Green Taxis of New York City

A report on trends green taxi usage for September, 2015

Green Taxi pickup locations during September 2015 [1]

By Ashiq Rahman (ashiqur@gmail.com)

CONTENTS

INTRODUCTION

Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely-populated areas of Manhattan (the yellow dotted area in the picture below). This report analyzes the trends of the Green Taxi usage per the guidance of Data Science Challenge. The instructions are provided in the appendix.

NYC Taxi and Limousine Commission provides taxi trip records (e.g. pick-up and drop-off dates, times, locations, trip distances, itemized fares, rate types, payment types, and passenger counts) from 2009 till now. The data (and picture above) were collected from NYC Taxi and Limousine Commission trip data website [5]. The data science challenge asked to analyze the trip data for September 2015. The data was located at [6]. The data dictionary describing all the fields in the dataset was located at [7].

All the questions raised in the Data Science Challenge is answered in the next few sections. The author used several tools (e.g. Mapbox [2], Tableau [3], Python and few libraries) as they are needed in various parts of the analysis. The code/tool/techniques are pointed out in the reference section where they are described in detail.

QUESTION 1

Programmatically download and load into your favorite analytical tool the trip data for September 2015. Report how many rows and columns of data you have loaded.

The data was downloaded with a python script [8]. There were 1,494,926 rows and 21 columns.

QUESTION 2

Plot a histogram of the number of the trip distance ("Trip Distance"). Report any structure you find and any hypotheses you have about that structure.

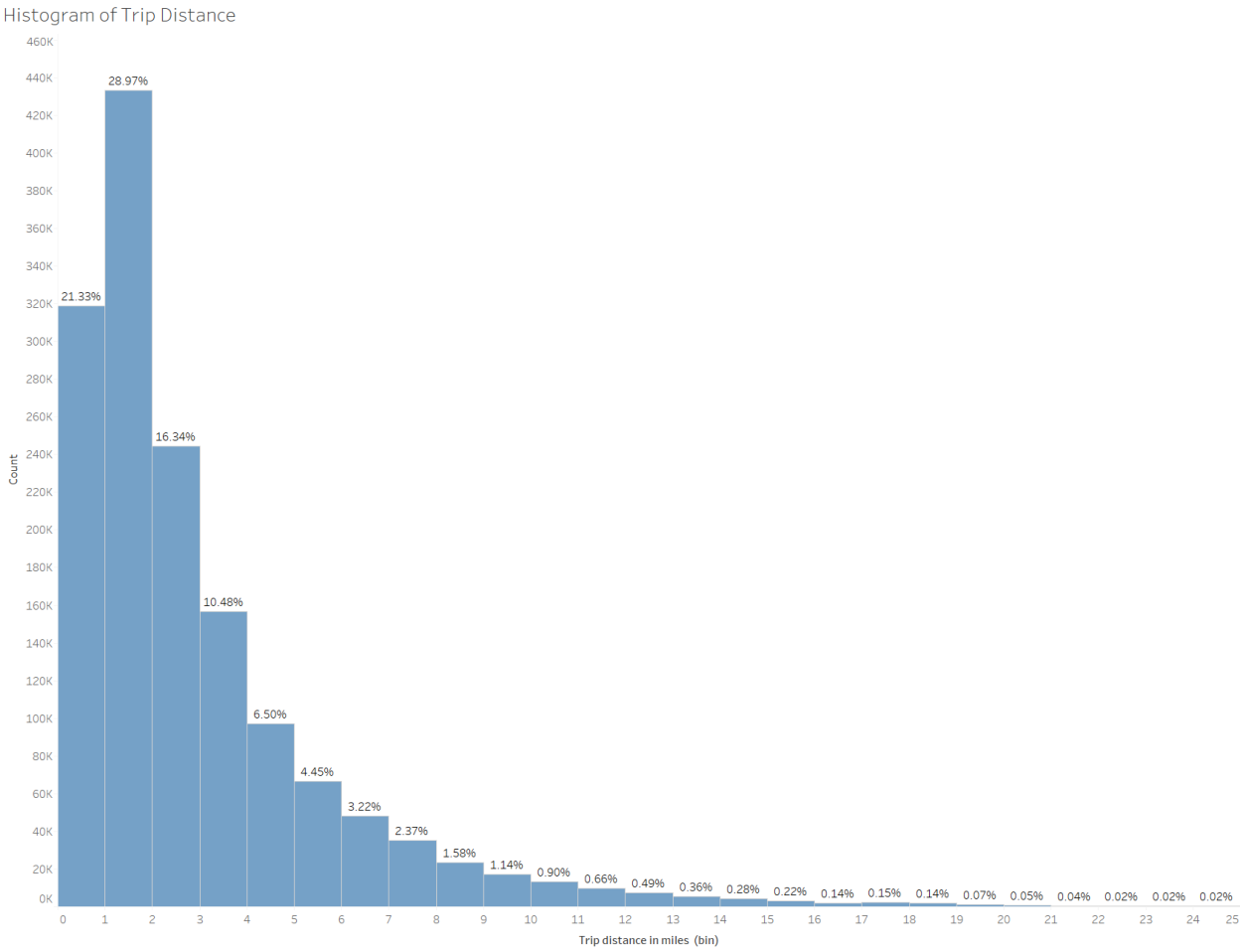


Figure Q2-1. Histogram of trip distance count with bin size 1 mile. Zoomed to 0-25 mile range

Figure Q2-1 above shows the histogram with 1 mile bin size. 29% of the total trips were between 1 to 2 mile range and 83.6% of the trips were less than 5 miles. The distribution falls off sharply (faster than linear) beyond 5 mile. The Tableau platform [10] was used to answer this question.

To observe finer granularity, the histogram is plotted with a reduced bin size (0.5 mile) and shown in the next page.

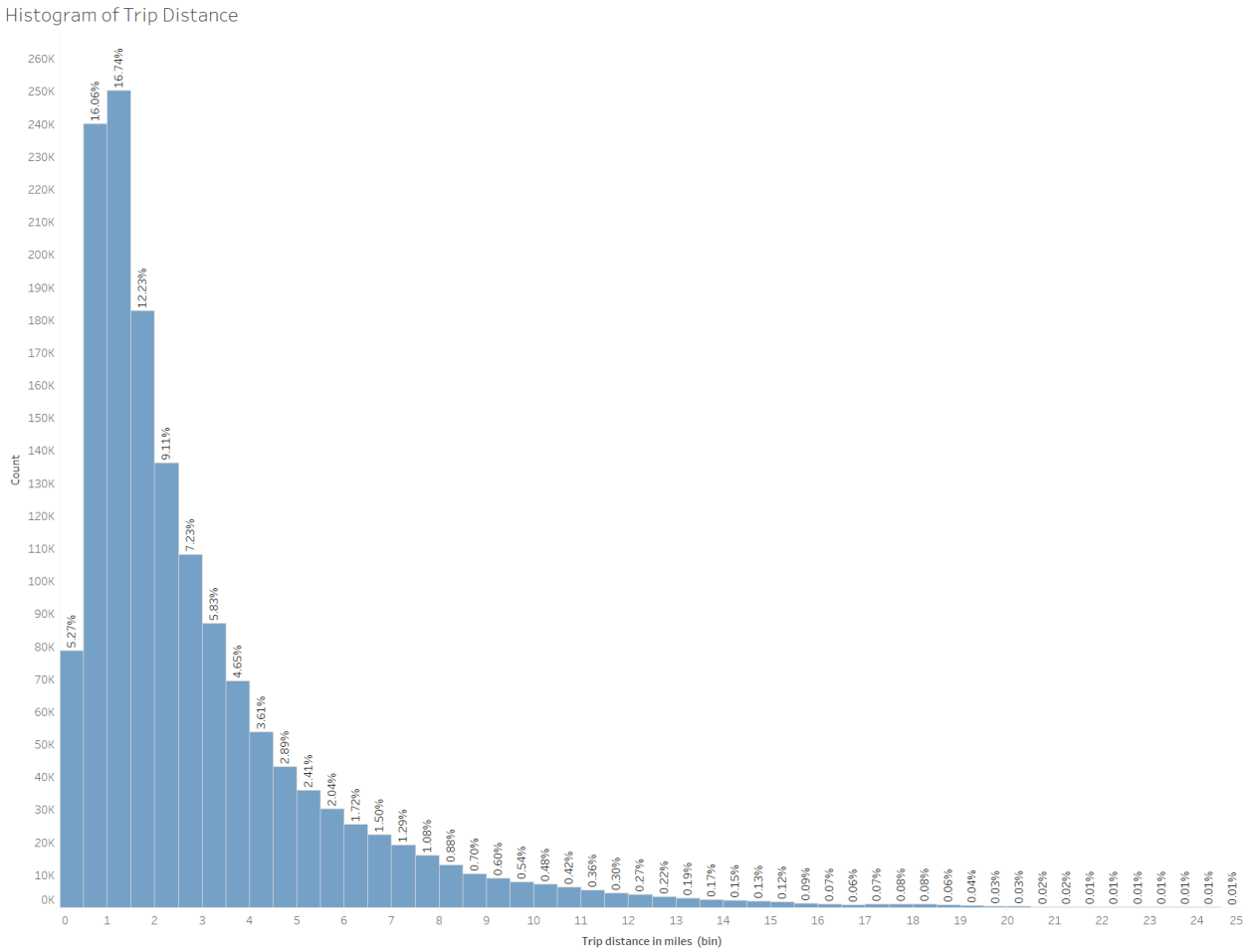


Figure Q2-2. Histogram of trip distance count with bin size 0.5 mile. Zoomed to 0-25 mile range

Figure Q2-2 above shows the histogram with 0.5 mile bin size. This smaller bin scale shows the distribution better. For example, 5.3% of the trips were less than 0.5 mile and a whopping 32.8% of the trips were between 0.5 to 1.5 mile range. The distribution peaked between 1 to 1.5 mile and sharply dropped from there.

The histogram is skewed to the right with a long tail. Also most of the trips were short distance – two thirds of the all the trips (66.6%) were within 3 miles.

The upper limit for trip distance was clipped at 25 mile for better visualization. Figure Q2-3 in the next page shows the histogram for the entire data range. Some large outliers were observed (one at 600 mile). The percentage of trips that were less than or equal to 25 miles was 99.1% [10].

Plotting the distribution (Figure Q2-4 in the next page) in even finer bin size (e.g. 0.1 miles) show additional structure but they were mostly due to noise (except for the 0 to 0.1 mile range, maybe).

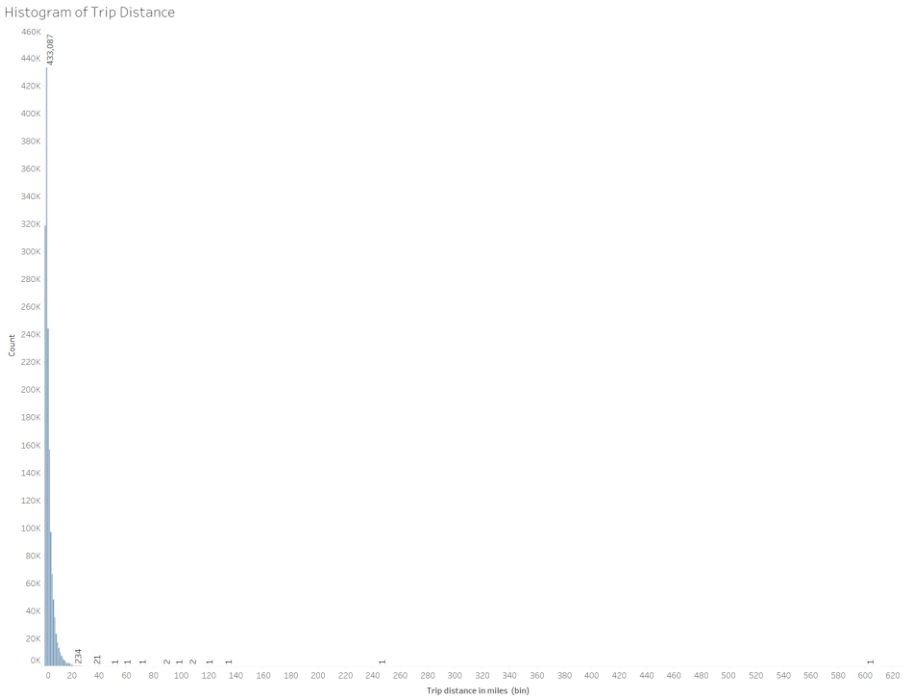


Figure Q2-3. Histogram of trip distance count with bin size 1 mile for the entire range of data

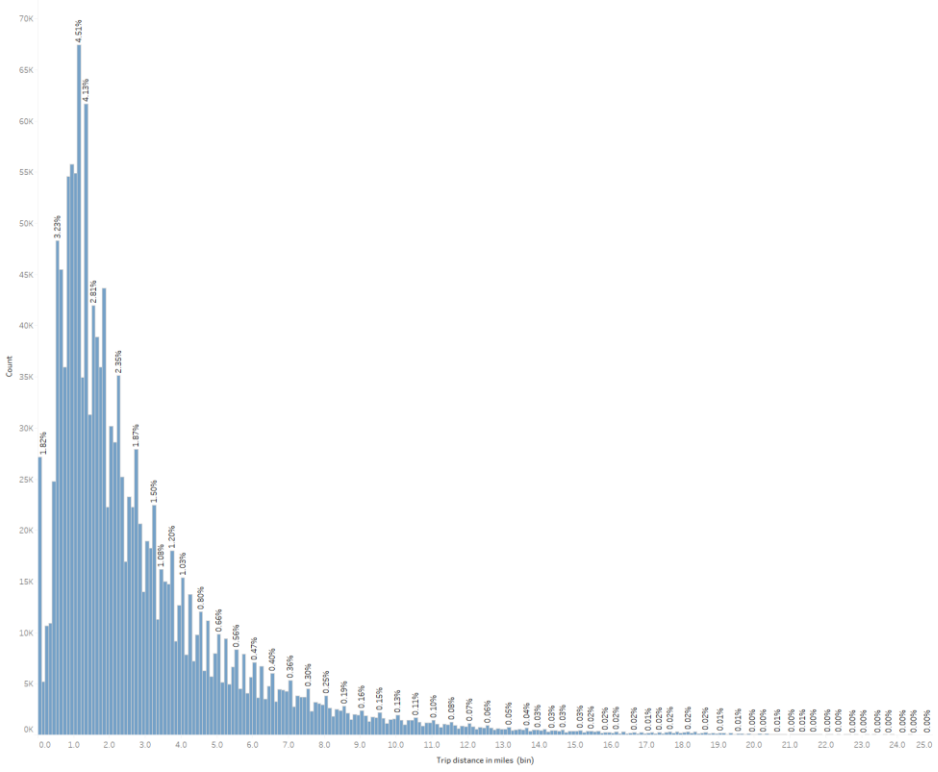


Figure Q2-4. Histogram of trip distance count with bin size 0.1 mile. Zoomed to 0-25 mile range

QUESTION 3

Q3 a. Report mean and median trip distance grouped by hour of day.

Figure Q3-1 (top) in the next page shows the mean and median trip distance grouped by the hour of the day [10]. Some observations are:

- Mean is greater than the median. This simply suggests that the distribution is skewed to the right (e.g. a tail stretching toward the right). This is consistent with previous observations in figures for question 2. Most of the trips are clustered within the short distances.
- Both average and median trip distance during the early morning hours (5-6am) are higher than the rest of days – a distinct 'peak' is observed in these hours. Why?

The author was curious about the 'the early morning' hour traffic and the next few paragraphs and figures examine this from several perspectives. These were not part of question 3a.

Trip count by hour: Figure Q3-1 (bottom) in the next page, shows the percent of total rides grouped by the hour of the day. The percentage of the trips in the early morning hours (5-6am) is the lowest compared to the rest of the day (e.g. 1.1% at 5am vs 6.5% at 6pm). Less number of trips but they go longer in average.

Trip distance box plot: Figure Q3-2 in page 9, shows the box plots of the trip distances grouped by hour. The box plots of the trip distances during 4-6am shows not only higher median compared to other hours, but also high percentiles (25, 75). The variances of trip distances during these hours are also higher. This means there are both shorter and longer travels during these hours.

Trip distance histogram: To better understand the trend, we compare the histogram of trip distances for 5am and 6pm in Figure Q3-3 in page 10. The frequency distribution at 5am is "wider" than the one for 6pm. There are more longer (distance wise) trips in proportion at 5am compared to 6pm.

The question arises, where were people going in these early morning hours? Could this be, people were going to work from home? Are there people living in say Brooklyn or Queens that take Taxis to go to work in Manhattan? Or are people are going to the airports to catch the first flights out of town? Or both?

Weekday vs weekend: To explore, if this is a work or airport related phenomenon, figure Q3-4 in page 11 shows the average and median trip distance grouped by the hour for each day of the week. The figure notes the max and min numbers for each plot.

During the weekdays (Monday to Friday), distinct peaks occur during 4-6am whereas during weekends (Saturday & Sunday), the peaks are relatively more gradual during 5-9am. Could this be the rush of arriving at work before cutoff time during weekdays vs flexible time during weekends?

Pickup drop-off locations in Map: To explore, exactly where people are going in the early hours, figure Q3-5 in page 12 shows the pickup and drop-off locations during 5am on weekdays (Mon-Fri).

The pickup locations show clusters in upper Manhattan, North-West Brooklyn and North-West Queens. The drop-off locations were more diverse. But there is a significant cluster in middle of Manhattan – specifically in between 30th to 59th street and 2nd to 8th avenue (oval highlight in figure). Also, there are clusters around La Guardia and JFK airport (circular highlights in figure).

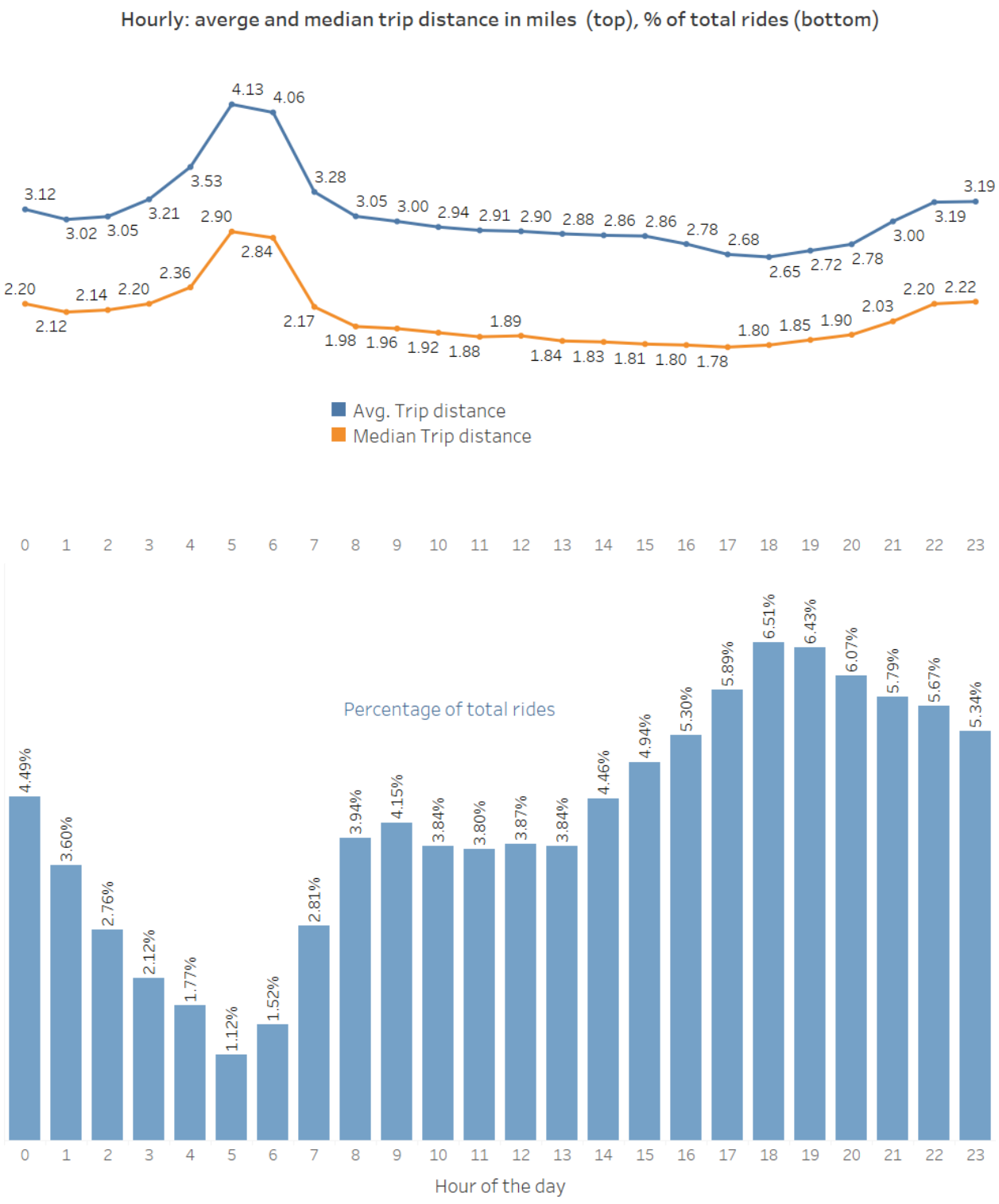


Figure Q3-1. (Top) Mean and median trip distance grouped by the hour of the day, (Bottom) Percent of total rides grouped by the hour of the day

