NYC Yellow Taxi Tip Amount Prediction Based on Hourly Weather Forecast

August 2021

Gilbert Putra

Abstract

This composition aims to develop **Penalized Regression model** for which it has the capacity to predict tip amount based on widely available hourly weather forecast. Furthermore, 2018 Yellow Taxi (YTX) data taken from Taxi and Limousine Commitee (TLC)[9] and Global Hourly Climate (GHC) data from National Centers for Environmental Information (NCEI)[4] will be analysed. From there, the model will be trained, validated and be utilised to predict 2019 with further critical analysis on its performance. With that said, through Geospatial Analysis we mainly recommend to pickup passengers in LaGuardia Airport. Furthermore, the model performs approximately **0.3071** for $\mathfrak{R}^2$, **1.007** for **MAE** and **3.171** for **MSE** when predicting 2019 data.

1 Introduction

New York City (NYC) is known for its iconic Yellow Taxi Cab which was highlighted throughout movies, comics and other media of entertainment. It has become reminiscent of New York City when it was mentioned. Despite the growth of Uber, the Yellow Taxi industry managed to tightly grasp 0.8 Million trips (0.1%) of its annual trips in 2015 according to Correa et al.(2015)[1]. With that in mind, the thought of how to support Yellow Taxi Industry drivers with the increasing number of competition materialized. The answer to it revolves around what the significant population carries daily, which is smartphones and how it can help.

As a significant part of the world is progressing towards digitalisation, **Weather Forecast** apps are frequently integrated with smartphones. In consideration of that, a research study conducted in New York City 2015 which states that 79% of its population possess smartphones[5]. Therefore, significant populations can be said to have access to weather forecast apps. By using hourly weather conditions that are conveniently accessible through smartphones, the **Penalized Regression model** will be trained upon it. Furthermore, **Yellow Taxi** data combined with **Global Hourly Climate** from 2018 as its training, validation set and a sample taken randomly from 2019 (after combined) respectively will serve as unseen instances of the future which the weather data assumed to be forecasted. Together the 2018 Yellow Taxi data holds approximately about **103 million instances** and after preprocessing, it has reduced to approximately **31.5 million instances** with a total of **17 features**.

This model intends to predict **Yellow Taxi Tip Amount** and significantly towards aiding Yellow Taxi Drivers in New York City, NY, USA. For the purpose of training the model, it has been assumed data that does not follow the data dictionary [6] are considered to be inaccurate and will be removed.

2 Data Preprocessing

Due to the scale of the dataset, Data Preprocessing will be split into two phases. During **Phase 1**, TLC Yellow Taxi dataset from 2018 and 2019 will be cleaned as per the Data Dictionary respectively with the specifications below. Furthermore at **Phase 2**, Yellow Taxi data will be merged with Global Hourly Climate data with closest hour as the key. Moreover, 2018 merged data will be sampled 10% and 10% will be taken from 2019 as a sample of the future. For example, 2018-01-01 20:51:00 closest hour will be 2018-01-01 20:00:00 as it is still a part of 20:00:00 technically according to the GHC data.

2.1 Phase 1

The cleaned features are as following:

- VendorID : Select only VendorID 1 or 2.
- trip_duration (Engineered Feature) : Cannot be negative or zero.
- passenger_count : Minimum of zero and maximum of six.
- trip_distance : Cannot be negative or zero.
- avg_speed (Engineered Feature) : Cannot be negative or zero.
- RateCodeID : Select RateCode from 1 to 6.
- PULocationID : LocationID 264 and 265 are unknown, therefore dropped.
- DOLocationID : LocationID 264 and 265 are unknown, therefore dropped.
- payment_type : Select only payment type 1.
- fare_amount : Cannot be negative or zero.
- extra : Surcharge of $0.5 and $1.0 are considered accurate including the negatives.
- mta_tax : Tax are always $0.5 or -$0.5.
- improvement_surcharge : Select surcharge with $0.3 or -$0.3.
- tip_amount : Cannot be negative or zero.

Moreover, those features are cleaned with a data dictionary [6] as its specification. Additionally, features such as trip_duration, trip_distance, avg_speed cannot be less than or equal to zero as being zero or less means that the cab is not moving. Therefore, it is considered as inaccurate. Negative fare_amount with payment_type 1 (credit) and 2 (cash) are also removed, as they are not a dispute or unknown. Although in this case only payment_type 1 are selected for the purpose of this study. Moreover, trips with less than minimum fare amount are considered to inaccurate and removed based on derived equation (see Appendix Equation 2) from Taxi Fare page [8]. Furthermore, trips traveling with average speed greater than 30 mph will be removed as the maximum speed limit as of 7^{th} November 2014 has been reduced towards 25 mph as part of Vision Zero Initiative [10]. Despite the reduction in speed limit, there are still 3% of streets that have a speed limit of 30 mph and it has been assumed that some trips may have taken the maximum speed limit.

