Midtown to JFK Trip Duration

MAST30034: Applied Data Science—Assignment 2

Abstract

New York is well known for its yellow taxi cabs. However, it is also the 4th most congested city in the United States, as well as the 14th most congested in the world [1]. This paper aims to model the trip duration along the frequently travelled route of Midtown Manhattan to JFK Airport in a Yellow Taxi cab.

Stakeholders

Given the existence of the JFK Airport flat rate, the ability to predict the duration of a trip falls within the interest of taxicab drivers. On trips between JFK and Manhattan, NYC taxis offer a flat-rate fare of $52 exclusive of tolls, tips, and extra surcharges [2]. It thus makes sense that drivers want to complete the trips in the fastest possible time in order to maintain profitability, and the interpretation of a trip duration model may offer insights as to what conditions are inductive of a profitable trip.

Likewise, passengers are likely to prefer shorter trips to and from the airport. While this may not ultimately be a factor falling entirely within their control (since flight times are usually fixed), it is still advantageous to have access to the predictive output of a trip duration model for their planning.

The data

The data examined are a combination of the New York City Taxi and Limousine Commission (TLC) trip record data that has been made freely available to public [3], and climate data collected from JFK Airport, also made public by the National Oceanic and Atmospheric Administration [4]. The TLC dataset describes attributes of individual taxi trips spanning 2009 – 2020. The climate data describes the observed daily weather conditions.

This paper focuses on the portion of trips between Manhattan and JFK Airport made by medallion taxis in 2019. It is chosen for its recency, completeness and relevance. The more recent 2020 dataset has not been considered. In 2020, border restrictions in response to the COVID-19 pandemic severely crippled flight demand in the aviation industry, and road-traffic volumes themselves decreased due to stay-at-home measures. Thus, the data available to explore in 2020 was itself limited.

Features

From the TLC trip record data, the *tip amount, passenger count, pickup datetime, drop-off datetime, pickup location ID, drop-off location ID, and congestion surcharge* features were extracted. The duration of the trip is computed by the difference in pick-up and drop-off time.

Fare amount and *trip distance* features are not selected, despite displaying a promising relationship with the computed trip duration (Pre-processing

Data cleaning

Every effort is made to try make the data as clean as possible. The climate data is very clean with few missing values, which are simply imputed with the mean value for the month. Most of the cleaning effort is thus focused on the TLC data, which is cleaned on two levels:

1. In the full dataset, eliminate instances (individual trips) which clearly are not actual trips;

2. In our focussed subset of the data, investigate for any patterns that do not make sense.

At the first level of cleansing, negative values in features that should only take positive values are eliminated. Negative fares were likely refunds or disputes, instead of actual trips. Negative trip durations (whatever the generation mechanism: some due to spanning midnight, or by a change of system date) should not be used for training

the model, so were discarded. Negative trip distances are rare to encounter, but these instances are similarly discarded. The lower bound for all these three features are effectively set to zero.

Figure 1 and

Figure 2). These two attributes would be meaningless as input to a predictive model of duration since their exact value would not be known until a trip is completed (and thus the trip duration would also already be known).

The values known at the commencement of a trip are *pickup* and *drop-off location*, and *number of passengers*, and this is the reasoning behind the selection of these features from the TLC dataset. Furthermore, the tip amount (which is not definitively known at the beginning of a trip) was selected as a feature, based on the hypothesis that passengers could be willing to commit to tipping more for being transported to the airport in a timely manner. This feature then may act as a very loose proxy for how much the customer values a fast trip to the airport (among, of course, other things).

From the climate data, the features extracted are listed in Table 1. They are chosen due their ability to adversely affect driving conditions, and the plausibility that each feature could affect demand for transportation in general.

