NYC Taxis Fare Prediction Analysis Report

Han Wang, Jam Lin, Jiayi Li, Yibin Long

Preface

This report was developed as a deliverable for the term project in DSCI 522 (Data Science Workflows), a course in the Master of Data Science program at the University of British Columbia.

The overall objective of this project was to automate a typical data science workflow. This report summarizes the results of a series of automated Python scripts that handle tasks such as data retrieval, data cleaning, exploratory data analysis (EDA), creation of a predictive machine learning model, and interpretation of the results. The report provides a detailed explanation of each step, applying it to the specific context of the dataset in question. It assumes the reader has a basic understanding of machine learning terminology and concepts.

Summary

This report presents a linear regression model developed to predict NYC taxi fare amount. Using data from 30,000 taxi trips in January 2024, we build a linear regression model using trip distance feature, with each additional mile increasing the fare by approximately $3.54. The model was evaluated using a test dataset, where predicted fare amounts were compared to actual fares. While the model performed well overall, some outliers were identified, suggesting the need for further data cleaning and additional features to improve accuracy. Future steps include incorporating more features, experimenting with other regression models like KNN and Lasso regression, and addressing hyperparameters for better generalizability and performance.

Introduction & Background

Taking a taxi in New York City can be overwhelming for first-time visitors, especially for tourists unfamiliar with the city's layout and fare system. With over 200,000 taxi trips taking place daily, yellow cabs remain an essential mode of transportation for both locals and visitors.

However, the lack of transparency often leads to concerns among tourists about overcharging or being taken on unnecessarily long routes.

To address these concerns, we use data from 30,000 Yellow Taxi trips recorded in January 2024, provided by the NYC Taxi and Limousine Commission (TLC). Our analysis examines how to predict taxi fare amounts, offering valuable insights to help tourists better understand what they should expect to pay for a taxi ride in NYC. By leveraging data-driven analysis, we aim to provide a clearer and more predictable fare structure, enabling tourists to make more informed decisions during their travels in the city.

Analysis Question

How do we best predict NYC yellow taxi fare amounts?

Dataset

The data set used in this project is the TLC Trip Record Data, which includes taxi and for-hire vehicle trip records collected by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). This data is sourced from the NYC Taxi and Limousine Commission (TLC) and can be accessed from the NYC Taxi and Limousine Commission's website (Taxi and Commission 2024). Yellow and green taxi trip records contain fields capturing pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Similarly, For-Hire Vehicle (FHV) trip records include fields such as dispatching base license number, pick-up date and time, and taxi zone location ID. It is important to note that the TLC did not create this data and makes no representations regarding its accuracy. FHV trip records are based on submissions from dispatching bases and may not represent the total volume of trips. Additionally, the TLC reviews these records and enforces necessary actions to improve their accuracy and completeness.

Methods & Results

Methods

In order to address our research question, we will begin by selecting the appropriate features from the dataset through exploratory data analysis (EDA) and by consulting the data dictionary to better understand the instances within the dataset. Since this is a linear regression modeling problem, we will use the LinearRegression model as our model of choice. Linear regression is a fundamental and widely used method for predictive modeling. The dataset used in this study includes daily fare amounts and trip distance data from January 2024, covering

30,000 observations. The data will be split into training and test sets, with 70% allocated for training and 30% for testing.

The analysis will be conducted using the Python programming language (Van Rossum and Drake 2009), with the following Python packages: numpy (Harris et al. 2020), pandas (McKinney 2010), scikit-learn (Pedregosa et al. 2011), and altair (VanderPlas 2018).