During Phase 1, it has been found that the 2018 Yellow Taxi dataset has no missing values and 2019 Yellow Taxi dataset contains some missing values which are dropped as a sample of it will only

be taken. As of phase 1, features will be dropped, namely VendorID, passenger_count, RateCodeID, payment_type, store_and_fwd_flag, extra, mta_tax, improvement_surcharge, total_amount. Most dropped features can be said to be irrelevant such as payment_type as we have selected only type 1 and mta_tax is the tax surcharge when entering several states.

2.2 Phase 2

During this Phase, ten weather stations closest to New York City will be selected with Global Hourly Climate data. Amongst ten of them, two are located in New Jersey which is adjacent to Staten Island and to be taken into consideration (see Appendix Figure 7). From there, the average of the five stations and eight main observations were selected out of 58. Those eight observations contain nested values for which some will not be available through regular weather forecast apps, hence will be dropped. The resulting features are: DATE, wind_angle, wind_speed, air_temp, dew_point, atp, vis_distance and sky_ceil.

In the Global Hourly Climate dataset, there are some missing values which have been imputed with its respective mean. As significant instances of weather data is desirable to create an accurate model, dropping missing values is not preferred. Furthermore, mean imputation over the columns can be said to be desirable as it takes the overall results throughout the year. Accordingly, as records of GHC dataset are taken approximately three times an hour, the average of the results within each hour are taken. Therefore, it creates hourly weather data throughout the specified date.

3 Preliminary Analysis

3.1 Geospatial Visualisation

Figure 1 shows the number of trips taken in 2018. From Figure 1, it can be seen that the trend gradually decreases from 0 (24:00) until 10:00 as its lowest. Furthermore, it significantly increases as it is approaching late office-hours (from 16:00 to 20:00) and declines towards midnight. This may be caused by significant populations that prefer to ride private taxis after-work to night-outs continuing to classical NYC night-life, which may end after midnight as a significant amount of rides can be seen.

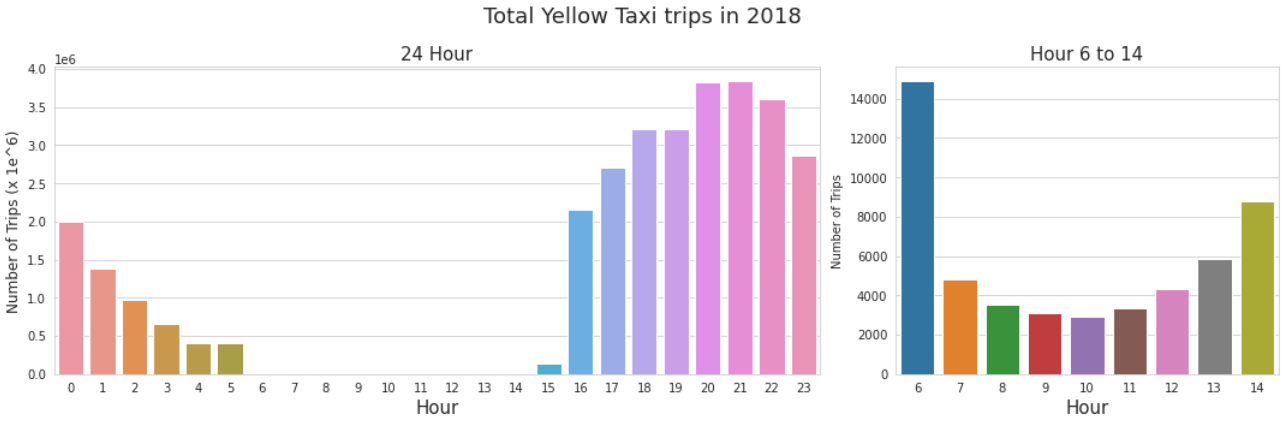

Figure 1: Number of Pickup Trips per Hour

For spatial factors, the total number of pickup trips are depicted in Figure 3. Moreover, it can be seen that two main areas, specifically Manhattan and La Guardia Airport, are reported to have a