Table 1—Climate features

Feature	Description
AWND	Average wind speed (mph)
PRCP	Precipitation (in)
SNOW	Snowfall (in)
SNWD	Snow depth (in)
TAVG	Average daily temperature (°F)
TMAX	Maximum daily temperature (°F)
TMIN	Minimum daily temperature (°F)
WT01	Fog
WT02	Heavy fog or heavy freezing fog
WT03	Thunder
WT04	Ice pellets, sleet, small hail
WT05	Hail
WT06	Glaze, rime
WT08	Smoke, haze
WT09	Drifting snow, blown snow

Pre-processing

Data cleaning

Every effort is made to try make the data as clean as possible. The climate data is very clean with few missing values, which are simply imputed with the mean value for the month. Most of the cleaning effort is thus focused on the TLC data, which is cleaned on two levels:

3. In the full dataset, eliminate instances (individual trips) which clearly are not actual trips;

4. In our focussed subset of the data, investigate for any patterns that do not make sense.

At the first level of cleansing, negative values in features that should only take positive values are eliminated. Negative fares were likely refunds or disputes, instead of actual trips. Negative trip durations (whatever the generation mechanism: some due to spanning midnight, or by a change of system date) should not be used for training the model, so were discarded. Negative trip distances are rare to encounter, but these instances are similarly discarded. The lower bound for all these three features are effectively set to zero.

Figure 1—Comparison of fare amount and trip duration

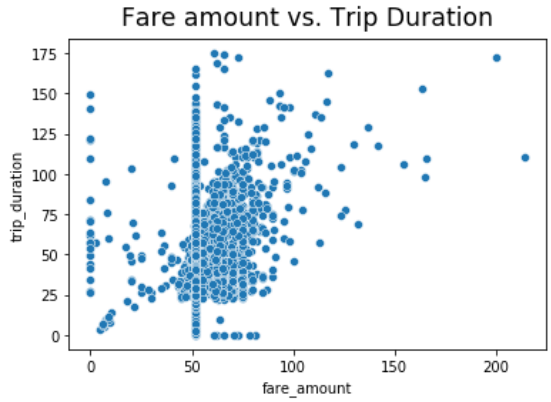
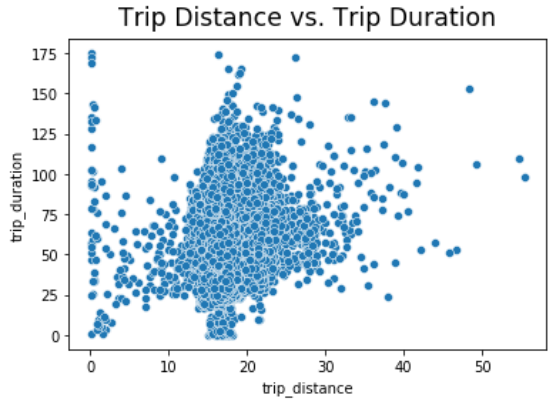


Figure 2—Comparison of trip distance and trip duration



To determine the upper bound of these features, we define the scope of analysis to pertain to direct trips with endpoints within New York City and Newark Airport only (however, trips may not necessarily be entirely contained within NYC e.g. Manhattan to Staten Island via New Jersey). By routing between the two (approximately) furthermost locations within NYC on Google Maps—Tottenville to City Island, a distance of 50.9 miles [5]—a conservative upper bound for the travel distance of a direct taxi trip is set at 60 miles. Furthermore, picking a rush-hour travel time (3:15pm Friday 8 Nov 2019), an upper bound on the travel time was set at 3 hours. Instances with feature values greater than these two bounds are considered safely outside the scope of analysis and discarded.

A few very high fare trips were found, which may have been entry errors or other kinds of correction. An arbitrary upper threshold of $350 (seems reasonable for a long trip) was taken and instances with fares above this discarded.

Feature engineering

This analysis does not attempt to model the time-series data continuously. As such, the pickup time was converted to a day, month, and hour feature—each classed as a category.

The trip duration was computed from the difference in pickup and drop-off time, and the drop-off time feature discarded (obviously, drop-off times will never be known in advance, and, predicting duration with both pickup and drop-off features present is a trivial task).

Lastly, because the decision tree regressor implementation used from *SciKitLearn* does not yet support categorical variables, the categorical variables (*day*, *hour*, *month*, *WT01* through *WT09*) are one-hot encoded so they can be fitted properly by the estimator.