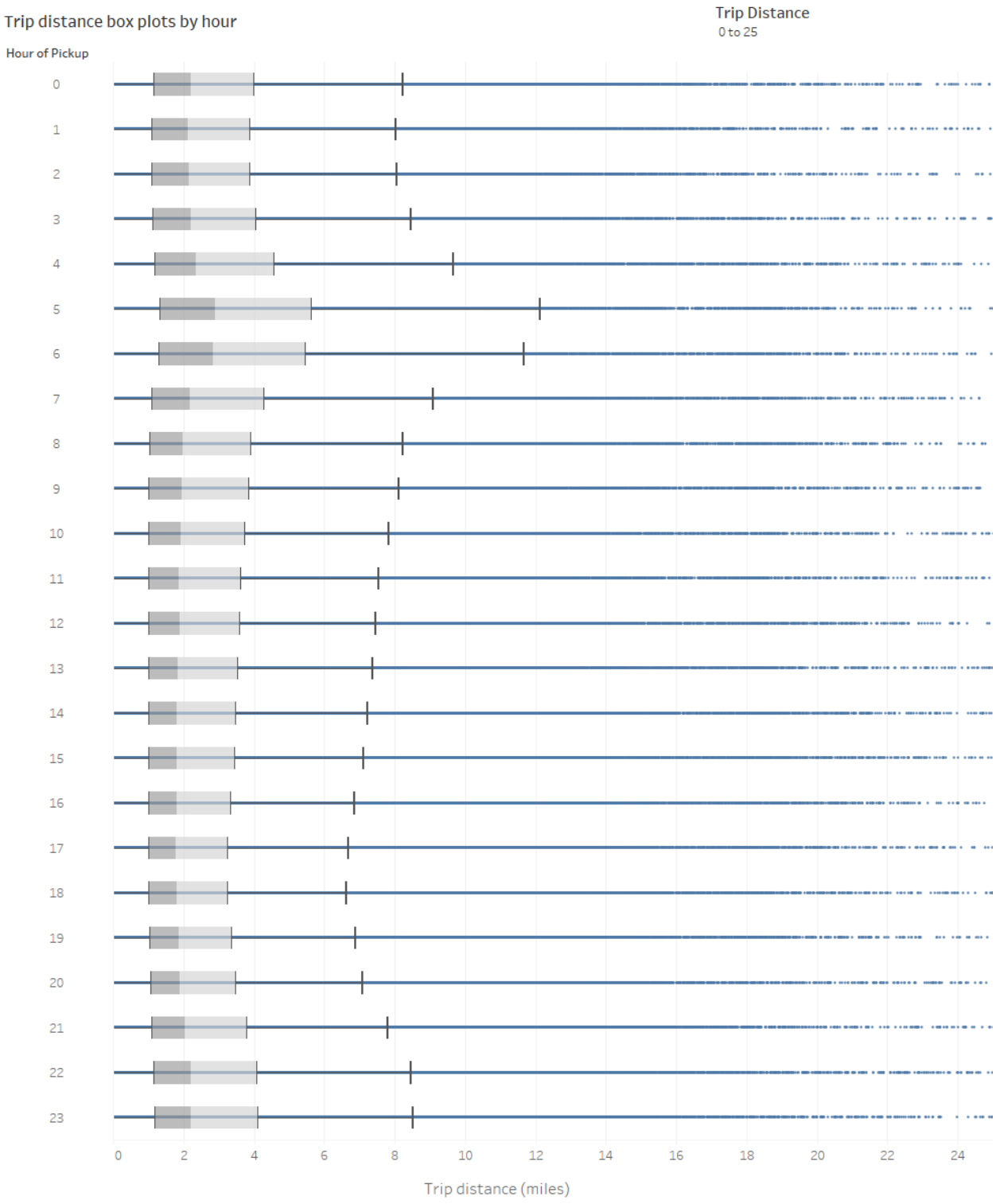


Figure Q3-2. Trip distance boxplot by hour

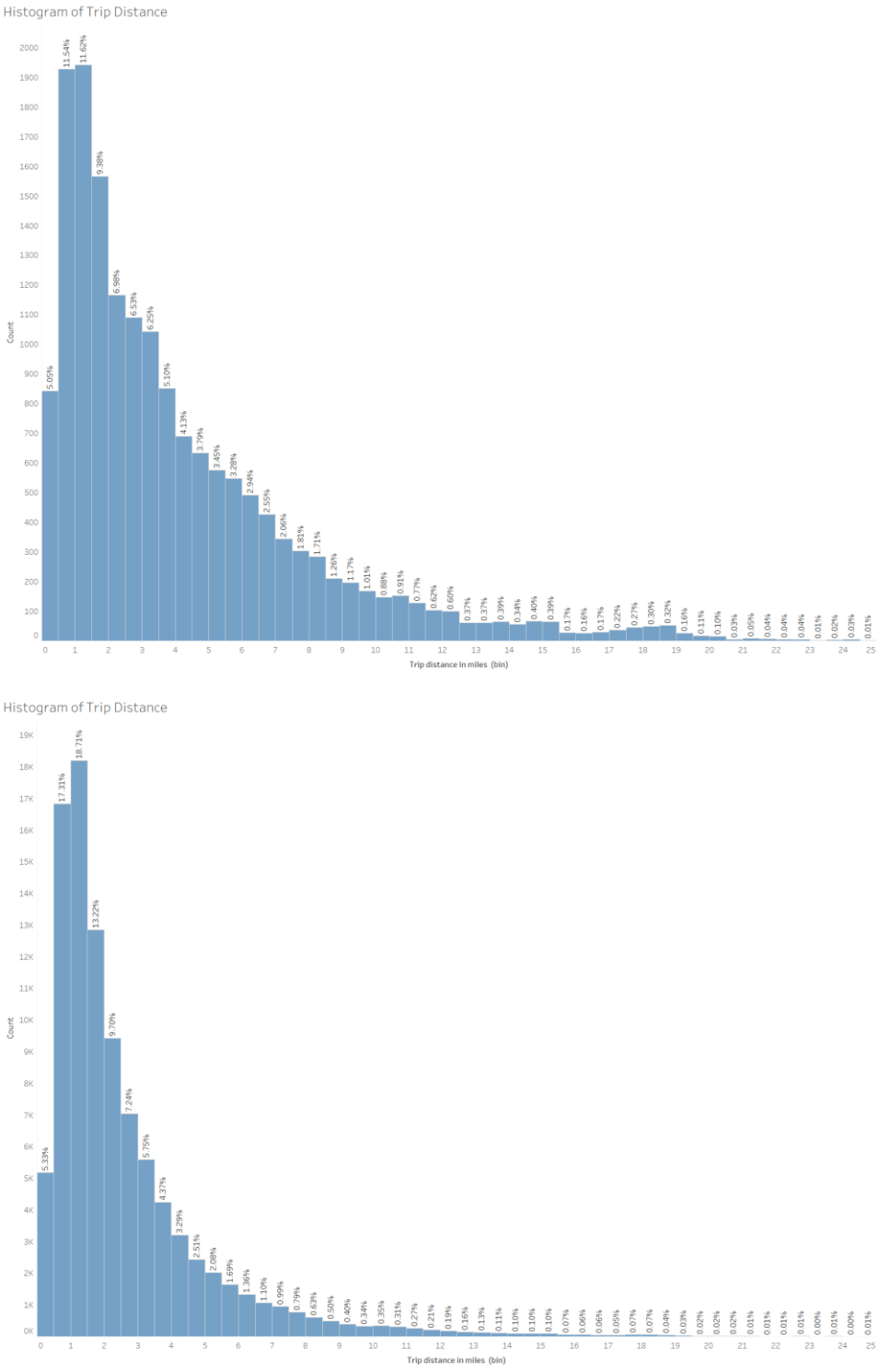


Figure Q3-3. Histogram of trip distance at 5am (top) and 6pm (bottom)

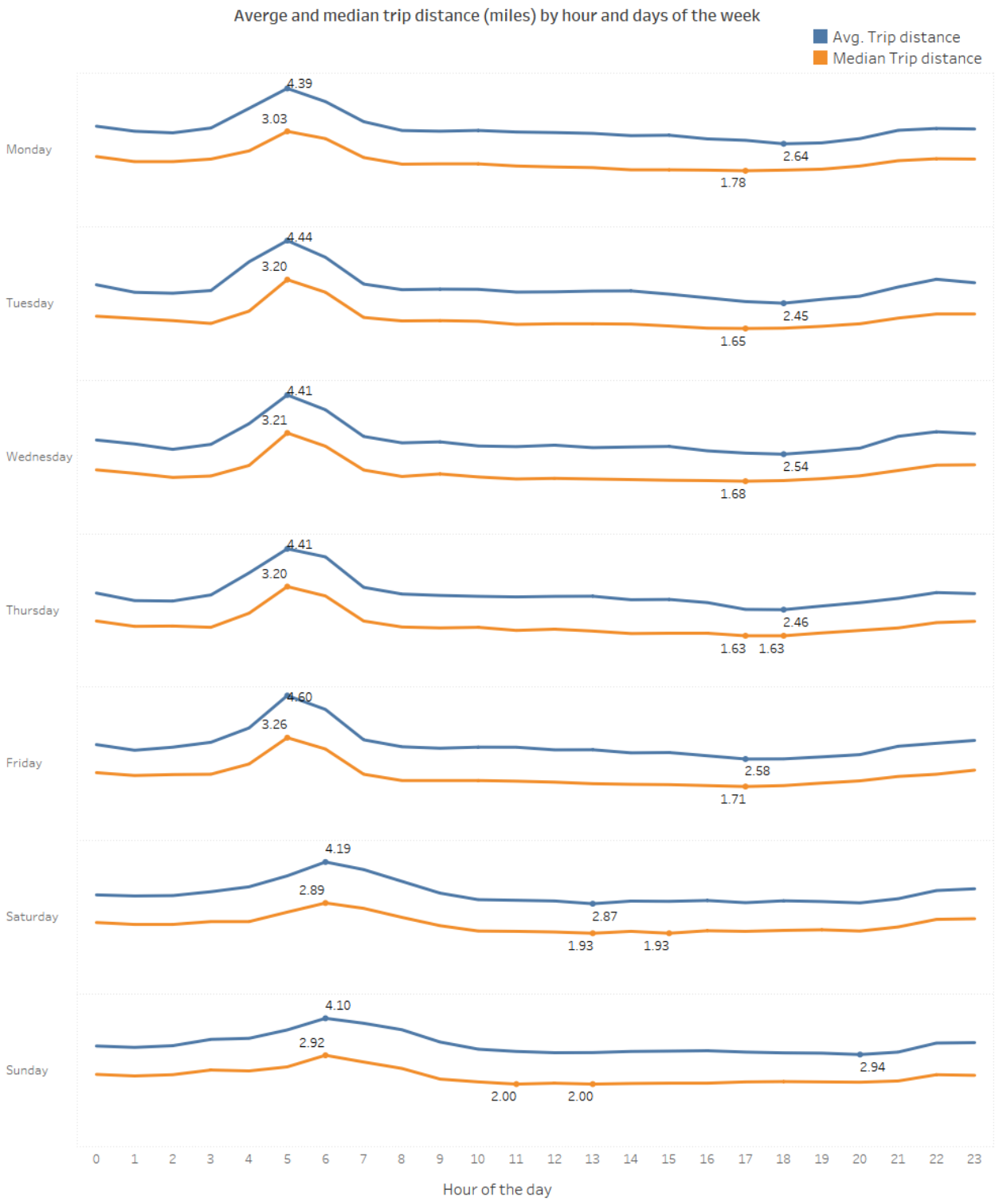


Figure Q3-4. Average and median trip distance grouped by the hour for each day of the week

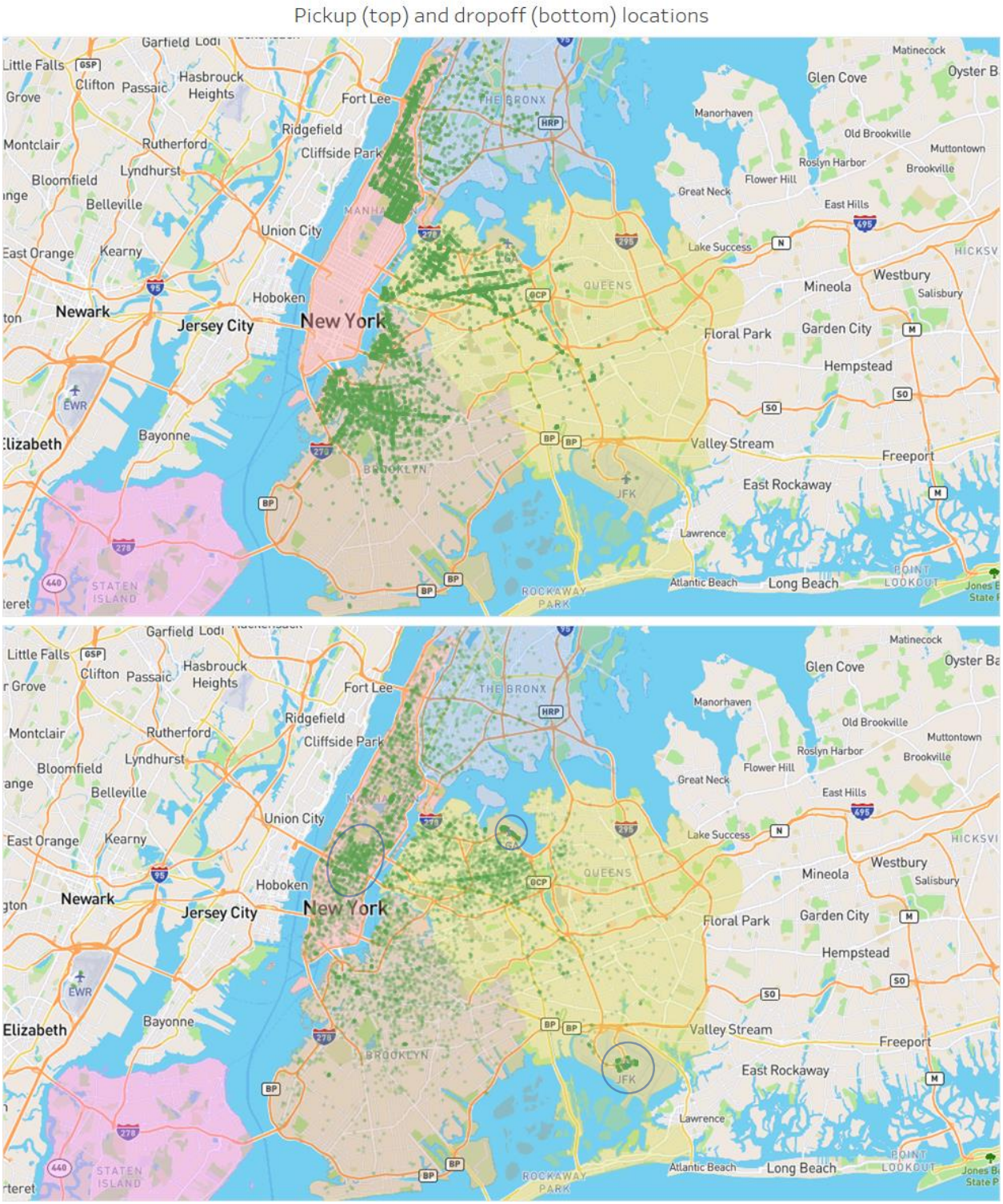


Figure Q3-5. Pickup (top) and drop-off (bottom) locations during 5am on weekdays (Mon-Fri)

Q3 b. We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

Motivation

Two different methods were used to classify the airport transactions – one gave inconsistent ('wrong' and interesting) results and the other more logical one. The methods had to do with the question: How do we find trip that originate/terminate at one of the NYC airports?

There are three airports around NYC:

- John F. Kennedy International Airport (JFK)
- LaGuardia Airport (EWR)
- Newark Liberty International Airport (LGA)

Since we have the drop off and pick up latitude longitude coordinates, we can calculate the distance from each of the airports and classify the trips - originated at airport, ended at airport, non-airport (**Distance method**).

Since the question asked to get a *'rough sense'* for trips that originate *or* terminate at *one* of the NYC area airports, we can use the field called "Rate Code ID" that can identify airport trips. The NYC Taxi and Limousine site [5] has detailed fare information for these airport trips.

- Rate Code ID = 1 Standard rate
- Rate Code ID = 2 To/From JFK and any location in Manhattan
- Rate Code ID = 3 To Newark Airport

Rate Code ID = 2 (JFK) can be used to get a rough sense for the trips between JFK and Manhattan (**Rate Code method**).

The 'Rate Code method' was used first since it's a simple one but the analysis showed it was not a good approach. However, it showed some interesting results. For example, for some of these JFK Rate code trips neither the pick up or drop-off were next to airport - the trip distance is less than 0.5 miles but fare paid on average was $46. These findings are presented in the next few pages.

The distance method showed consistent results but it also showed another interesting fact – there were more drop-offs to airports compared to pickups from the airport. Again, the results are discussed in detail in the next few pages.

For both methods, a new field 'Trip duration' in minutes (difference between pickup time and drop-off time) was calculated. The calculations and visualizations are done with the Tableau platform [10].

Rate Code method

The table below shows the distribution by the description of Rate Code ID.

Ride count distribution by Rate Code ID

Rate Code ID	Number of Records	
Standard rate	1,454,464	97.3%
JFK	4,435	0.3%
Newark	1,117	0.1%
Nassau or Westchester	925	0.1%
Negociated fare	33,943	2.3%
Group ride	36	0.0%
??	6	0.0%

Q3 Table 1

4,435 (0.3% of total) trips were JFK rate coded during the month of September 2015. Also, the table below show additional metrics (average trip distance, trip duration, fare amount) by Rate Code ID.

Distribution by Rate Code ID

Rate Code ID	Number of Records	Avg. Trip distance	Avg. Trip duration (min)	Avg. Total fare amount	Total fare amount
Standard rate	1,454,464	2.9	20	$14.74	$21,439,333
JFK	4,435	10.2	30	$56.53	$250,718
Newark	1,117	10.9	30	$59.90	$66,903
Nassau or Westchester	925	14.9	44	$68.77	$63,610
Negociated fare	33,943	3.4	24	$19.19	$651,210
Group ride	36	1.3	1	$2.74	$99
??	6	0.0	160	$12.18	$73

Q3 Table 2

The trips with JFK rate code were on average 10.2 mile and 30 minutes. The average fare for these trips was $56.53.

Since summary statistics like average, does not tell the whole story, the trip distance histogram is plotted in the next page.

The trip distance distribution below shows a peak around 20 mile which is expected (Manhattan is around 20 miles from JFK). But 41.1% of JFK coded trips were less than 0.5 mile with $46 on average [10].

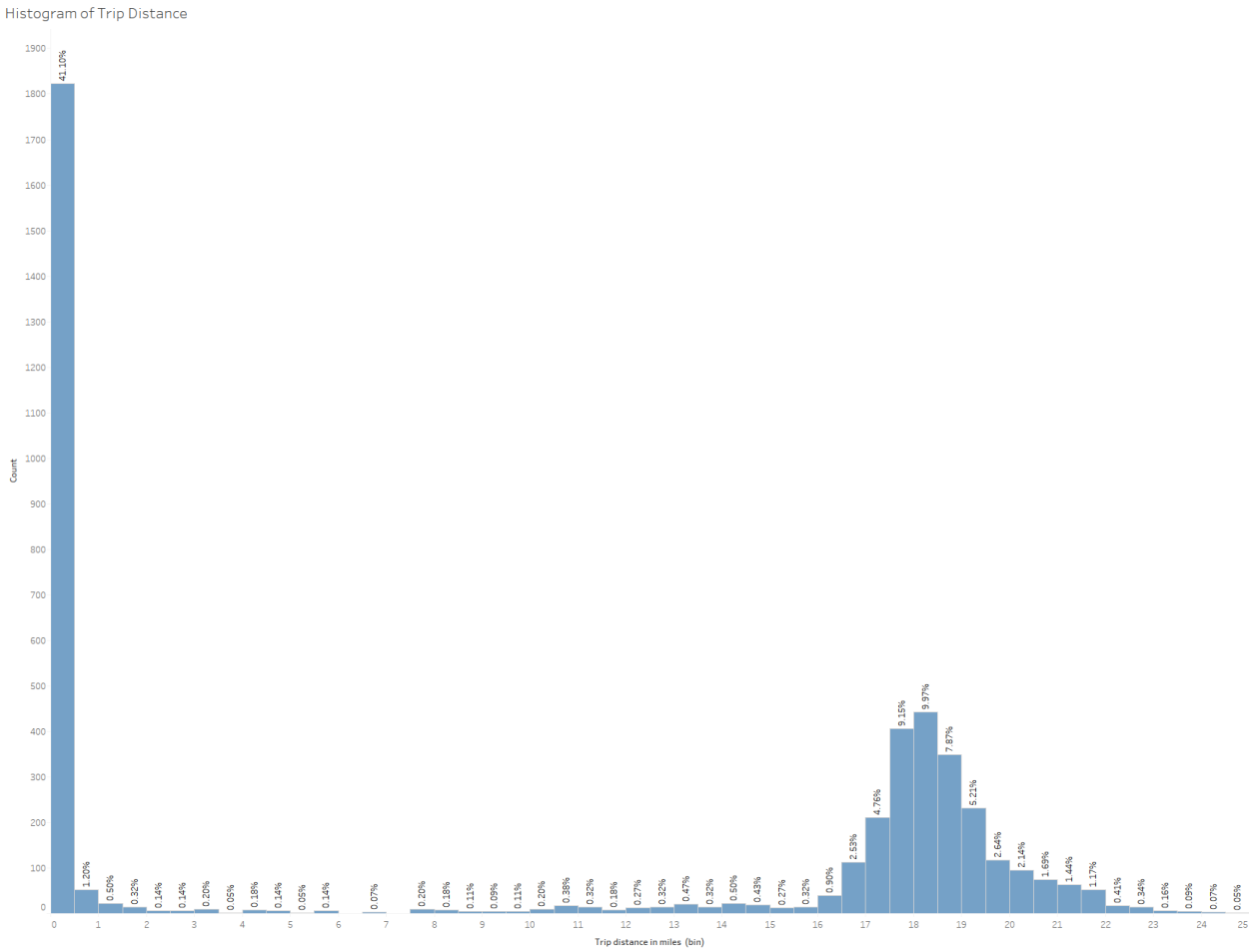


Figure Q3-6. Trip distance distribution for JFK Rate Coded trips

To investigate further, the trip distance histogram group by hour is plotted in the next page. There were hardly any 'normal' (20 mile) airport traffic during the middle of night hours of 11pm – 3am but there were several 'less than half mile' trips during this time.

