**Deliverable 2: Project Proposal**

Cross Zeigler  
Ewin Hong

Neethu Battula  
Sagar D. Goswami  
Suchada Hapikul

George Mason University

AIT 614: Big Data Essentials

Dr. Liao

March 20, 2022

**Deliverable 2: Project Proposal**

**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 traffic often leads to congestion that can delay travel and cause people to be late, waste gas, and feel dissatisfied with personal and public transportation. One of the prevailing questions for NYC, then, is how to improve upon the traffic situation there? What changes in architecture can people make to increase the satisfaction with transportation in the city amongst other qualities like improved safety and decreased travel times. There are also other issues that can be solved by solving traffic issues, such as pollution levels, improving public transportation, and more.

The main goal of this project is to study one or more datasets of traffic patterns of NYC and to create a model that predicts when traffic occurs the most hourly, daily, weekly, monthly, and annually if possible. 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. Perhaps this could 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 is becoming more important than ever to conserve gas not just for the environment but for personal use too as the cost of living, including gasoline prices, rises.

There are a few additional goals as well. One is to create a predictive dashboard that updates weekly or monthly with data from an API tracking NYC traffic. This dashboard would display the most current data from the past few weeks and could tell users how traffic patterns may have changed over this time. Perhaps a holiday occurred which people traveled extensively on, leading to higher volumes of traffic. Maybe a particularly bad accident occurred that hindered traffic flow for a few hours on a particular day. Anything from construction to parades can impact traffic severely, leading to increased congestion that the average person must deal with. The dashboard could also serve to figure out where or when bottlenecks occur based on the average traffic speed for the current or neighboring locations. Knowing where traffic is coming from is important to determine where it is going as well. Perhaps one area is congested for several hours in a day and then moves slowly to another area. If people knew how the traffic shifted in their area, they would not need that strict of a schedule to ensure they got to their destination on time.

Another goal is a broader geospatial analysis. Our initial analysis focuses mostly on *when* traffic is heaviest, not *where*. The where is just as important as the when, as a general idea of when traffic is heavy in NYC only helps so much when one is unaware of how traffic works in your specific area. Traffic may be heavy in one district at one hour of the day and light in another district in the same hour, so living in either area would heavily affect how one would approach getting to your destination during the same time. Having a geospatial analysis, perhaps in the form of a heatmap or several graphs based on the location for the dashboard, could help even more with specific situations.

Prediction dashboard (figure out bottlenecks by feature extraction of the average traffic speed for neighboring locations), clustering locations and sensors, geospatial analysis (see where most traffic is by region), heatmap/bubble map

The main motivation for accomplishing these goals is to help give suggestions to better NYC’s infrastructure. A minor stepping stone is improving the personal schedule of everyone, but adjusting the overall infrastructure could help to improve everyone’s situation without needing them to constantly check for traffic in one of the most congested cities within the U.S. These suggestions would not only improve traffic and congestion, but also help with pedestrian safety and ease of access for public transportation. Maybe the changes in infrastructure lead to more pedestrian-friendly roads and help people access public transportation easier, leading to less cars on the road over time. To help with this, we would look over specific points of congestion and how many people within an area use public transportation. The first point could be used to reroute traffic depending on where it is coming and going, and the second point could be used to determine if more public transport is needed in a district and what routes that vehicle should follow to maximize access to it. Overall, we would be advocating for public policy to refine infrastructure.

Many audiences could benefit from these goals and motivations. We have targeted several audiences from individuals to organizations. Our 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.

Our 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.

The desire to improve NYC’s infrastructure and traffic issues are not new. They have been studied extensively by dozens if not hundreds of people 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. They specifically dealt with air traffic, but 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 our own models.

**Project Goals**

The primary goal of this project is to provide valuable insights into the city's traffic system to improve transportation within NYC for increased satisfaction, safety, and decreased travel times. There are several focus areas covered in the study. The first focus area is the actual data we are dealing with and what we hope to extract from it, that being the times and areas with the greatest amount of congestion. As a result of this study, the authors hope to provide the aforementioned audiences with this data so they can either improve their own commuting schedule and potentially increase their own safety or to affect policy to improve infrastructure in NYC.