Table 2—Feature summary statistics

	count	mean	std	min	25%	50%	75%	max
fare_amount	124235	52.160589	2.29511819	0.01	52	52	52	214
tip_amount	124235	8.43406431	6.10649824	0	0	11.06	12.28	120
passenger_count	124235	1.66410432	1.24398863	0	1	1	2	7
trip_distance	124235	17.0834711	1.54367475	0.01	16.4	16.9	17.43	55.4
congestion_surcharge	124235	2.29373767	0.6878339	0	2.5	2.5	2.5	2.5
trip_duration	124235	49.4556966	17.2712226	0	36.3166667	47.2166667	60.1333333	174.616667
wind_speed_avg	124235	11.5292557	4.24485616	4.25	8.5	10.51	13.87	29.3
precipitation	124235	0.1447349	0.31875336	0	0	0	0.1	2.07
snowfall	124235	0.02759045	0.20817829	0	0	0	0	3.2
snow_depth	124235	0.03773011	0.24279372	0	0	0	0	3.1
temp_avg	124235	55.0423069	16.1634797	9	41	56	69	89

temp_max	124235	62.2621564	17.167961	15	48	63	77	99
temp_min	124235	48.0879865	16.1062223	3	34	49	62	80

Descriptive analysis

Value distributions

The final Midtown Manhattan to JFK dataset consisted of 124,559 instances. A statistical summary of the final features used are described in **Error! Reference source not found.**. Visualisations of the value distributions of each individual feature are included in Appendix A—Feature distributions.

The *trip duration* is bell shaped and centred around 50 minutes, as expected since the origin and destination are fixed.

The most common *tip* is between $12-14, but many travellers also do not tip. This yields a bimodal distribution. Higher tips (not shown in figure) do exist but are rare.

Passengers tend to travel to the airport alone.

Figure 3—Mean trip duration by pickup time across the week

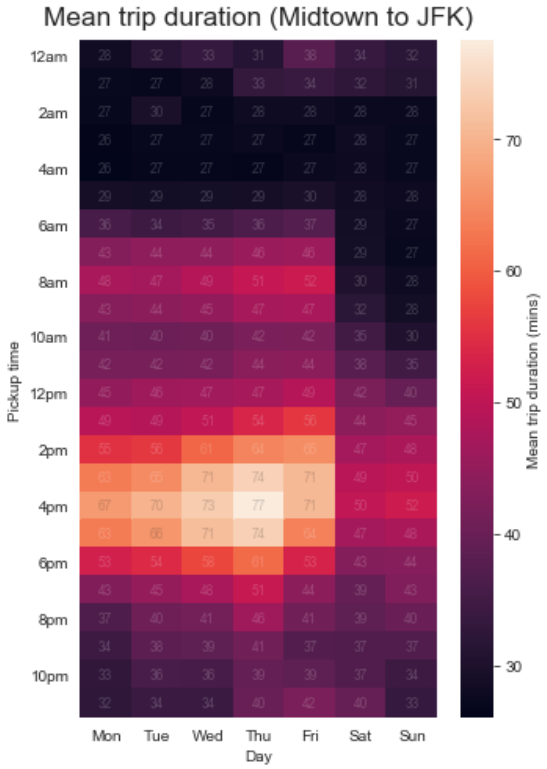
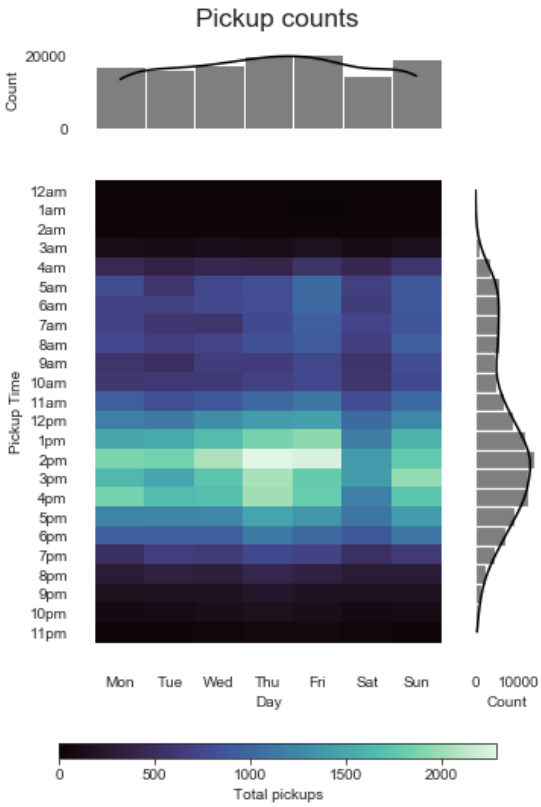


Figure 4—Total pickups at each time period across the week



The number of trips to JFK roughly increases through the work week, peaking Friday. Saturday is the least popular day to travel to JFK.

