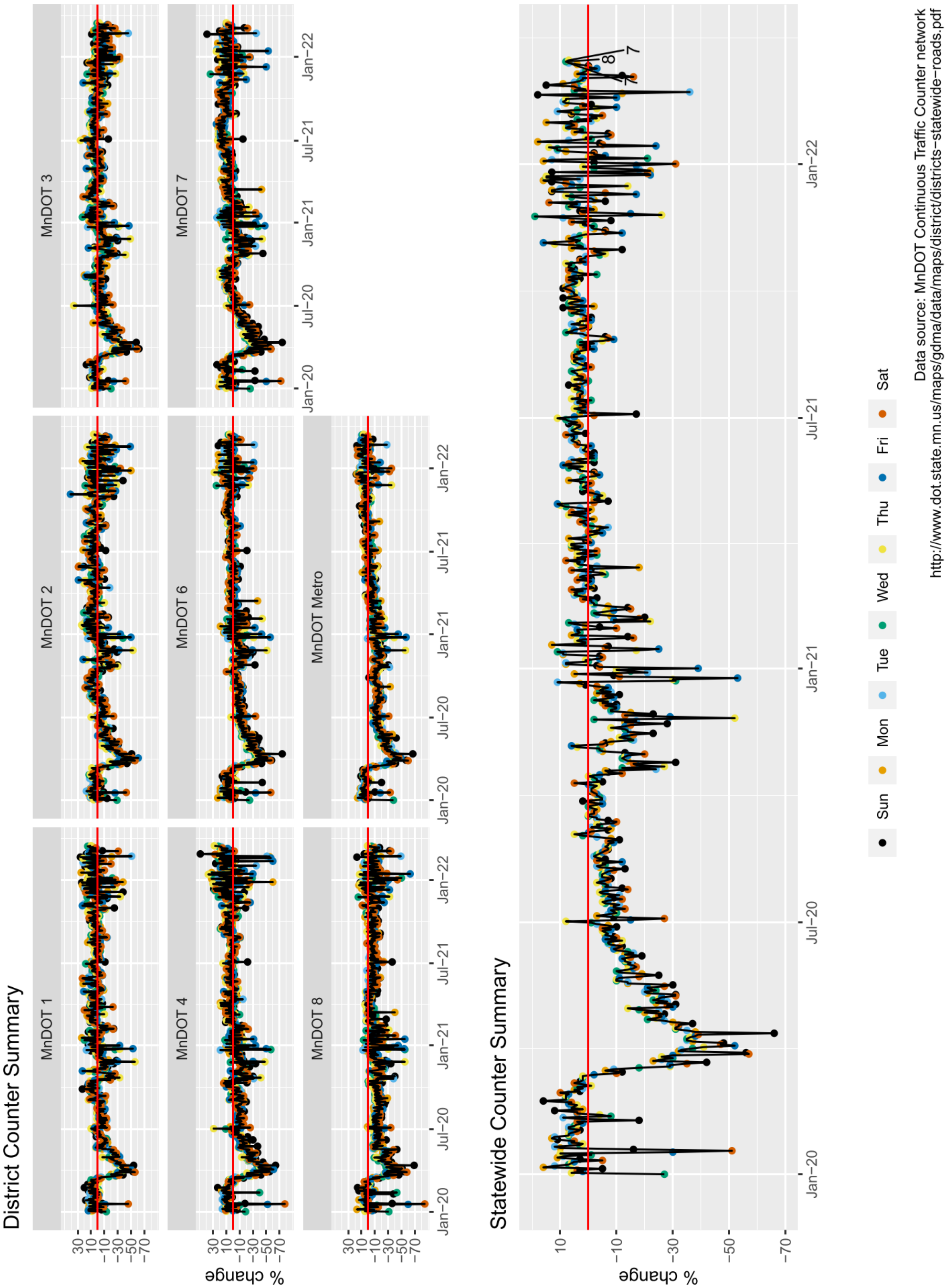
Average Train RMSE: 0.03648380780603501
Average Test RMSE: 1.0891085991999927