The second focus area is creating a machine learning model based on previous NYC average traffic speed data. This is sourced from the New York City Department of Transportation and can help determine times/dates and locations that observe a relatively high amount of congestion. The aim is to make this model as accurate as possible to determine how traffic works in NYC, and hope that the aforementioned dashboard weekly update can assist in this task. Perhaps this machine learning model could have an API similar to the dashboard or even adjust its data based on the board’s output, adjusting itself based on the conclusions the board “makes”. It would be dynamic, almost like a person was manually updating the model weekly for audiences. The authors would monitor to ensure both the board and model were working as intended, but hopefully the maintenance would be minimal.

**Project Requirements**

The project has several functional requirements. The requirements for analysis include several tools from programming languages to visualization tools. One such tool is Excel. The primary use for Excel at the moment is basic exploratory analysis and getting a feel for what the data is. Figuring out what data actually represents traffic and the geospatial data is important to determining what variables need to be used in the advanced analysis. Our team has done basic visualizations of the data to solve these problems and looked towards other visualizations of traffic in NYC to figure out if the data matches up alongside becoming familiar with good visualization techniques. In the future, we may use it for our finalized visualizations as a more direct way of presenting our data, but for the moment we are mainly using Excel to get an initial feel for the data. We may also use it to clean the data as well if we do not use the following tools to do so.

Our next tool is R. R is good for doing various analyses from linear regression to Chi-Square tests. This combined with its ease of use is vital to the health of our project, as we need a tool with a good amount of complexity combined with a low entry bar to help with our initial analysis. We will likely be using R for both data exploration and deeper data analysis as well due to it affording us the proper tools to do so. It may not be a great tool, perhaps useless even, for other tasks such as creating the dashboard and streaming data. R also is not as good at visualizations as Excel or other tools are, but it does provide clear enough graphs and outputs that help to create a data story for projects.

The last main analysis tool we plan to use is Python with specific libraries such as Pandas. Python combines many of Excel’s and R’s strengths from good visualizations for data stories, a great breadth of data analysis tools, and a low bar for entry, especially if one is familiar with Java or C. The language has many libraries that assist with these tasks, including the aforementioned Pandas which helps with extracting, processing, and cleaning data. As we move further into the project, Python will likely be our main tool, especially in the final deliverable given its heavy usage with APIs and its ability to weave beautiful data stories.

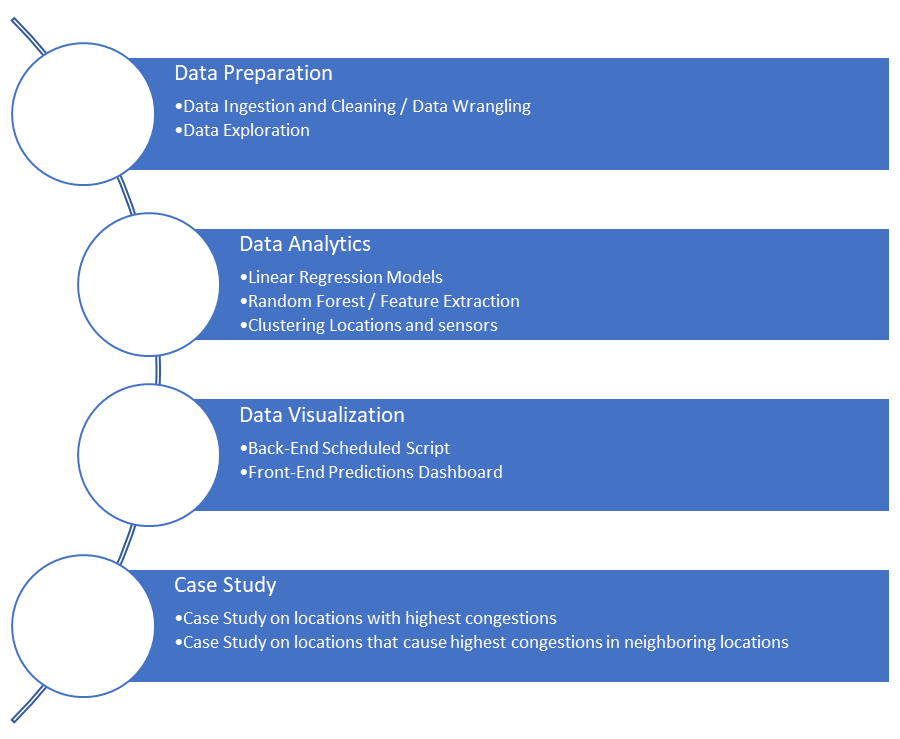
Our main NoSQL database will likely be MongoDB. This is the database most of our group members are familiar with and supports the use of R and Python. We will be using Apache Spark for similar reasons.