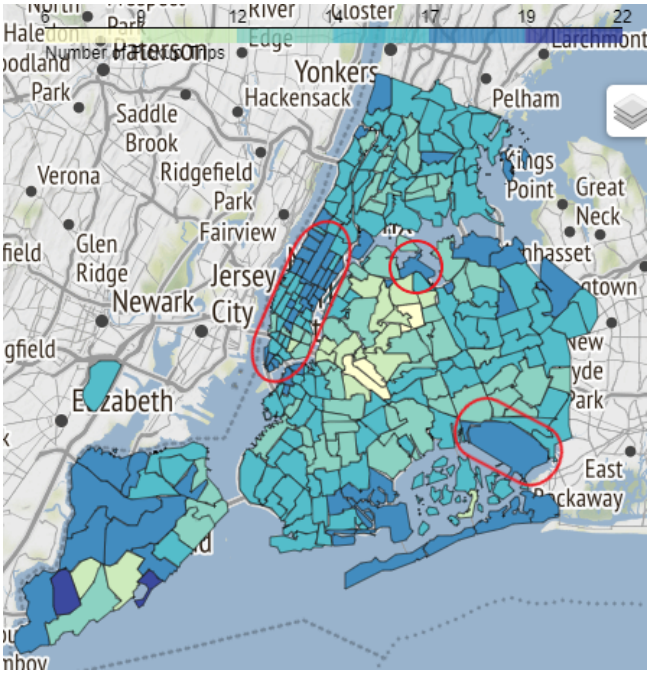

Figure 2: Location Active Hours
Circled areas are Manhattan, La Guardia and JFK airport

significant number of pickup trips in 2018 especially between 17:00 to 19:00 (see Figure 2). This might correlate with the infamous night activity of Manhattan, as predominantly populated in Downtown and Midtown Manhattan. Although there are two main airports in NYC, namely JFK airport and La Guardia airport, it can be seen that domestic La Guardia airport passengers commute with Yellow Taxi cab relatively more frequently than JFK airport passengers. This may be caused by the fact that JFK airport is substantially occupied by international flights rather than domestic, and international passengers might be bewildered with riding a Taxi in a foreign country.

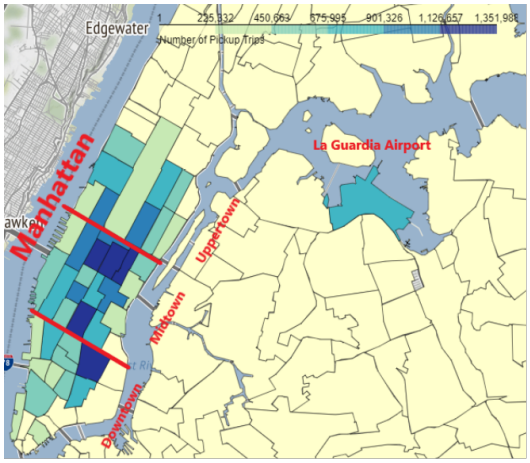

Figure 3: Number of Pickup Trips based on Pickup Zone

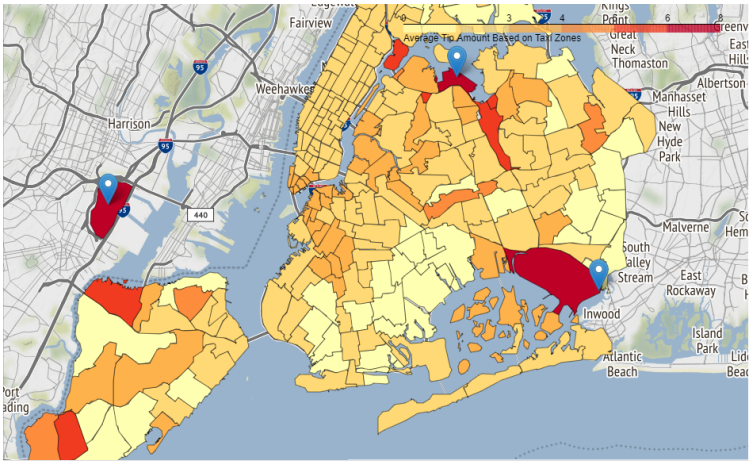

Figure 4: Average Tip Amount based on Pickup Zone

With that said, figure 4 shows three significant areas which are pinpointed that tip relatively higher than other areas. Those three areas have similarities which happened to be Airports, namely JFK, La Guardia and Newark Airport. Through only geospatial analysis, optimally drivers should pick up

passengers in LaGuardia airport as it has significant trips and highest tip amount on average during peak hours (17:00 to 19:00).

