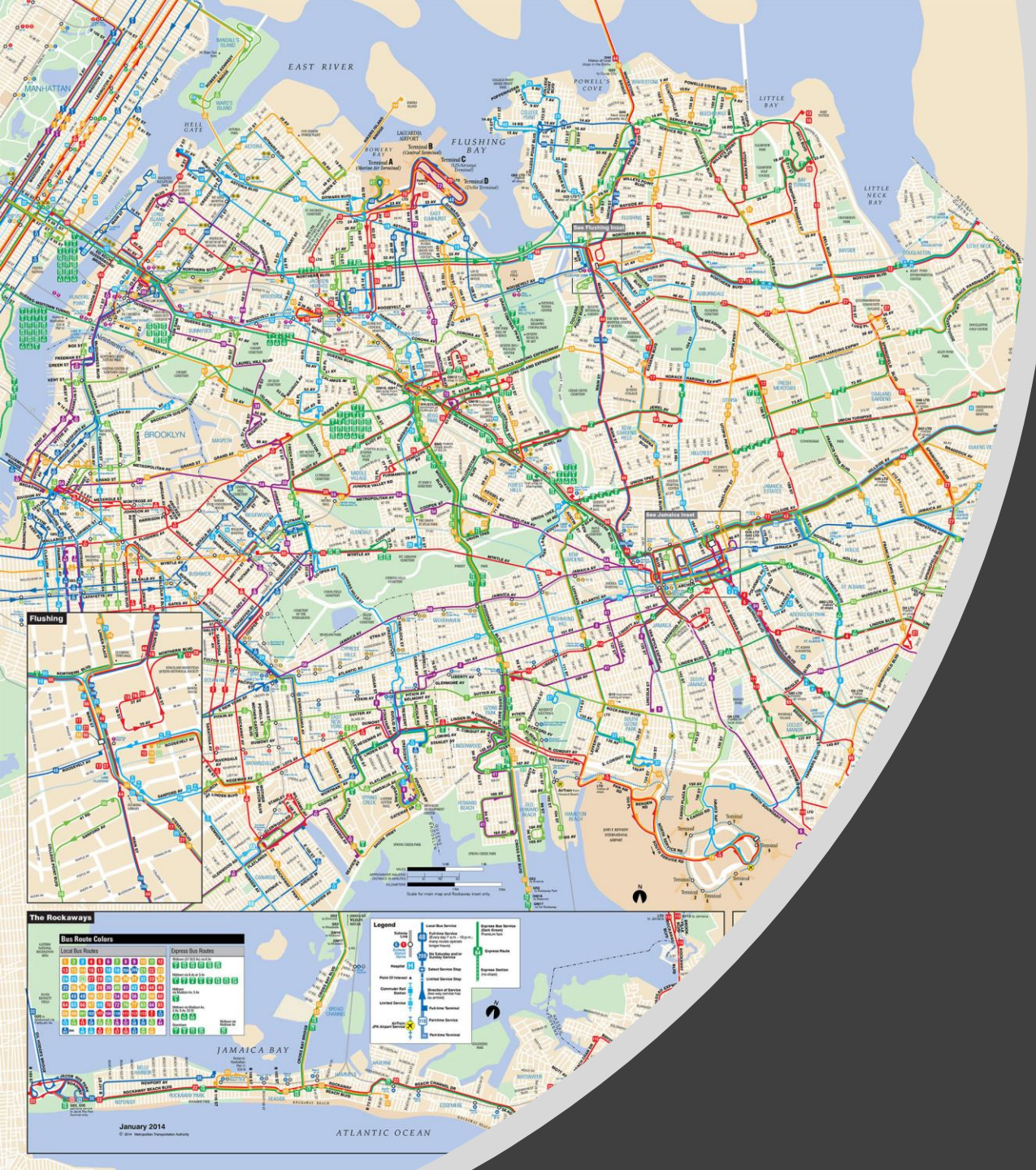


Reliability of the New York City Bus System

Purpose of project

- Gain insights from live traffic data.
- Build multiple machine learning classification algorithms to predict whether a bus in New York City will be on-time or not on-time.

THE DATASET

Latitude and Longitude coordinates

Names of origins, destinations, and bus lines.