**Proposed Selected Dataset**

The proposed dataset comes from the New York City real time traffic speed found at: https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9. As of March 9th, 2022, the dataset has 58.8 million records starting on April 17, 2017. The dataset is continuously updated with real time data being provided by the traffic sensors. There are 13 features listed found as Table 1 in Appendix. With these 13 features, eight features are useful for our research. These eight features are id, speed, travel time, data as of, link points, owner, borough, and link name.

Speed in this dataset is the average speed between all of the link points. Travel time is time spent through the link points. The link points are a group of latitude and longitude points. The link point can be used to calculate the distance of the starting and ending points of a single link. Additionally, distance could be calculated by using speed and travel time as means of validating the calculated distance by link points. These link points can be evaluated over time and their evolution. Then with the link points, the Boroughs can be added for grouping for additional analysis. Each link point has at least 2 points and could go to 10 points. Due to the dataset, there is some evidence there might be more than 10, but are cut off due to the original dataset’s database type constraint of 256 characters.

**Description of Proposed System**



1. Data Preparation
   1. Data Ingestion and Cleaning  
      24 GB worth of CSV data is fetched from the NYC Department of Transportation open data portal. The file will be imported as a data frame inside a python module running on the Databricks portal. Additionally, NYC DOT also has an API that enables fetching live data regarding traffic information in a batch as well as streaming data format. This allows performing streaming data analytics on the traffic congestion data. Later, records will be filtered to remove any null/garbage data. Appropriate data wrangling steps will be carried out to bring the data in a tidy format.
   2. Data Exploration  
      Basic data analysis will be carried out to understand the scope of advanced analytics on the data. Locations with lower average traffic speeds will be identified to concentrate on the most affected regions. Additionally, basic visualizations will be performed to understand the data distribution and range better.
2. Data Analytics
   1. Linear Regression Models to predict Congestion  
      Machine learning models will be trained to perform linear regression and predict the frequency, interval, severity, and location of the congestion for a given time. Multiple Regression models will be trained to predict the congestion a few hours, days, weeks, months, and years ahead. These models can provide pivotal insights to understand the nature and seasonality of occurrence for congestion. Further, even GWR (geographically weighted regression) can be performed to train models and make predictions based on latitude and longitude information.
   2. Random Forest / Feature Extraction  
      To study the effect of congestion at one place on neighboring localities, the Random Forest technique can be used to identify locations that are potential bottlenecks or chokepoints. A dependency map can be created based on the most important features.
   3. Clustering locations and sensors  
      Clustering can be performed on the data points to group the traffic speed sensors with similar proportionate delta values for change in average traffic speeds. Such clusters can provide insights regarding the directional flow of traffic for a given time and location. Features like Average Traffic Speed and Congestion Duration can be used for clustering.
3. Predictions Dashboard
   1. Back-end Scheduled Script  
      A back-end script will run on a schedule to fetch the data from the department of transportation website, run machine learning models based on the new data, and generate new predictions. The coefficients for the new model will be logged and stored on a file. Even Near-Real time Predictions can be made using Spark, instead of scheduled batch processing.
   2. Front-end Predictions Dashboard   
      A dashboard will be created that predicts congestion based on the latest model and data. Various predictions will be made to predict congestion in a few hours, days, weeks, and months in the future.
4. Case Studies
   1. Case study on localities with the most congestion  
      Locations with the highest congestion will be identified in the Linear Regression stage. Literature Review will be carried out for these locations to validate the findings and understand the independent factors that cause congestion.
   2. Case studies on localities with the highest dependency for congestion in neighboring localities  
      Locations that have the highest dependency for congestion in neighboring localities will be identified in the Random Forest stage to understand how a particular location causes congestion in surrounding locations. Again, a literature review will be performed to validate the findings and outcome of using this method and determine the factors that cause such situations.