As for the particular time of day to travel, most commonly people start their trip to the airport in the afternoon, peaking at 3pm. However, there also exists a smaller early morning peak at 5 or 6am. This could possibly be explained by an influx of airport staff arriving for work, but the data cannot confirm the cause of this.

Interestingly, comparing the distribution of *wind*, *precipitation* and *snow* data recorded for every trip (blue), to the weather distribution over the year, we see a higher number of trips occurring when there is good weather (low wind speed, low precipitation, low snowfall and snow depth) as indicated by the increased right skew of the blue density curve (trip weather distribution) compared to the yellow curves (day weather distribution). As the number of trips being taken at any point in time is related to the future trip duration (see below), this suggests there is some sort of correlation between good weather and favourable traffic conditions.

Temperature displayed a similar pattern. More trips happened in less extreme temperatures, as the blue density curve is less flat than the yellow curve.

Relationships

Relationships between the features were also explored.

Pairwise relationships amongst the continuous and categorical variables were considered but the plots were not very insightful (see Appendix B—Pair plot). As a general rule, there appears to be no clear relationship between each of the variables and the trip duration, however, the variance of the trip durations increases with the number of samples at each feature level. Only the pickup time and day showed any signs of a clear relationship. Thus, different plots were made to try explore these relationships.

One hypothesis is that the duration of a trip depends strongly on the time of departure. Figure 3 shows that for each day, the hour of 4pm is usually the most congested time to travel to JFK, with travel times reducing before and after that. Thursday is the most congested day, and interestingly average travel times during the weekend are much shorter.

Another hypothesis is that taxi pickup activity may be a good predictor of future trip duration. Figure 4 shows the number of pickups follows a similar trend to the mean trip duration, with increases of pickups translating into a pickup in trip duration with 1-2 hours' delay.

Model

The data seems to violate a few key assumptions of the linear regression model. The above section shows a lack of a clear linear relationship demonstrated between the features and target variable. Additionally, many of the weather variables will clearly be dependent on each other. Therefore, a linear regression was ruled out as a model for trip duration.

A preliminary linear model was produced and found to perform poorly (see Appendix C—Linear Regression Model).

We therefore turn to a regression tree to model the trip duration.

Regression Tree

The decision tree regressor works similarly to a decision tree classifier: by learning decision rules inferred from data features to predict the target variable. At each node of the tree, these decisions rules are chosen to increase the quality of split measured by the mean squared error i.e. the L2 loss using the mean of each terminal node.

$$MSE(x) = \frac{1}{n}\Sigma_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

The model makes few assumptions: the response is normally distributed (so that we can apply the mean square error criterion to reduce variance) [6], and that instances are distributed recursively on the basis of attribute values. No assumption of linearity or smoothness is required [6]. The first of these assumptions is satisfied by looking at the distribution of trip duration, which is approximately normal (Figure 9). The second of these is assumptions is harder to presents a challenge for decision trees if faced with an XOR problem, but this is not the case in this instance.

The decision tree model is simple to understand and to interpret, so it is hoped the structure of the tree sheds light on which factors influence the trip duration.

However, the disadvantage of decision trees is they can generate over-complex models that overfit the data, and thus yielding poor generalisation results. The generalisation performance of a decision tree may be improved by limiting its complexity: train it only on discriminatory features, limit the depth, or limit the purity of leaf nodes. Another method of addressing generalisation performance is to use ensemble methods e.g. Random Forest.

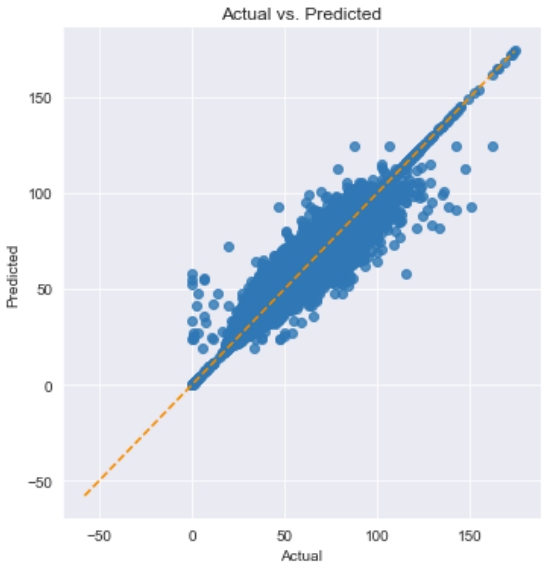
This problem of generalisation performance is addressed by model optimisation described below.