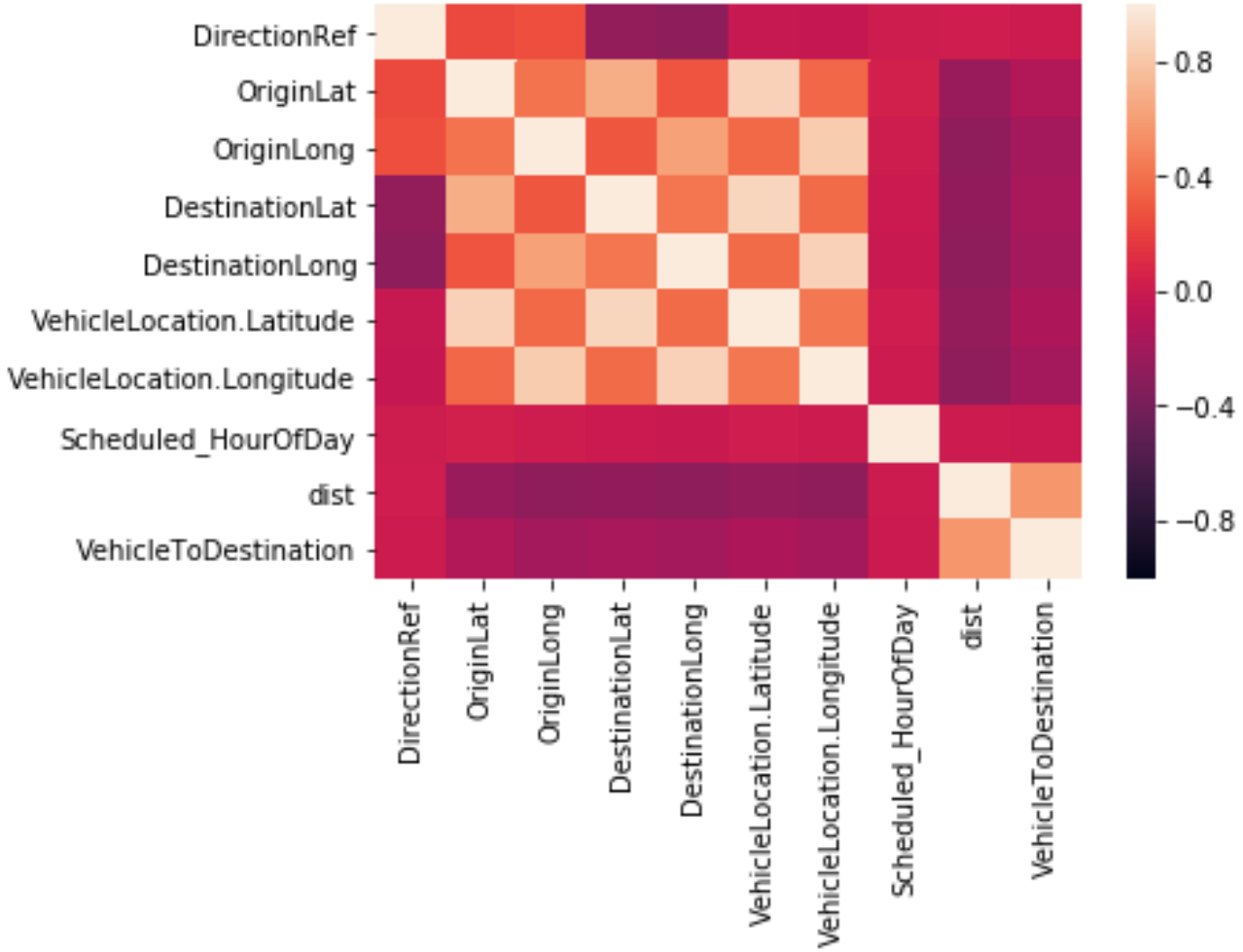
Datetime columns

Columns for minutes and hours.

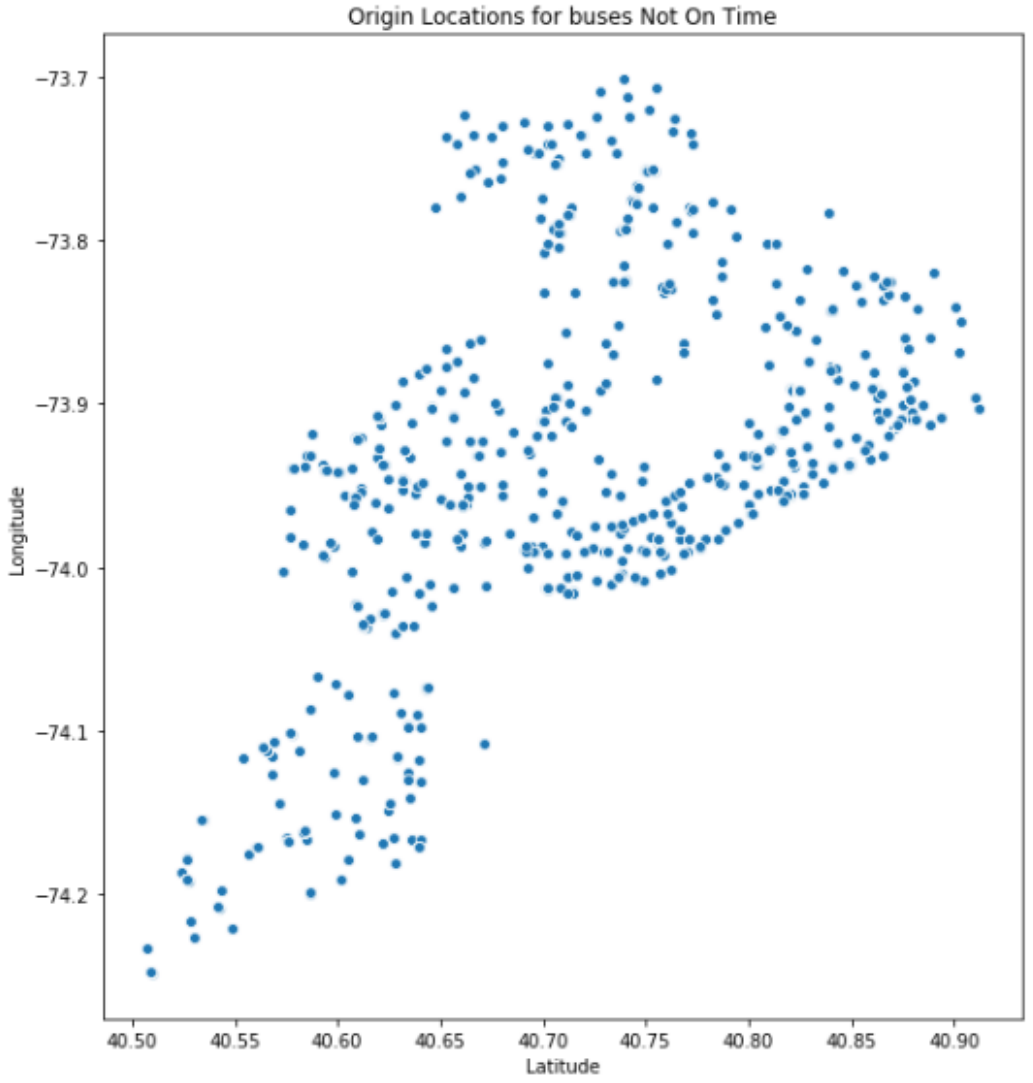
Columns representing distance.