Lastly, the pickup and drop locations of these '*less than half mile*' JFK rate coded trips are plotted in a map view in page 17. The pickup and drop-off locations are all over Bronx, Manhattan, Queens and Brooklyn. The JFK airport itself is more than *one mile* across. There are few possibilities:

- *the trips were mislabeled*
- *the trips were round trips to airport*
- *the trips did not originate or terminate at JFK airport but they are coded/charged as JFK fare.*

Since 41% data was uncertain due to Rate Code field, a decision was made to choose the distance method that did not use the Rate Code field.

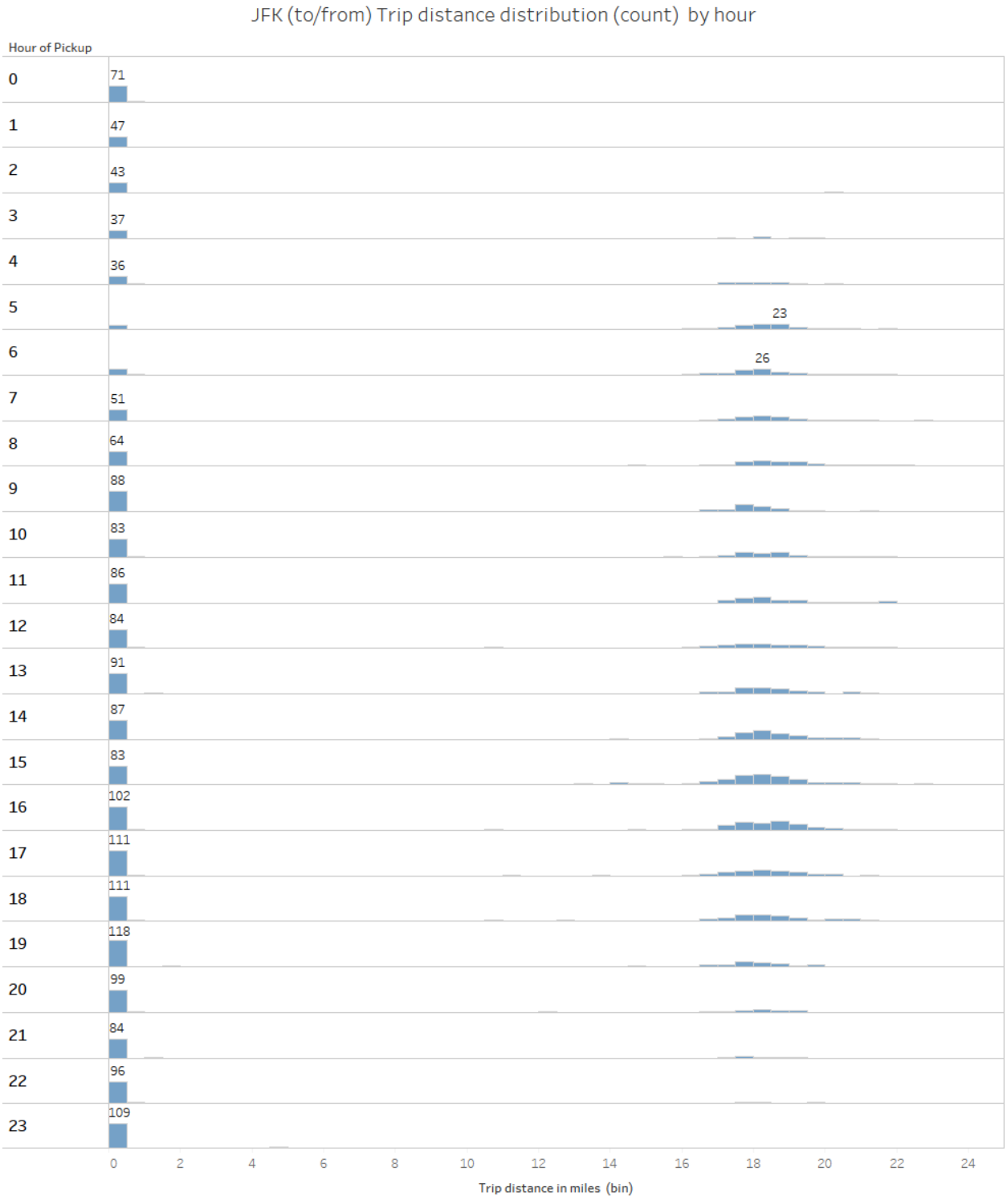


Figure Q3-7. Trip distance distribution for JFK Rate Coded trips

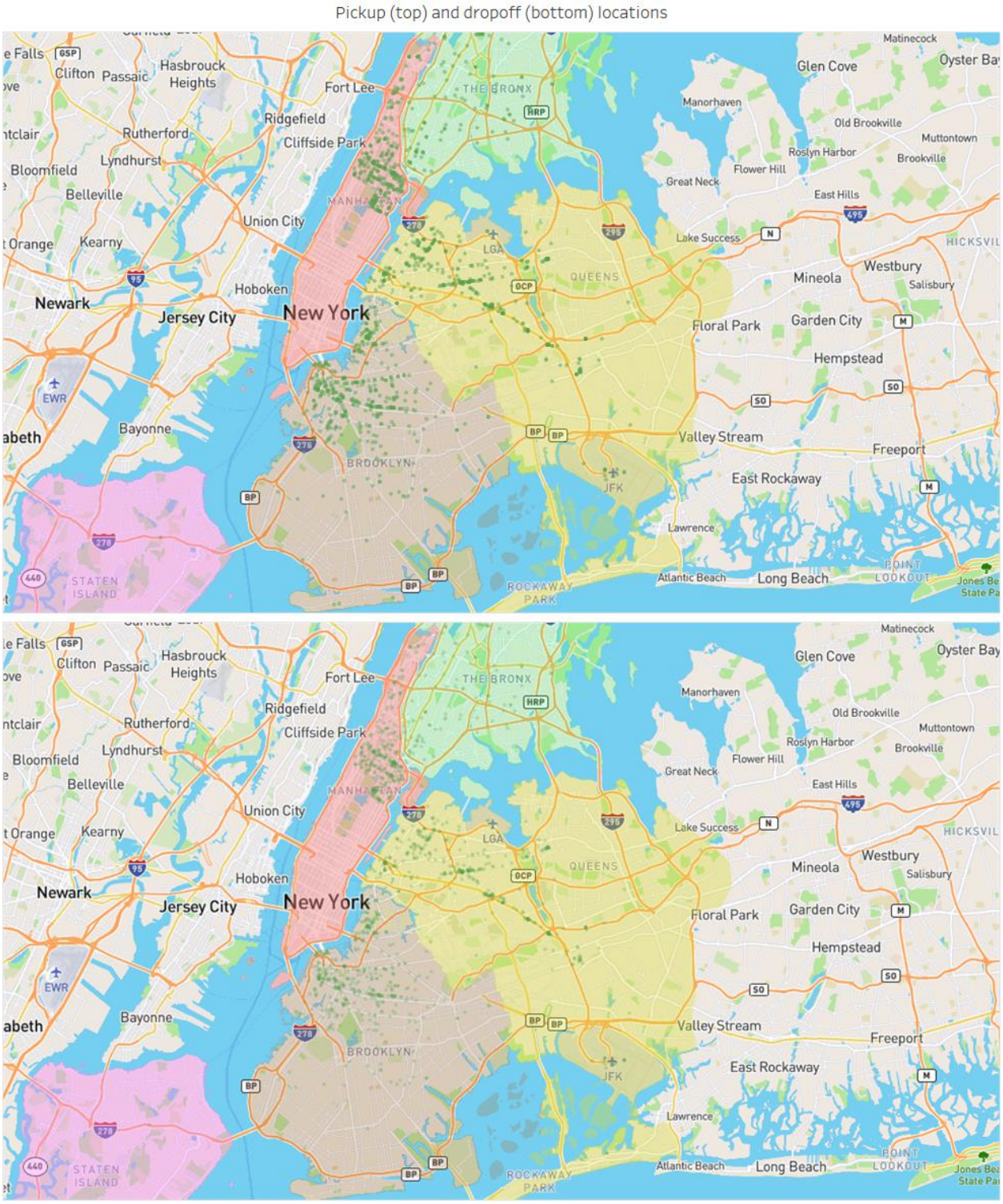


Figure Q3-8. Pickup and drop-off locations for JFK Rate Coded trips that are less than half a mile. JFK airport itself is more than one mile across.

Distance method

Using the distance from the airports, each of the trip was classified whether if it was picked/dropped at the airports or not. First using google maps, latitude and longitude of the airport location and a radius distance was determined such that all points that falls with the radius from that location will be tagged for the airport.

The calculations were done in the python notebook [11]. The python package pyproj [12] was used to calculate the distance between the locations (latitude and longitude pairs) by using the WGS84 standard [13]. If the pickup and drop-off locations are near the 3 major airports of NYC, the records were tagged accordingly (in the new variables called Pickup tag, Dropoff tag and Airport tag). The table below shows the airport related trip counts.

	Dropoff tag				
Pickup tag	JFK	LGA	EWR	Non Airport	Grand Total
JFK	200	18		87	305
LGA	20	209	1	145	375
EWR			39	2	41
Non Airport	12,738	20,736	669	1,460,062	1,494,205
Grand Total	12,958	20,963	709	1,460,296	1,494,926

JFK	John F. Kennedy International Airport
LGA	LaGuardia Airport
EWR	Newark Liberty International Airport

Q3 Table 3. Airport related trip counts.

The first observation is that Green Taxis were used mostly for dropping off to airport. For example, the table shows 12,958 trips were for dropping off to JFK whereas only 305 trips were for pickup! Why? From Wikipedia [14]

“Boro [Green] taxis are taxis in New York City that are allowed to pick up passengers (street hails or calls) in outer boroughs (excluding John F. Kennedy International Airport and LaGuardia Airport unless arranged in advance) and in Manhattan above East 96th and West 110th Streets.”

The author was unaware of this airport pickup restrictions and learned from observing and questioning the data. The 305 pickups from JFK mentioned above were prearranged then. Also, another interesting observation is that there were 448 airport to airport trips.

The tables (Q3 4a, Q3 4b) in the next page show the average fare and total revenue by airport related pickup and drop-off. Total revenue from airport related trips $1.25M is 5.57% of the total $22.5M revenue whereas the total number of airport related trips, 34,630 is only 2.32% of the total 1.5M trips.

As expected, the average fare to/from airport is higher than non-airport trips (e.g. $48.70 to JFK fare vs $14.53 non-airport fare).

	Dropoff tag				
Pickup tag	JFK	LGA	EWR	Non Airport	Grand Total
JFK	$45.35	$20.50		$35.08	$40.96
LGA	$30.56	$30.50	$89.55	$29.77	$30.38
EWR			$73.72	$66.05	$73.34
Non Airport	$48.78	$26.28	$99.05	$14.53	$15.02
Grand Total	$48.70	$26.31	$97.65	$14.53	$15.03

Q3 Table 4a. Average fare by airport related pickup and drop-off

	Dropoff tag				
Pickup tag	JFK	LGA	EWR	Non Airport	Grand Total
JFK	$9,071	$369		$3,052	$12,492
LGA	$611	$6,374	$90	$4,316	$11,391
EWR			$2,875	$132	$3,007
Non Airport	$621,352	$544,869	$66,267	$21,212,569	$22,445,056
Grand Total	$631,034	$551,612	$69,231	$21,220,069	$22,471,946

Q3 Table 4b. Total revenue by airport related pickup and drop-off

For the rest of the analysis, we focus on one airport – JFK and explore the trip to and from JFK in detail. Figure Q3-9 in the next page shows the trip distance, fare and duration histograms for pickup and drop-off trips to JFK and it also shows the pickup/drop-off locations in the map. The trips to/from JFK were on average 13.7 miles, 34 mins and $48.57.

The trip distance histogram shows three distinct regions (5-8, 10-16, 17-19.5 miles) and figures Q3-10, Q3-11, Q3-12 in pages 21-23 focus on these regions respectively. The figures are self-explanatory – the histograms and location maps, average metrics all updated with the location selections.

Figure Q3-13 in page 24 shows details for the trips that originated or terminated within 1 mile from JFK. There were 234 such trips – most of them are terminal to terminal short transfers (trip duration were 5 minutes or less). The fares for such short trips were diverse – from $5 to $85. And some trips 40 min and longer.

The fare distribution in Figure Q3-9 (page 20) shows the most common fare (in $2.5 bin size) was $57.5 - $60 (9.4% of total). Figure Q3-14 in page 25 shows the details for 1230 such trips. Most of trips are from Manhattan and 773 of them using the Rate Code = 2 (JFK).

Figure Q3-15 in page 26 shows the average and median trip distance and trip count for the JFK trips by the hour of the day. The night seems relatively quiet and trip to airport starts to build up from 3am with the morning peak during 5-6am. The afternoon peak for JFK traffic happens during 3-4pm.

Lastly, figure Q3-16 in page 27 shows the median trip distance and trip count for the JFK trips by the hour and day of the week. Both of the metrics pattern by hour is similar for each day (i.e. weekday or weekend).

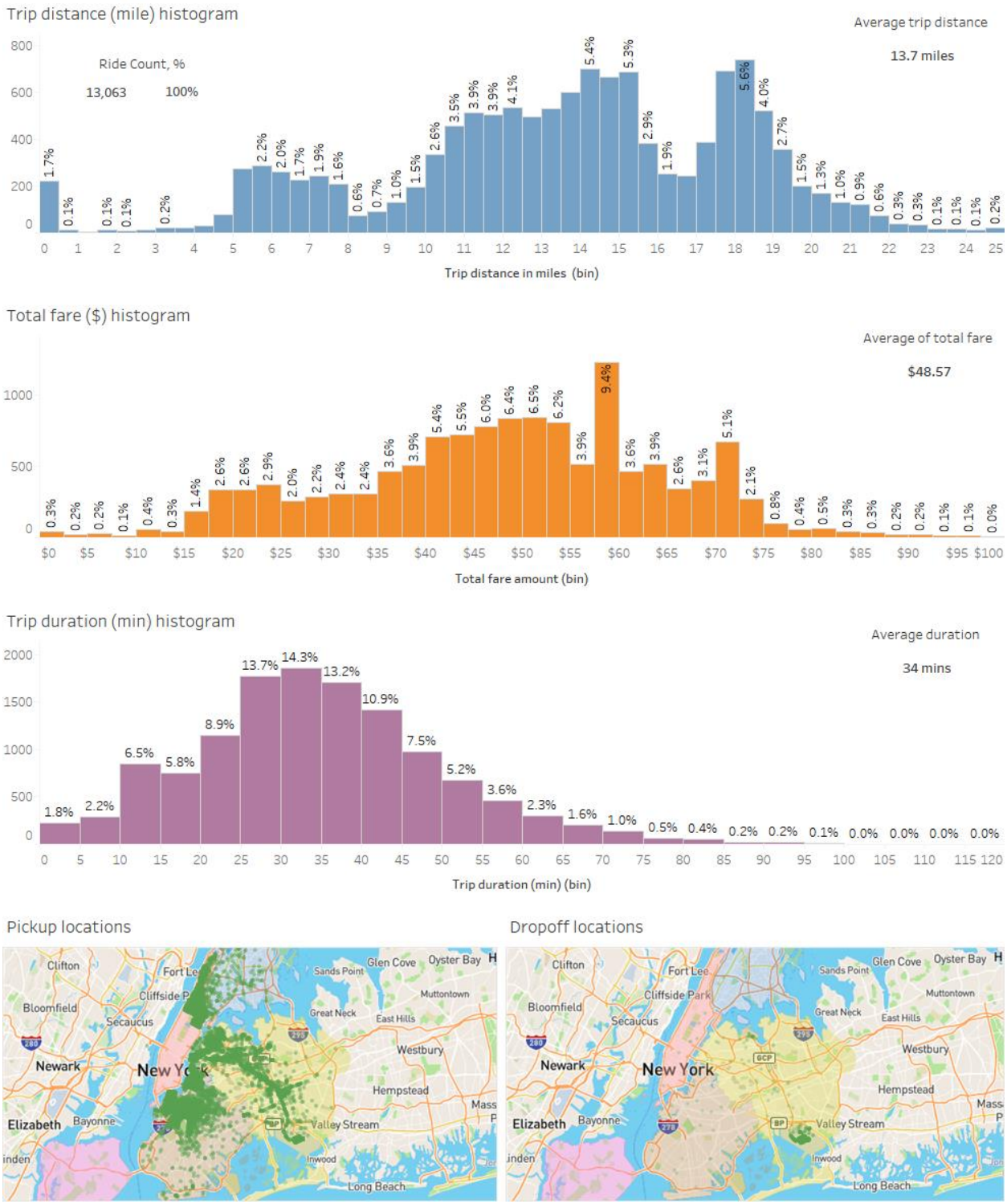


Figure Q3-9. Trip to and from JFK airport – trip distance, fare, duration histograms and locations.

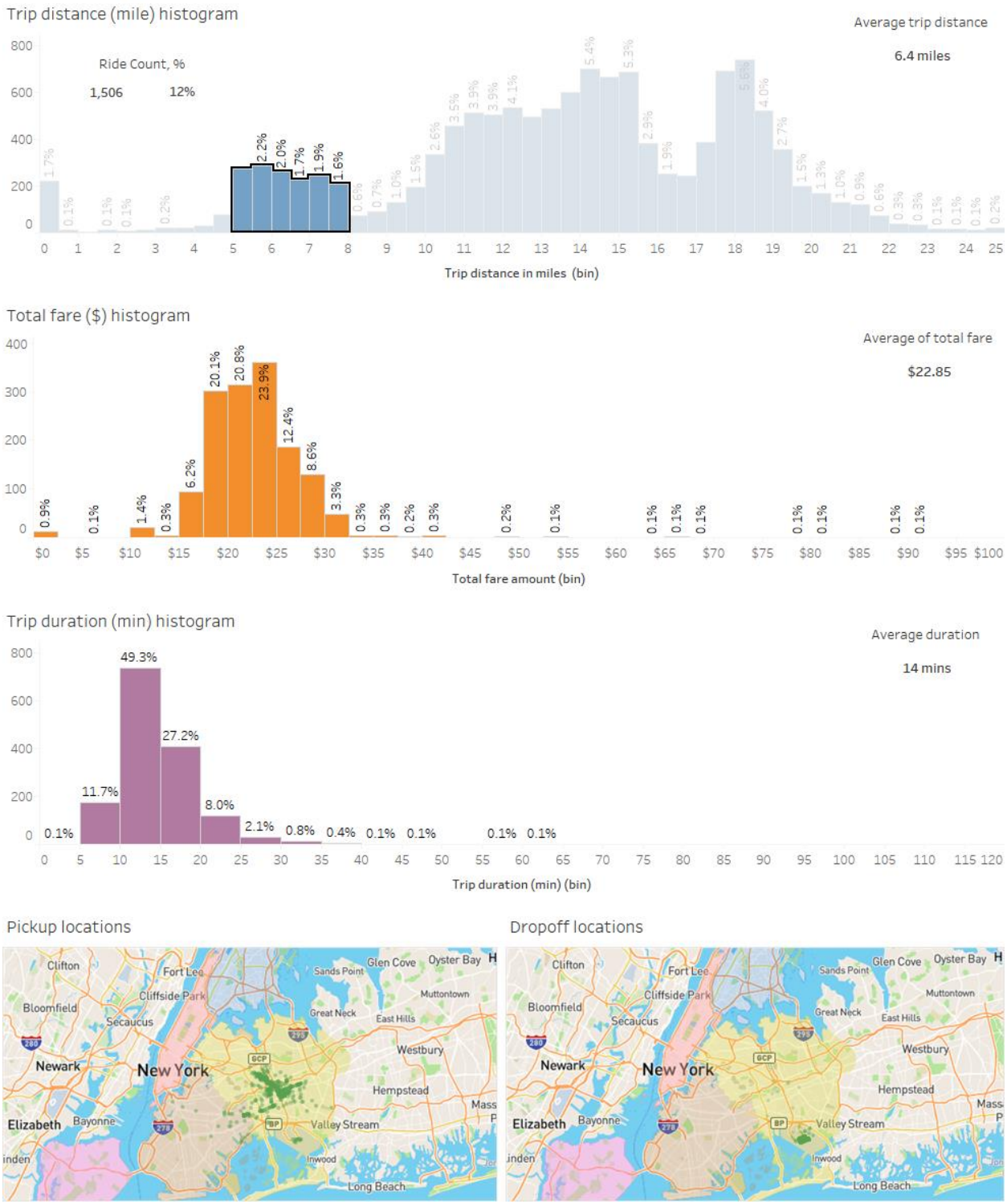


Figure Q3-10. Trip to and from JFK airport – trip distance, fare, duration histograms and locations for trips that originated or terminated 5-8 miles from JFK.

Pickup/drop-offs mostly from mid Queens area.