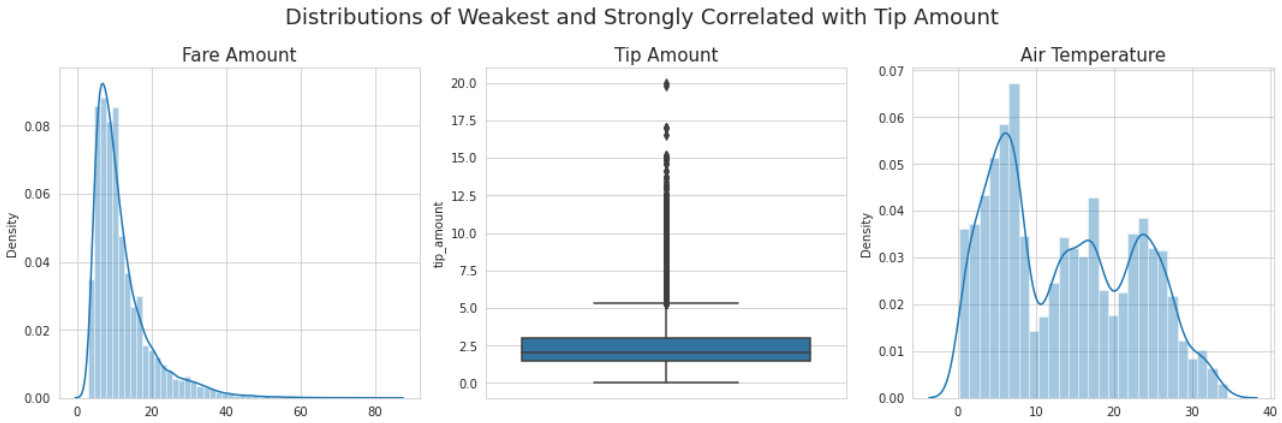

Figure 5: Highly and Weakest Correlated Features with Tip Amount

3.2 Features Analysis

As it is depicted from Figure 5, Yellow Taxi drivers were frequently tipped around $0 to $5 on Average, except for a few trips that tip around $20. Furthermore, those trips with higher tip amounts than average are predominantly Airport Pickup Trips (see Figure 4). To continue, Fare amount in Figure 5 appears to have some trips that pay more than average. Those trips are correlated with the Airport Pickup Trips as mentioned before. Moreover, Figure 5 shows the strongly and weakly correlated features, respectively Fare amount and Air Temperature, which is shown in figure 6. Along with that, multi-modal distribution of Air Temperature and other weather features (see Appendix Figure 9) could be seen which could be explained by seasons.

Despite being weakly correlated with weather data, tip_amount strongly correlates with trip_distance, trip_duration, fare_amount. This might be caused by the culture of the US populations which are familiar with tipping. Nonetheless, correlation between features will be investigated further in Statistical Modelling.

4 Statistical Modelling

4.1 Model

In this section, predictive modelling will be performed aiming to be robust towards future real-world data. The model chosen will be Ridge Regression as several features appear to be highly correlated with each other as seen in Figure 7. This may be caused by the variance-bias tradeoff which Ridge Regression reduces variance but may increase the bias. Furthermore, features will be standardized and transformation is unnecessary as non-linearity is not present (see Appendix Figure 10).

Ridge Regression estimators can be denoted as such:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

where Y is the predictor variable, X is the independent features and λ as the penalty parameter.

