Predict Trip Duration for Yellow Taxi Rides in each Borough of NYC

Quynh Phuong Le
Student ID: 1288599
Github repo with commit

August 25, 2024

1 Introduction

The TLC Yellow Taxi Trip Records in New York City can provide great insights into the traffic volumes and travelling times around New York City, which are useful not only for taxi drivers and customers, but also for any citizens and tourists coming to New York. Since it is a one of the most famous tourist attraction in the world, the travelling times around New York could drastically change at different times in a day, different days in a week and different seasons in a year due to differences in traffic volumes as well as other factors such as weather and public holidays.

In this project, two machine learning models which are Linear Regression and Random Forest Regression are built for each borough of New York City including Queens, Manhattan, EWR, Staten Island, Bronx and Brooklyn to predict trip duration for Yellow Taxi rides. The predictor variables used to predict trip duration are pickup time, pickup location, dropoff location, trip distance and various other factors such as public holidays as well as weather elements. Since trip durations can be strongly correlated with trip distances, throughout this project we will visualise most of our data using the average speed in a trip, which serves as a ratio between trip distance and trip duration.

Being able to precisely predict trip duration is particularly useful for tourists and independent travelers who are the main customers of taxis, as they are not familiar with New York's traffic volumes in a day or what other factors could affect their travelling times. Therefore, predicted travelling times can help customers to plan their vacations or trips in New York City more effectively.

1.1 Datasets

The main dataset being analysed is the **TLC Yellow Taxi Trip Records** [1], which is retrieved from the NYC Taxi & Limousine Commission webpage. The Yellow Taxi type is chosen as the subject of interest because it is the only type of taxi that is allowed to pickup passengers anywhere in NYC [2], which makes the predictions from this taxi type more useful to all customers and tourists. The timeline chosen for the training dataset is from Dec 2022 to May 2023. This time period is chosen to capture a variety of climates, from winter to the start of summer. The training dataset is used to predict future testing dataset which is consisted of 30 randomly chosen days from Jan 2024 to March 2024. This time period is also chosen to capture a transition from winter to warmer weather. This is a large-scale dataset with a mixture of categorical and continuous features including the pickup, dropoff locations, pickup, dropoff time, trip distance, fare amount, etc.

The external dataset used is the hourly weather data within the timeframe of the training and testing timeline, retrieved from the query builder of Visual Crossing Corporation [3]. The weather datasets are consisted of a mixture of continuous and textual features showing the hourly measurements of weather elements such as temperature, snow depth, wind gust, visibility and a short description of the weather.

Datasets	Number of instances	Number of features
TLC Yellow Taxi Trip Records from Dec 2022 to May 2023 (Train set)	19,585,935	19
Visual Crossing Weather Data from Dec 2022 to May 2023 (Train set)	4367	9
TLC Yellow Taxi Trip Records from Jan 2024 to Mar 2024 (Test set)	9,554,778	19
Visual Crossing Weather Data from Jan 2024 to Mar 2024 (Test set)	2183	9

Table 1: Summary of dataset size

Note: the number of instances for the training test set is calculated before randomly choose 30 days, the actual number of testing instances used for each borough will be presented later.

In addition to these datasets, a list of NYC public holidays in 2022, 2023 and 2024 is also used to flag any day as public holiday as well as the holiday scale. The list indicates the day and name of the holiday and whether it is the main public holiday, the observed main public holiday, floating holidays and observed floating holidays. Observed holiday is usually the day before of after the main holiday and the floating holidays only affect a proportion of eligible employees.

A taxi zone dataset which includes the borough name for each pickup location ID is also used to split the dataset based on borough.

The external weather dataset and list of public holidays are used with the expectation that certain weather elements might affect the road condition and influence the number of people being outdoors.

2 Preprocessing

2.1 Schema Integration

The whole set of data from Dec 2022 to May 2025 is combined from monthly TLC Yellow Taxi Trip Records dataset, therefore inconsistent schema for the same feature is observed between the datasets. In response to this, the schema from Dec 2022 dataset is corrected with regard to the TLC Yellow Taxi Trip Records's data dictionary [1] and then integrate this schema to all other datasets before combining them into one.

2.2 Feature Selection and Data Wrangling

2.2.1 TLC Yellow Taxi Trip Records

The only valid predictor variables for the trip duration are the variables that are known before the trip ends, such as pickup location or time. According to the TLC Yellow Taxis Data Dictionary [1], all the payment-related features are known and recorded when the trip ends, which makes them irrelevant to predict trip duration, so we decide the remove them from the dataset.