Baseline regressor

As a baseline, an initial decision tree model was fitted on all features of a 90% held-out training set with no limitation on the model complexity. The tree grew to a depth of 57 with 95,962 leaf nodes.

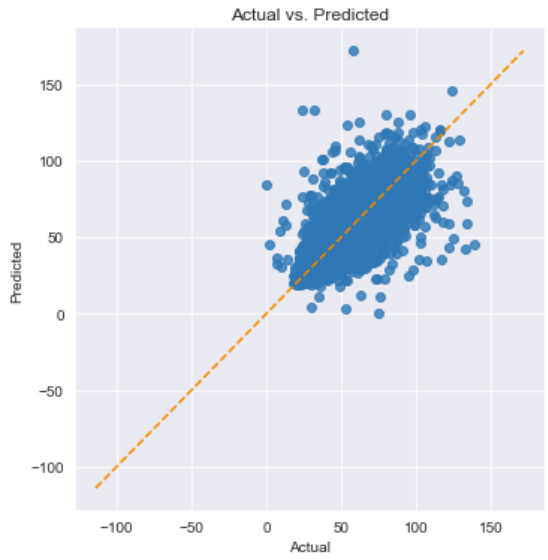
As expected, the tree fitted the training data extremely well and yielded a R2 score of 0.9726.

Figure 5—Baseline tree overfits on training set



However, this tree generalised poorly. On the held out test set, the model scored 0.6085. 10-fold cross validation gives an average test R2 score of 0.0375, which means the model is only marginally better than guessing the mean trip duration every time.

Figure 6—Poor quality of fit of baseline tree on held out test set



Feature selection and hyper-parameter tuning

To optimise the model, we perform a 5-fold nested cross-validated grid search across a series of candidate hyperparameter values.

First, a set of trees are fitted using all features for feature selection. This feature selection tree is hyperparameter optimised with 5-fold cross validation over candidate values of max tree depth of 5, 10, 20, 40 and no limit.

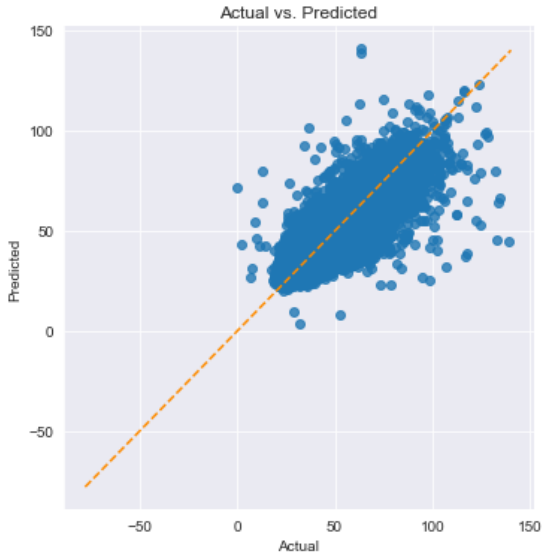
The best feature selection tree is then used to select from features where the importance of each feature used to train this tree is ranked above 0.5, 1 and 1.5 times the mean importance.

Each set of selected features are then used to fit another tree which is again optimized over the depth hyper-parameters.

Performance evaluation

The final regressor selected via the optimisation process used only features above average importance and limited the tree depth to 20 nodes. It was able to achieve a good generalisation with a mean validation R2 score of 0.6546. Testing on the 10%-held out test set yielded an R2 of 0.6682.