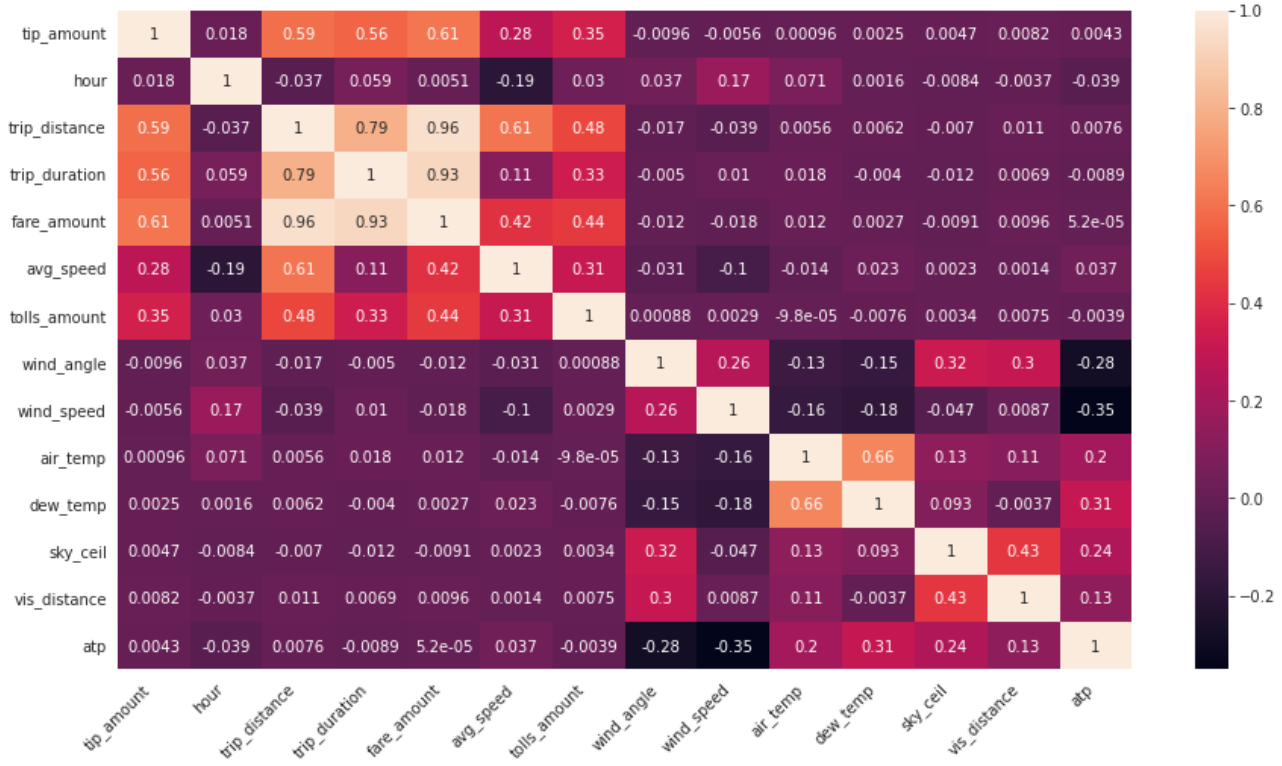

Figure 6: Pearson Correlation Matrices

Furthermore, training will be applied with approximately **2.516 Million** instances and validation set of **629 Thousand** instances from preprocessed Yellow Taxi 2018 dataset. Moreover, approximately **1.746 Million** instances from Yellow Taxi 2019 dataset will be taken as a prediction real-world set.

4.2 Results and Discussion

As model fitting is applied, Hyperparameter λ has been found to be **0.9639** with parameters coefficients to be:

$$\begin{aligned}\text{Tip Amount} =& 2.4876(\text{Intercept}) + 0.3409(\text{Trip Distance}) - 0.3603(\text{Trip Duration}) + \\ & 0.3797(\text{Fare Amount}) + 0.0938(\text{Average Speed}) + 0.2181(\text{Tolls Amount}) + \\ & 0.0020(\text{Wind Angle}) + 0.0078(\text{Wind Speed}) - 0.0020(\text{Air Temperature}) - \\ & 0.0063(\text{Dew Point}) - 0.0014(\text{Sky Ceiling}) - 0.0017(\text{Visibility Distance}) - \\ & 0.0006(\text{Atmospheric Pressure})\end{aligned} \tag{1}$$

As it can be seen from the equation above, Ridge Regression almost eliminates both Atmospheric Pressure, Air Temperature and most of the weather parameters. This results aligned with findings in Features Analysis (Section 3.2), as weather data are weakly correlated against tip_amount. Moreover, the four strongest correlated features that were found in the Ridge Regression model suggest similar findings with Preliminary Analysis, namely trip_distance, trip_duration, fare_amount and tolls_amount.

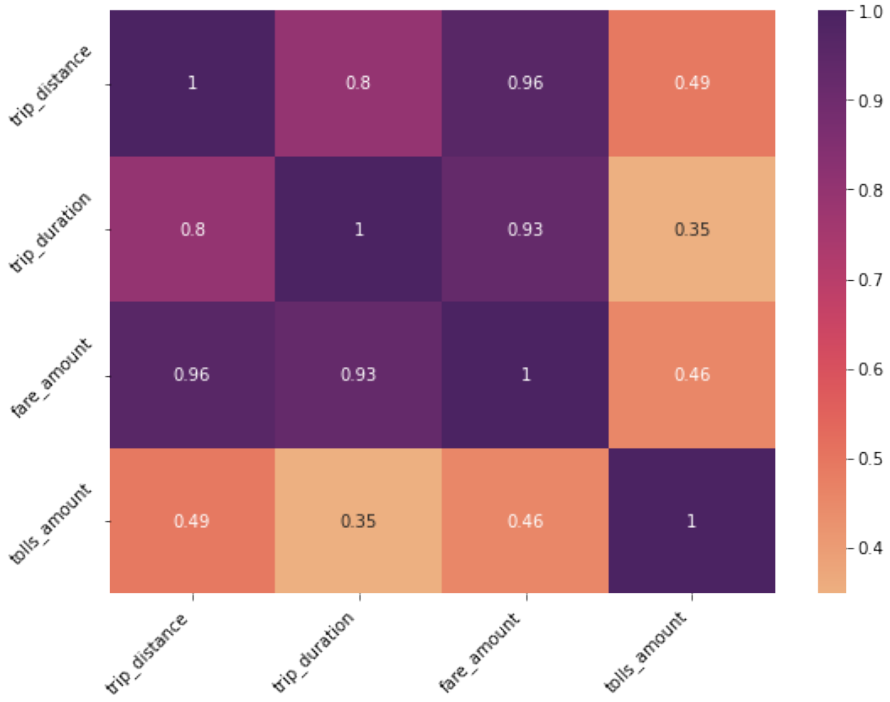

Figure 7: Multicollinearity of Independent Features

	$\Re^2$	MAE	MSE
2018 Training	0.4864	0.7408	1.9818
2018 Validation	0.4669	0.7410	2.127
2019 Prediction	0.3071	1.007	3.171

Table 1: Evaluation Metrics

On top of that, the fitted model (Table 1) reported R^2 metrics of **Training** was **0.4864** while it score **0.4669** on **Validation** set which is a slight advance. Although, when evaluated against **2019 Prediction** sample it scores **0.3071** in terms of R^2.