EDA

Lets do some data validation by checking the target variable fare_amount's distribution:

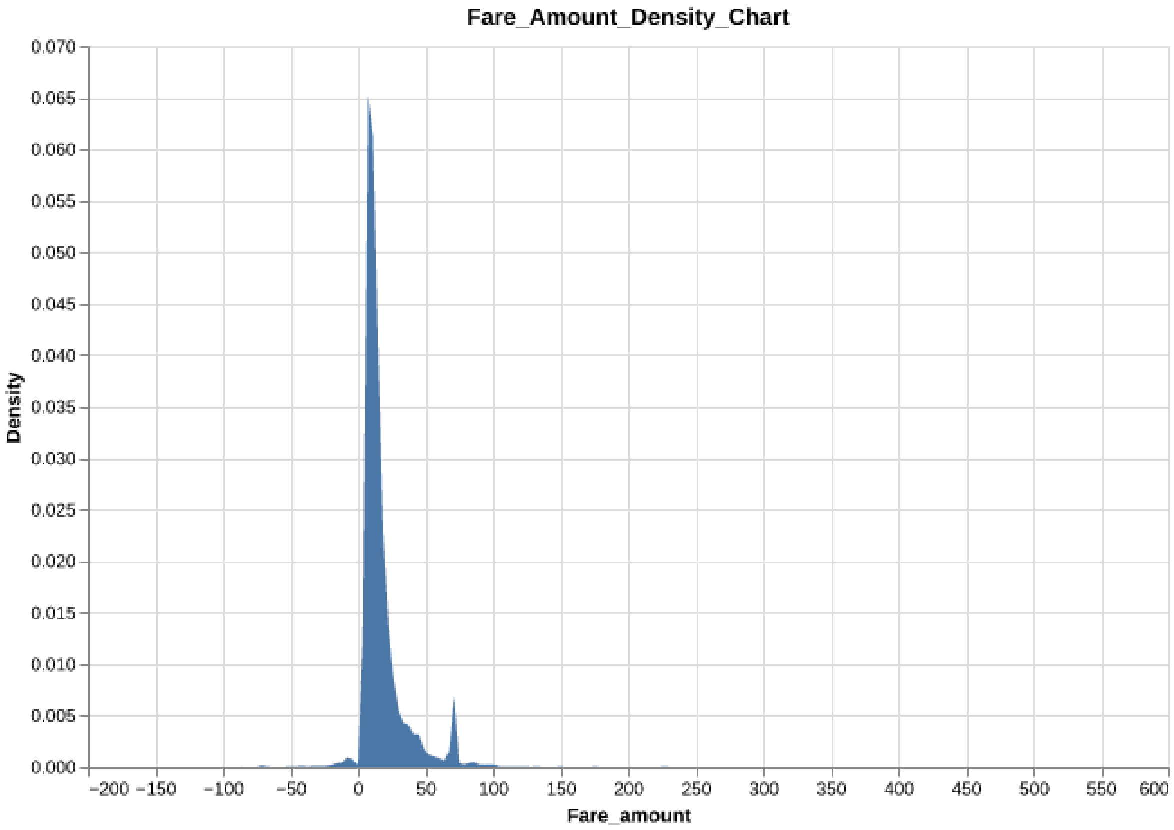

Figure 1: Fare Amount Density Chart

Based on Figure 1, we see that most values of the fare amount make sense - they are mostly under $50, with a cluster of fare amounts at around 65 USD, likely a longer trip from the airport. However, we see fare amounts that are lower that 0 which are not explained by the documentation, that "The time-and-distance fare calculated by the meter." They might be simply refunded trips, but since we cannot confirm, we will choose to drop rows where the fare amount is negative.

We'll split the dataset into training set and test set.

Now, we'll perform a summary of the data set that is relevant for exploratory data analysis related to our regression analysis. We'll check out the summary statistics for each column in the dataset.

