**Deliverable 3: Project Progress Report**

Cross Zeigler  
Ewin Hong

Neethu Battula  
Sagar D. Goswami  
Suchada Hapikul

George Mason University

AIT 614: Big Data Essentials

Dr. Liao

April 24, 2022

**Deliverable 3: Project Progress Report**

**Re-Introduction**

New York City (NYC) is notorious for having heavy traffic at nearly all times of day, heavily impeding motor traffic in the city even at the best of times. This often leads to congestion which impacts travel and causes delays, wastes gas, and can leave people feeling dissatisfied with personal and public transportation. A prominent question for NYC, then, is how to improve the traffic infrastructure within NYC, leading to changes and improvements in transportation, pollution levels, and more.

The main goal of this project was to study one or more datasets of traffic patterns of NYC and to create a model that predicted when traffic occurs the most hourly, daily, weekly, monthly, and annually if possible. The thought was that this could help to consolidate which times of the year are “problem” times and allow people to adjust their commute accordingly before major infrastructure changes are made. Another thought the authors had was this may even help change traffic patterns based on people’s ideal personal schedule, allowing someone to leave right on time to make traffic as smooth as possible for them and others. It became more important than ever this year to conserve gas not just for the environment but for personal use too as the cost of living, including gasoline prices, rises.

The main motivation for completing these goals is to give insights into the NYC traffic system to give suggestions to improve it. Improving everyone’s personal schedule is a small stepping stone to accomplish this goal, but that is useless and a short-term solution without improving the actual infrastructure of NYC. These changes could lead to more pedestrian-friendly roads and easier access to public transport, reducing the number of cars on the road over time. To accomplish this, the experimenters attempted to look over the specific areas of congestion along with the times there is little to high congestion. The thought was that by looking over the areas of high congestion they could find a way to reroute traffic to make those areas less busy. Looking over times of congestion helped us to see how traffic changed over time and allowed us to theorize why these changes occurred due to holidays, construction, or other events. The end goal was to advocate for public policy to refine infrastructure.

Several audiences could benefit from this analysis. The primary audiences are the stakeholders in the urban transportation industry and transportation agencies. For both, if people have an easier time moving around, they may be more inclined to utilize public transportation, which could drive revenue up. The latter in particular may like this, as they have a lot more to gain directly from people using their services more.

Two other audiences are emergency services and NYC residents. Ambulances and firetrucks need to get to their destinations quickly to maximize the amount of lives they save and the chances of saving said lives. With NYC as it is, even if people try to maneuver out of their way, they still have to deal with traffic regularly that could impede their progress. NYC residents would not just appreciate this ease of access emergency services have to save their lives, but also in day-to-day life. Someone may struggle to get to work on time or meet up with friends if they constantly have to battle heavy traffic. Changing infrastructure could help improve satisfaction with living in NYC, improve work ethic and punctuality, and lead to increased happiness with one’s social life by ease of access alone.

Attempting to improve NYC’s traffic situation is not new. People have studied this issue extensively in hopes of understanding *why* traffic is so bad in the city and how to improve it over time. Nibareke and Laassiri (2020) used various machine learning models to model traffic flow over time to predict traffic effectively. While they specifically dealt with air traffic their model allowed people to see how accurately one could predict delays and traffic in transportation with the correct model. Vasudevan (2016) presented a technical approach that combined Apache Spark’s open-source data analytics and machine learning techniques to predict traffic flow patterns using simulated connected vehicle messages. The study reported that connected vehicle data can be processed rapidly using Big Data analytics to generate precise predictions of traffic flow regimes. Other researchers reviewed had similar results. As the project progresses, the authors are likely to research more relevant topics to help with their own models.

**Implemented Solutions, Preliminary Results & Analysis**

The researchers implemented several solutions to help study NYC’s traffic patterns in order to provide feedback on how to improve the transportation infrastructure of NYC. This included a K-means clustering model, time series models, a decision tree, and Multivariate Adaptive Regression Splines (MARs).