Despite having relatively high correlation with selected features through our discovery in Preliminary Analysis, it performs comparably moderate as those selected features suggest multicollinearity (Figure 7). Moreover, weather features were expected to be dropped to 0 with the penalty applied as it has inadequate correlation to tip_amount. Through our model findings, it can be suggested that tip_amount may be affected by other factors other than chosen features, specifically mood and driver behaviour perhaps. According to Jahan N. in 2018, **Tipping Behaviours** are generally influenced by **Demographic** and **Cultural Differences** making it a social norm instead of rational behaviour which is independent of Households or Individuals income [3].

5 Recommendation

As suggested from both Preliminary Analysis and Statistical modelling, Tip Amount is highly correlated with **Fare Amount**, **Trip Distance**, **Trip Duration** and **Tolls Amount** propose to take trips that are longer than average. Furthermore, trips from the airport generally Tip more than other areas

specifically **La Guardia Airport** between **17:00 and 19:00** as it also has significant Pickup Trips in 2018. There is a little no less correlation with weather forecasting. Therefore, Taxi Drivers are suggested to Pickup Passengers from LaGuardia Airport and in between 17:00 and 19:00 regardless of the weather.

6 Conclusion

In conclusion, weather data may be lacking in correlation with tip amount or the weather data was insufficient meaning rather than taking closest hour as a key, it should be based on closest minute or perhaps minutely readings. However, ridge regression performs comparably moderate dealing with this problem although it could still be improved by adding more relevant features or increasing weather readings. In the future, time series forecasting should also be considered when dealing with this kind of problem as it is more of a predicting future values through previously observed values.

References

[1] Correa, D.; Xie, K.; Ozbay, K. Exploring the Taxi and Uber Demands in New York City: An Empirical Analysis and SpatialModeling. In Proceedings of the Transportation Research Board's 96th, Annual Meeting, Washington, DC, USA, 8–12 January 2017.

[2] Python Glmnet Package. https://glmnet.stanford.edu/articles/glmnet.html

[3] Jahan, N. "Determinants of Tipping Behavior: Evidence from US Restaurants" (2018). Electronic Theses and Dissertations. 2633. https://openprairie.sdstate.edu/etd/2633

[4] National Centers for Environmental Information (NCIE). National Oceanic and Atmospheric Administration. Accessed: 11/08/2021
https://www.ncei.noaa.gov/data/global-hourly/archive/csv/2018.tar.gz
https://www.ncei.noaa.gov/data/global-hourly/archive/csv/2019.tar.gz

[5] New York City Mobile Services Study Research Brief. New York City Department of Consumer Affairs, 2015.
https://www1.nyc.gov/assets/dca/MobileServicesStudy/Research-Brief.pdf

[6] NYC Taxi & Limousine Commision. Data Dictionary.
https://www1.nyc.gov/assets/tlc/downloads/pdf/data$_d ictionary_t rip_r ecords_y ellow.pdf$

[7] NYC Taxi & Limousine Commission. Chapter 54 - Drivers of Taxicabs and Street Hail Liveries.
https://www1.nyc.gov/assets/tlc/downloads/pdf/rule_book_current_chapter_54.pdf

[8] NYC Taxi & Limousine Commission. Taxi Fare - TLC. Accessed: 06/08/2021
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[9] NYC Taxi & Limousine Commission. TLC Trip Record Data. Accessed: 06/08/2021
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page .

[10] Vision Zero Initiative. New York City Department of Transportion.
https://www1.nyc.gov/html/dot/html/motorist/motorist.shtml

Appendix

Derived Formula

Equation 2: Taxi Minimum Fare $2.50 initial charge Plus 50 cents per 1/5 mile when traveling above 12mph or per 60 seconds in slow traffic or when the vehicle is stopped[8]. It can be interpreted mathematically as such:

$$f(n) = \begin{cases} \$2.5 \text{ per mile} & \text{if speed} > 12 \text{ mph} \\ \$2.5 \text{ per mile} = \$0.5 \text{ per minute} & \text{if speed} = 12 \text{ mph} \\ \$0.5 \text{ per minute} & \text{if speed} < 12 \text{ mph} \end{cases}$$