**Proposed Development Platforms**

Hardware/OS By Person:

* Cross: Mac 8GB RAM, 3.1 GHz, Dual-Core Intel Core i-5
* Ewin: Mac 6core CPU at 3.2GHZ Intel i7 64GB of ram
* Fon: Mac M1 chip with 8-Core CPU and 7- Core GPU
* Neethu: Mac 16GB RAM, 2GHZ,Quad-Core Intel Core i5
* Sagar: Windows Machine: 16 GB RAM, AMD Ryzen 4800H 8-cores 16-threads, Nvidia RTX 3060 Mobile GPU

Platforms & Software:

* Databricks (Apache Spark)
* Python, version 2.7 or above
* R, version 3.0 or above

**Project Tasks & Deadlines**

Our project involves several main tasks and subtasks as listed below:

**Data Preparation:** To prepare our data, we’ll focus on extracting it from the source then cleaning and exploring it. This will run from approximately 3/01/2022 to 3/20/2022.

Members Responsible:

* Ewin, Fon, Sagar

**Data Analytics:** After preliminary research, we’ll conduct deeper research involving linear regression modeling and clustering locations and sensors in NYC amongst other tasks. This will run from approximately 3/20/2022 to 4/10/2022.

Members Responsible:

* Cross, Fon, Sagar, Neethu
  + Linear Regression Model: Cross, Neethu
  + Clustering: Sagar

**Data Visualization:** For our data visualization, we plan on having a back-end script that extracts data from the API weekly or monthly to load to the front-end dashboard with traffic predictions. This will go on from around 4/05/2022 to 4/15/2022.

Members Responsible:

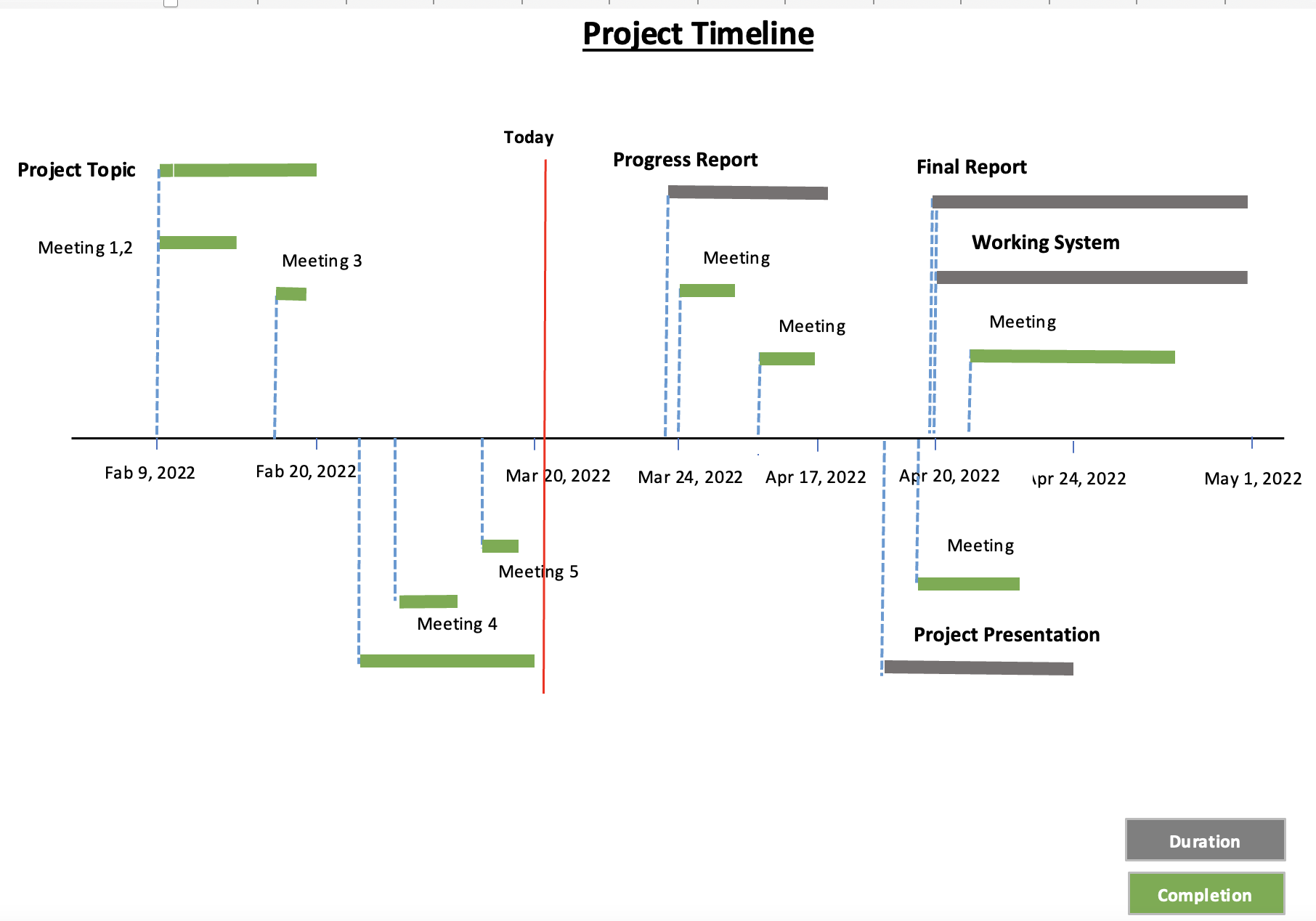
* Everyone
  + Back-End: Everyone
  + Front-End: Everyone