The researchers performed a K-means and Principal Component Analysis on the dataset to determine a cluster analysis for the 139 link points by grouping the hour of the day relative to average speed. A single link point has 24 features to represent each hour of the day and an average speed respective to the hour. The authors performed a clustering evaluation by performing two to ten clusters and based on the cluster evaluation chart (Figure 1), the selected k was three clusters. With three clusters, a scatter plot was generated but did not provide significant insight (Figure 2). The authors decided to correlate the clustering results back to the features of the dataset. In performing this cross analysis, the smallest cluster, cluster 1, was determined to be the slow cluster where the average speed was consistently below an average of 30 miles per hour for 14 link points. Figure 3 illustrates the various areas of concern such as those close to bridges and within the Manhattan borough. Figure 4 provides a contextual representation of the average range of speeds for the 14 link points in cluster 1. Figure 5 demonstrates the average speeds of consistently faster link points to have a larger range where the interquartile range starts at 30 miles per hour.

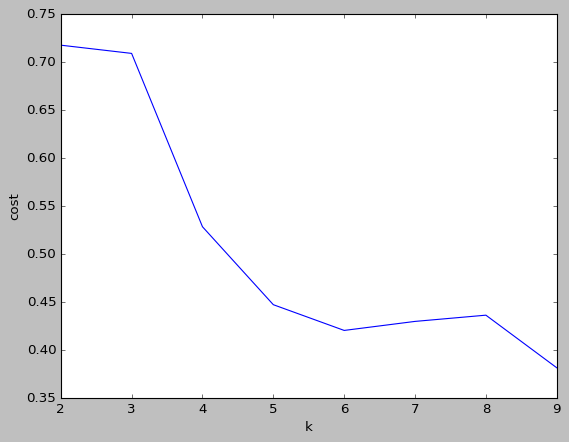


Figure 1: Clustering Evaluation