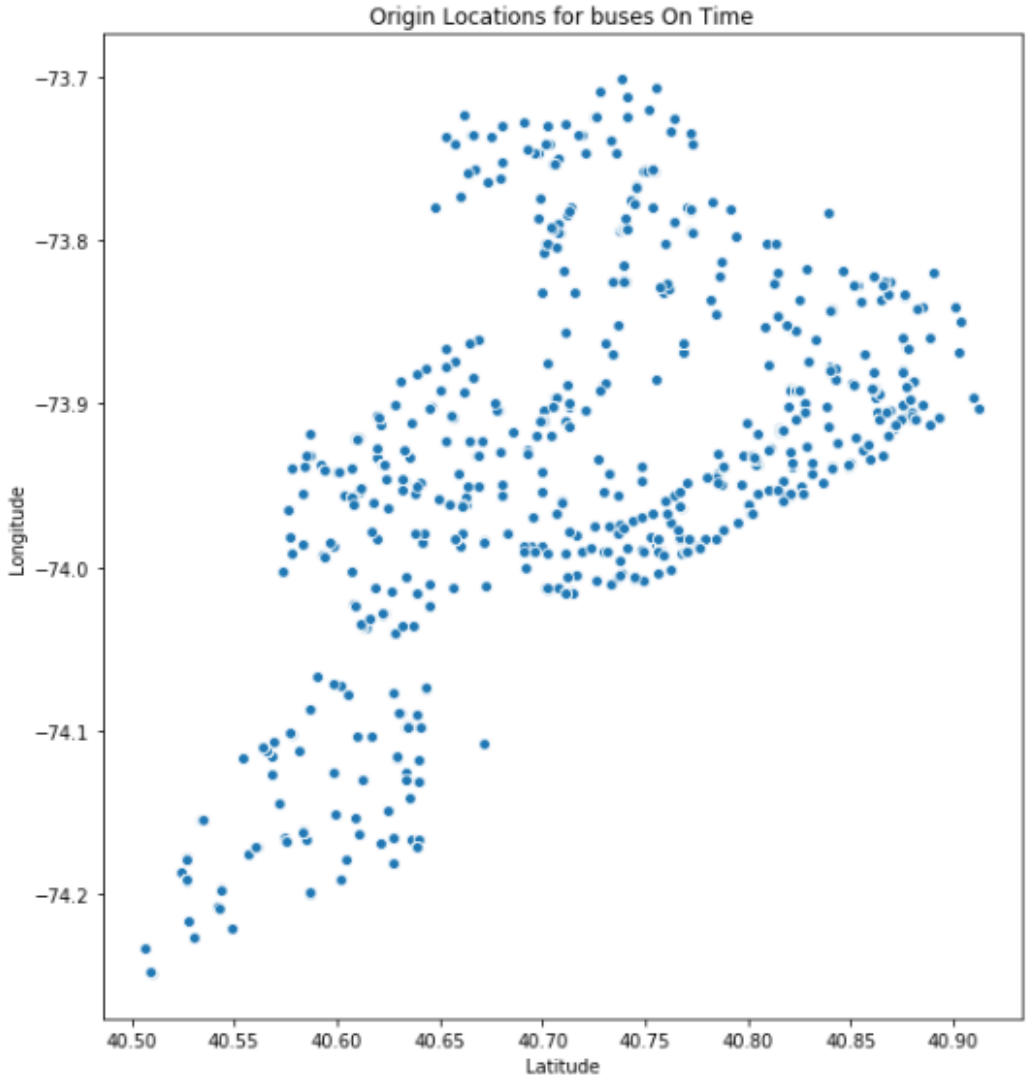
Output variable “on time”.