From there, we can derive the minimum fare as:

$$\text{minimum fare} = \$2.5 + max(0.5 \cdot minute, 2.5 \cdot distance) \tag{2}$$

Figures

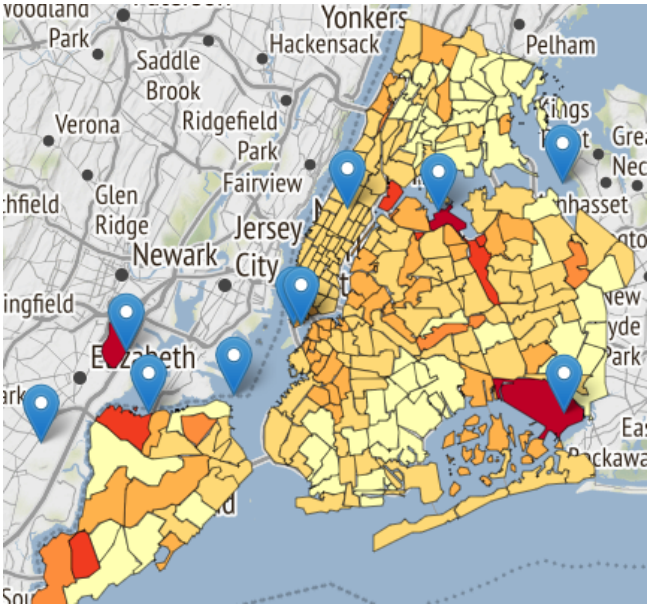

Figure 8: Weather Station Location