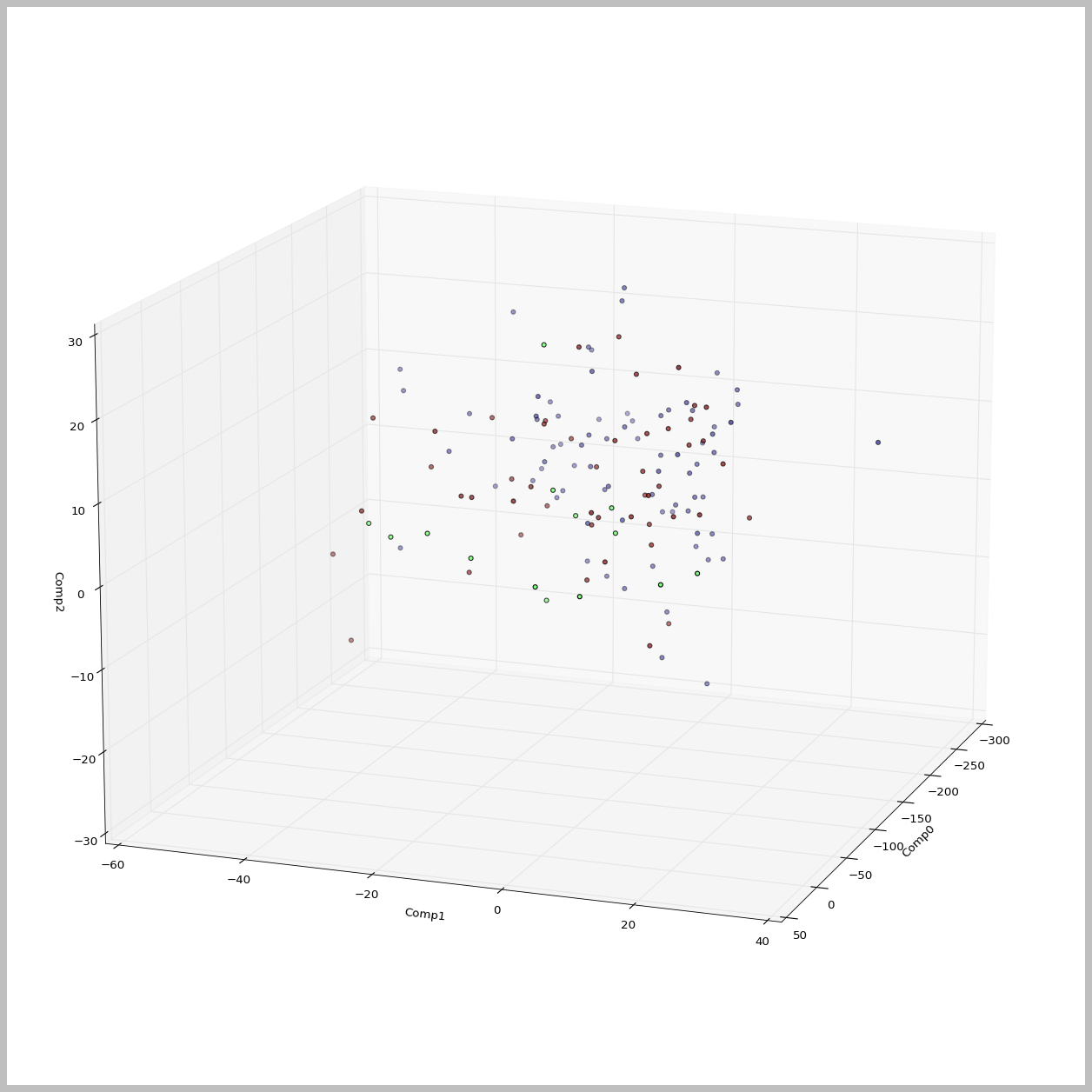


Figure 2: K-means cluster and Principal Component Analysis

Map

Description automatically generated

Figure 3: Map with Cluster 1 link points

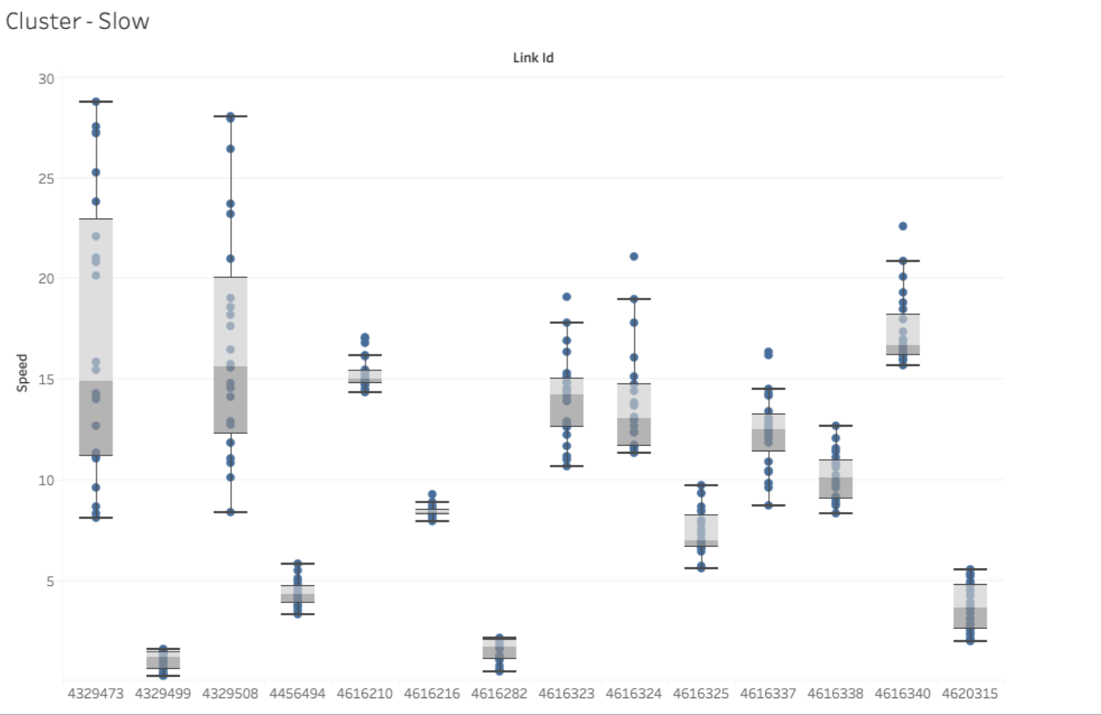


Figure 4: Cluster 1-Speeds below average of 30 mph

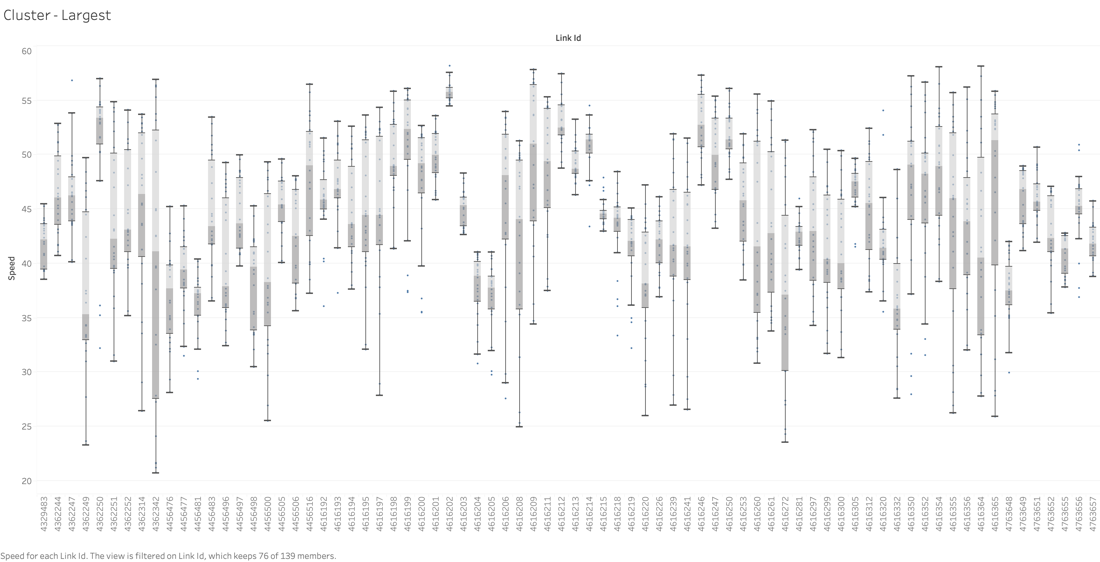


Figure 5: Largest Cluster

The next analysis the researchers performed was a Time Series Analysis of daily to annual plots. These plots typically visualized average speeds throughout the select time period and gave researchers insight into how the speed shifted over time. Some plots even sought to predict the speed into the year 2024 and the variation this observers may see in said speed. For the daily plot (see Figure 6), the observable trend shows that the speed drops dramatically from 4:00am until 8 or 9am. The speed then levels off for an hour until viewers see another drastic drop from 10am to 4pm. After 4pm, however, the average speed rises steadily until 4am. This may be an indication of how traffic may clear up during the afternoon and night, leading to people being able to drive faster with more desolate roads.

For the weekly plot (see Figure 7), the average daily speed decreases from 40 mph to 37.5 from Monday to Friday, and then increases on the weekend up to nearly 42 mph. This could be due to traffic becoming heavier during the week as more people go to work and may need to travel the same routes, while the weekends see a reduced number of cars on the road since there may be less people in need of motor transportation.

Months (Figure 8) demonstrated that generally, the average speed in a month stayed between 38 and 40 mph until a sudden jump on the 24th day to 41 mph and then a severe drop to below 38 mph after the 25th. Why there’s a sudden jump and then drop is unknown, but there could be several reasons. Many holidays may occur towards the end of the month, leading to people staying inside for a bit before going out to shop or for recreation. The data may also be heavily influenced and skewed by the 2020 data since there was a notable jump in average annual speed for 2020, but this likely should have affected the whole month, not a select few days.

For the annual average speed (Figure 10), this was consistently between 35 and 40 mph with the exception of the year 2020. The average speed during this year was 42 mph. During 2020, the COVID-19 pandemic hit the U.S., leading to shutdowns that left roads clearer than ever before. As a result, people likely drove faster, leading to a higher average travel speed. As lockdowns let up in 2021 and 2022, the average speed decreased to between 35 and 40 mph.