**Case Study:** The case study will answer the question of what areas and times have the highest amount of congestion. The group will study the case study from around 4/11/2022 to 4/20/2022.

Members Responsible:

* Everyone

**Deliverables:** The deliverables are the final parts of the project we will need to turn in, from the presentation, working system and final report. We’ll make the deliverables between 4/17/2022 to 5/01/2022.

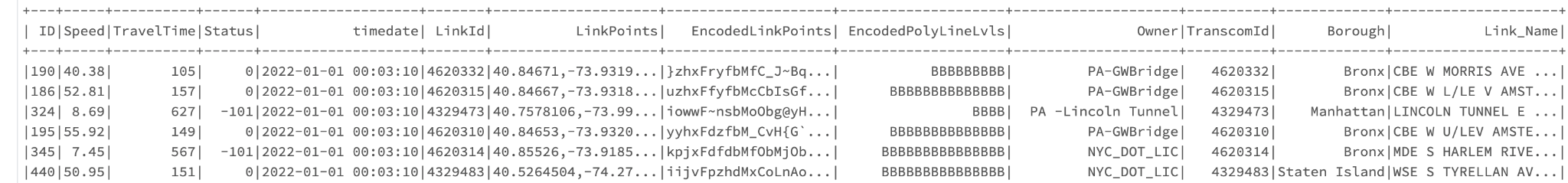
~~~~

**Appendix**

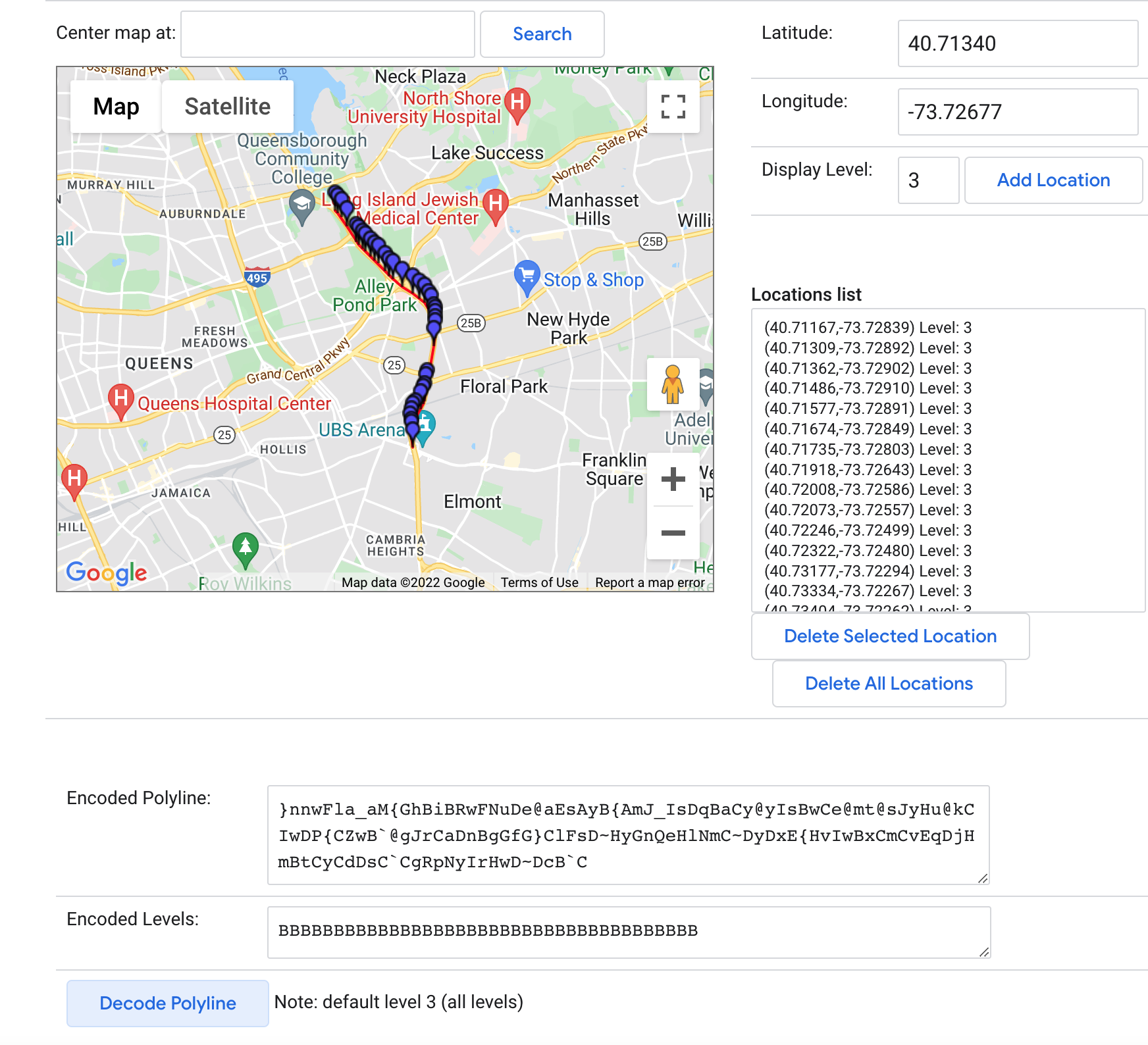
**Table 1 - Dataset Features, Description, and Data Type**

| Name of Feature | Description | Data Type |
| --- | --- | --- |
| ID | Unique Identifier for Sensor within dataset | integer |
| Speed | Average Speed traveled between the link points origin and destination | double |
| TravelTime | Time Travel in seconds | integer |
| Status | Artifact (not useful) | integer |
| Data\_As\_Of | Date and Time of Day for Sensor Data | datetime |
| Link\_Id | TRANSCOM Link ID | integer |
| Link\_Points | Group of Latitude and Longitude points of Sensor data | List of 2 double (latitude and longitude) points |
| Encoded\_Poly\_Line | Link\_Point representation of Google compatible poly line | string |
| ENCODED\_POLY\_LINE\_LVLS | Encoded representation of Poly Level | string |
| Owner | Owner of Sensor | string |
| TRANSCOM\_ID | Artifact (not useful) | string |
| BOROUGH | Name of Borough Sensor exists | string |