For the remaining features, these are the observations and issues detected as well as actions that were taken to resolve these issues:

- Pickup and dropoff date times are not within the desired timeline: remove the instances where the date recorded is outside of the specified timeline.
- The minimum trip distance is 0 mile and maximum distance exceeds 300,000 miles: The whole perimeter of NYC is around 600 miles [4] therefore the maximum trip distance is unreasonable. For this project, we aim to predict for trips that are at least 0.5 miles and at most 50 miles so only instances within this range are retained.
- Pickup and dropoff locations at unknown borough: instances where the pickup and dropoff location IDs are not well defined in the taxi zone (not in 1-263) were removed.

After these steps, the training data set has 15,068,351 instances and the testing dataset has 7,152,387 instances. Both has 5 columns left which are the pickup and dropoff time, location as well as the trip distance.

2.2.2 Weather Dataset

Most of the features in the weather dataset are believed to have an effect on the road condition or the number of people outdoors such as temperature or snow. However some features represent the same weather element such as wind gust and wind speed.

- Wind gust and wind speed all represent wind condition: wind gust is retained because it is often more associated with severe weather such as thunderstorms [5].
- Snow and snow depth: According to the weather dataset's dictionary [6], snow is the amount of new snow and snow depth is the thickness of snow at that time. Therefore values in these two features are added and stored as snow depth.

After these steps, the weather dataset has 5 features left with the same number of instances as stated above.

2.3 Feature Engineering and Data Aggregation

2.3.1 TLC Yellow Taxi Trip Records

- Trip durations are computed in minutes by taking the dropoff time minus the pickup time.
- Negative trip durations or unreasonably high durations: We decide to only retain trips that last for at least 5 minutes and at most 2 hours. Any trips outside of that range have several more factors affecting the trip duration such as resting time for very long trips.
- For visualisation purpose, average speed is computed by taking the trip distance over trip duration. Because trip durations can be highly correlated with trip distance, the average speed serves as the ratio of them which can be compared and visualised easier.
- Impossibly high speeds: only instances where speeds are below 65 mph were retained since this is the highest speed limit in New York [7].
- Weekends and holidays are flagged and a holiday scale is added to indicate the extent to which people are affected by this holiday, with 4 as the highest effect and 0 for non-holidays.

After these steps, the training data set has 14,766,455 instances and the testing dataset has 8,407,957 instances. Both has 10 columns.

2.3.2 Weather Dataset

- Rainy hours are flagged based on the *description* feature in the dataset and the weather data dictionary [6]. Then the *description* is discarded because it only contains irrelevant information such as clear or overcast day. The dataset still has the same number of instances and columns.

The TLC Yellow Taxi Trip Records dataset is aggregated with the weather dataset using an inner join based on day and hour.

3 Analysis and Geospatial Visualisation

3.1 Outliers

The values of trip distance and trip duration have already been filtered to be within a reasonable range for the interest of this project. These fixed ranges also ensure consistency in both training and testing dataset so that we do not extrapolate and use out-of-range values of predictor variables (such as trip distance) to predict trip duration. Therefore we do not look for outliers in these features but instead, we removed outliers for speed using the IQR rule that flag outliers if $(\sqrt{\log N} - 0.5) \times IQR$ away from IQR. After this step, the training dataset has 14,593,487 instances and testing data has 8,057,911 instances left.

3.2 Log-transformation of Trip duration

The distribution of trip duration is right skewed, therefore we decided to apply a log-transformation to normalise trip duration.