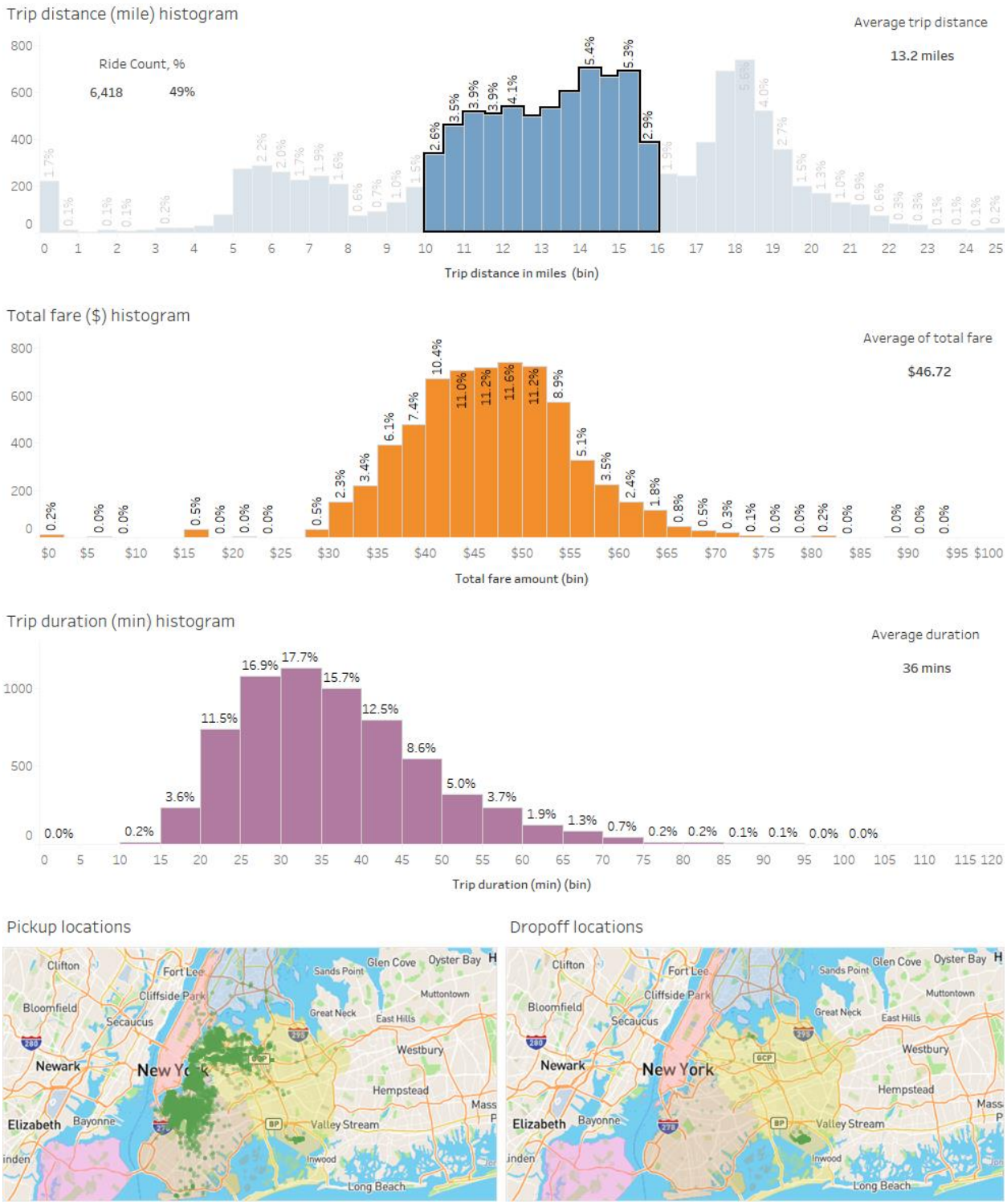


Figure Q3-11. Trip to and from JFK airport – trip distance, fare, duration histograms and locations for trips that originated or terminated 10-16 miles from JFK.

Pickup/drop-offs mostly from north-west Brooklyn/Queens area.

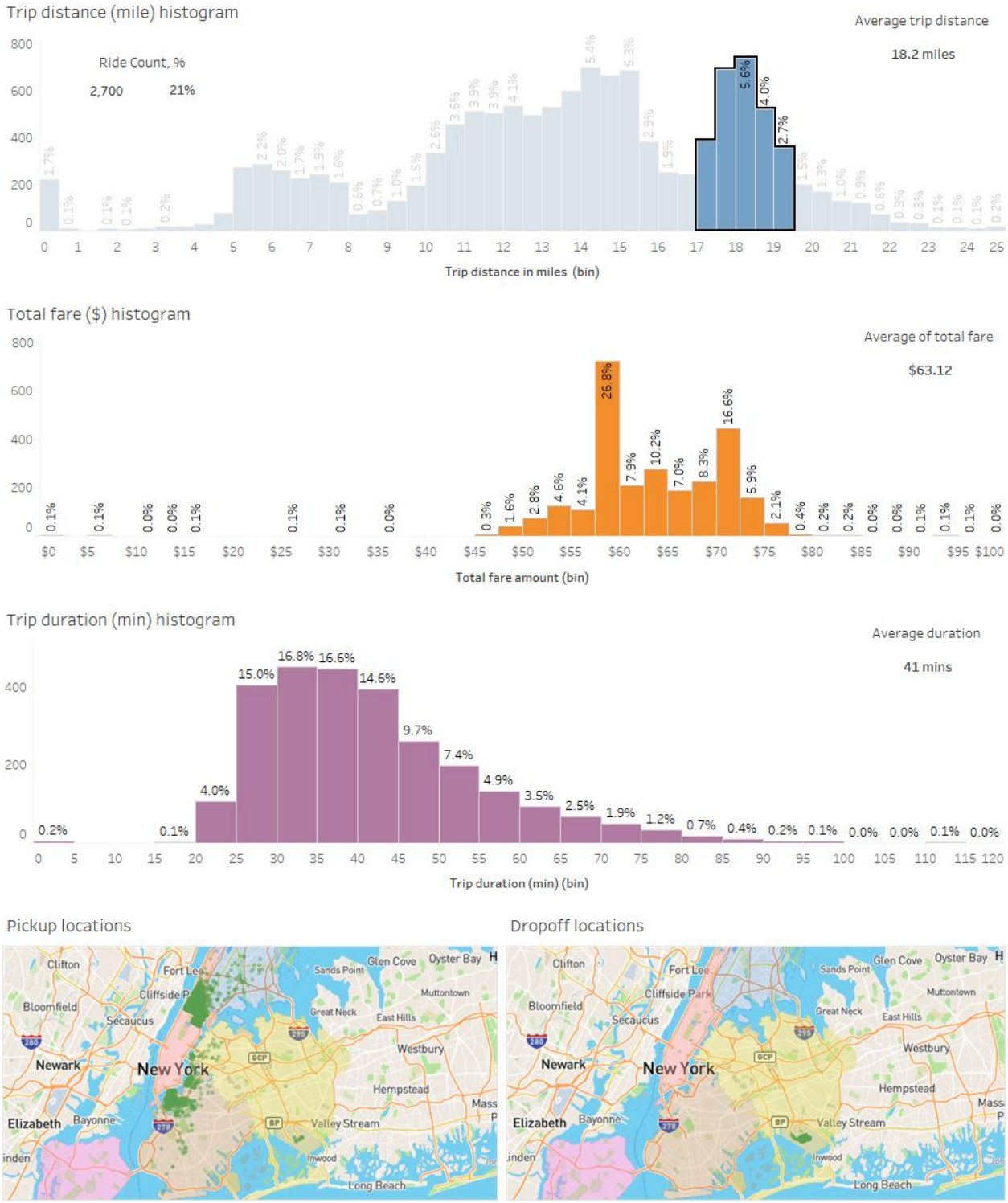


Figure Q3-12. Trip to and from JFK airport – trip distance, fare, duration histograms and locations for trips that originated or terminated 17-19.5 miles from JFK.

Pickup/drop-offs mostly from north-west Brooklyn and upper Manhattan area.

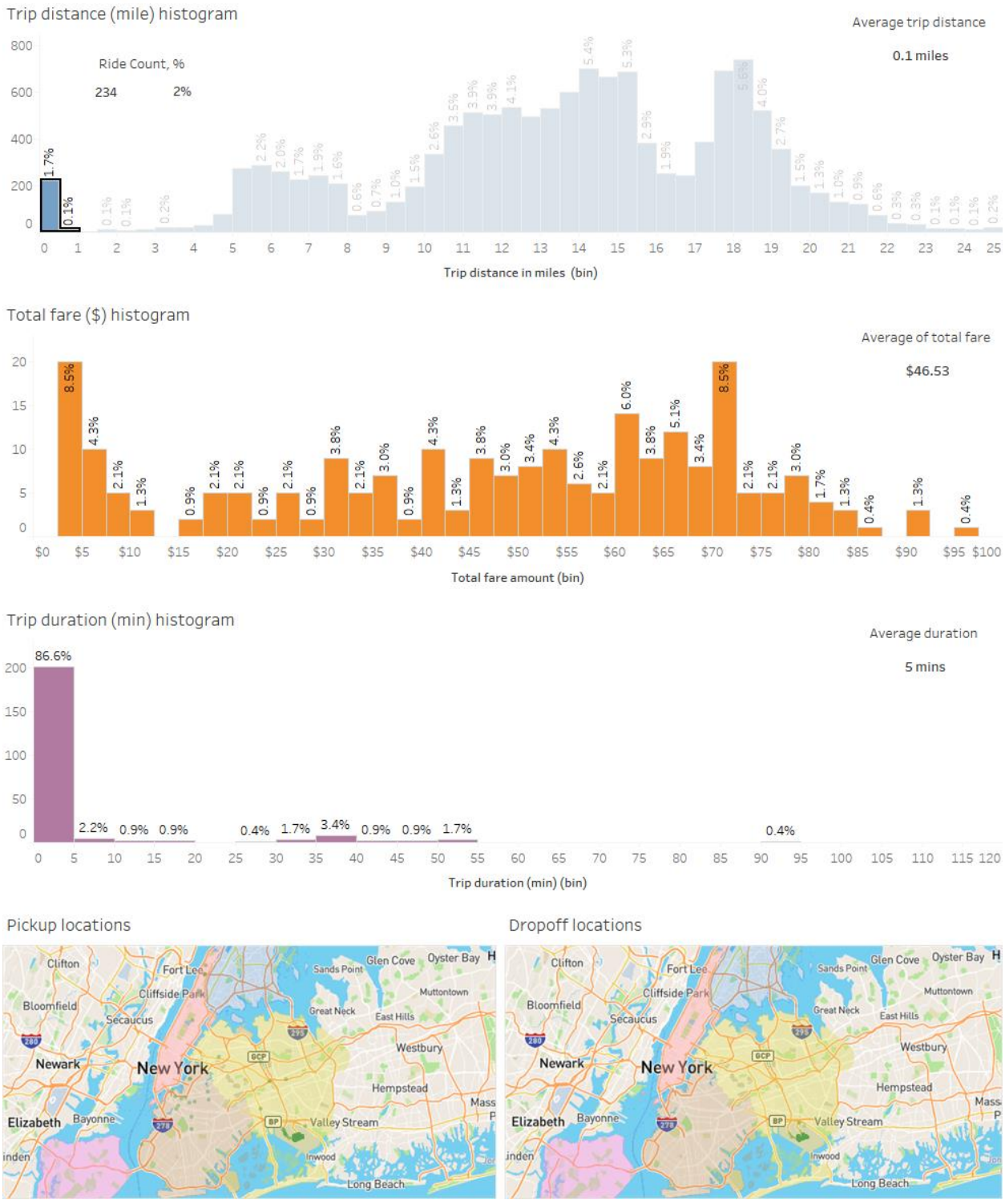


Figure Q3-13. Trip to and from JFK airport – trip distance, fare, duration histograms and locations for trips that originated or terminated within 1 mile from JFK

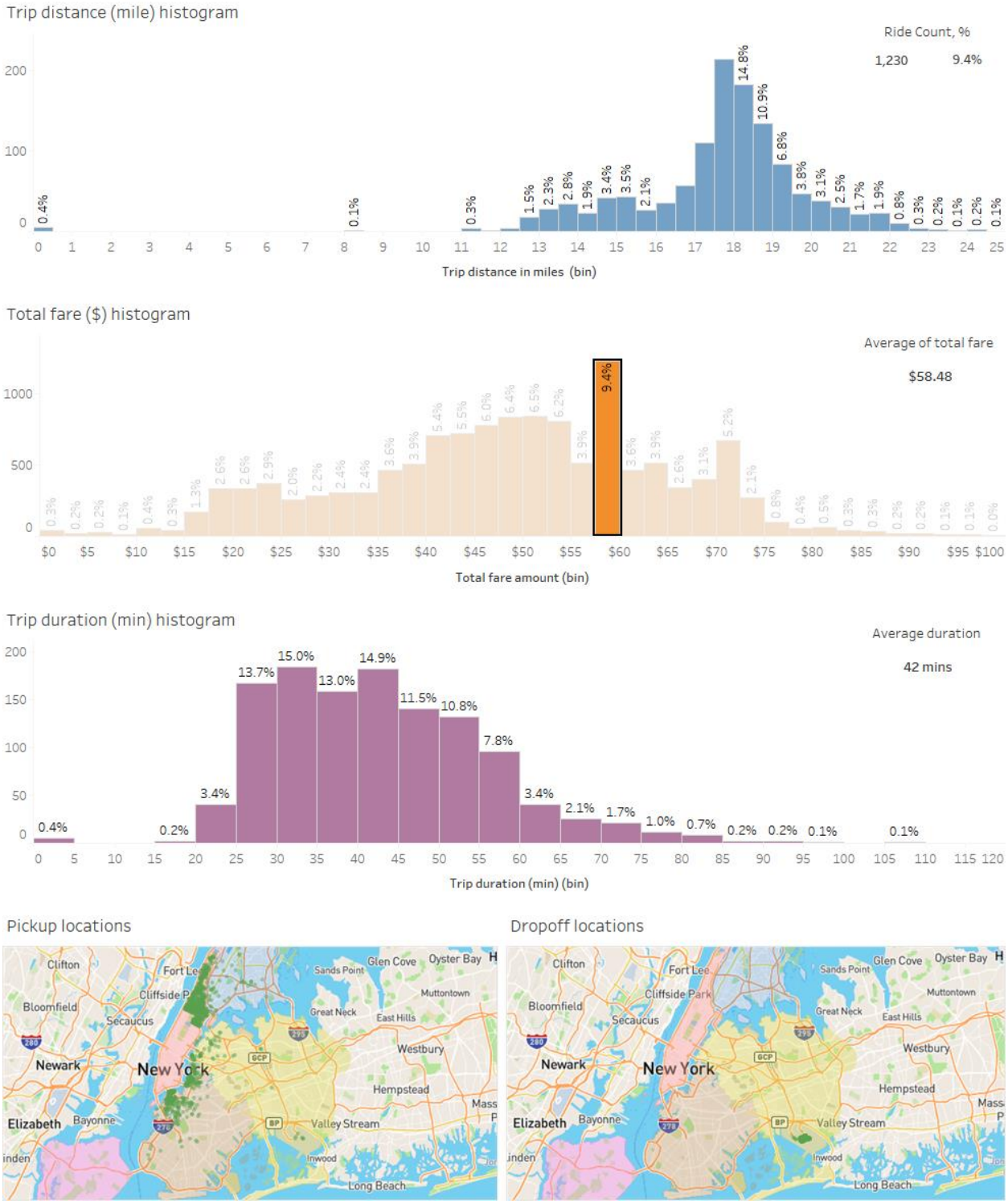


Figure Q3-14. Trip to and from JFK airport – trip distance, fare, duration histograms and locations for trips whose fare were $57.5 - $60

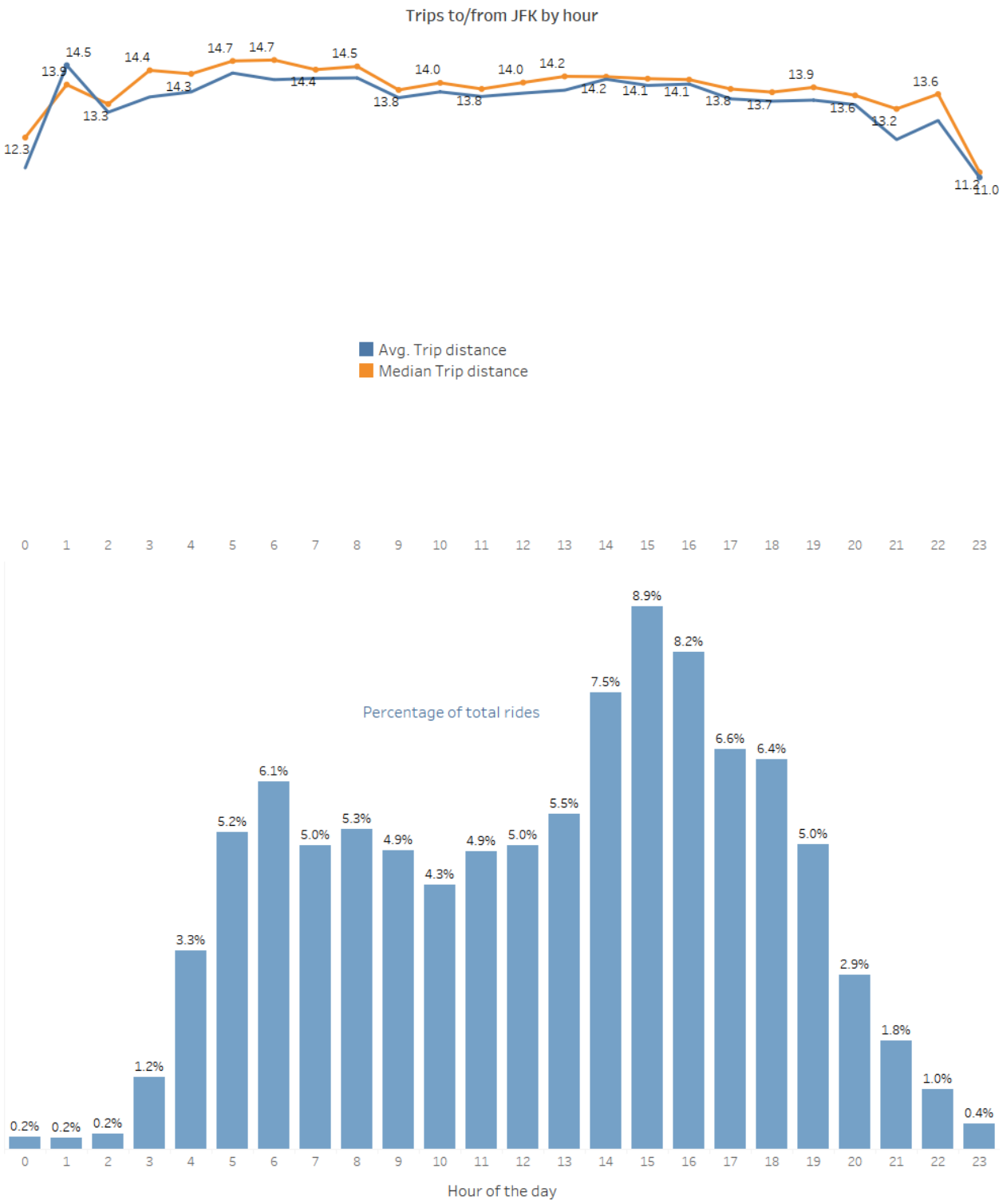


Figure Q3-15. Average and median trip distance and trip count distribution for trips to and from JFK airport by the hour of the day.

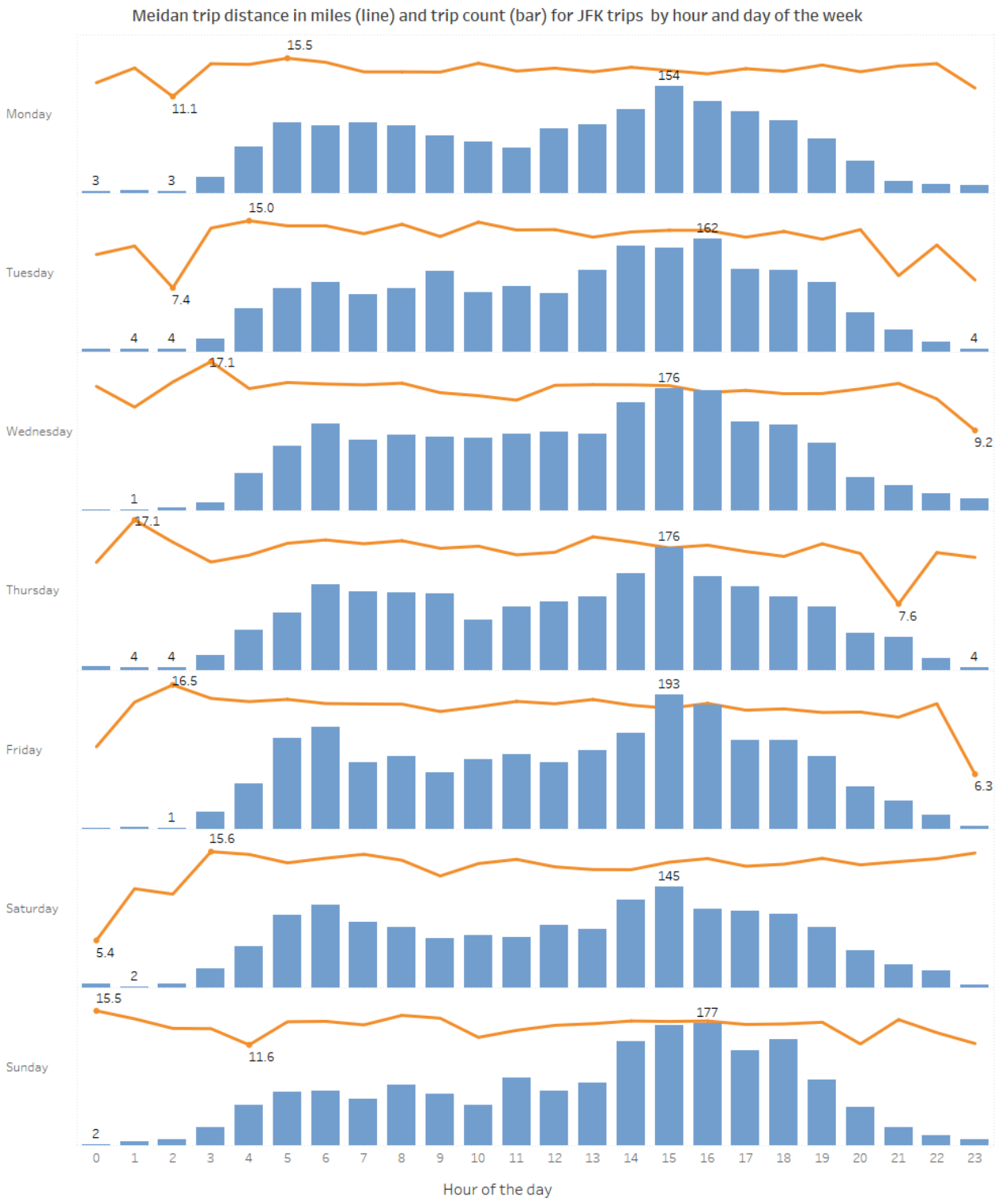


Figure Q3-16. Median trip distance and trip count distribution for trips to and from JFK airport by the hour of the day and the day of the week.

QUESTION 4

Build a derived variable for tip as a percentage of the total fare. Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). We will validate a sample.

Introduction

The purpose of this analysis is to define a trip percentage of the taxi trips and build a predictive model. This was achieved by training few models, selecting one through cross-validation and finally providing confidence by scoring accuracy on a test data set.