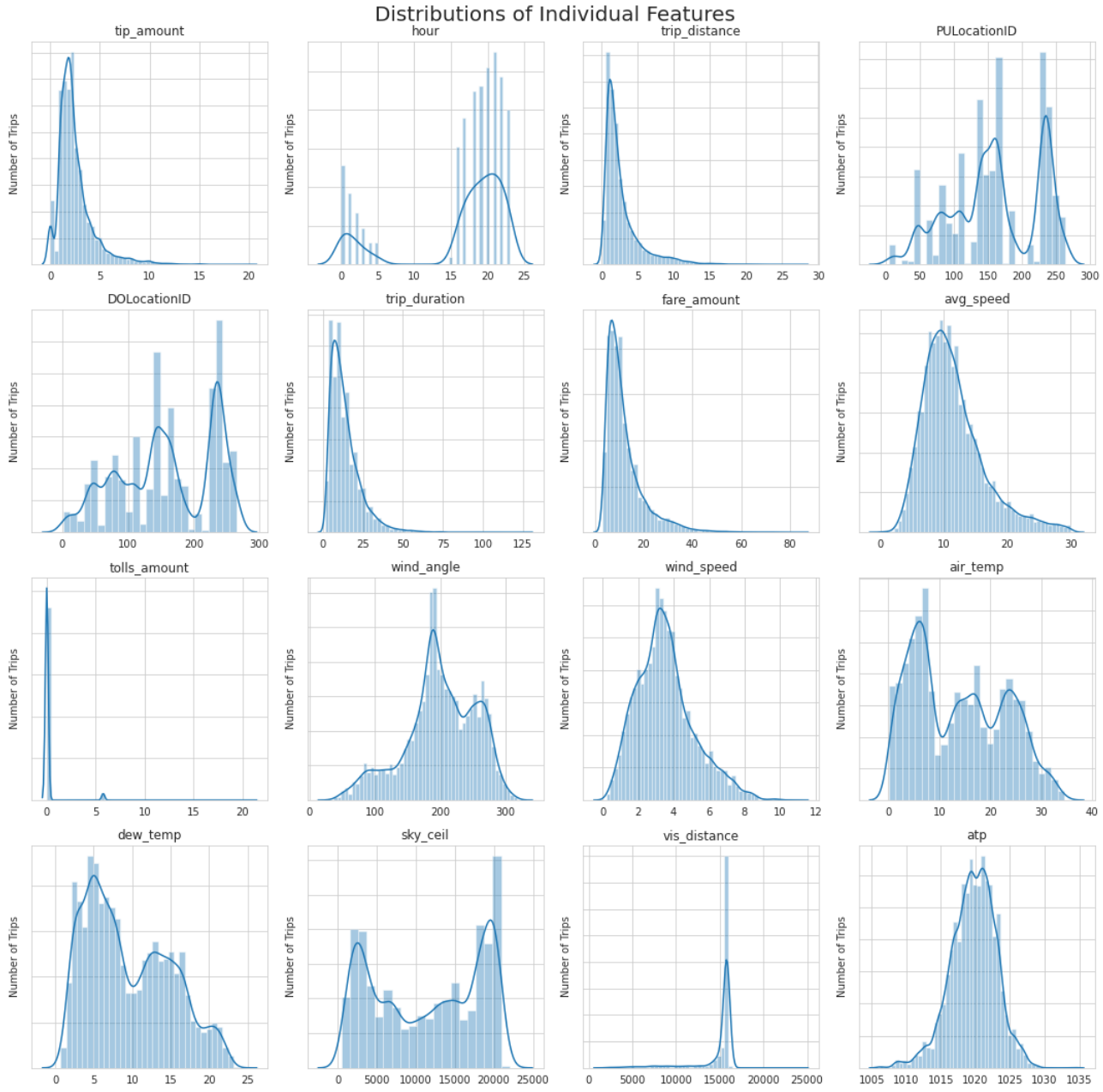

Figure 9: Distribution of Individual Features

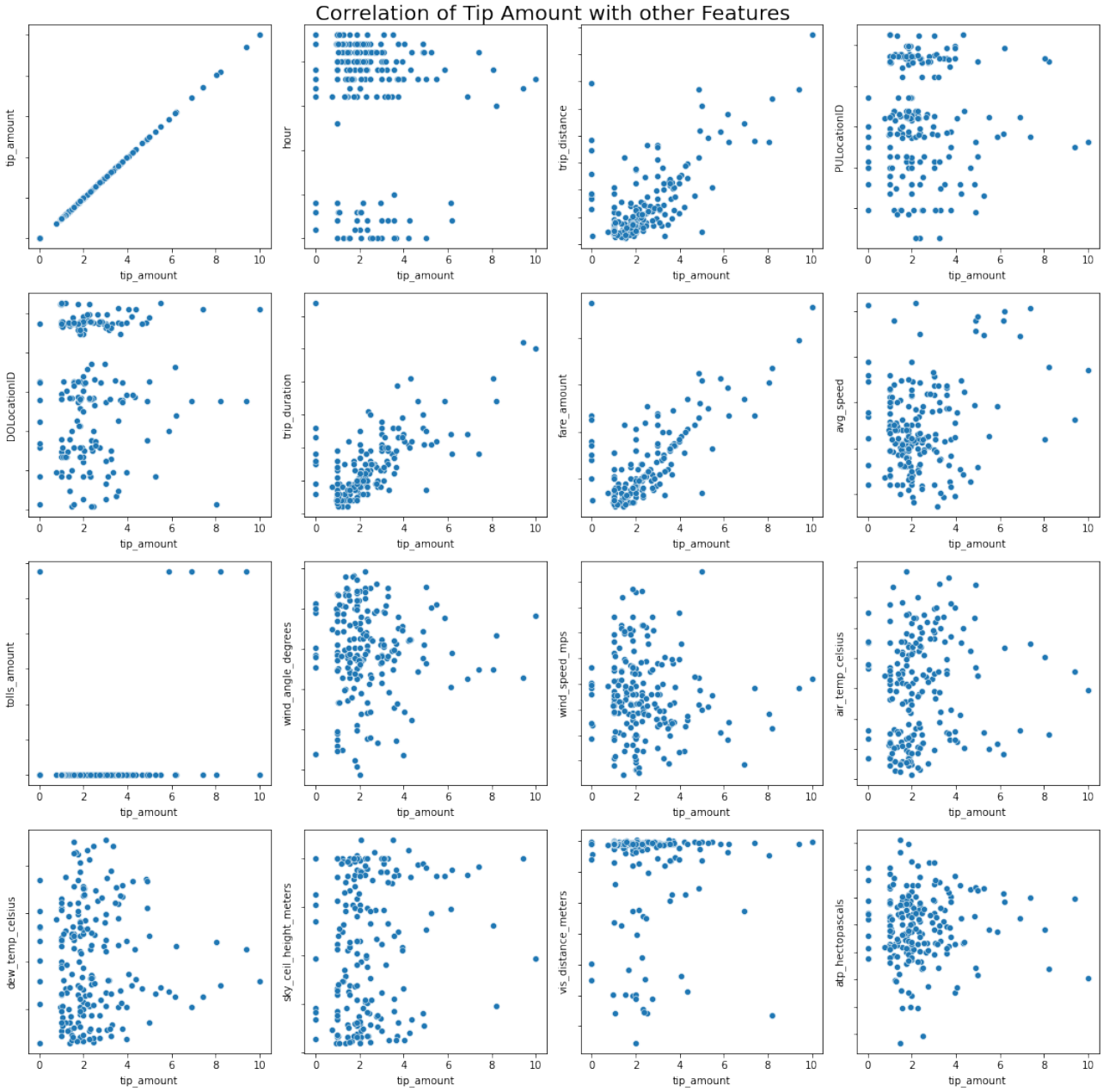

Figure 10: Correlation of Tip Amount with other Features