Table 1: Data Set Summary

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge	Airport_fee
20720	20720	20720	19690	20720	19690	20720	20720	20720	20720	20720	20720	20720	20720	20720	20720	19690	19690
1.749276	2024-01-17 00:46:01.501838	2024-01-17 01:03:32.636196	1.33873	3.19747	1.70340	166.39	165.39	1.13436	18.58	1.4825	0.4953	3.3770	0.52501	0.999628	27.3623	2.31060	0.143448
1	2024-01-01 00:06:10	2024-01-01 00:10:03	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
1	2024-01-09 16:07:21.750000	2024-01-09 16:26:40.250000	1	1	1	132	114	1	8.6	0	0.5	1	0	1	15.48	2.5	0
2	2024-01-17 10:28:41	2024-01-17 10:46:02	1	1.68	1	162	163	1	12.8	1	0.5	2.8	0	1	20.16	2.5	0
2	2024-01-24 17:47:47	2024-01-24 18:01:19.500000	1	3.1	1	234	236	1	20.5	2.5	0.5	4.2	0	1	28.6	2.5	0
6	2024-01-31 23:56:14	2024-02-01 00:15:56	8	68.48	99	265	265	4	600	11.75	4	99	62.75	1	601	2.5	1.75
0.434275	nan	nan	0.844333	4.31892	6.5663	63.626	69.466	0.52711	17.8716	1.79912	0.0549	3.8902	2.0960	0.019002	22.1385	0.56614	0.480072

Based on the summary statistics in Table 1, for the fare amount, the mean fare amount is 18.59 USD, while the median is 12.8 USD, which again points to a right-skewed distribution for this column. This confirms what we saw with the density chart.

Next, we'll create a correlation plot of all of the columns against one another, to check out the strength and direction of associations between columns.