Exploratory Analysis

Exploratory analysis was used to (1) verify the quality of the data, (2) impute missing data, (3) transform some of the data, and (4) divide the data into training and cross-validation and test sets.

Statistical Modeling

To identify prediction models, we looked at linear regression models (e.g. Ridge, Elastic Net), traditional ensembles models (e.g. Random Forest Regressor) and gradient boosted ensemble models (Gradient Boosting Regressor). We split the data into training and validation sets, trained the models on the training set and observed the accuracy through 5-fold cross-validation. After choosing the models, we fine-tuned the hyper-parameters on the validation set.

The python packages numpy, pandas, datetime, dateutil (data processing), matplotlib, seaborn (plotting) and scikit-learn (statistical modeling) were used for this part of the project [15].

Reproducibility

All analyses/results in this report can be reproduced with attached ipython (ipynb) notebooks and tableau files. See the ReadMe.txt folder on how to run the training [16] and testing [17] code.

Data processing and model building

Train-test data separation:

The dataset contained 1,494,926 records with 21 variables. Before proceeding further, dataset was divided in two parts randomly. One with 80% of the original dataset for training the predictive model. The remining 20% of the data was used for testing the accuracy of the model. The test dataset was save in a file for later processing [17] and was removed from training process [16].

Train data selection:

Below is a list of variables in the given dataset:

```
VendorID,
lpep_pickup_datetime,
Lpep_dropoff_datetime,
Store_and_fwd_flag,
RateCodeID,
Pickup_longitude,
Pickup_latitude,
Dropoff_longitude,
Dropoff_latitude,
Passenger_count,
Trip_distance,
Fare_amount,
Extra,
MTA_tax,
Tip_amount,
Tolls_amount,
Ehail_fee,
improvement_surcharge,
Total_amount,
Payment_type,
Trip_type
```

The data dictionary [7] contained descriptions of all these variables. Only some of given variables are used for building the predictive model for tip percentage. Also the tip information was recorded only for credit card transactions (47% of the given data set), rest of the records (e.g. ones with cash payment method) has no tip information. So only the records with credit card payment were chosen. Details of this process is documented in the code [16].

Variable to predict - Trip percent:

The derived variable Trip percent was calculated simply by diving the tip amount by the total amount [16]:

```
df["Tip_percent"] = df.apply(lambda r: r["Tip_amount"]*100/r["Total_amount"],
axis =1)
```

To gain insight of the prediction variable tip percent, few plots were created (histogram, box/violin, tip percent vs total amount scatter). Figure Q4-1 in the next page shows the tip percent histogram and the box plot. Other than the peak at 0, there are three distinct peaks at 16 (highest count as expected), 20, 23 (could these be the default 3 options for tip in credit card slip/receipt?) The tip percent had min 0, max 100, average 14.14, standard deviation 7.84. The 25, 50, 75 percentiles were 10.77, 16.67 and 16.67 respectively! More than one fourth of the tips were at 16.67%.

Figure Q4-2 also in the next pages, plots Tip percent vs Total fare amount. The peaks mentioned above are observed. The plot also shows several “y = 1/x” type lines – these were the points for which the tip was a constant amount (e.g. $20 tip irrespective of the Total fare amount).

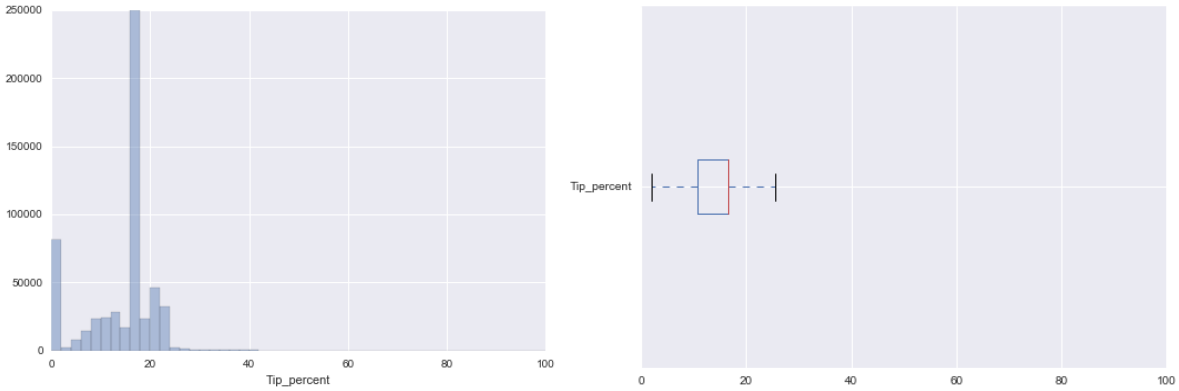


Figure Q4-1. (Left) Histogram of tip percent. (Right) Box plot of tip percent

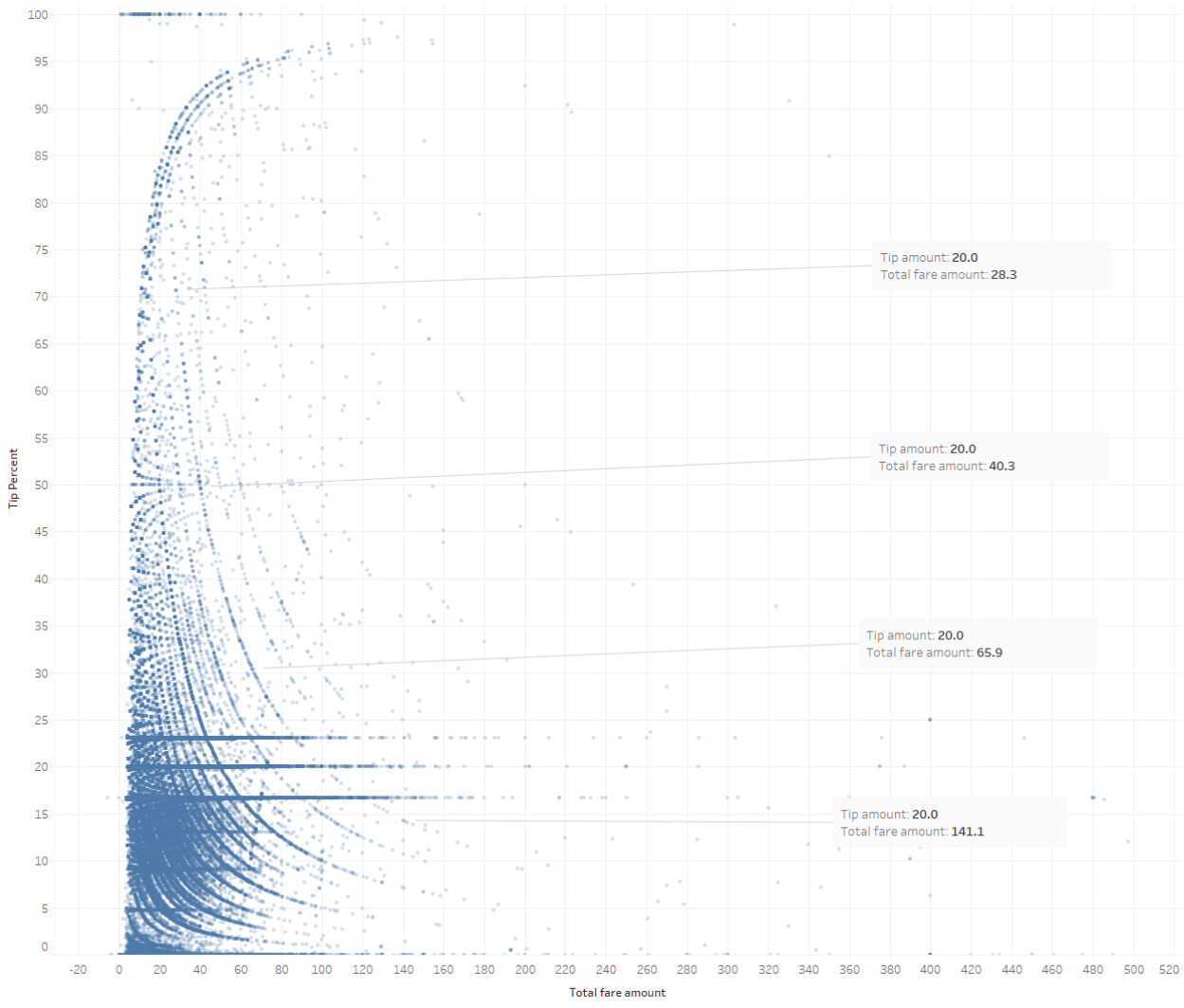


Figure Q4-2. Tip percent vs Total fare amount scatter plot.

Variable selection and exploration

For the first phase of predictive model creation, the following 10 variables were selected.

```
RateCodeID
Passenger_count
Trip_distance
Fare_amount
Extra
MTA_tax
Tolls_amount
improvement_surcharge
Total_amount
Trip_type
```

Data cleansing and plotting

We explored all the chosen variables one by one – computed statistics and visuals. The model_train.ipynb file contains detailed analysis (median/mode, frequency histograms/violin plots) for all the variables. The missing value locations were identified for each variable and 'median' for imputation.

Encoding categorical data

We performed numerical encoding and one hot encoding for the categorical variables before model building. Before the encoding we had 10 variables (both numerical and categorical). After the one hot encoding, we end up with 25 variables.

Some of these variables could be confounding. We let the model (Random Forest for example) choose the most important variables.

Training, cross validation and holdout test set

We explored multiple models and chose one tree-based ensemble model with the highest score. We did 80-20 split on our 'model building' data set – used 80% of the data for model training with 5-fold cross validation and the rest 20% for hyper parameter tuning/validation. Once we a model and its hyper parameters, we used the whole training dataset for the final model

Analysis of models

We explored both linear and tree based ensemble methods and evaluated the performance with the coefficient of determination, $\mathbf{R^2}$ score [18]. A $\mathbf{R^2}$ score of 1 means perfect prediction, $\mathbf{R^2}$ score of 0 implies model performance is equal to a model that predicts mean value for everything.

Linear Models

A linear model with L2 regularization (Ridge regression) and a model with both L1 and L2 regularization (Elastic Net regression with L1_ratio =0.5). The scores for 5-fold cross validation were:

```
Ridge scores: 0.5294, 0.5549, 0.4382, 0.5078, 0.5537
Ridge accuracy: 0.52 (+/- 0.09)

ElasticNet scores: 0.4574, 0.4711, 0.4253, 0.4451, 0.4668
ElasticNet accuracy: 0.45 (+/- 0.03)
```

where, accuracy = mean score (+- 2 standard deviation).

Both of the linear models had poor performance.

Ensemble models

Two variations of ensemble models (e.g. RandomForestRegressor, ExtraTreesRegressor) and the gradient boosted model (GradientBoostingRegressor) were explored. The scores for 5-fold cross validation were

```
RandomForestRegressor scores: 0.9972, 0.9944, 0.9962, 0.9963, 0.9938
RandomForestRegressor accuracy: 1.00 (+/- 0.00)

ExtraTreesRegressor scores: 0.9953, 0.9934, 0.9939, 0.9954, 0.9925
ExtraTreesRegressor accuracy: 0.99 (+/- 0.00)

GradientBoostingRegressor scores: 0.7835, 0.7825, 0.7759, 0.7777, 0.7713
GradientBoostingRegressor accuracy: 0.78 (+/- 0.01)
```

Random Forrest Regressor had better $\mathbf{R^2}$ scores and this model was chosen

Model tuning

A small amount of model tuning was performed by varying hyper parameter number of estimators for Random Forrest Model. Model built with 80% of data set and validated in remaining 20%. Below are the scores as number of estimators varied:

```
Number of estimators:  10,   RandomForestRegressor score: 0.99690
Number of estimators:  15,   RandomForestRegressor score: 0.99717
Number of estimators:  20,   RandomForestRegressor score: 0.99721
Number of estimators:  25,   RandomForestRegressor score: 0.99728
Number of estimators:  30,   RandomForestRegressor score: 0.99724
```

Result and analysis

Model Selection

Upon observation of the scores, we selected RandomForestRegressor with 20 number of estimators. Score on our validation set: 0.997. Lastly, we trained the model on all of the training data and saved the models and related data in the disk [19].

Variable Importance

Not all the features contributed equally in the prediction model. Below are feature importance ranking (printed the first 12) and a plot (Figure Q4-3) of their relative significance.

```
 1. Feature 24:  Trip_type  index 1: (62.04 %)
 2. Feature 19:  Extra: (29.09 %)
 3. Feature 20:  MTA_tax: (5.74 %)
 4. Feature 22:  improvement_surcharge: (1.61 %)
 5. Feature  4:  RateCodeID index 5: (1.17 %)
 6. Feature 18:  Fare_amount: (0.07 %)
 7. Feature 17:  Trip_distance: (0.06 %)
 8. Feature  0:  RateCodeID index 1: (0.06 %)
 9. Feature 21:  Tolls_amount: (0.05 %)
10. Feature 16:  Passenger_count index 9: (0.04 %)
11. Feature 23:  Total_amount: (0.04 %)
12. Feature  7:  Passenger_count index 0: (0.00 %)
…
```

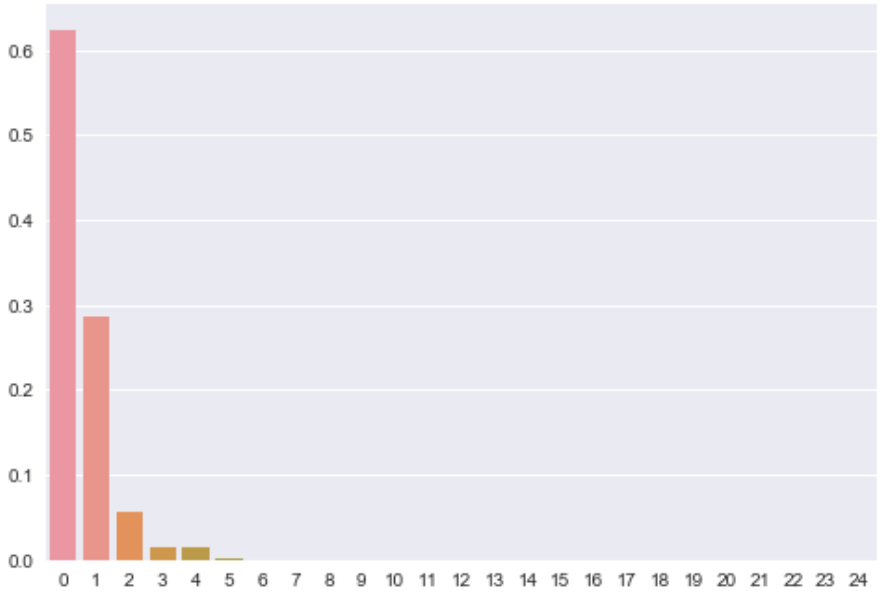


Figure Q4-3. Variable importance

The findings were interesting – the top 5 ranked features could predict the model well. The top ranked feature Trip type was [7]:

"A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver.

1 = Street-hail
2 = Dispatch"

And the trip type had 62% relative importance! The plot below (Figure Q4-3) show the tip percent distribution for both trip types. The plot shows when people hail taxis from the street, they tip more!

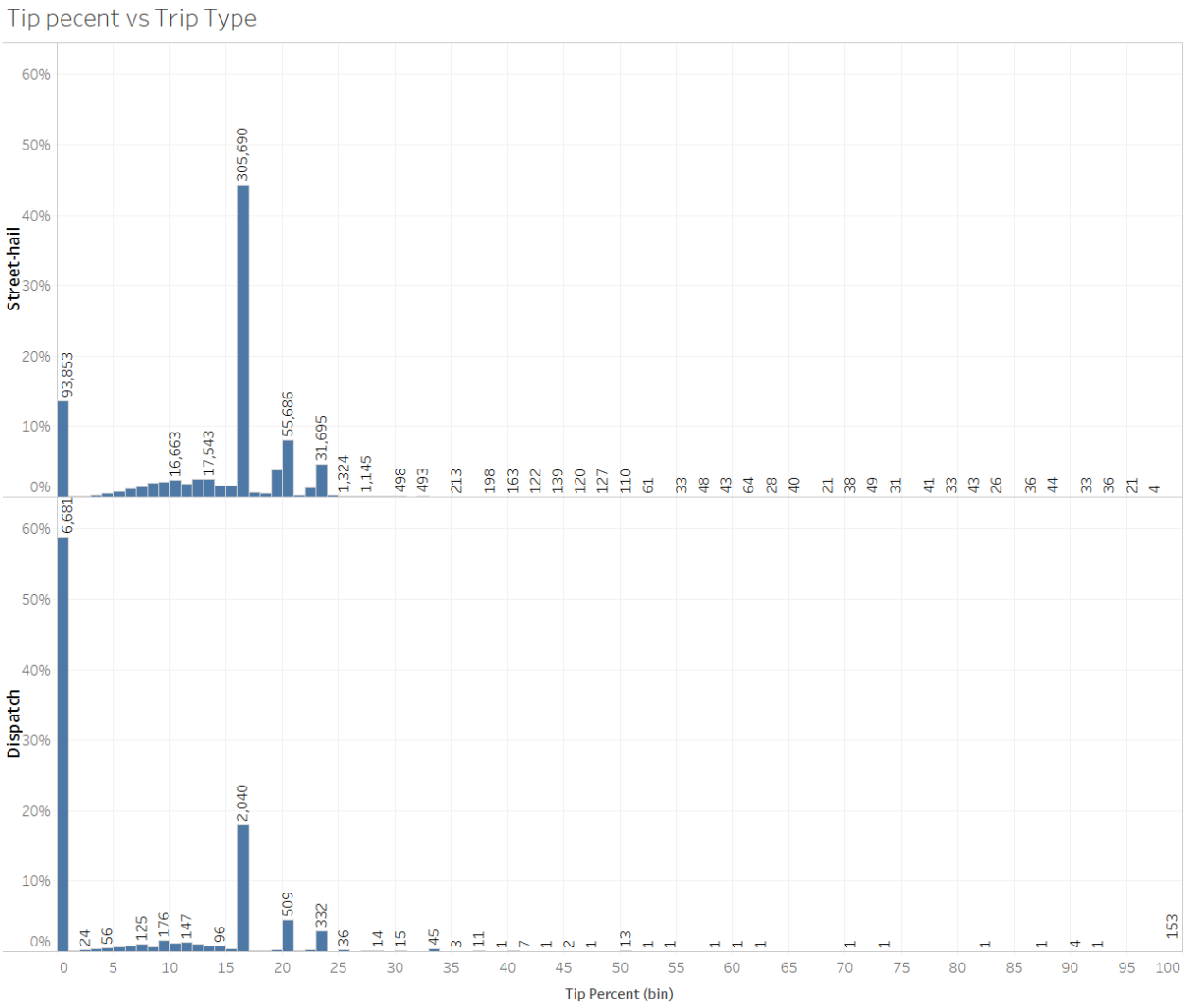


Figure Q4-4. Tip percent histogram by trip type (Street-hail vs Dispatch)

The next important feature (29% relative importance) was “Extra” [7]:

“Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges”

The plot (Figure Q4-5) below shows the tip distribution by the common values of “Extra”. It not the extra charge but the fact that these rides are during the rush hour and people tip a bit more when they are taking the taxi during rush hour.

Next important features were:

- MTA_tax (5.74% relative importance) [7]: $0.50 MTA tax is automatically triggered based on the metered rate in use

- improvement_surcharge (1.61% relative importance) [7]: $0.30 improvement surcharge assessed on hailed trips at the flag drop

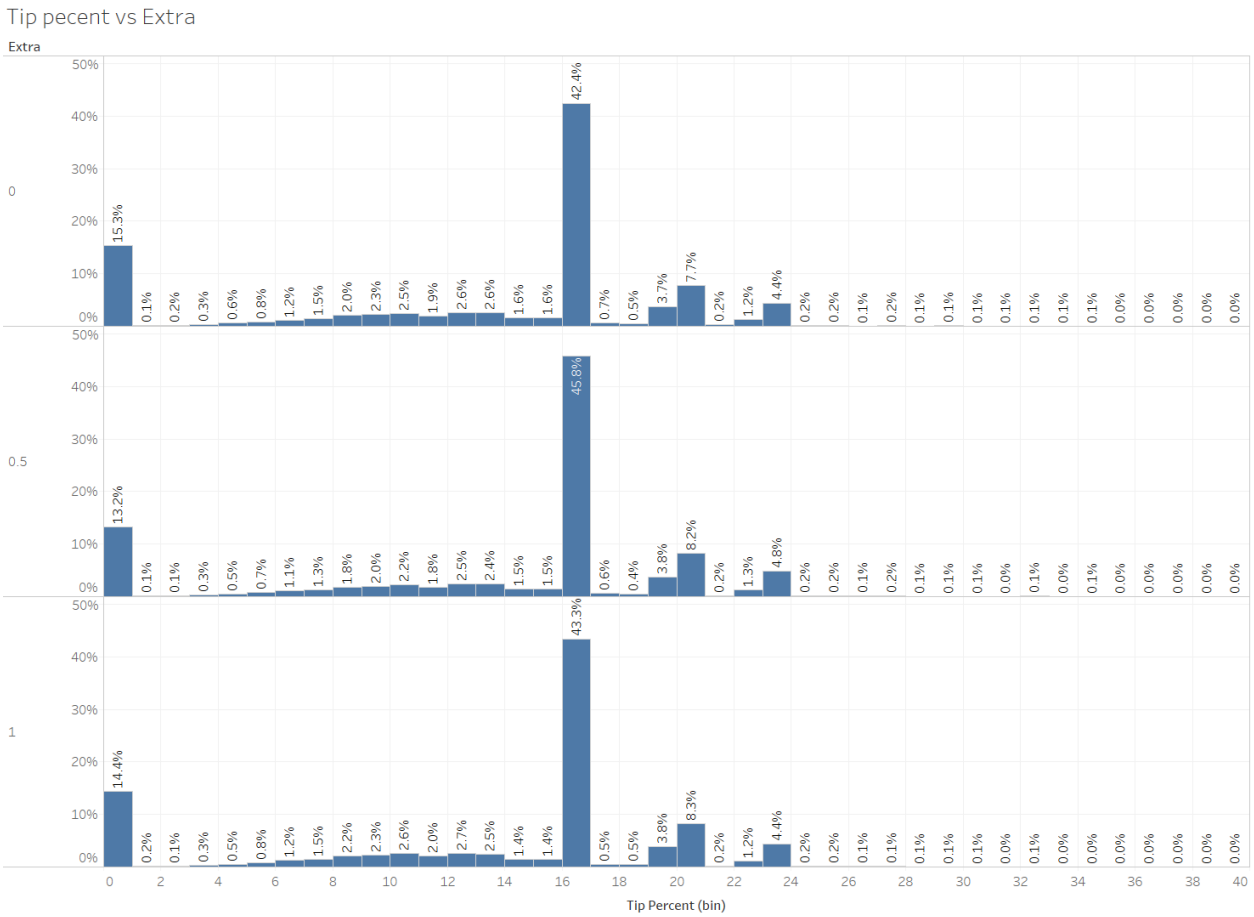


Figure Q4-5. Tip percent histogram by “Extra”

RateCodeID index 5 which is for “Negotiated fare” was the 5th important feature (1.17% relative importance). The plot (Figure Q4-6) shows the tip percent histogram by Rate Code. It seems for “Negotiated fare”, people tip less.