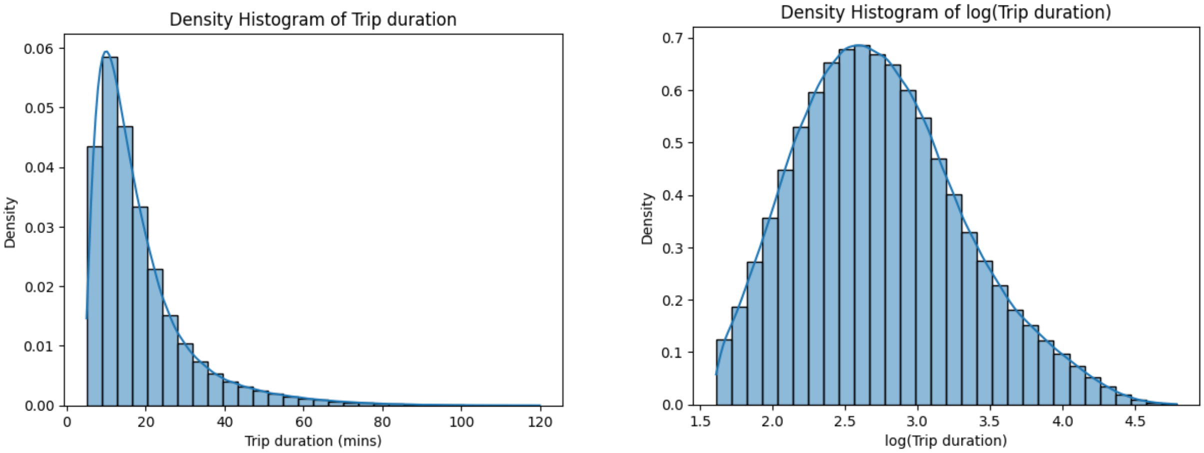

Figure 1: Histogram of Trip duration before (left) and after (right) log transformation

3.3 Missing values

- No missing values were found in the remaining features in both training and testing set of the TLC Yellow Taxi Trip Records.
- One row of NULL values is detected in the testing weather dataset, because this is a small proportion of the dataset, we decide to remove this instance.

3.4 Average Speed by Time

From Figure 2, we can see that travelling time in a day and week clearly has an effect on the average speed. Ideally for customers who would like to use taxis in the day (from 8am to 6pm), it is more time-efficient to travel in the weekends rather than weekdays. This is due to the increased in traffic volumes in weekdays when people are heading to and back from work.