Figure 7—Optimised model performance



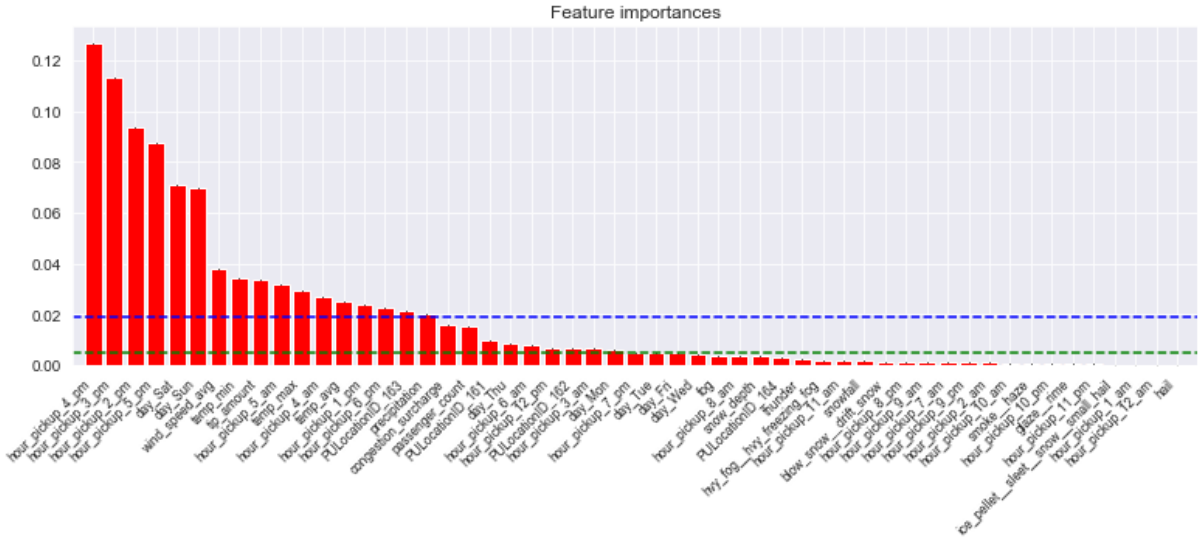
Interpretation

To see which features were selected, we can plot the feature importance of the selection tree (Figure 8). The selected features are those with above mean performances (above the blue line).

In line with the hypotheses above, the afternoon hours of pickup (2pm-5pm) are very predictive of the trip duration, comprising the first 4 most important features. The weekend days are also predictive, having low travel times. Surprisingly, wind speed and the temperature were much more predictive of the duration of a trip compared to the precipitation. Furthermore, tip amount may also be used to help predict the trip duration.

The individual decision rules that were used for the regression are 20 layers deep, this is quite difficult to interpret.

Figure 8—Feature importances of the optimised model



Summary and recommendations

This initial modelling helps to refine some of the relationships that may be influencing the duration of a taxi trips between Midtown Manhattan and JFK airport. By optimising the model and selecting features, it is determined that afternoon times, weekends and select weather features (wind, temperature and precipitation) show a relationship with the trip duration.

Such a model may explains some variation of the trip duration and may be used predict the trip duration to a good accuracy.

The exact influence of these features on the duration is not known by the outcome of this modelling, and may be explored deeper and at greater granularity using time series modelling.

References

1] INRIX, "New York City, NY," 2019. [Online]. Available: https://inrix.com/scorecard-city/?city=New%20York%20City%2C%20NY&index=14. [Accessed Oct 2020].

2] New York City Taxi and Limousine Commission, "Taxi Fare," 2020. [Online]. Available: https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page. [Accessed 1 October 2020].

3] New York City Taxi and Limousine Commission, "TLC Trip Record Data," 2020. [Online]. Available: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

4] National Oceanic and Atmospheric Administration, "Climate Data Online: Dataset Discovery," 2019. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND. [Accessed October 2020].

5] Google, "Directions Tottenville to City Island," [Online]. Available: https://www.google.com/maps/dir/Tottenville,+Staten+Island,+NY,+USA/City+Island,+The+Bronx,+NY,+USA/@40.6906086,-74.6057664,9z/data=!3m1!4b1!4m18!4m17!1m5!1m1!1s0x89c3caa227f45491:0x76be106492aa3540!2m2!1d-74.2355404!2d40.5083408!1m5!1m1!1s0x89c28c7f7705f. [Accessed Oct 2020].