An intersecting observation is that fare amount or total fare amount had little importance in predicting the tip percentage.

Figure Q4-6. Tip percent histogram by Rate Code

Test and conclusions

The predication model built in the previous step was saved in disk. In this step, a separate code file [17] was used to load test data (again saved in disk from previous step), and saved model.

Just like the training data set, data with only positive total fare and credit card transactions were chosen. The test data was checked against missing values (if there were, they would be imputed with the median data saved during the training time) and histograms/box plots were created for test data quality check.

Using the save model (Random Forecast Regressor), prediction was made for the tip percentage using the test data and various accuracy metrics were checked.

On the test data set, our prediction model gave an R2 score of 0.997, mean absolute error of 0.037 and mean squared error of 0.186. Since there were no guidance on how much (or even what kind of accuracy metrics) is expected, this appears to be a general 'good' model. With guidance from the intended use of prediction model, the accuracy and model can be modified.

In sprit of the agile methodology [20] ('deliver initial results as quickly as possible and iterate with feedback'), we put together this models in a quick manner and there is room for improvement. The following can be taken as next steps:

1) For simplicity, this model omitted any explicit dependence of time and locations. *Does the tip percent depends on what time of the day*? Does people from Manhattan tip more than people from Brooklyn? Adding explicit variables for time and place could boost the accuracy.

2) Only few models were explore in this study. Other types of regression models (e.g. Logistic, Bayesian, Nearest Neighbor, xgboost) could be explored and compare the performances.

3) Hyper parameter tuning: more extensive hyper-parameter tuning can be explored – especially with the gradient boosted models. These models have shown superb performance in the industry recently.

4) Sensitivity analysis with respect to data size can be done (i.e. compare models trained with 10K, 50K, 100K, 200K, 300K, 400K, 500K records). It could be that our models are not complex enough and increased data does not give better results. In that case, deep neural network models (e.g. with python based Keras wrapper on top on TensorFlow or Theano backend [21]) can be explored. Once again the deep learning based models have shown record-breaking performances in certain class of problems in recent years.

5) Lastly for a production system, if we can collect more data that a single node computer cannot keep up with, we can explore cluster based computations with Spark [22] for example.

QUESTION 5

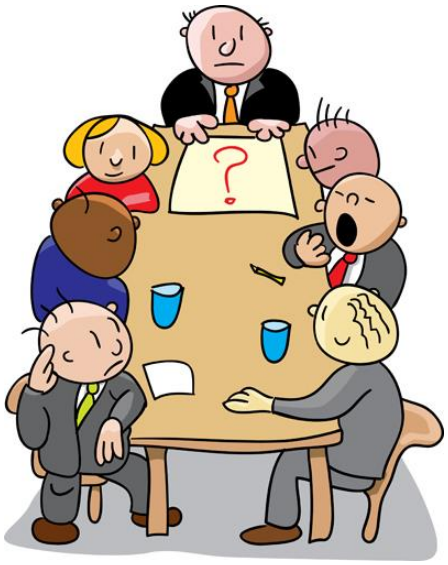
Choose only one of these options to answer for Question 5.

Option B: Visualization
Can you build a visualization (interactive or static) of the trip data that helps us understand intra- vs. inter-borough traffic? What story does it tell about how New Yorkers use their green taxis?

Option E: Your own curiosity!
If the data leaps out and screams some question of you that we haven't asked, ask it and answer it! Use this as an opportunity to highlight your special skills and philosophies.

The author choose option E and used elements from option B as example.

Background and motivation



Most of the business questions to data can be broken down to two main elements:

Context
Which case/scenario are we interested in? For this data set, some examples of contexts are: trips to the airport, trips from Queens, "short" trips, trips during weekend vs weekday, trips in early morning from Brooklyn to Manhattan, etc.

Metrics
What numbers would we would like to know for your scenario? For this data set, some examples of metrics are: the number of trips, trip distance, fare, duration (average/median, distribution), taxi speed etc.

In business settings, the questions on the same metrics are asked again and again for various different contexts. For example, metrics like revenue, profit, expense, ROI, customer satisfaction are quite popular and these questions need to be answered for various context (business units, customer segments, etc). In this data set, for example, the author was curious to find out the fare and duration as well as the trip distance for every question studied.

The author solves this kind of repetitive questions on the metrics by building answers in an interactive portal ('Decision Support System' [23]). The way the business person can ask several context combinations (e.g. profit trend for business unit A for customer segment X in quarter 3 of the last 5 years).

In this spirit of this 'Data science as a service', the author built a simple interactive portal to answer questions related to intra- vs. inter-borough traffic (as indicated in Option B). A table was showing the traffic pattern was created first and maps and other metrics were used to enrich it.

Analysis

At first, each pickup and drop-off location were tagged with the borough they belong to. The details of coding are described in [24].

The top table in Q5 Table 1 below shows the trip counts by pickup and drop-off borough. For example, 97,077 trips originated at Brooklyn and terminated at Manhattan. The middle table shows the percentages of these trip counts relative to the total 1,494,926 trips.

The bottom table also shows the percentages of these trip counts but relative to the total number of trips originating at respective boroughs. For example, 84,740 trips were picked up at Bronx and 63,082 trips were both picked up and dropped off at Bronx – this gave the intra-traffic ratio of 74.4% for Bronx. This table shows intra-traffic ratio of 87%, 85.2%, 76.3% and 86.3% for rest of the boroughs (Manhattan, Queens, Brooklyn, Staten Island) respectively. ***So most of the traffic is really within the borough traffic***.

Pickup drop-off distribution by borough

Trip count

	Dropoff						
Pickup	Bronx	Manhattan	Queens	Brooklyn	Staten Island	Outside NYC	**Grand Total**
Bronx	63,082	18,487	1,561	445	1	1,164	84,740
Manhattan	37,388	370,736	13,152	3,138	15	1,679	426,108
Queens	2,245	36,854	351,538	20,303	34	1,846	412,820
Brooklyn	760	97,077	34,870	433,631	177	1,455	567,970
Staten Island		3	1	9	88	1	102
Outside NYC	451	528	528	578		1,101	3,186
Grand Total	103,926	523,685	401,650	458,104	315	7,246	1,494,926

Percentage of total trips

	Dropoff						
Pickup	Bronx	Manhattan	Queens	Brooklyn	Staten Island	Outside NYC	**Grand Total**
Bronx	4.2%	1.2%	0.1%	0.0%	0.0%	0.1%	5.7%
Manhattan	2.5%	24.8%	0.9%	0.2%	0.0%	0.1%	28.5%
Queens	0.2%	2.5%	23.5%	1.4%	0.0%	0.1%	27.6%
Brooklyn	0.1%	6.5%	2.3%	29.0%	0.0%	0.1%	38.0%
Staten Island		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Outside NYC	0.0%	0.0%	0.0%	0.0%		0.1%	0.2%
Grand Total	7.0%	35.0%	26.9%	30.6%	0.0%	0.5%	100.0%

Percentage of trips within a borough

	Dropoff						
Pickup	Bronx	Manhattan	Queens	Brooklyn	Staten Island	Outside NYC	**Grand Total**
Bronx	74.4%	21.8%	1.8%	0.5%	0.0%	1.4%	100.0%
Manhattan	8.8%	87.0%	3.1%	0.7%	0.0%	0.4%	100.0%
Queens	0.5%	8.9%	85.2%	4.9%	0.0%	0.4%	100.0%
Brooklyn	0.1%	17.1%	6.1%	76.3%	0.0%	0.3%	100.0%
Staten Island		2.9%	1.0%	8.8%	86.3%	1.0%	100.0%
Outside NYC	14.2%	16.6%	16.6%	18.1%		34.6%	100.0%
Grand Total	7.0%	35.0%	26.9%	30.6%	0.0%	0.5%	100.0%

Q5 Table 1. Distribution of trip count and percentages by borough

Figures Q5-1 to Q5-9 in pages 41-49 show interactive plots with pickup and drop-off locations in maps and user selectable boroughs. They also show the average and the distribution of trip distance, total fare and trip duration. Figure Q5-1 in page 41 shows the traffic for all the boroughs. Figure Q5-2 in page 42 shows trips that were picked up only at Bronx. The bar chart on the top right show that 74% of the Bronx picked up trips were dropped-off also at Bronx. The maps also provide the visual clues. *The plots show that the trips were mostly within the borough*. Similar trend can be seen for Manhattan pickups in Figure Q5-3 (page 43), for Queens pickups in Figure Q5-4 (page 44), for Brooklyn pickups in Figure Q5-5 (page 45) and for Staten Island pickups in Figure Q5-6 (page 46). The trips that originated outside the NYC boroughs are tagged as "Outside NYC" and are shown in Figure Q5-7 (page 47).

These interactive plots also show the inter-borough traffic pattern. For example, Figure Q5-8 (page 48) shows trips that were picked up at Manhattan and dropped off at Queens (Manhattan in the Pickup box and 3.1% bar next to Queens are selected). The average trip distance, total fare and duration increased accordingly. From the trip distance histogram peaks and the location density on the maps, we can infer some of these trips were going to JFK and LaGuardia airport. Figure Q5-9 (page 49) shows another example for trips picked up at Brooklyn and dropped off at Queens.

These interactive charts could be served from a web portal and any user can see and 'play' with the interactive plots without the need for asking a data scientist. That way, the user can ask any combination of questions and the plots with insights would be generated on the fly. The plots could be even more customized with additional menus for other context (e.g. time period). This is an example of 'Self-service analytics' the author has implemented and popularized in his current workplace.

As we explore these plots and the data, we keep on finding additional interesting facts:

Morning traffic

The time effect on the intra-borough traffic is explored in Figure Q5-10 (page 50) with both intra borough traffic percentage and trip counts plotted against the hour of the day. All boroughs show a drop in intra borough traffic percentage (which is increase in inter borough traffic) in the morning hours (especially Brooklyn). It could be people are going to work in other boroughs in those morning hours.

Brooklyn to Manhattan in the morning

Figure Q5-11 in page 51 tabulates and plots the drop-off traffic percentages by hour. The pickup borough is user selectable and Brooklyn pickup was chosen in the view in Figure Q5-11. During 6-9am in the morning, 30+% traffic from Brooklyn is going to Manhattan (the financial district of Manhattan is close to the West Brooklyn).

Midnight traffic from Brooklyn

The trip counts in Figure Q5-10 (page 50) show another interesting pattern for Brooklyn. The Brooklyn traffic was highest in the midnight hours! Why?

Since Figure Q5-10 was a cumulative sum for all the days (by hour), Figure Q5-12 in page 52 brings the effect of the day of week in picture. We observe that increased midnight traffic at Brooklyn were mostly during the weekends. Seems like the people in Brooklyn really likes to party! Or its just, they take the green taxis more while they go out for fun in the middle of the night.

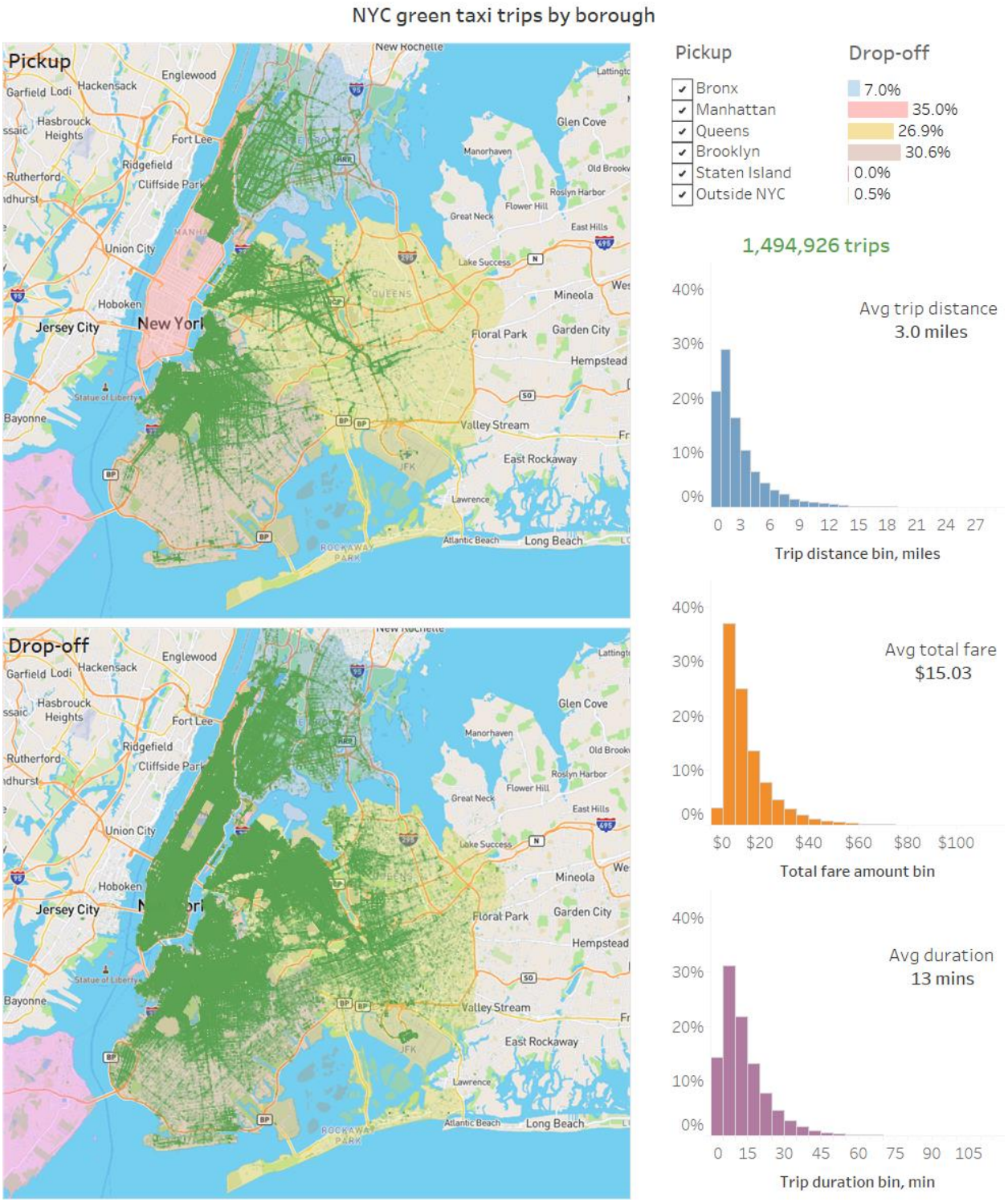


Figure Q5-1. Locations, trip distance, total fare and duration by pickup and drop-off borough

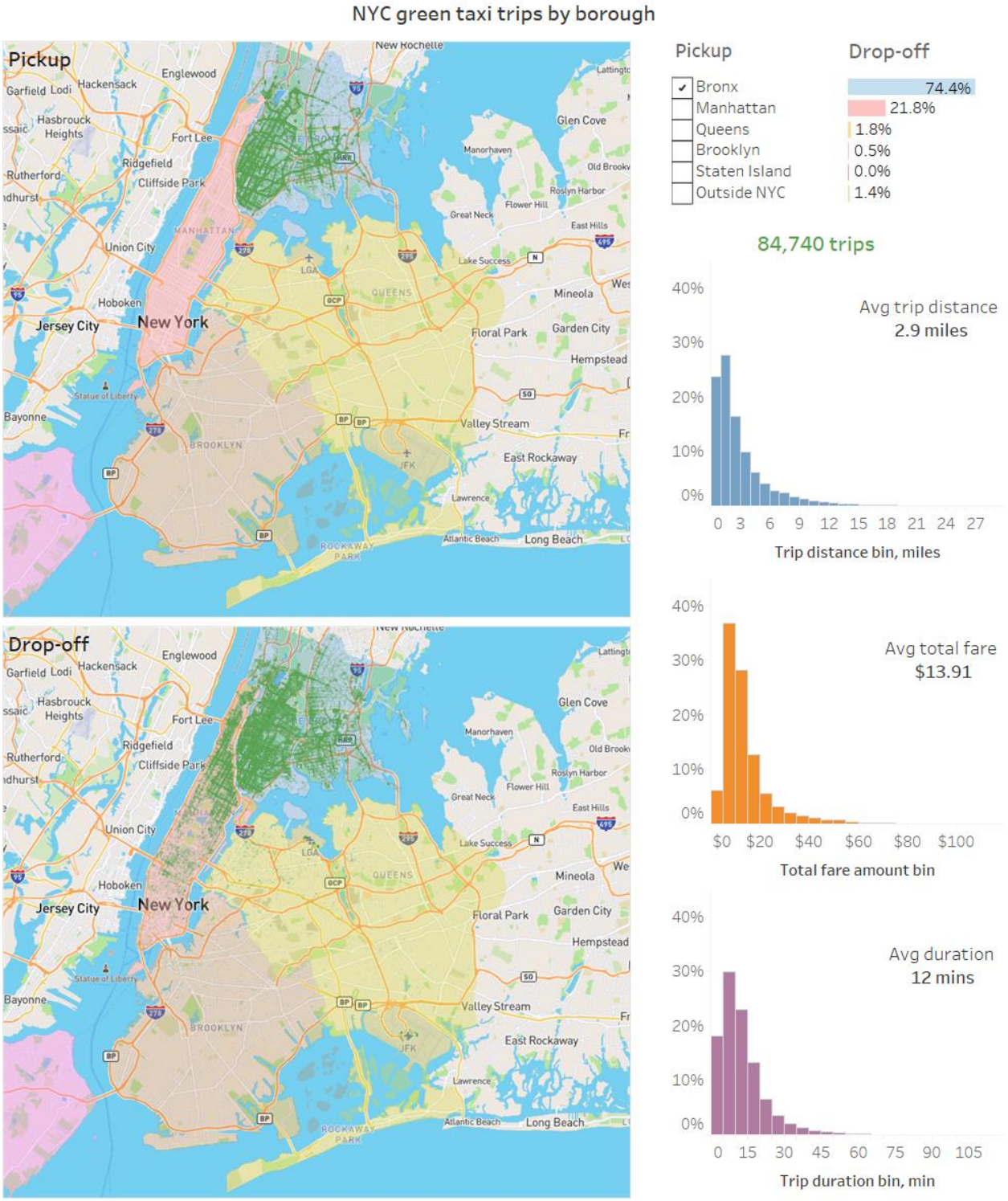


Figure Q5-2. Locations, trip distance, total fare and duration for pickups in Bronx

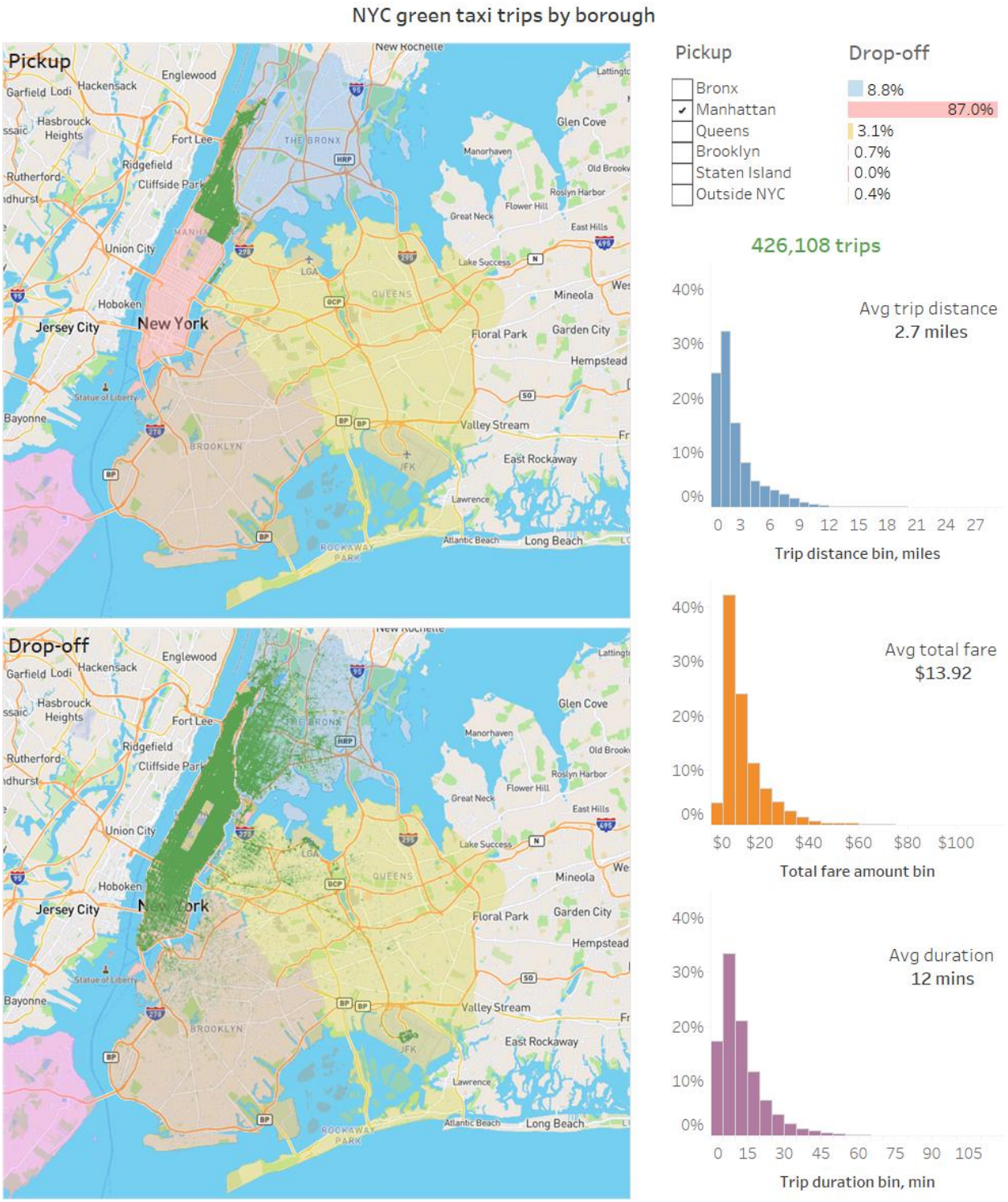


Figure Q5-3. Locations, trip distance, total fare and duration for pickups in Manhattan