Figure 11 shows how the time-series model segregates the overall variations into trend, yearly, weekly, and daily patterns. The trend can be observed fairly straight, with a bump in speed during 2022 (possibly due to lockdowns imposed during COVID-19 pandemic). The daily and weekly components correspond to the Data Exploration graphs, as they follow a very similar pattern. The yearly component however was observed to have a lot of variations, which suggests a lack of discerning pattern overall. However, it can only be confirmed once more data is analyzed, as some of the data might be outliers, caused by COVID-19 pandemic.

Chart, line chart

Description automatically generated

Figure 6: Daily Time Series Analysis

Chart, line chart

Description automatically generated

Figure 7: Weekly Time Series Analysis

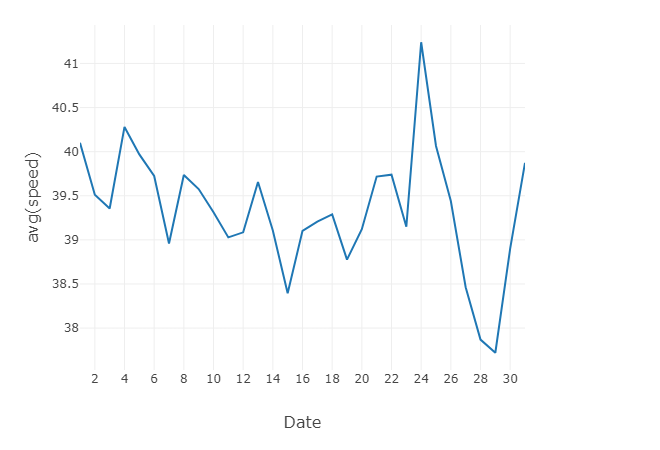


Figure 8: Monthly Time Series Analysis

Chart, line chart

Description automatically generated

Figure 9: Monthly Time Series Analysis by Year

Chart, line chart

Description automatically generated

Figure 10: Annual Time Series Analysis

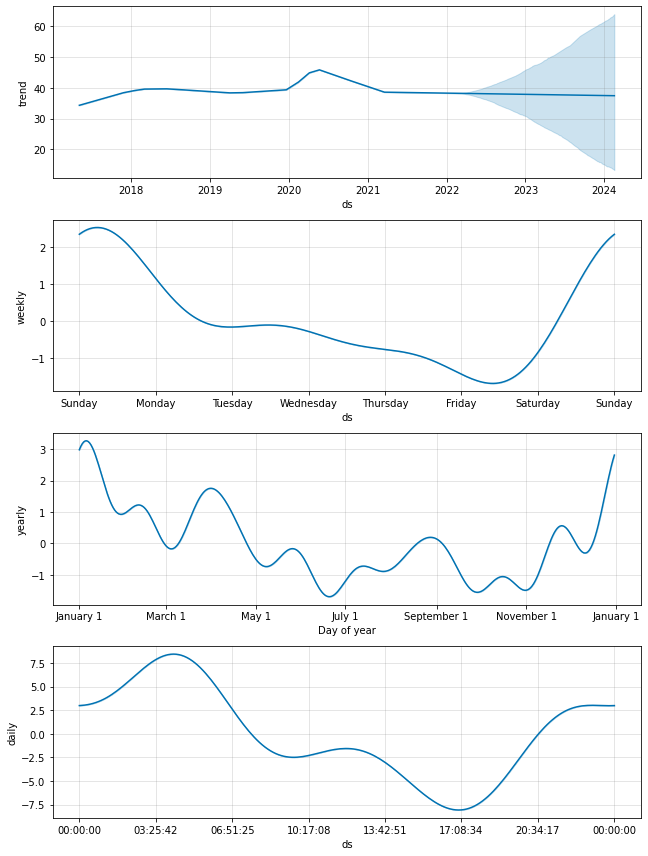


Figure 11: Time-Series Model (Component Plots)

**Overall Progress**

At the moment, the primary goal is mostly on track. The experimenters have nearly completed the proposed solutions and are going over their final analysis of the solutions’ results to provide proposals to address NYC's traffic concerns. There have been notable deviations from the project’s original scope, including a shift in goals and solutions to implement.