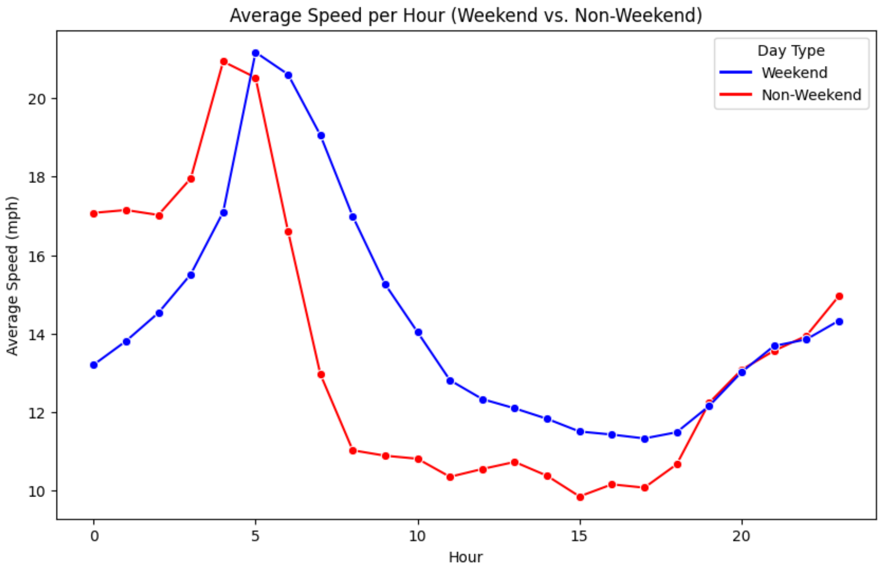

Figure 2: Average speed by Hour in Weekdays and Weekends

3.5 Average Speed and Weather

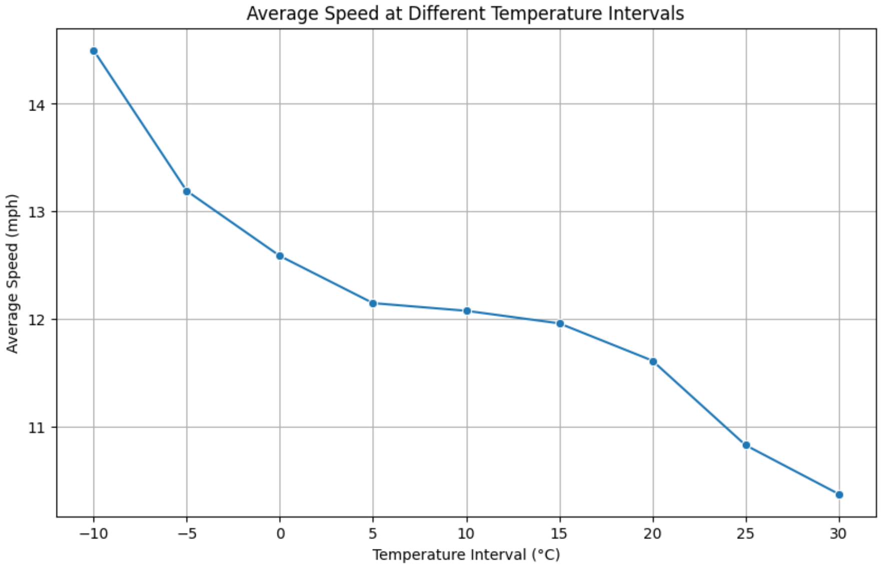

Figure 3: Average speed by Temperature

From Figure 3, we can see that in warmer weather, the traffic volumes increase (average speed decreases). This is probably because warmer days influences more people to spend time outdoors.

This suggests that when travelling in New York in the summer, customers and tourists should allocate more travelling time on the roads due to an increased in traffic volumes, which leads to slower travelling speed.

For weather elements apart from temperature, we found that only snow depth and wind gust are shown to have an effect on travelling speed but not visibility and rain.

According to the weather data dictionary [6], visibility is low when there is smoke or haze, in which we expect a slower driving speeds. However, through inspection in the data, we observed that this usually occurs at midnight or earlier in the morning, when the traffic volume is very low, which allows drivers to drive faster and compensate for the visibility effect.

Similarly for rain, we found that there is no large difference in average speed, this is probably because rainy weathers influence people to stay indoors, and thus reduce traffic volumes, which compensates for the slower driving speeds in rainy weathers. Therefore we decided to remove visibility and rain feature from the dataset.