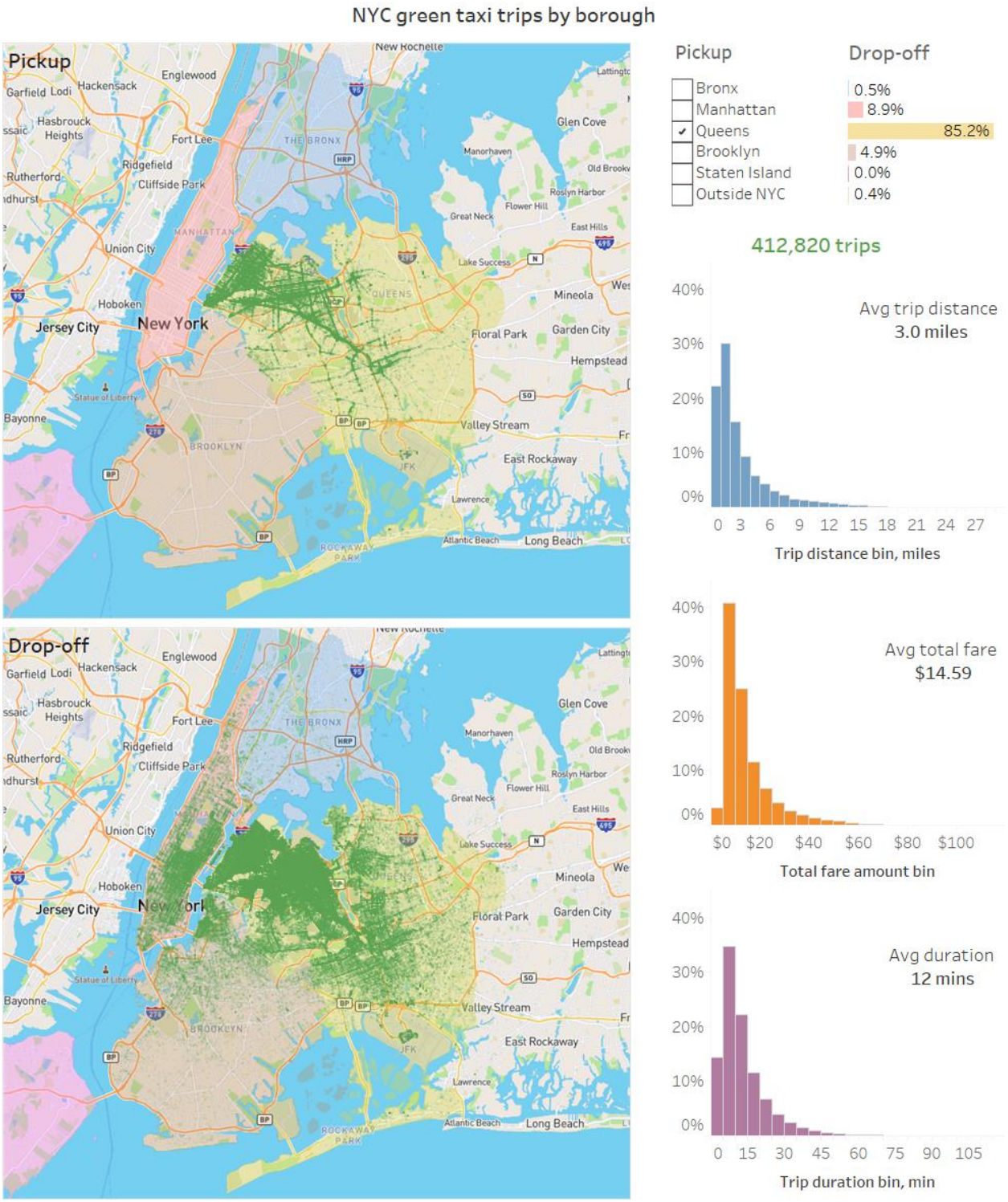


Figure Q5-4. Locations, trip distance, total fare and duration for pickups in Queens

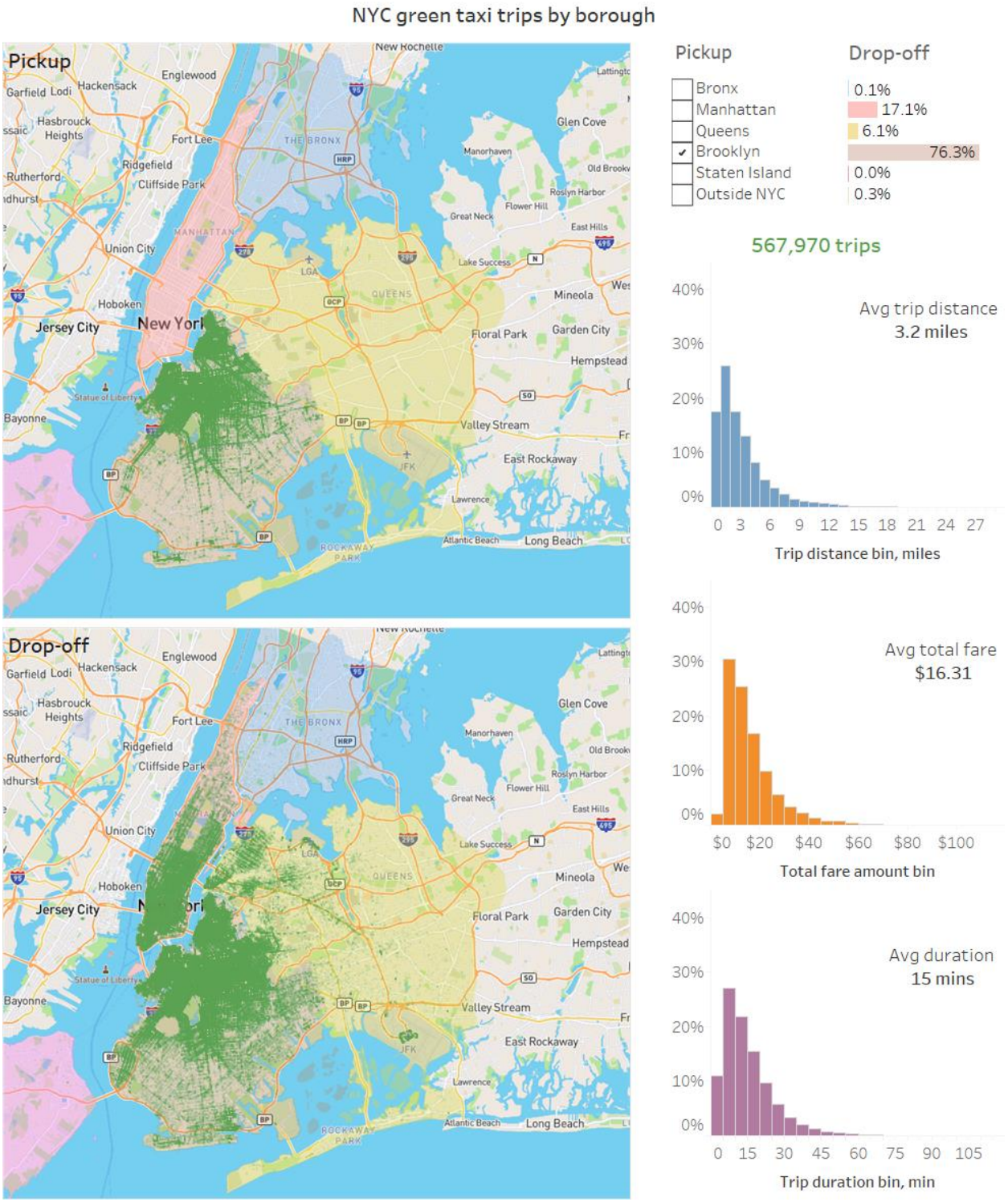


Figure Q5-5. Locations, trip distance, total fare and duration for pickups in Brooklyn

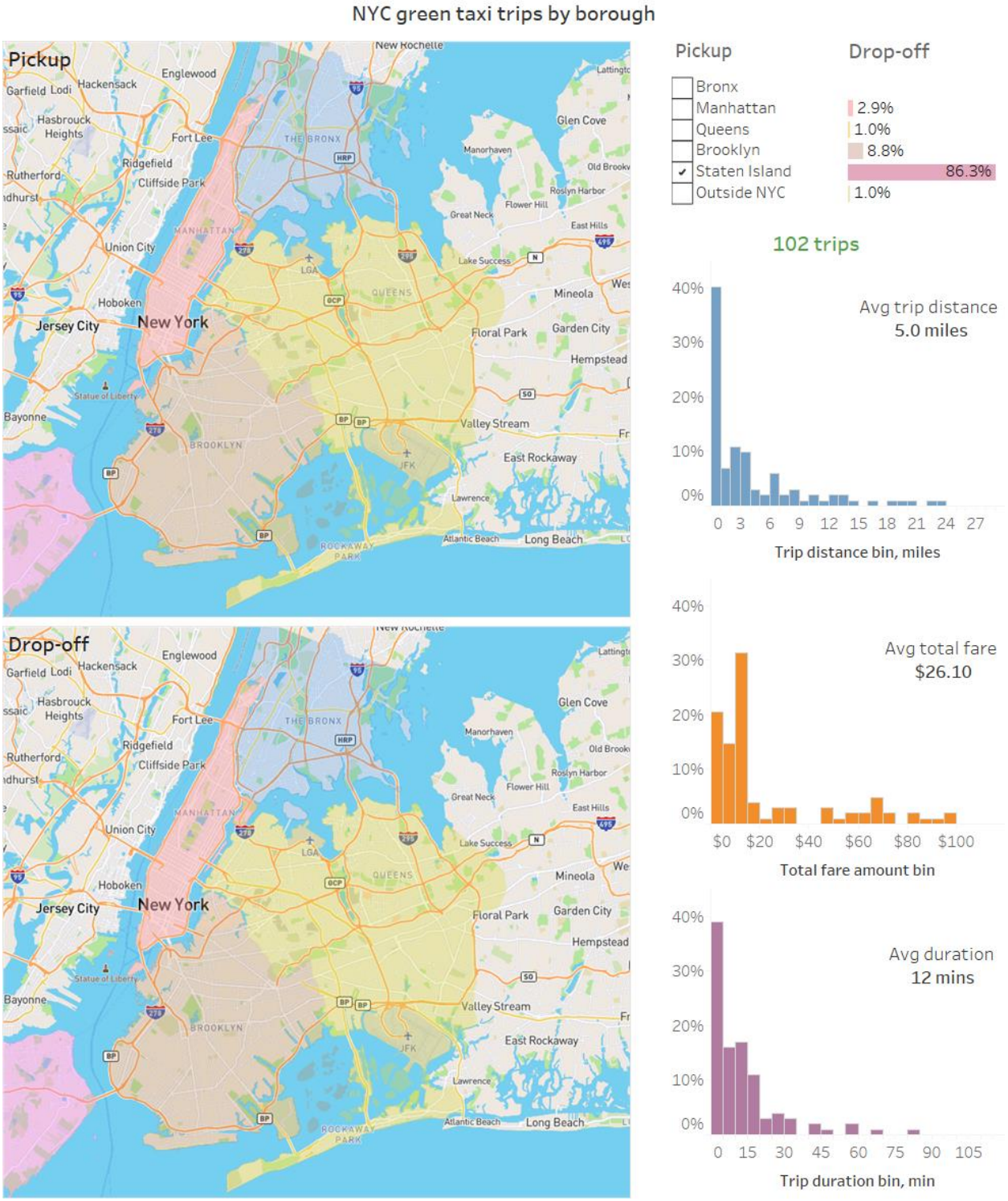


Figure Q5-6. Locations, trip distance, total fare and duration for pickups in Staten Island

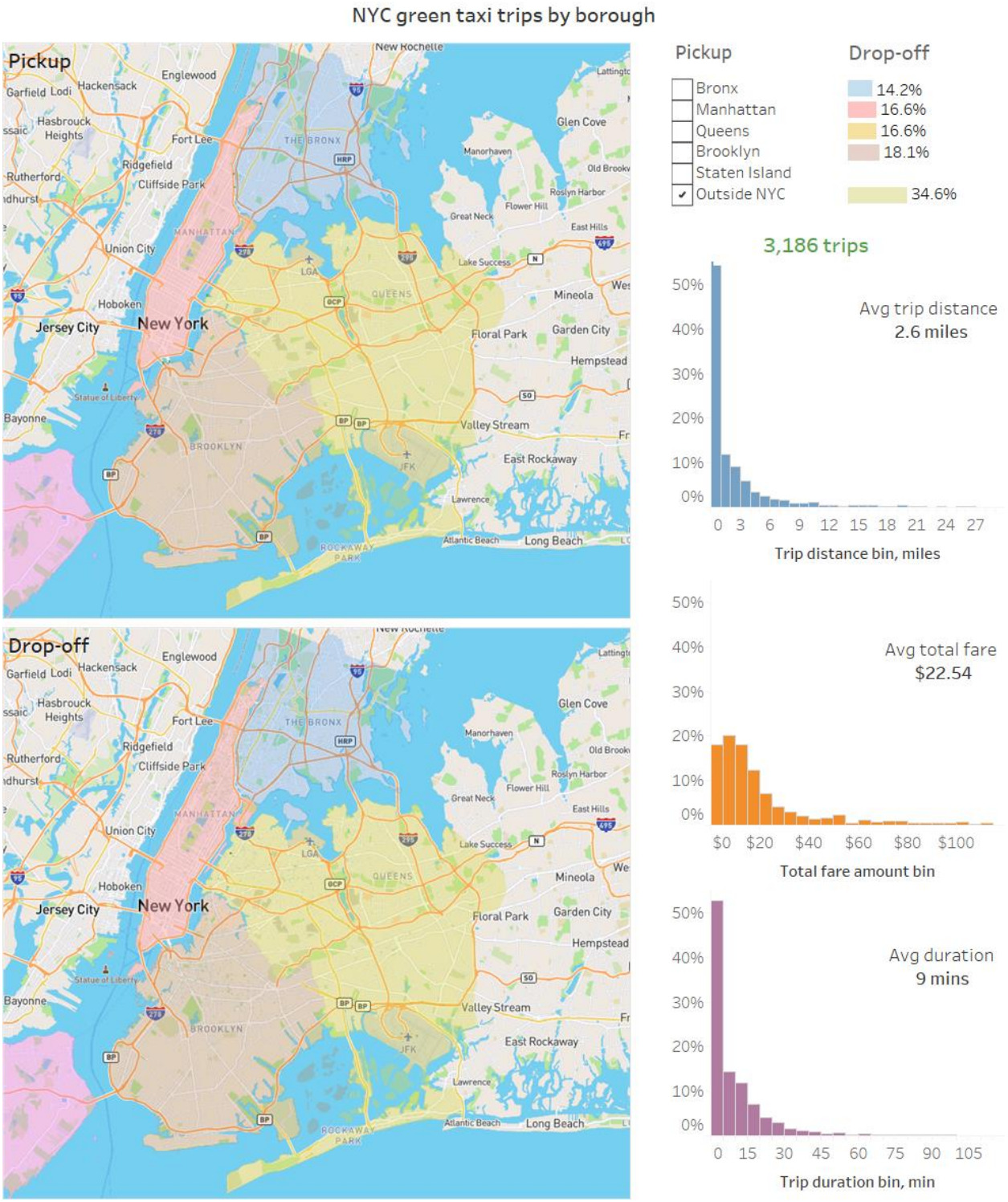


Figure Q5-7. Locations, trip distance, total fare and duration for pickups in outside NYC

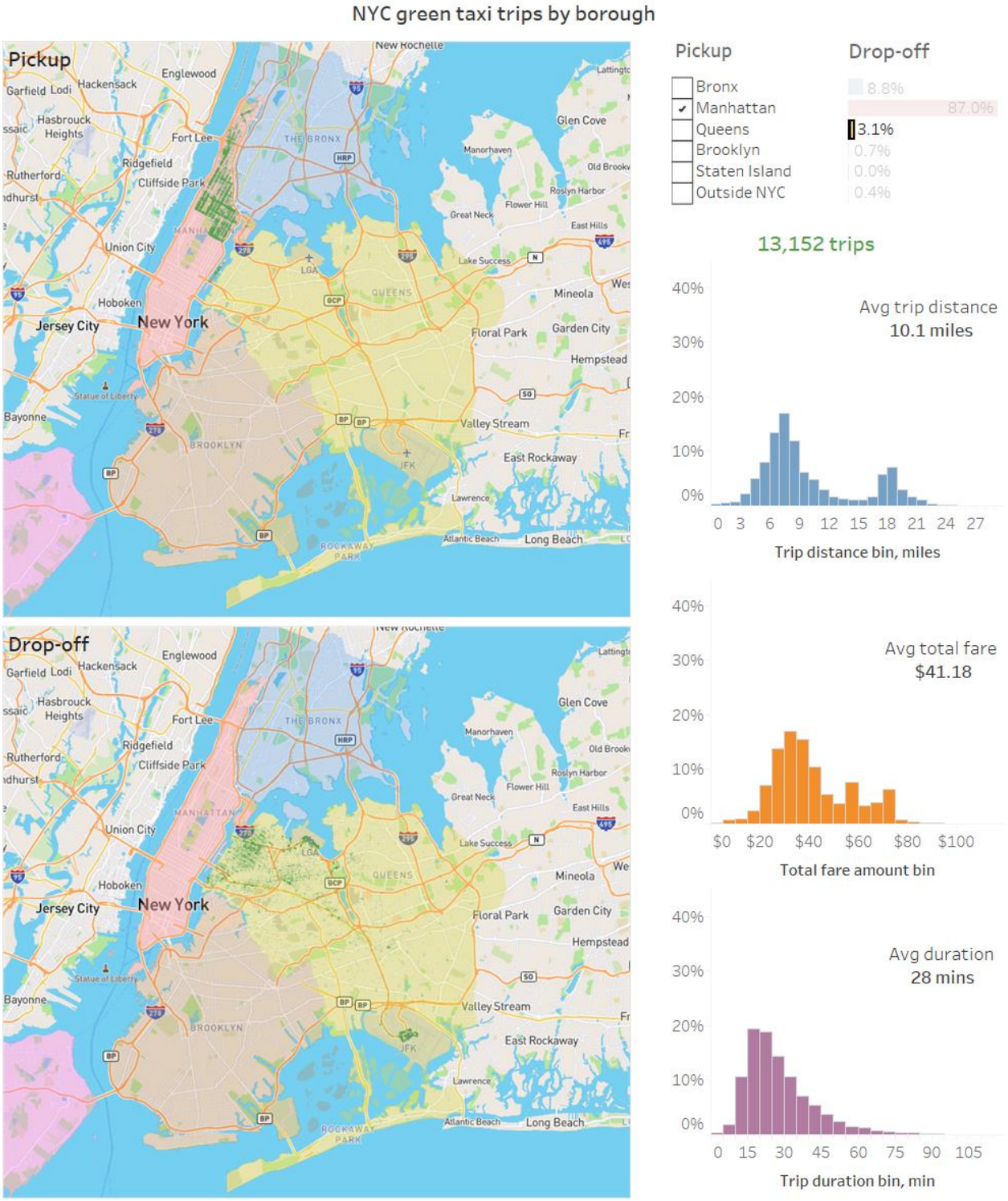


Figure Q5-8. Locations, trip distance, total fare and duration for pickups in Manhattan and drop-offs at Queens