Multicollinearity

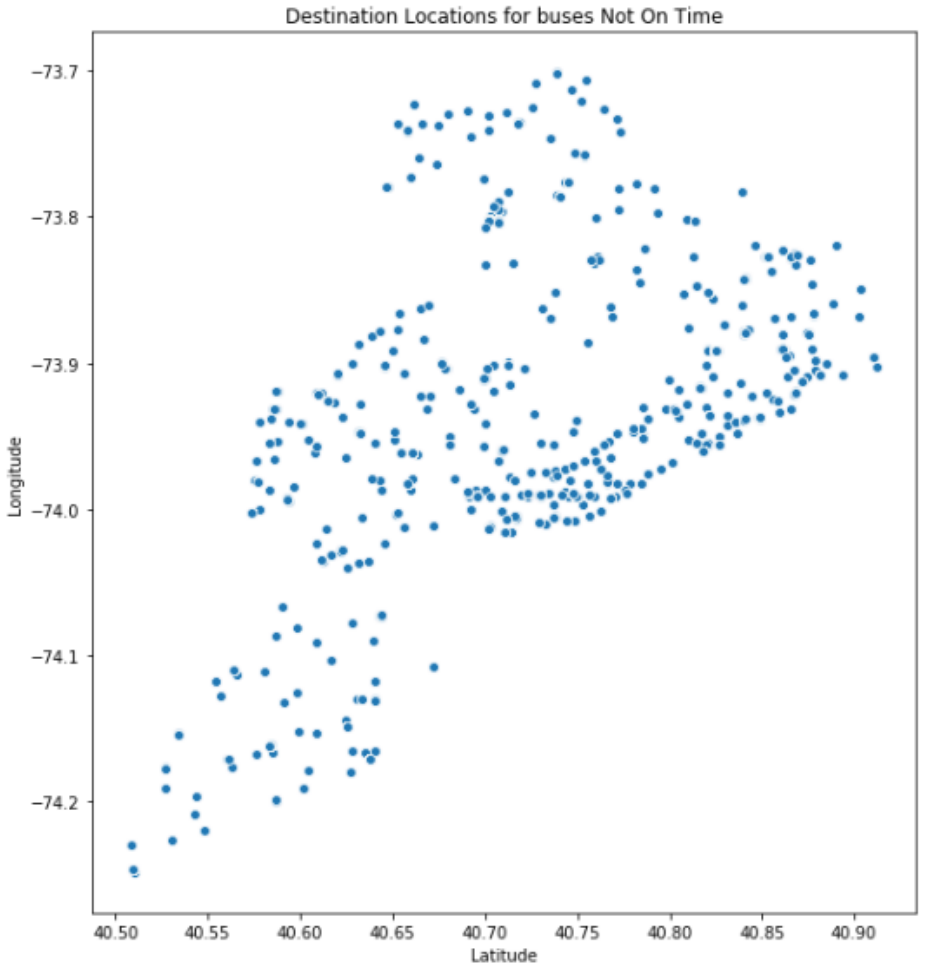


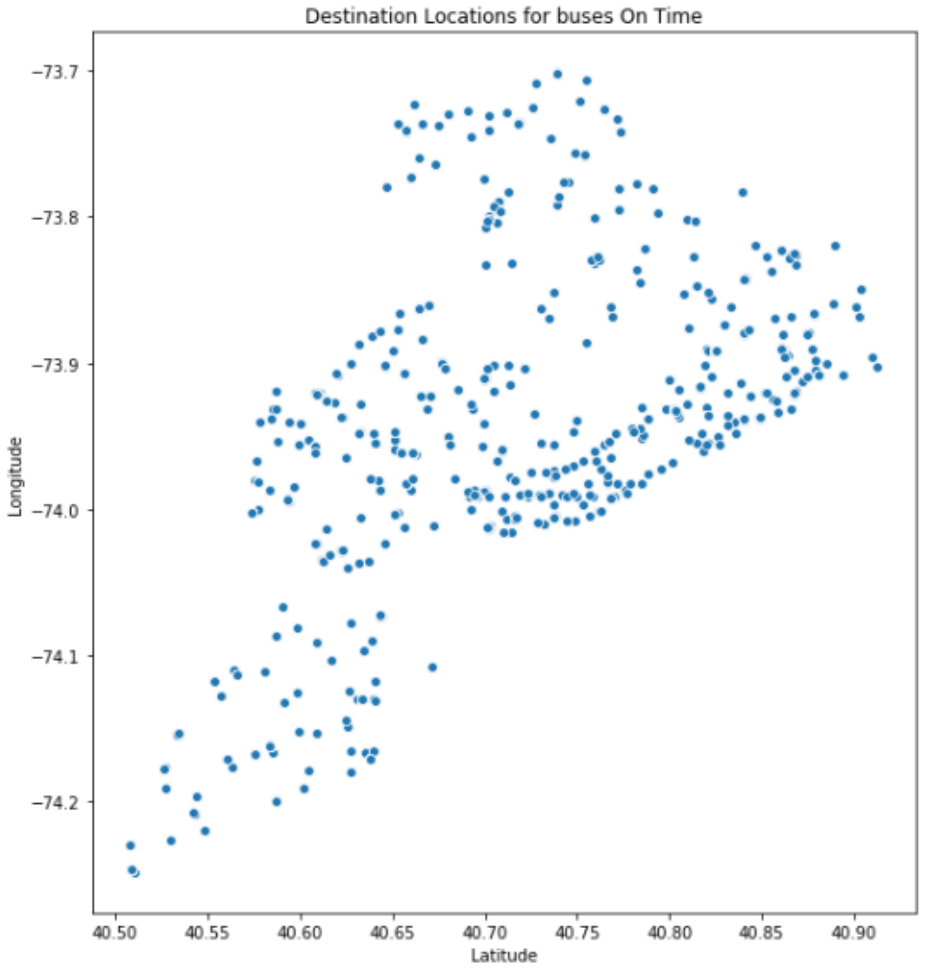
ORIGIN LOCATIONS





DESTINATION LOCATIONS



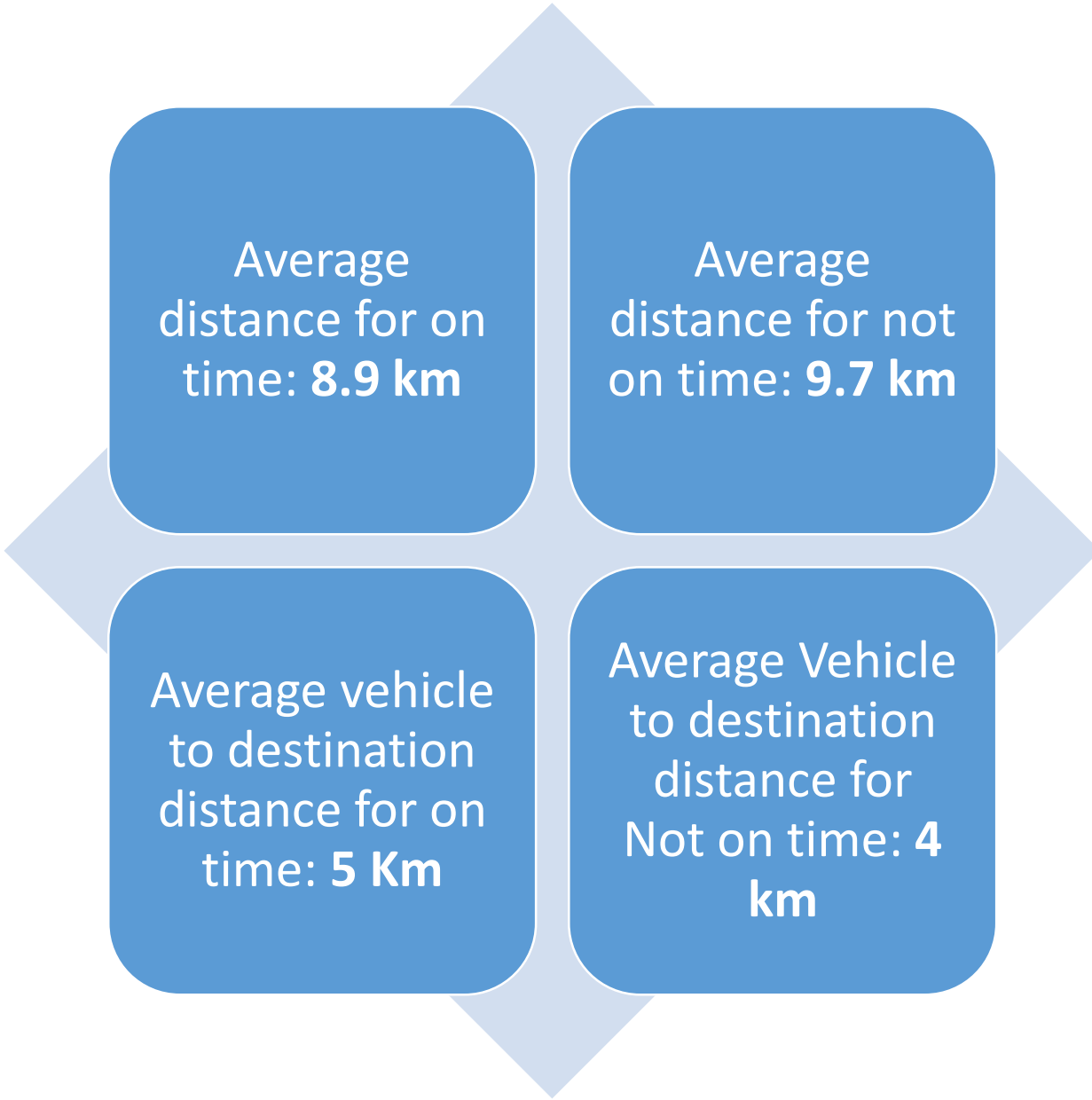


INTERESTING FACTS

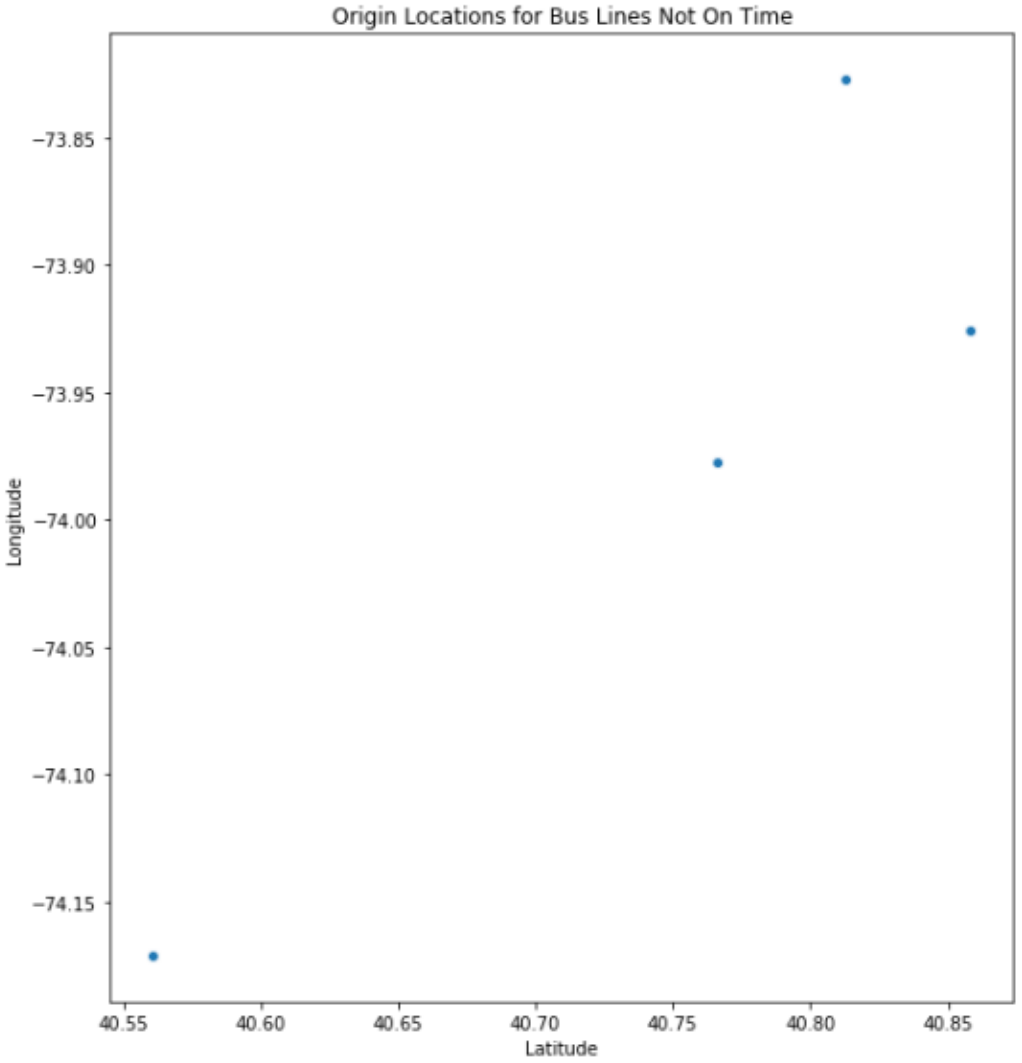