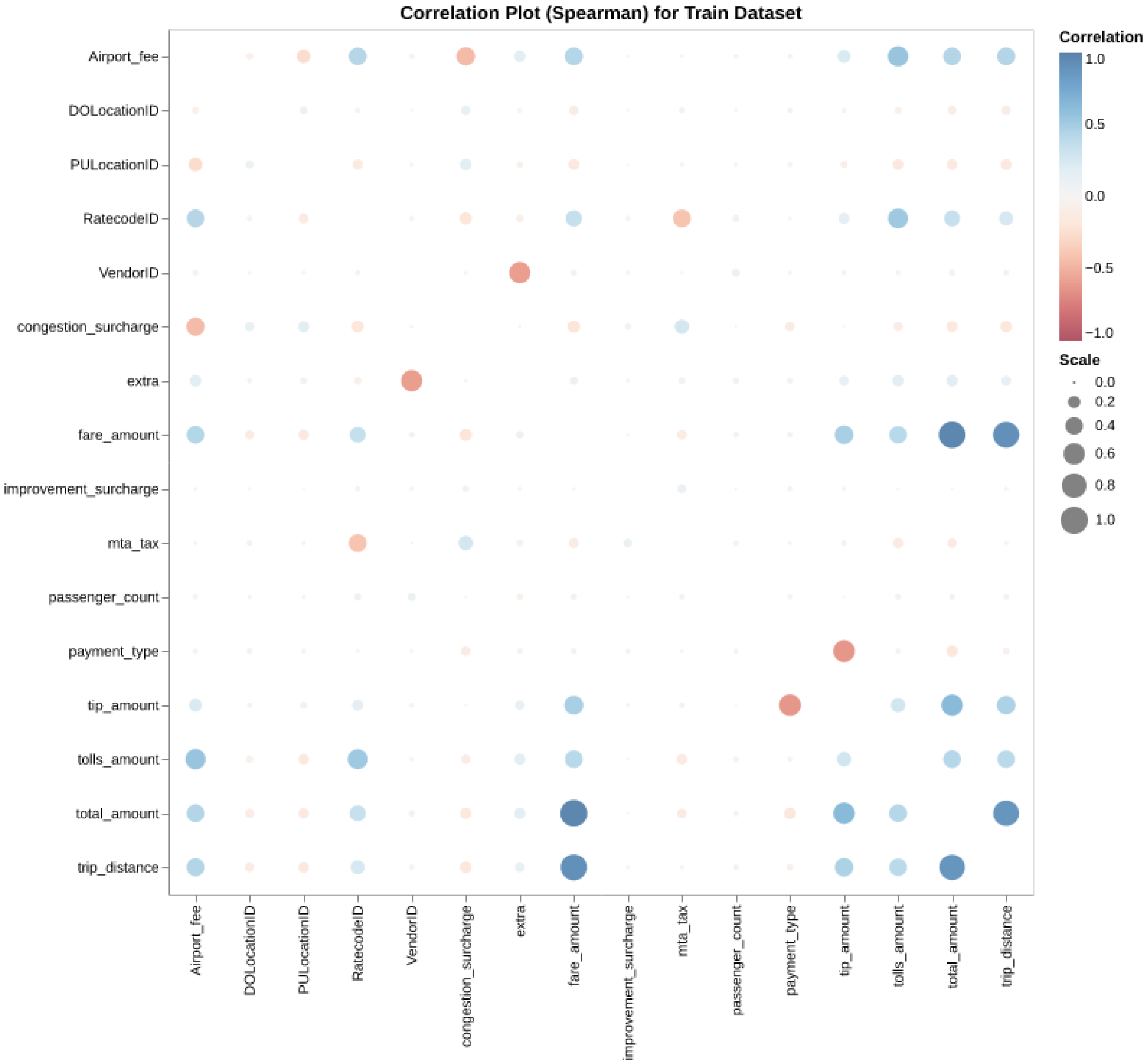

Figure 2: Correlation Plot

Based on Figure 2, we see that “trip_distance” and “fare_amount” have a fairly high positive correlation, which may indicate that they are fairly positively associated with each other. We decide now to use “trip_distance” to try and predict “fare_amount”. Since there are no null values in “trip_distance”, no null values need to be dropped for “trip_distance”.

Modeling

As discussed in the Methods summary, we will now build and test our Linear Regression Model.

The regression line formula is: y_hat = 3.62 * trip_distance + 7.02

Finally, let’s calculate some error metrics and perform a visualization of the result of the regression model in the form of a scatter plot with the regression line.

Table 2: Regression Performance Metrics

Metric	Value
RMSE	6.78282
R^2	0.848052
MAE	3.23701

Lastly in Figure 4, here is the regression line for the predictions.

Results

The objective of the analysis was to predict the fare amount using a linear regression model. The regression plot, generated from the model, demonstrates a positive linear relationship between trip distance and fare amount for NYC Yellow Taxis in January 2024. This confirms that as the trip distance increases, the fare amount also increases, as expected.

The linear regression model shows moderate predictive power with an R-squared value of 0.85, which explains about high percent of fare variance. The low MAE of 3.24 USD indicates good accuracy for typical rides, although the higher RMSE (6.78 USD) suggests sensitivity to outliers. Figure 3 shows that the model performs best for the 10-50 USD range, but tends to underestimate fares above 100 USD. The actual NYC taxi fare is calculated through initial fare, per-mile charge, and additional fees for peak hours, nighttime, tolls, etc that this model cannot predict, but can be built further using more complicated multilinear regression or ML models.

Based on the estimated result from the regression line formula derived from the model, this suggests that for each additional mile traveled, the fare increases by approximately 3.62 USD, and the base fare (when the trip distance is zero) is around 7.02 USD.