One of the first changes the experimenters made to the project’s scope was to the solution implementations, specifically the predictive dashboard. The original scope of the project included a dashboard that could predict traffic up to a week or more in the future based on the past few weeks of traffic data, potentially pointing out the events which led to drastic changes in speed and congestion to an area such as holidays, accidents, or new construction. This would be accomplished by pulling data via an NYC traffic API and would have been one of the aforementioned short-term solutions by affording everyone a way to study congestion for their immediate area. This would have allowed drivers to plan their driving trip to adjust how they moved throughout NYC, changing their departure times and mode of transportation. However, the authors didn’t implement this solution for several reasons.

One, other solutions were prioritized since they gave us greater analysis tools to show to the stakeholders to influence policy. This analysis was mainly meant to provide an insight into historical NYC traffic patterns, with predictions being made to emphasize the problem of NYC traffic. A personal predictive dataset may have been useful to the average person, but it would have been a short-term solution that may not have been as accurate as desired. When designing a system meant to be used in real-time by a large population, accuracy is key, as it can mean the difference between life and death in some cases.

Additionally, there wasn’t enough time to implement the API dashboard. Again, other solutions were prioritized that were A) more important to the goal, and B) realistic to implement. The other solutions provided better results to analyze and come to conclusions on NYC traffic, and were much simpler to implement. They also provided the most stable outputs and ways to give analysis, as each solution wasn’t subject to data that updated so constantly and in such a large volume relative to the entire given data that the output would change drastically with several or fewer changes in the data.

Another change in scope was for some parts of the geospatial analysis. The authors originally considered several other columns that may have allowed them to see the direction of traffic flow to consider as preliminary geospatial analysis. One of these columns were called Link Points, with each link point having a list of latitude and longitude points. However, a large issue in analyzing the link points was that link points for one direction were usually identical to the opposite direction. If this issue could be resolved, then the authors could have performed some analysis in which directions traffic flowed fastest in and could have predicted why; i.e. at certain points of the day people may be going to and from work, leading to heavier traffic in either direction. Unfortunately, there was no known reliable way to analyze the data, and instead the researchers focused on other proposed solutions. It was considered too costly to implement the geospatial analysis within the project’s timeline, so other solutions were prioritized. This may be a future point of study to consider.

During the study, the authors experienced minor hurdles in understanding certain implementations of the proposed algorithms and understanding the new platform. Databricks was the selected collaborative platform, however the authors still needed to become familiar with its capabilities. As such, it took some additional time to learn the full potential of Databricks to work together on the project. The authors also had to spend some time setting up GitHub repositories, but most if not all of them were familiar with the platform, so this delay took less time due to varying levels of git experience. Data cleaning was time-consuming so the authors performed preliminary analysis on uncleaned data to continue progressing forward while awaiting the data cleaning.