Appendix A—Feature distributions

Figure 9—Trip duration distribution

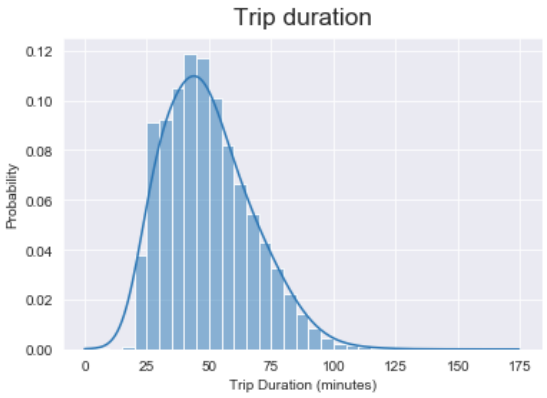


Figure 10—Wind speed distribution

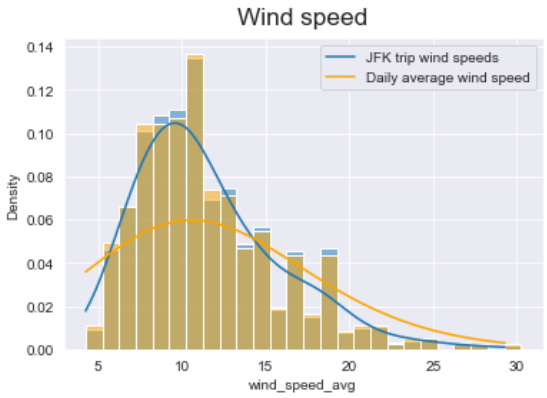


Figure 11—Tip distribution

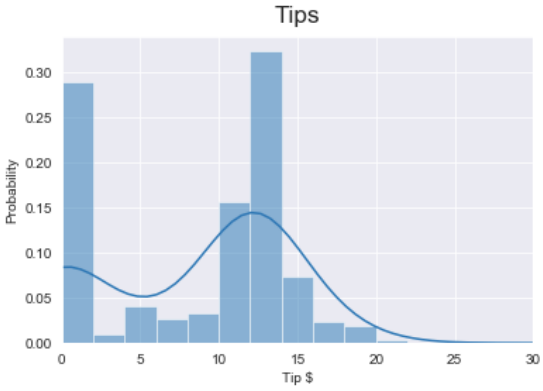


Figure 12—Precipitation distribution

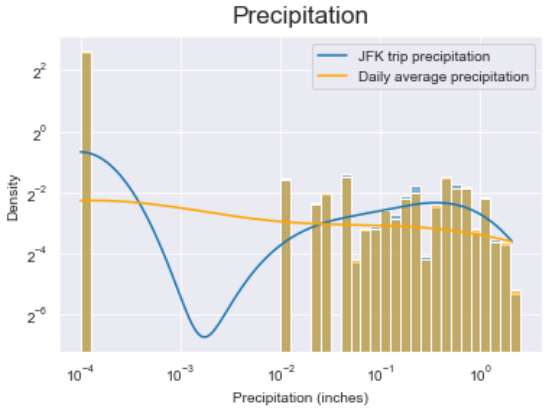


Figure 13—Passenger distribution

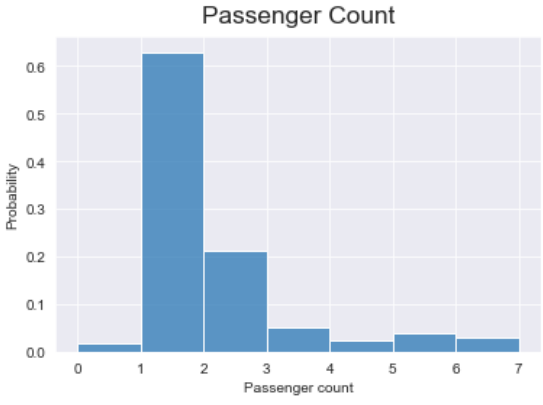


Figure 14—Snowfall distribution

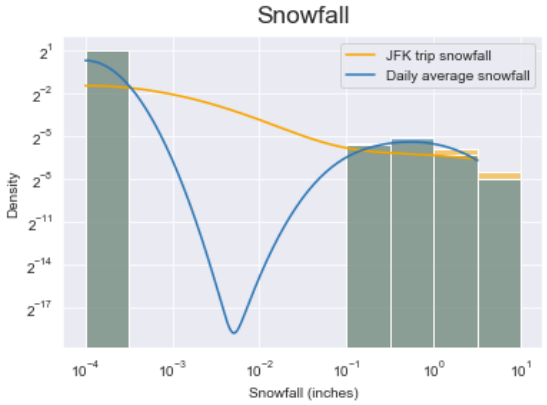


Figure 15—Trip day distribution