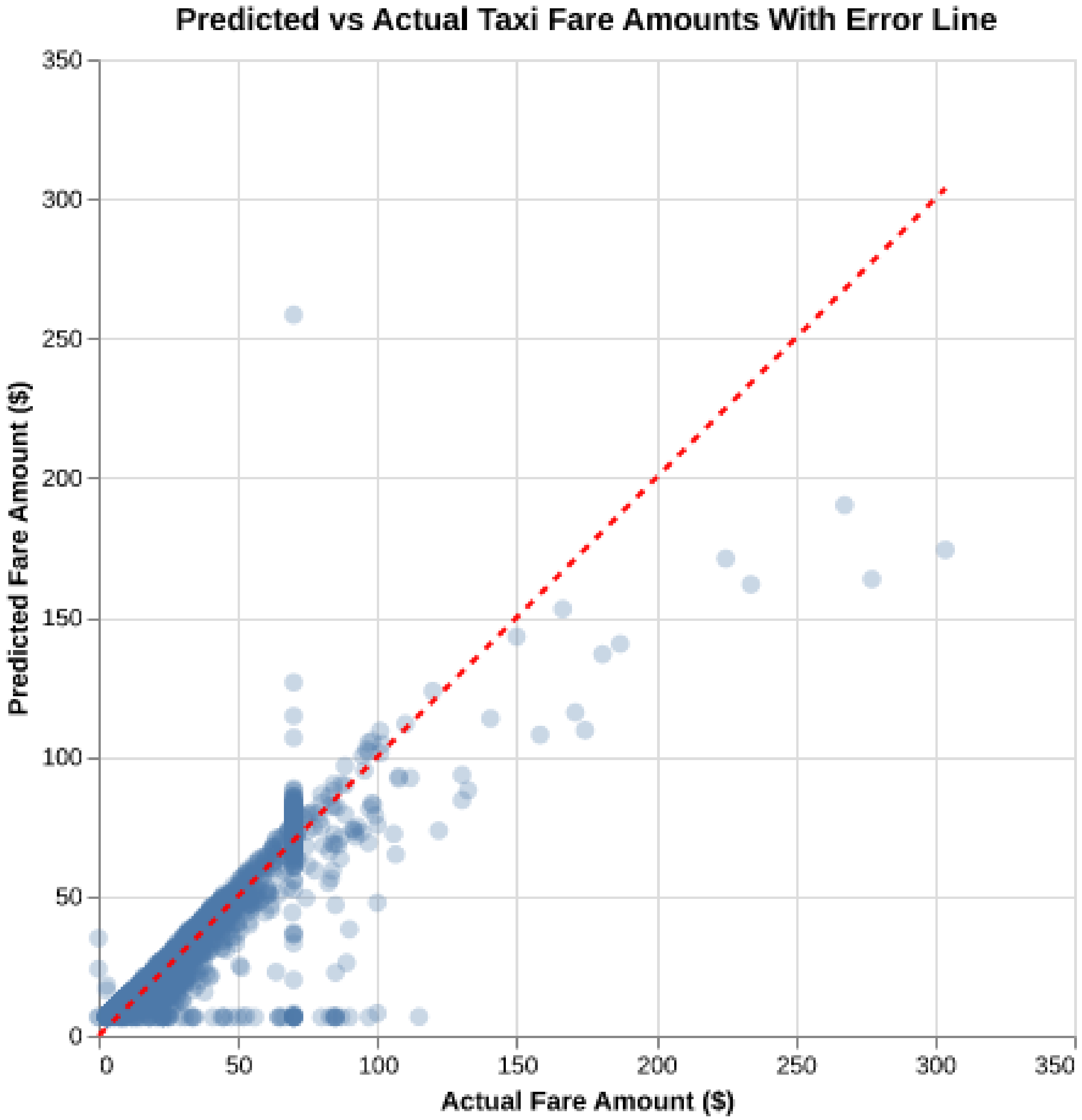

Figure 3: Predicted vs Actual Taxi Fare Amounts With Error Line

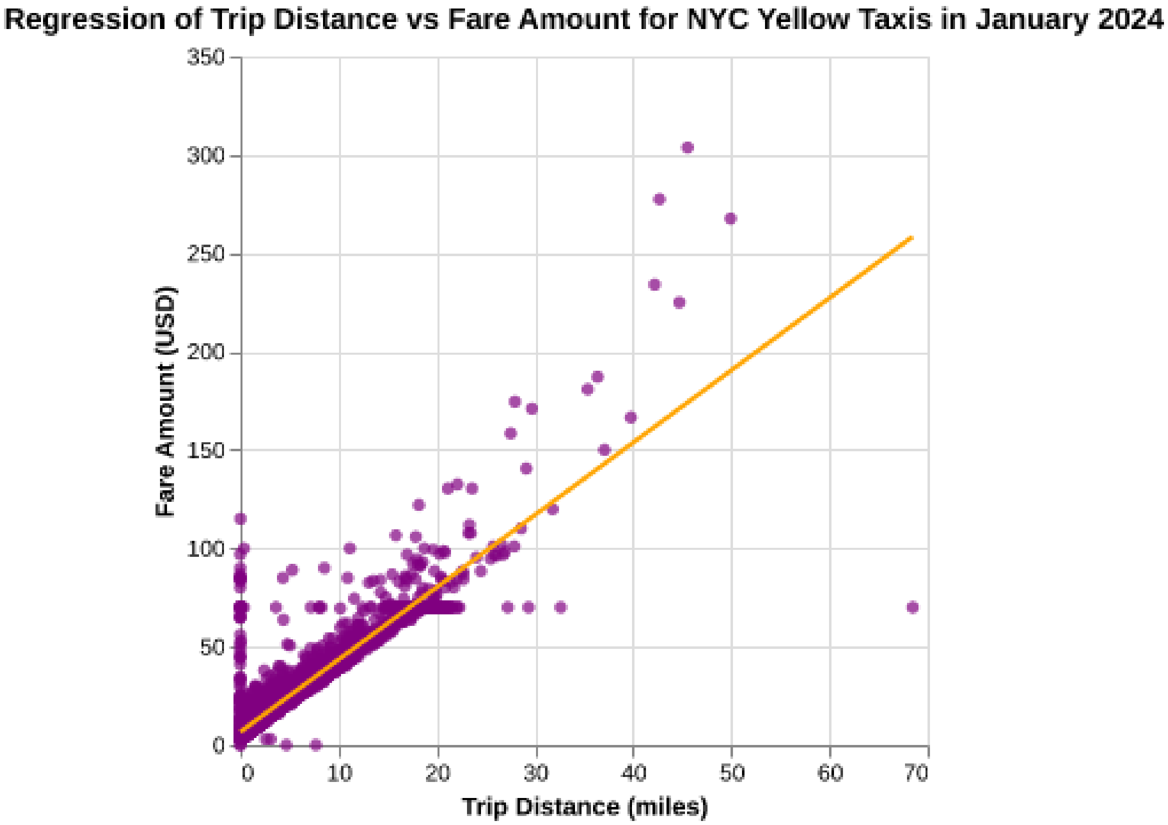

Figure 4: Regression of Trip Distance vs Fare Amount for NYC Yellow Taxis in January 2024

To evaluate the model's predictive performance, we applied it to the test data, which was split from the original dataset (70 percent training, 30 percent testing). The predicted fare amounts were compared against the actual fare amounts in the test set. The predicted values were calculated using the formula derived from the model, where trip distances from the test set were used as input to generate the corresponding fare predictions.

A scatter plot of the actual fare amounts versus trip distances, along with the regression line, was generated to visualize the results. Figure 4 showes that the model fits most of the data well, but there were some outliers where the predicted fare deviated significantly from the actual values. These discrepancies could be due to errors or special cases in the fare data.

Limitations & Next Steps

Overall, our model may be useful in the initial analysis of tourists interested in making informal predictions about NYC taxi fare amounts. However, there are several areas where this work can be improved. The linear regression model used in this analysis assumes a linear relationship between trip distance and fare amount. If this assumption doesn't hold, the model may miss important patterns, potentially leading to biased predictions. Additionally, the model assumes that residuals (errors) are independent and that the variance of residuals is constant across the range of trip distances. If these assumptions are violated, the accuracy of the model could be compromised. While multicollinearity is not an issue with the current model, it could arise if additional features are incorporated in future versions, leading to unstable coefficient estimates. We also did not include regularization parameters, such as , which are essential for controlling overfitting and improving the model's generalizability. Without these parameters, the model may be prone to overfitting, which reduces its ability to generalize to new data and may introduce bias.

In the next steps, we plan to address these limitations by incorporating additional features such as pick-up location and time of day, which may help capture more complexity in the fare prediction. We will also experiment with other models, like KNN regression to handle non-linear relationships and Lasso regression to improve regularization and model generalizability. These steps will improve the model's robustness and predictive power.

References

Harris, Charles R, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. “Array Programming with NumPy.” *Nature* 585 (7825): 357–62.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, =51–56.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12: 2825–30.

Taxi, New York City, and Limousine Commission. 2024. “TLC Trip Record Data.” https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jake. 2018. “Altair: Interactive Statistical Visualizations for Python.” *Journal of Open Source Software* 3 (7825, 32): 1057. https://doi.org/10.21105/joss.01057.