3.6 Distribution of Average Speed by Pickup location

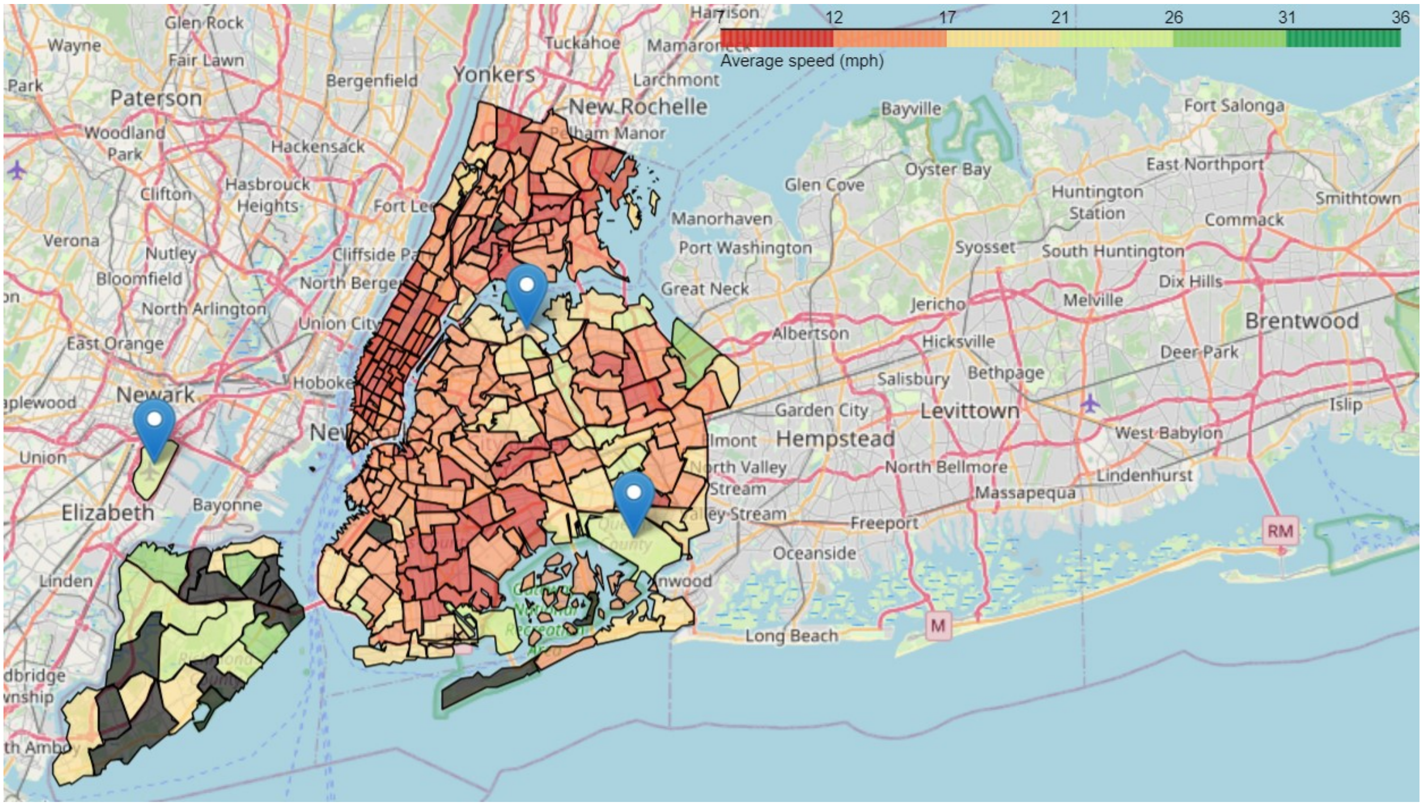

Figure 4: Average Speed by Pickup Locations

From Figure 4, we can see that the traffic volumes are most densely packed in Manhattan based on the map of boroughs in NYC [8]. This suggests customers and tourists to allocate more travelling times to or from Manhattan. The three markers on the figure indicate airports, we can see that travelling speeds near the 3 airports are quite high. This is probably because there are always highways facilitate access to and from the airports. The area in the LaGuardia Airport near the centre of the city has a lower travelling speed compared to the other ones, therefore customers and tourists should allocate more time travelling to this airport.

4 Statistical Modelling

From Figure 4, we can see a difference in distribution of speed and traffic conditions in NYC, therefore we decide to train the models for each borough in NYC separately. Two regression models which are Linear Regression and Random Forest Regression are built for each borough to predict trip duration in the testing dataset.

4.1 Linear Regression

This model is chosen as it is suitable for predicting a continuous response variable. In addition, Linear Regression can handle both categorical and continuous predictor variables. We assume no interactions between the predictor variables, for instance the weather conditions are independent of the hour in the day.

Therefore Linear Regression model fitted to predict trip duration is based on the effect of the following predictor variables: hour of pickup, pickup location, dropoff location, trip distance, weekends, holiday scales, temperature, snow depth and wind gust.

4.2 Random Forest Regression

Random Forest Regression is also a suitable model to predict continuous response variable using a mixture of categorical and continuous data. This is an ensemble tree-based model in contrast to the linear-based model like Linear Regression with higher complexity, therefore it can capture more complex data patterns and perform well on non-linear data. The similar predictor variables as in LR were used to train RFR model to predict trip duration.

4.3 Results

To compare the performance of two models for each borough, we present the Mean Absolute Error (MAE) because this evaluation metric is less sensitive to outliers and allows straightforward interpretation. MAE is particularly useful when comparing the performance of two models on the same dataset, since MSE or RMSE can be affected by the scale of the errors, MAE provides a consistent basis for comparison. The table below summarises the MAE of each model as well as the number of training instances in each borough.

Borough	Queens	EWR	Bronx	Brooklyn	Manhattan	Staten Island
N.o training instances	1,588,880	70	19,478	104,351	12,879,723	985
MAE of LR	6.76	15.0	11.6	9.45	3.76	15.8
MAE of RFR	8.71	9.53	13.0	11.0	4.16	14.9

Table 2: Summary of models' performance

We can see that both models perform best on datasets with the highest training instances (Manhattan). Overall, LR is shown to outperform RFR on more popular boroughs with large training datasets, whereas RFR helps to reduce errors in small datasets such as EWR. This is probably because RFR is an ensemble model which uses bootstrap sampling to *create* more training datasets and introduce random variation among each decision tree. In addition, the normal assumption in LR generally works well for larger sample size, however for smaller datasets this assumption might fail and the model can easily overfit.