Figure 10: COVID test results

Lessons Learned

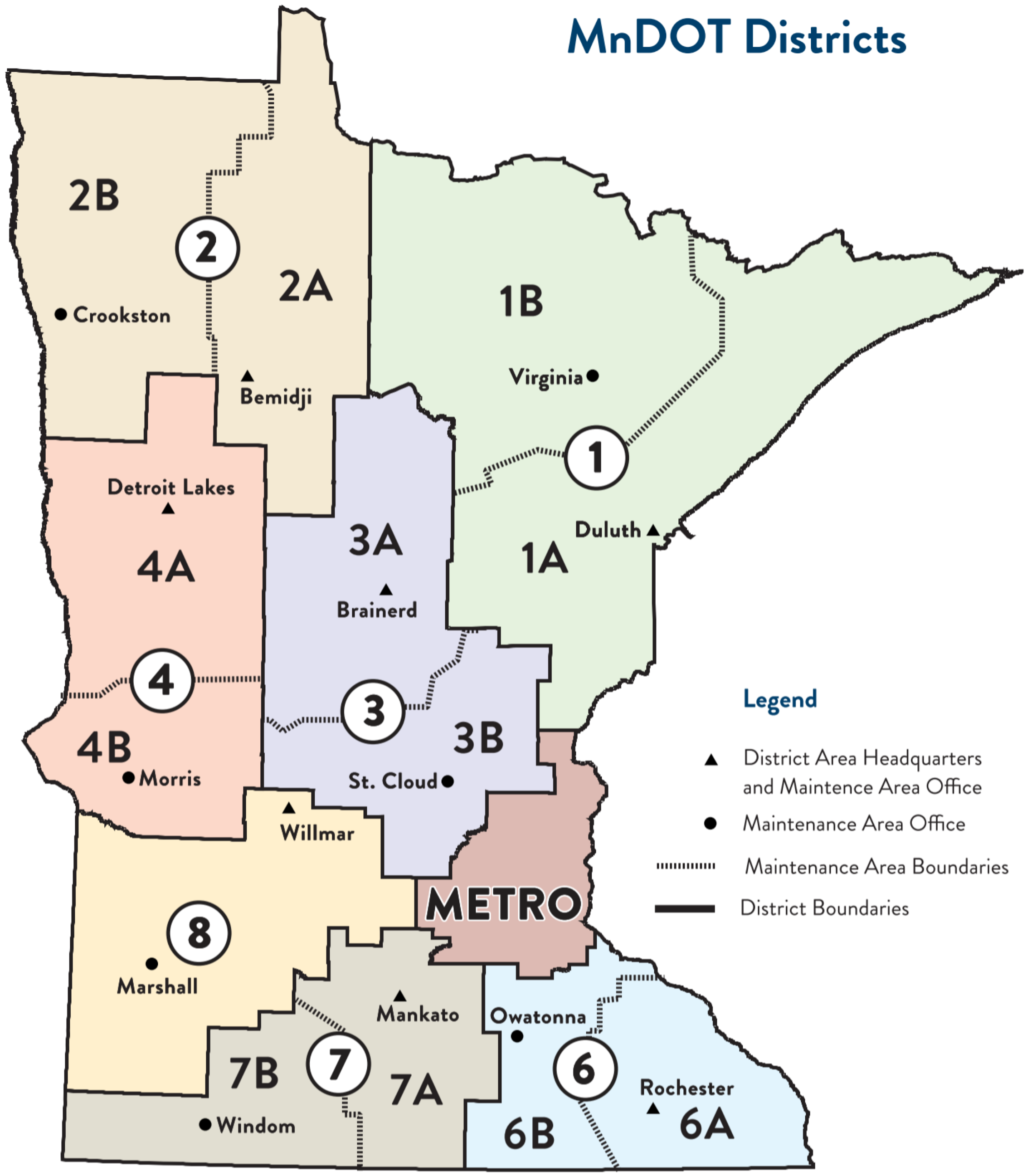
The primary lesson from this project is that data gathering, preparation, and visualization takes a LONG time. I was shocked at how quickly I could lose an hour just trying to decide the right way to filter and join DataFrames and especially how hard it was to get time series data into useful formats. The amount of data that is trimmed when doing preparation was also surprising as I started with millions of rows of fine-grained county and weigh station measurements which ultimately got aggregated into a little over 700 rows for the two-year COVID timelines. This made it hard to build a strong machine learning solution, and in future projects I think having that end goal of a model that is trained on enough data to be accurate would change the way I prepared the data a lot. In the NoSQL project I stated that fitting the tool to the problem and not the other way around was an important lesson. However, the opposite was also true in this project. Had I known that Prophet could only be trained on a single time series I would have looked for other models that can learn from many short time series. Lastly, having a clear definition of completion is important. Had I set clearer goals at the start of the project I would have been able to recognize milestones and stay on track to not need this late extension and penalty!

Appendix A – Original COVID Traffic Volume Comparison (https://edocs-public.dot.state.mn.us/edocs_public/DMResultSet/download?docId=12227832)



Appendix B – MN District Map (https://www.dot.state.mn.us/information/docs/district-map-with-sub-areas.pdf)

DEPARTMENT OF TRANSPORTATION



MnDOT office contact information can be found online at: **www.mndot.gov/information/locations.html**

Appendix C – 2019 traffic volume proportion for baseline testing