- Average distance for on time: **8.9 km**
- Average distance for not on time: **9.7 km**

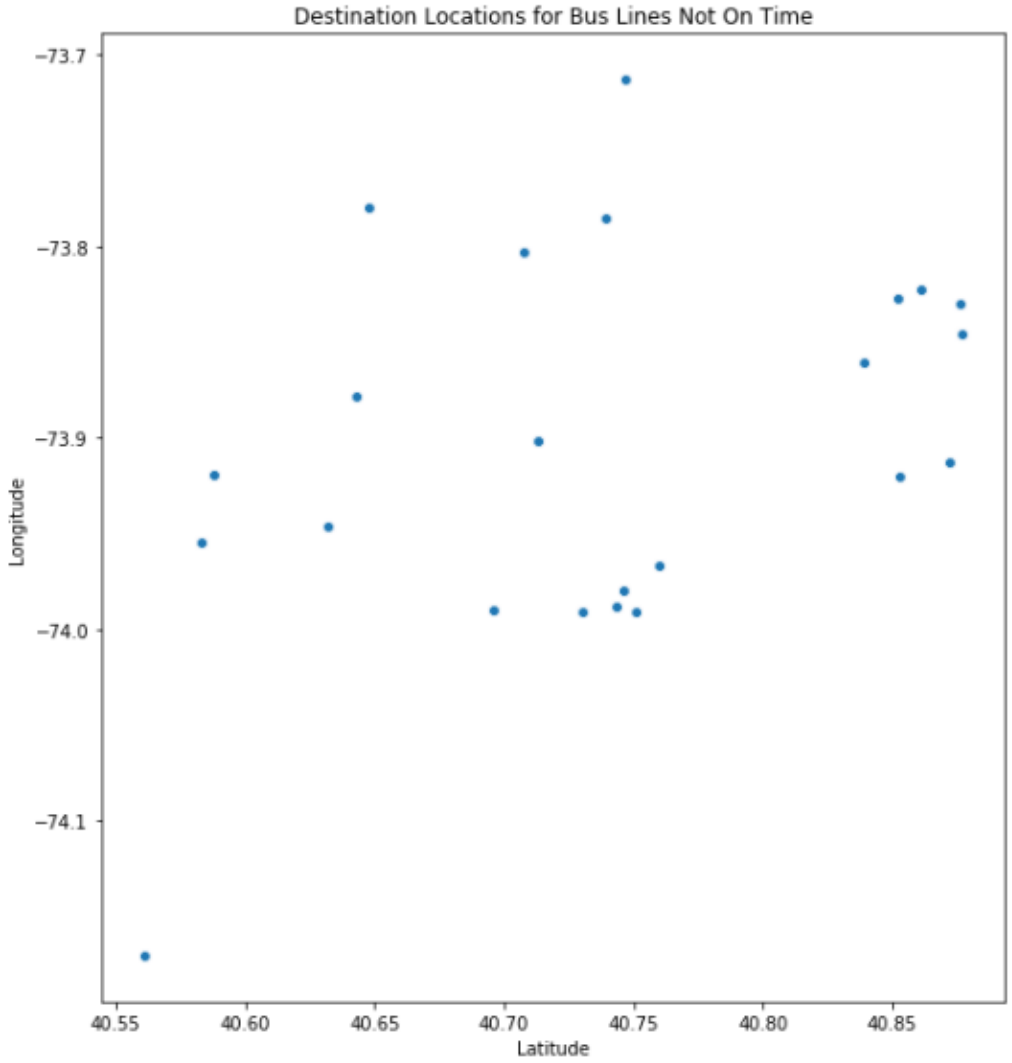
- Average vehicle to destination distance for on time: **5 Km**
- Average Vehicle to destination distance for Not on time: **4 km**