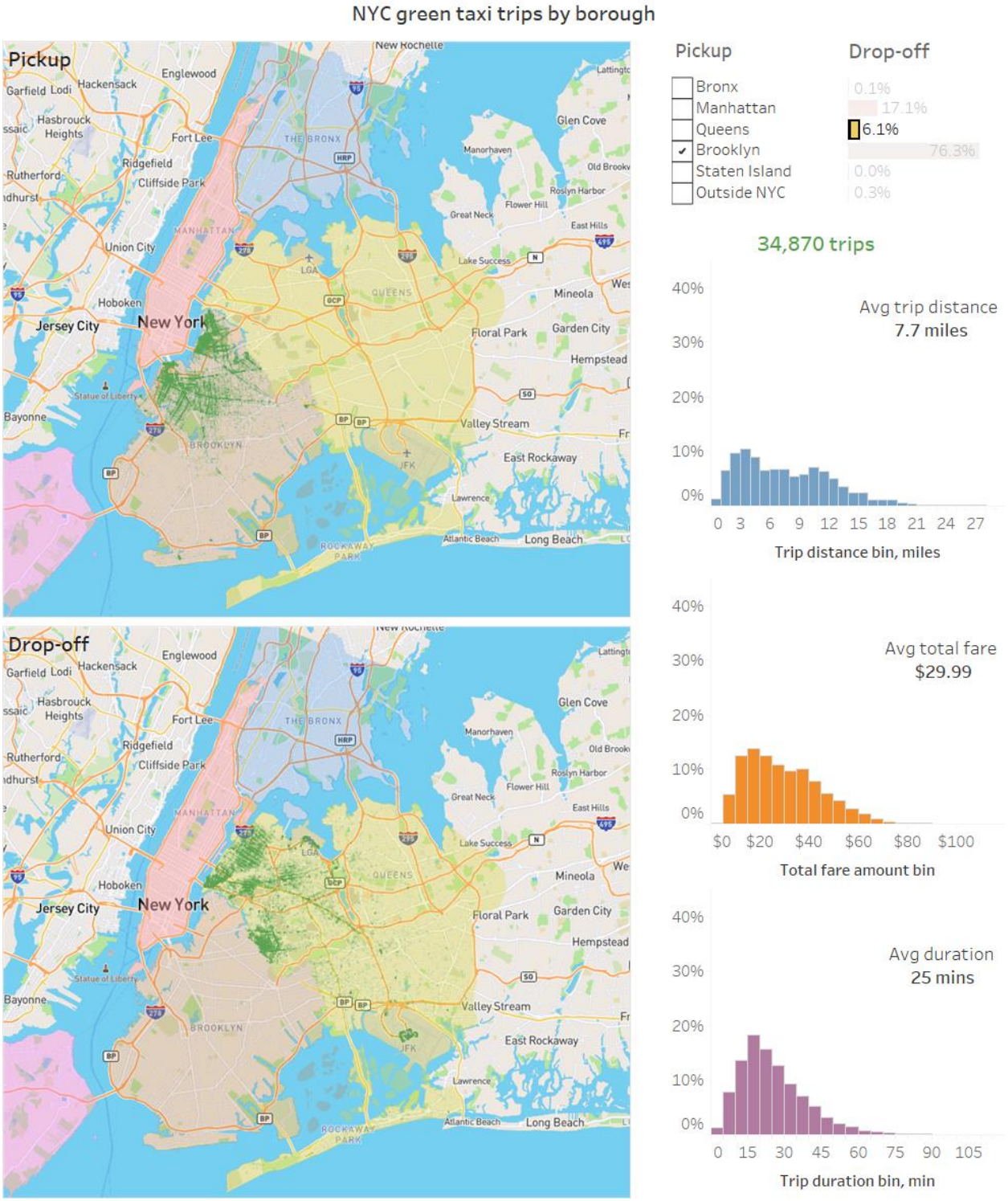


Figure Q5-9. Locations, trip distance, total fare and duration for pickups in Brooklyn and drop-offs at Queens

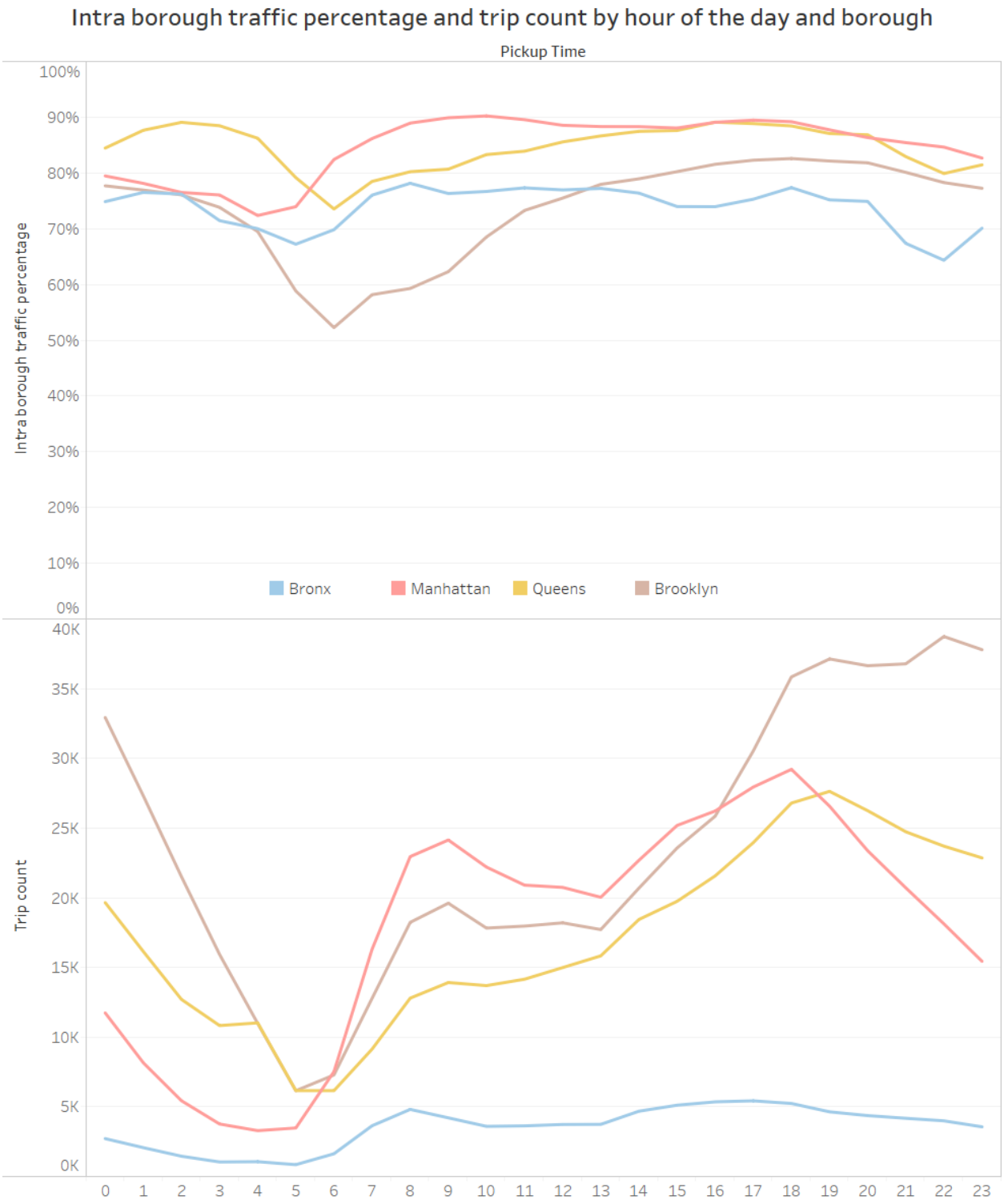


Figure Q5-10. Intra borough traffic percentage relative to the traffic in the borough and trip count by hour of the day and borough

Traffic percentage by hour and drop-off borough

Pickup: Bronx | Manhattan | Queens | **Brooklyn** | Staten Island | Outside NYC

Hour (Drop-off)	Bronx	Manhattan	Queens	Brooklyn	Staten Island	Outside NYC
0	0.2%	15.0%	6.9%	77.7%	0.0%	0.2%
1	0.2%	15.5%	7.1%	76.9%	0.0%	0.2%
2	0.2%	15.8%	7.6%	76.1%	0.1%	0.3%
3	0.3%	16.7%	8.9%	73.8%	0.0%	0.3%
4	0.4%	17.4%	12.2%	69.5%	0.1%	0.5%
5	0.3%	23.1%	17.0%	58.8%	0.1%	0.7%
6	0.3%	31.4%	15.6%	52.3%	0.1%	0.4%
7	0.2%	31.9%	9.2%	58.2%	0.0%	0.4%
8	0.1%	32.8%	7.5%	59.3%	0.0%	0.3%
9	0.1%	30.4%	6.8%	62.3%	0.0%	0.4%
10	0.1%	24.8%	6.2%	68.5%	0.0%	0.3%
11	0.1%	20.3%	5.9%	73.3%	0.0%	0.3%
12	0.1%	18.3%	5.7%	75.5%	0.0%	0.4%
13	0.1%	15.5%	6.1%	78.0%	0.0%	0.3%
14	0.1%	14.3%	6.3%	78.9%	0.0%	0.3%
15	0.1%	13.3%	6.0%	80.2%	0.0%	0.3%
16	0.1%	12.6%	5.4%	81.6%	0.0%	0.3%
17	0.1%	12.6%	4.8%	82.3%	0.0%	0.2%
18	0.1%	12.9%	4.2%	82.6%	0.0%	0.2%
19	0.1%	13.5%	4.1%	82.1%	0.0%	0.1%
20	0.1%	13.9%	4.1%	81.8%	0.0%	0.1%
21	0.1%	15.4%	4.2%	80.1%	0.0%	0.2%
22	0.1%	16.2%	5.2%	78.3%	0.0%	0.2%
23	0.2%	16.4%	6.0%	77.2%	0.0%	0.2%
Whole day	0.1%	17.1%	6.1%	76.3%	0.0%	0.3%

Figure Q5-11. Drop-off traffic percentages for trips picked up at Brooklyn by hour

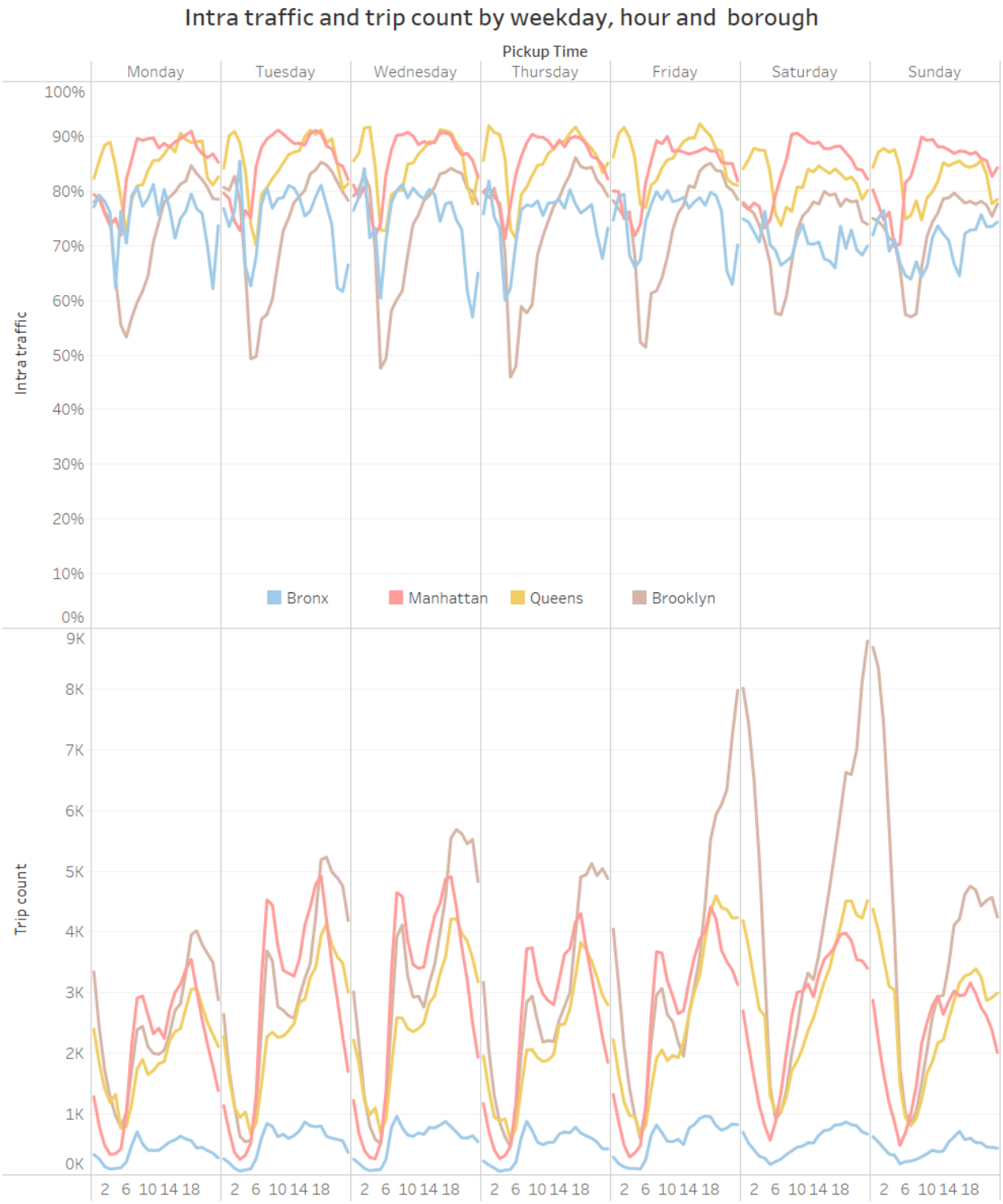


Figure Q5-12. Intra borough traffic percentage relative to the traffic in the borough and trip count by hour of the day/week and borough

REMARKS

The hidden content and patterns in this dataset is rich and more insight can be extracted given time. Also some of the analysis could be improved given time. Below are some of these improvement suggestions:

Q3-b. Categorizing airport pickup/drop-off locations

We used the distance from a central location of the airport to categorize pickup and drop-off locations – this works fine for JFK and Newark Liberty where the taxi locations are in a 'circular perimeter'. The two taxi location groups in LaGuardia are in a 'rectangular perimeter' and our circular distance method categorized few nearby non-airport pickups/drop-offs as LaGuardia airport pickups/drop-offs. Using a boundary shapefile for LaGuardia with geospatial libraries [25,26] would correct this.

Figure R-1. Green taxi pick/drop-off locations next to airports. JFK (top left), Newark Liberty (top right), LaGuardia (bottom).

Q4 Accuracy improvements

As mentioned in the conclusion of the Q4 chapter, more feature variables e.g. time of the pickup or location could improve the accuracy. Although the variable "Extra" accounts for the rush hour, explicit hours as variable could help. Also for the location aspect, we encoded the pickup and drop-off locations (latitude and longitudes) into geohash [27] buckets. The details of the code is in [28]. The geo hashed location buckets could be further grouped into bins (to reduce the number of categories) and used as location categories for model prediction. Also there are other model specific improvements noted in conclusion of Q4 chapter.

Geospatial code speed improvement

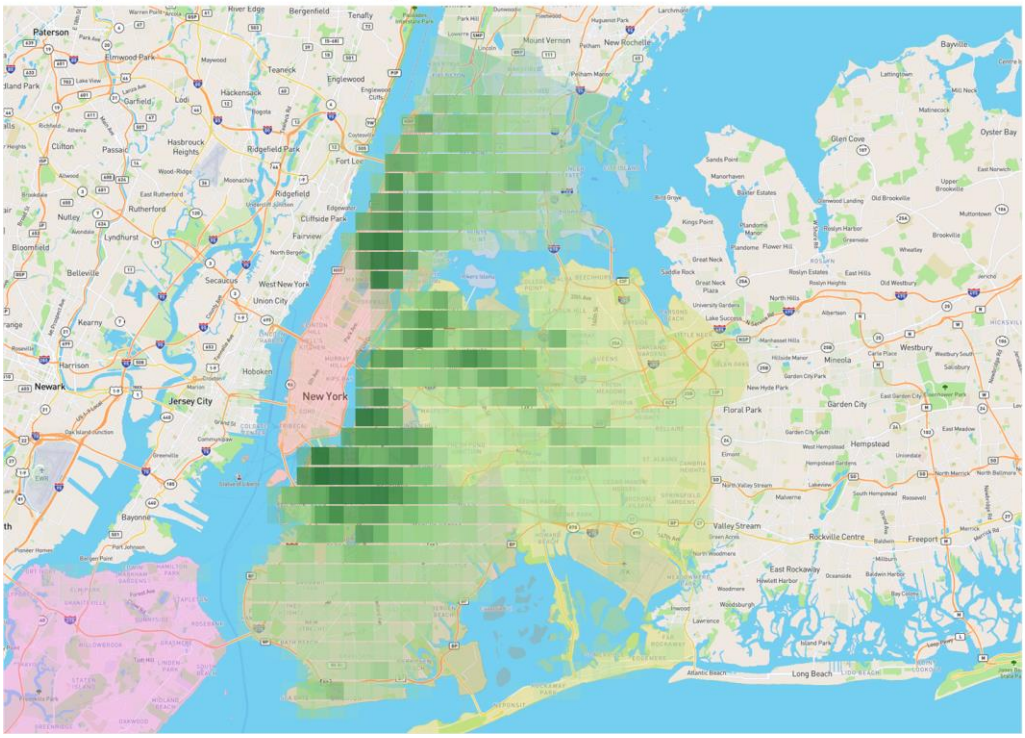
As mentioned in [24], the method for finding borough names for the pickup and drop-off locations took long (2 days on a computer with Intel Core i7-6850K 3.6GHz processor). Alternate methods could be investigated to speed up the computation – for example indexing the boundaries with some kind of geohash or a geospatial library like Magellan [30] on top of spark [22] running on a cluster [31].

Visualizations

Various geohashed tiles showing the taxi location density were explored. Figure R-2 in the next page shows examples of geo hashed tiles with precision 7 and 9 were explored. They did not look good or they were not useful in the context of our analysis. Tiles of different size could be explored.

Also, all the font size on label of the plots could be made larger for better readability.

Taxi density heatmap with precision 7 geohash grid



Taxi density heatmap with precision 9 geohash grid

Figure R-2. Examples of geohashed tiles with precision 7 (top) and 9 (bottom) showing taxi density.

REFERENCES

[1] The cover was created by using the mapbox [2] and Tableau [3] platform with the NYC borough boundary data from NYC OpenData site [4]. A custom mapbox background map was created first and the Visualizations-Q1-4.twb file in the code folder of the project created the cover image.

[2] http://www.mapbox.com

[3] http://www.tableau.com (Version 10.1.5)

[4] https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm/data

[5] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[6] https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015-09.csv

[7] http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf

[8] A python (3.5) environment with anaconda distribution [9] was used for this part. The file data_download.ipynb in the code folder contains the code.

[9] Python 3.5 Anaconda distribution from https://www.continuum.io/downloads

[10] The calculations and plots were created in the file Visualizations-Q1-4.twb in the code folder

[11] Python notebooks add_airport_tags.ipynb and helper.ipynb in the code folder

[12] https://jswhit.github.io/pyproj/

[13] https://en.wikipedia.org/wiki/World_Geodetic_System

[14] https://en.wikipedia.org/wiki/Boro_taxi

[15] These packages come standard with the anaconda distribution [9]

[16] Training code at model_train.ipynb file in the code folder

[17] Model testing code at model_test.ipynb file in the code folder

[18] https://en.wikipedia.org/wiki/Coefficient_of_determination

[19] prediction_model.p (a python cpickle file) in the data folder

[20] https://en.wikipedia.org/wiki/Agile_software_development

[21] https://keras.io, https://github.com/tensorflow/tensorflow, https://github.com/Theano/Theano

[22] http://spark.apache.org

[23] http://www.mckinsey.com/features/advanced_analytics

[24] To find which borough a pickup or drop-off location belong to, python Geospatial Libraries fiona [25] and shapely [26] were used. Both libraries were Linux/Max only in the anaconda distribution and the author used an Ubuntu Linux computer for this part of the code.

Package fiona was used to load the borough boundaries from a shapefile in the data directory (boundary shapefile downloaded from [4]). Then package shapely was used to locate lat,long point within the shape boundaries. This was a computationally intensive process - processing 5,000 locations took ~12 minutes in author's computer. About 3M locations (pickup and droff-off combined) needed to be resolved and the workload was split into two files: find_borough_pickup.ipynb, find_borough_dropoff.ipynb with a calculation function defined in the helper.ipynb (all in the code folder). It took about 2 days to process all the data points. Further investigation is needed to optimize this process (.e.g. indexing the boundaries and locations). The borough names are written into separate files for pickup and drop-off. The combine_data.ipynb file merged the new fields with source data (along with airport tags) into the modified file green_tripdata_2015-09_modified.csv. This files was then used with the Tableau platform to create the tables and interactive visualizations.

[25] https://github.com/Toblerity/Fiona

[26] https://github.com/Toblerity/Shapely

[27] https://en.wikipedia.org/wiki/Geohash

[28] The python package pygeohash [29] was used to geohash the pickup and drop-off locations in the add_geohash.ipynb file in the code folder. Three different precision levels (5, 7, 9) were used to get three different grid sizes. The one corresponding to precision 5 or 7 could be further grouped into bins and used in tip percent prediction model for Q4. The ones with precision 7 or 9 could be used in map visualizations in Q5.

[29] https://github.com/wdm0006/pygeohash

[30] https://github.com/harsha2010/magellan

[31] https://aws.amazon.com

APPENDIX

Instructions

This coding challenge is designed to test your skill and intuition about real world data. For the challenge, we will use data collected by the New York City Taxi and Limousine commission about "Green" Taxis. Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. We will use the data from September 2015. We are using NYC Taxi and Limousine trip record data: (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml).

Required Questions: Please answer completely all four required questions.

Question 1

- Programmatically download and load into your favorite analytical tool the trip data for September 2015.
- Report how many rows and columns of data you have loaded.

Question 2

- Plot a histogram of the number of the trip distance ("Trip Distance").
- Report any structure you find and any hypotheses you have about that structure.

Question 3

- Report mean and median trip distance grouped by hour of day.
- We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

Question 4

- Build a derived variable for tip as a percentage of the total fare.
- Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). We will validate a sample.

Question 5

Choose **only one of these options** to answer for Question 5. There is no preference as to which one you choose. Please select the question that you feel your particular skills and/or expertise are best suited to. If you answer more than one, only the first will be scored.

- *Option A*: Distributions
    - Build a derived variable representing the average speed over the course of a trip.
    - Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?
    - Can you build up a hypothesis of average trip speed as a function of time of day?
- *Option B*: Visualization
    - Can you build a visualization (interactive or static) of the trip data that helps us understand intra- vs. inter-borough traffic? What story does it tell about how New Yorkers use their green taxis?

- *Option C*: Search
    - We're thinking about promoting ride sharing. Build a function that given point a point P, find the k trip origination points nearest P.
        - For this question, point P would be a taxi ride starting location picked by us at a given LAT-LONG.
    - As an extra layer of complexity, consider the time for pickups, so this could eventually be used for real time ride sharing matching.
    - Please explain not only how this can be computed, but how efficient your approach is (time and space complexity)
- *Option D*: Anomaly Detection
    - What anomalies can you find in the data? Did taxi traffic or behavior deviate from the norm on a particular day/time or in a particular location?
    - Using time-series analysis, clustering, or some other method, please develop a process/methodology to identify out of the norm behavior and attempt to explain why those anomalies occurred.
- *Option E*: Your own curiosity!
    - If the data leaps out and screams some question of you that we haven't asked, ask it and answer it! Use this as an opportunity to highlight your special skills and philosophies.