| Link\_Name | Description of Sensor location | string |

**Figure 1 - Example of Dataset**



**Figure 2 - Example of Polyline**



Polyline value: }nnwFla\_aM{GhBiBRwFNuDe@aEsAyB{AmJ\_IsDqBaCy@yIsBwCe@mt@sJyHu@kCIwDP{CZwB`@gJrCaDnBgGfG}ClFsD~HyGnQeHlNmC~DyDxE{HvIwBxCmCvEqDjHmBtCyCdDsC`CgRpNyIrHwD~DcB`C

Explanation of polyline: https://developers.google.com/maps/documentation/utilities/polylinealgorithm

Example of polyline: https://developers.google.com/maps/documentation/utilities/polylineutility?csw=1

References

Chen, H., Rakha, H. A., & Sadek, S. (2011). Real-time freeway traffic state prediction: A particle filter approach. *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. https://doi.org/10.1109/itsc.2011.6082873

City of New York Department of Transportation. Retrieved on March 9, 2022 from:

https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9

Hofleitner, A., Herring, R., & Bayen, A. (2012). Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological*, *46*(9), 1097–1122. https://doi.org/10.1016/j.trb.2012.03.006

Kong, F., Li, J., Jiang, B., Zhang, T., & Song, H. (2018). Big data‐driven machine learning‐enabled traffic flow prediction. *Transactions on Emerging Telecommunications Technologies*, *30*(9). https://doi.org/10.1002/ett.3482

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

Yisheng, L., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2014). Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1–9. https://doi.org/10.1109/tits.2014.2345663