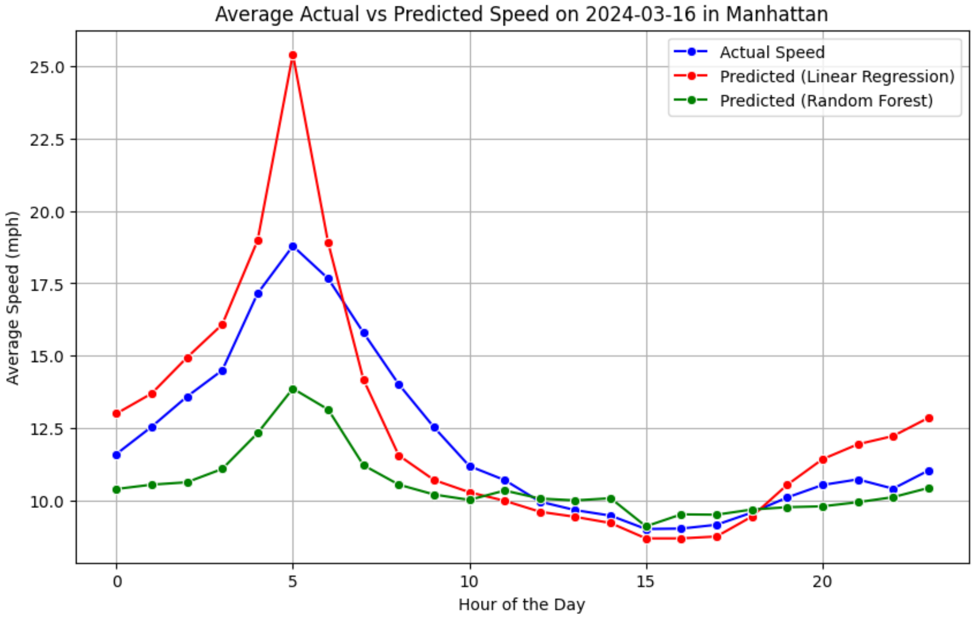

Figure 5: Average Actual vs Predicted Speed on a random day in Manhattan

A random day in the testing set is chosen to visualise the performance of models in a day. Overall, we can see that both models do not perform well in the earlier morning especially at around 5am. This is probably due to the lack of sufficient training instances in this hour leading to overfitting in both models, as we can see that this is when we have the lowest traffic volumes (highest speed), therefore there are limited trip records at this time. At peak hours when there are sufficient training instances, both models perform well especially LR as its predictions closely fit the actual speeds.

5 Recommendations

Taken into account the greater interpretability and lower complexity of the model, LR is shown to be more suitable for this task especially at popular boroughs and peak hours in NYC where there are a lot of training instances. However, for less popular boroughs or less busy hours, RFR is shown to be effective due to its ensemble nature that uses bootstrap sampling and the fact that it does not make the normal assumption which generally only holds for larger dataset like LR.

Therefore we recommend travelling services and agencies to spend more investigations and resources to refine these models for different areas and time in NYC which helps to produce a better and more efficient plan for your customers to travel in New York. This includes a detailed plan of depart locations and time to minimise the travelling times on the road and save time for other activities.

For customers and tourists, we recommend to prepare and allocate more time when travelling in warmer weather or days with good weather in general. This is because such weathers influence more people to be outdoors which increases the traffic volumes and thus increase travelling times as seen in Figure 3. In addition, customers and tourists should also allocate more time when travelling from or to Manhattan and the LaGuardia Airport, due to the dense traffic volumes in these areas as shown in Figure 4.

Overall, the findings and models from this project not only help to predict the trip duration for rides in NYC but also, shed light into the pattern of traffic volumes in NYC.

References

[1] NYC Taxi Limousine Commission. *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2024-08-24.

[2] NYC Taxi Limousine Commission. *Yellow Cab*. https://www.nyc.gov/site/tlc/businesses/yellow-cab.page. Accessed: 2024-08-24.

[3] City of Melbourne. *Weather Query Builder*. https://www.visualcrossing.com/weather/weather-data-services. Accessed: 2024-08-24.

[4] Britannica. *New York City*. https://www.britannica.com/place/New-York-City. Accessed: 2024-08-24.

[5] Bureau of Meterology. *Severe weather and coastal hazard warning services*. http://www.bom.gov.au/weather-services/severe-weather-knowledge-centre/warnings.shtml. Accessed: 2024-08-24.

[6] Visual Crossing. *Weather Data Documentation*. https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/. Accessed: 2024-08-24.

[7] John Ma. *Speed Limits in New York*. https://www.stateregstoday.com/living/traffic-and-driving/speed-limits-in-new-york. Accessed: 2024-08-24.

[8] World Atlas. *The Boroughs of New York City – NYC Boroughs Map*. https://www.worldatlas.com/articles/the-boroughs-of-new-york-city.html. Accessed: 2024-08-24.