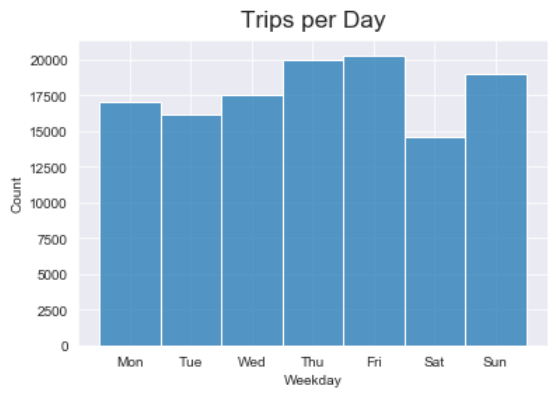


Figure 16—Snow depth distribution

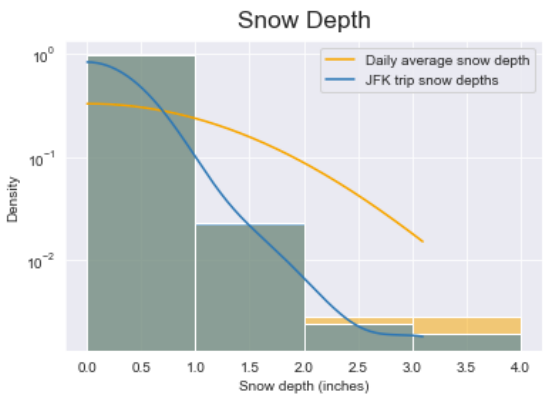


Figure 17—Trip hour distribution

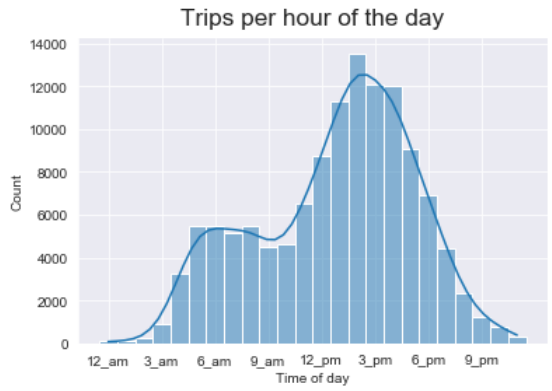
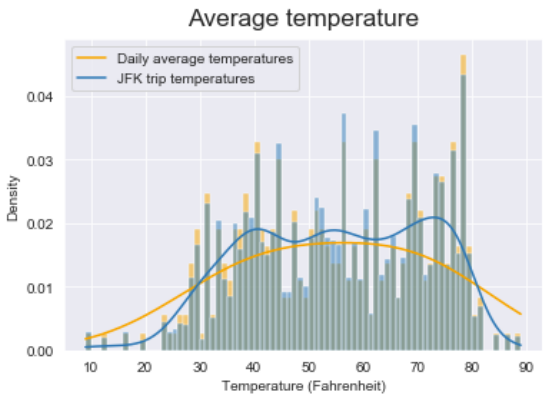


Figure 18—Average temperature distribution



Appendix B—Pair plot

Figure 19—Continuous feature pairs

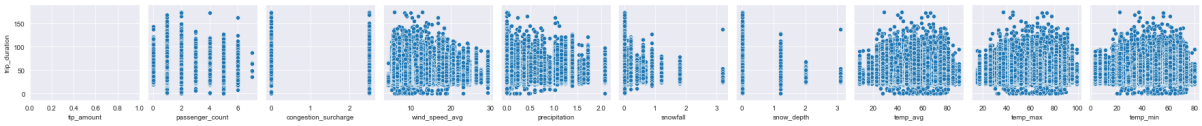
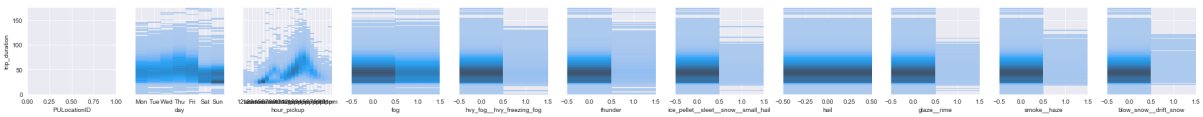


Figure 20—Categorical feature pairs



Appendix C—Linear Regression Model

Figure 21—Output of linear regression model