INTERESTING FACTS



BUS LINES NEVER ON TIME

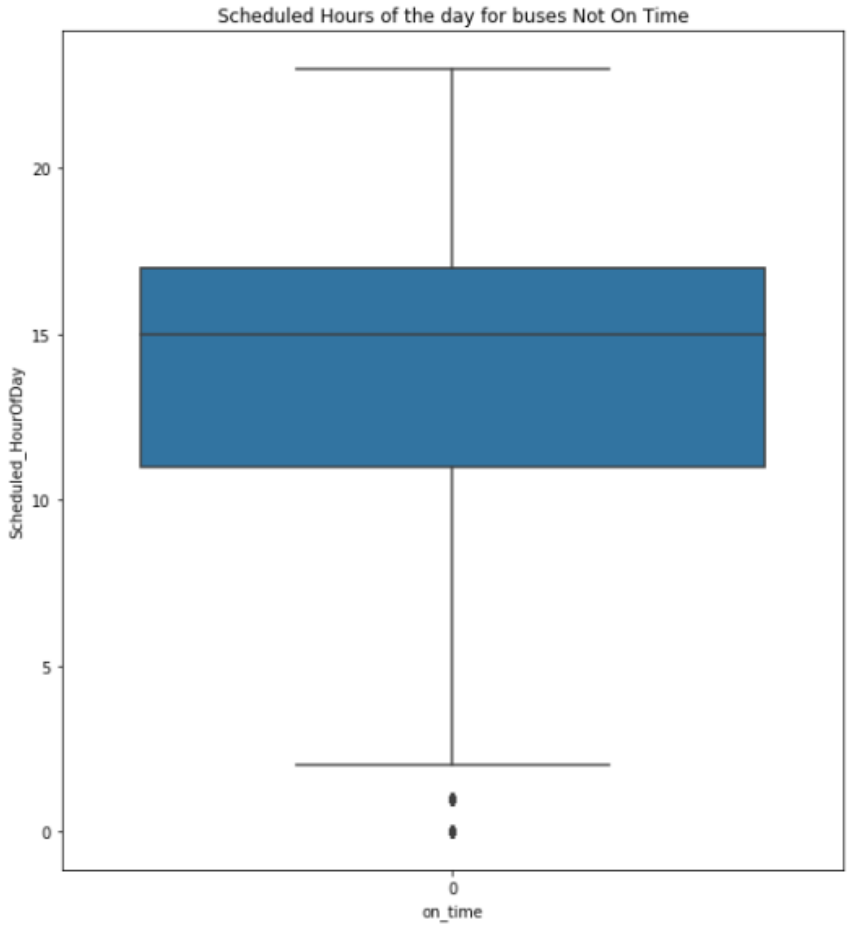


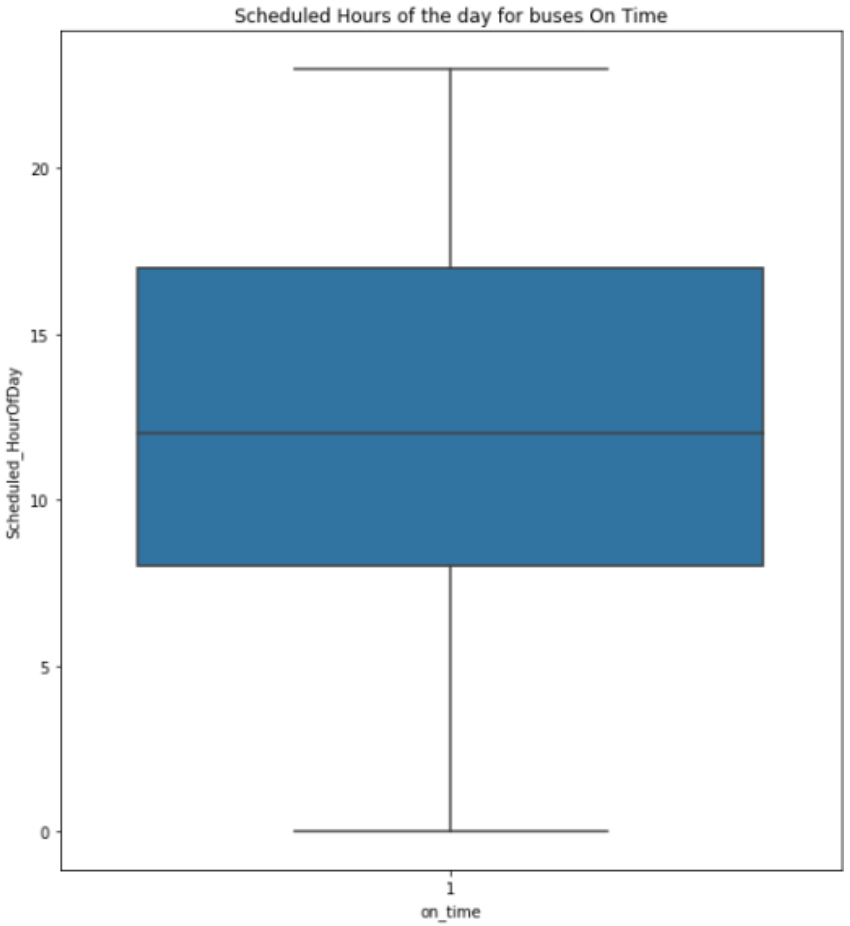


ORIGINS WITH A PROBLEM

- 40.56 and -74.18 (Staten Island)
- 40.77 and -73.97 (upper east side of Manhattan)
- 40.86 and -73.92 (Bronx, near Yankee Stadium and the Bronx Zoo)
- 40.81 and -73.82 (near the East River)