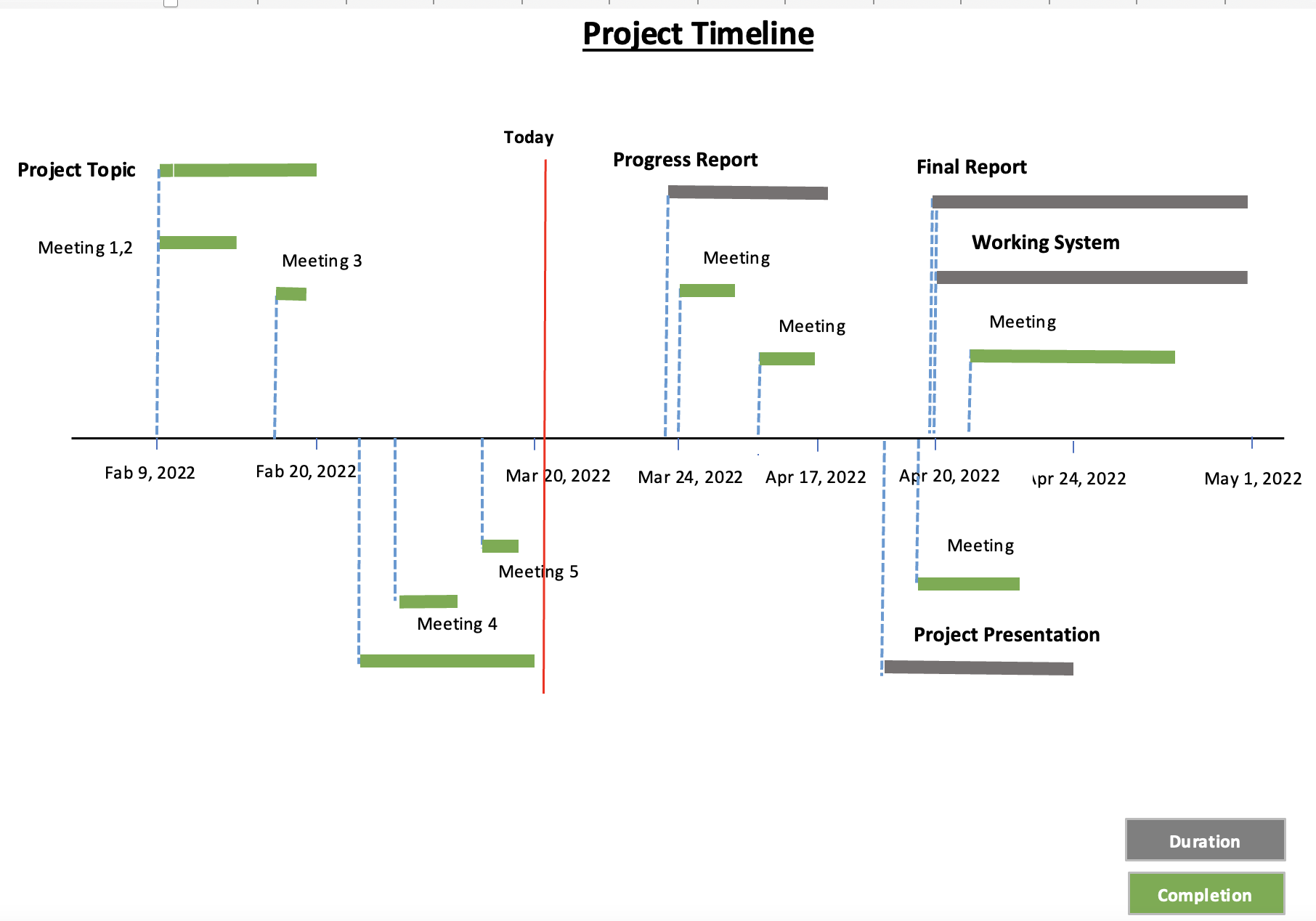
Several changes in tools were met with replacements for any technology dropped. Those replacements were specifically the decision trees and Multivariate Adaptive Regression Splines (MARS). While the original scope was going to rely on a simple linear regression model, the authors decided to replace it with the two aforementioned tools along with the time series analysis. This way they could have a more in-depth analysis by utilizing multiple ways to see solutions to the problem. The authors used all of these methods in particular to predict what traffic, specifically speed, would look like in the future to give recommendations to stakeholders on how to handle transportation infrastructure.

The researchers chose to implement decision trees to see how speed may change at a certain point in time during the day, week, month, or year. This way they could gauge what the most accurate prediction for future speeds was and make a more informed guess. The researchers chose to run a MARS algorithm for the same reason as well.