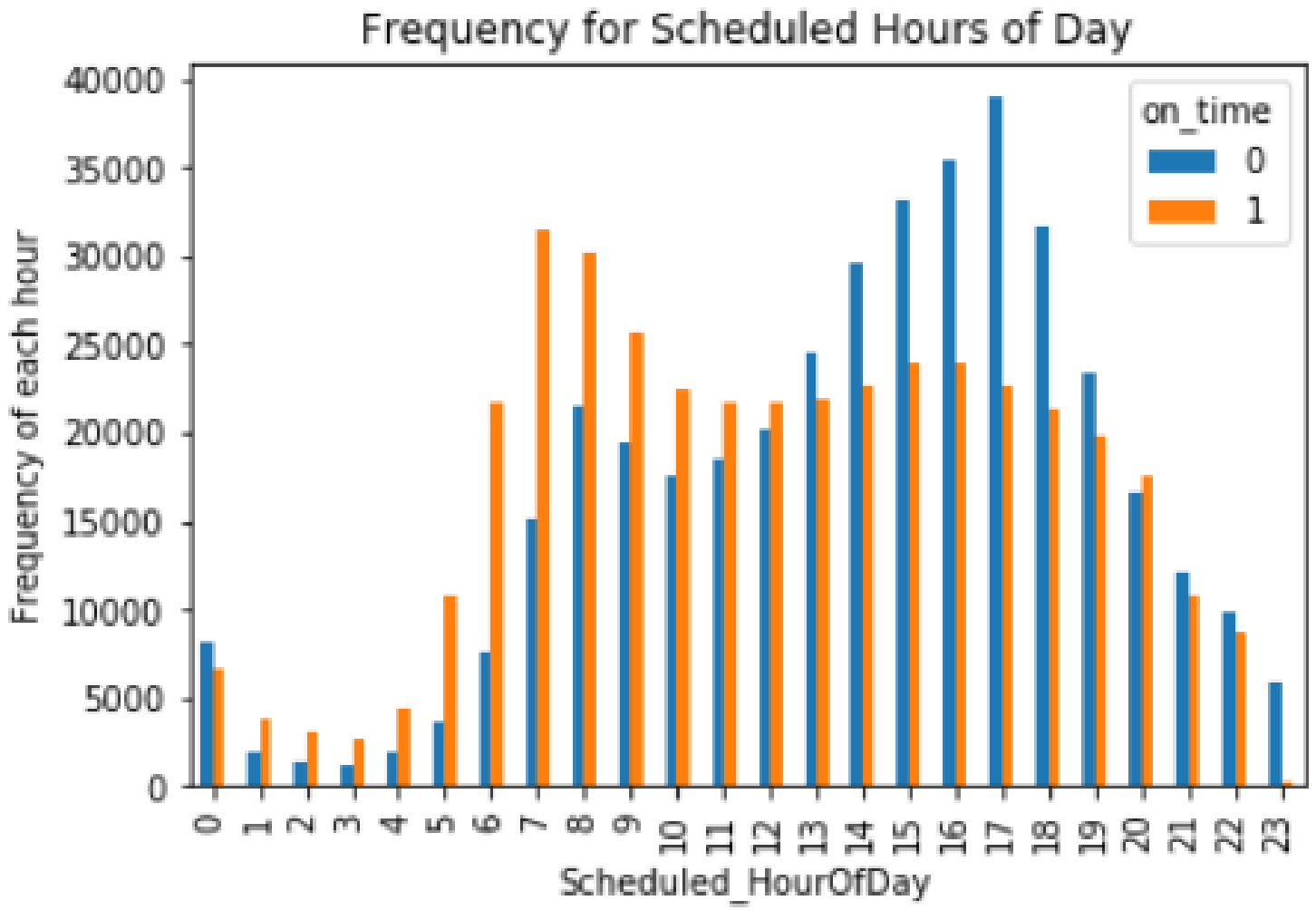
Scheduled Hours Of Day





Average

- On time: 12.25
- Not on time: 13.94

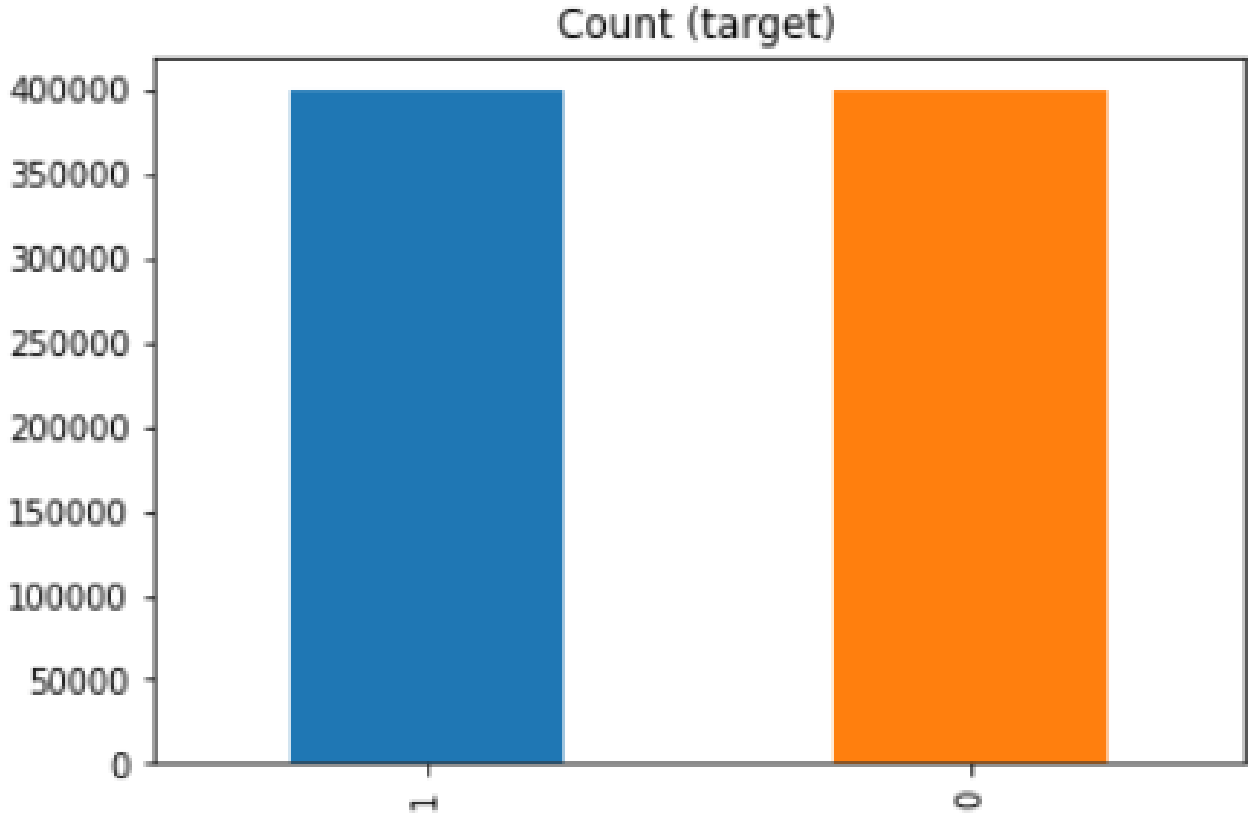


AVERAGE SCHEDULED HOURS OF DAY

- On time: 12.25
- Not on time: 13.94

BALANCED CLASSES

- Originally the data had a 60%/40% split between the classes, representing a slight imbalance.
- Performed Under sampling to have 50/50 split between the classes.

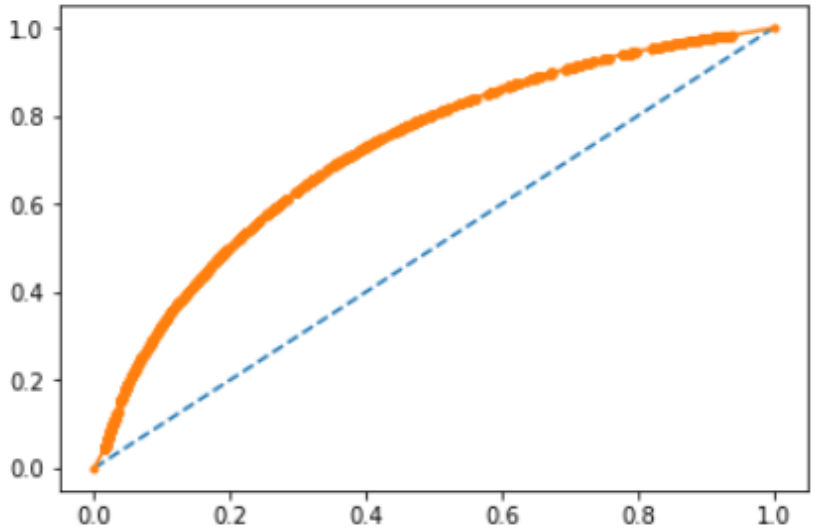


INITIAL DECISION TREE

- Created using full features
- Max depth set at 16.
- Used Entropy
- Max features set at 10.