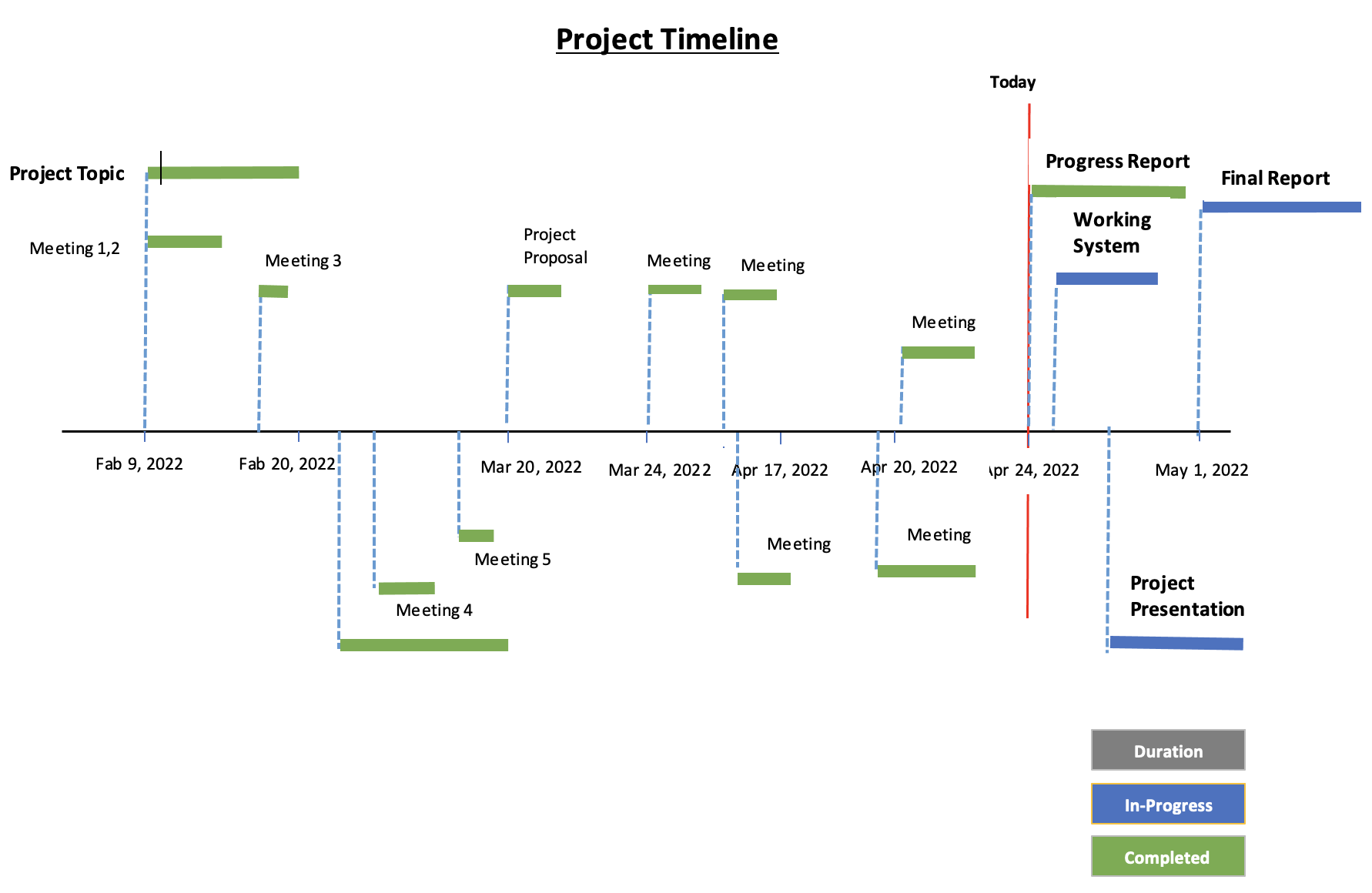
**Update the status of the project tasks and timeline**

For overall completion, only several main tasks left, including the presentation, final report, documentation, and implementing a working system. Currently the main focus is the presentation at the moment as that is due next, but have begun to shift focus to the documentation and final report due to how time consuming both are. As the project progressed, the researchers also slowly shifted their goals and how much time was dedicated to each one. Below are charts of the original and updated plans:

Original Plan:

~~~~

Updated Plan:



Works Cited

Nibareke, T., & Laassiri, J. (2020). Using Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00355-0

Vasudevan, M. (2016). *Big data analytics: predicting traffic flow regimes from simulated connected vehicle messages using data analytics and machine learning.* US Department of Transportation. Retrieved March 14, 2022, from https://rosap.ntl.bts.gov/view/dot/32616