- Average run time of 5.5 seconds.
- Accuracy Score of 67%
- 10-fold cross validation score of 67%

```
[0.66996067 0.67014854 0.66723447 0.66703407 0.66782315 0.66847445
 0.67044088 0.6694514  0.66684619 0.66822395]
Average cross validated score from our decision tree model is: 0.6685637772815237
```

ROC CURVE



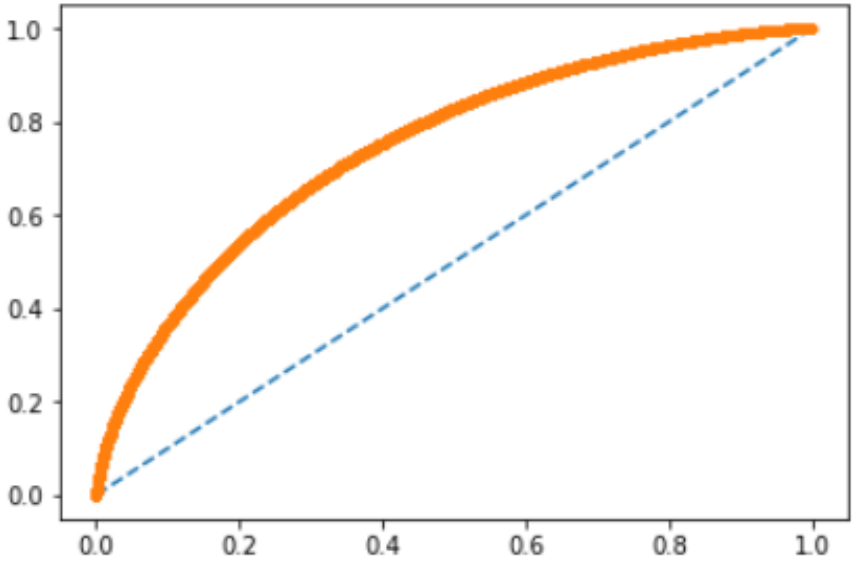
AUC Score of 72%

RANDOM FOREST

- 100 trees created using full features.
- Max depth set at 16.
- Using “gini”.
- Max features set at 6
- Average Run time of 3 minutes and 30 seconds
- Accuracy score of 68%
- 10-fold cross validation score of 69%

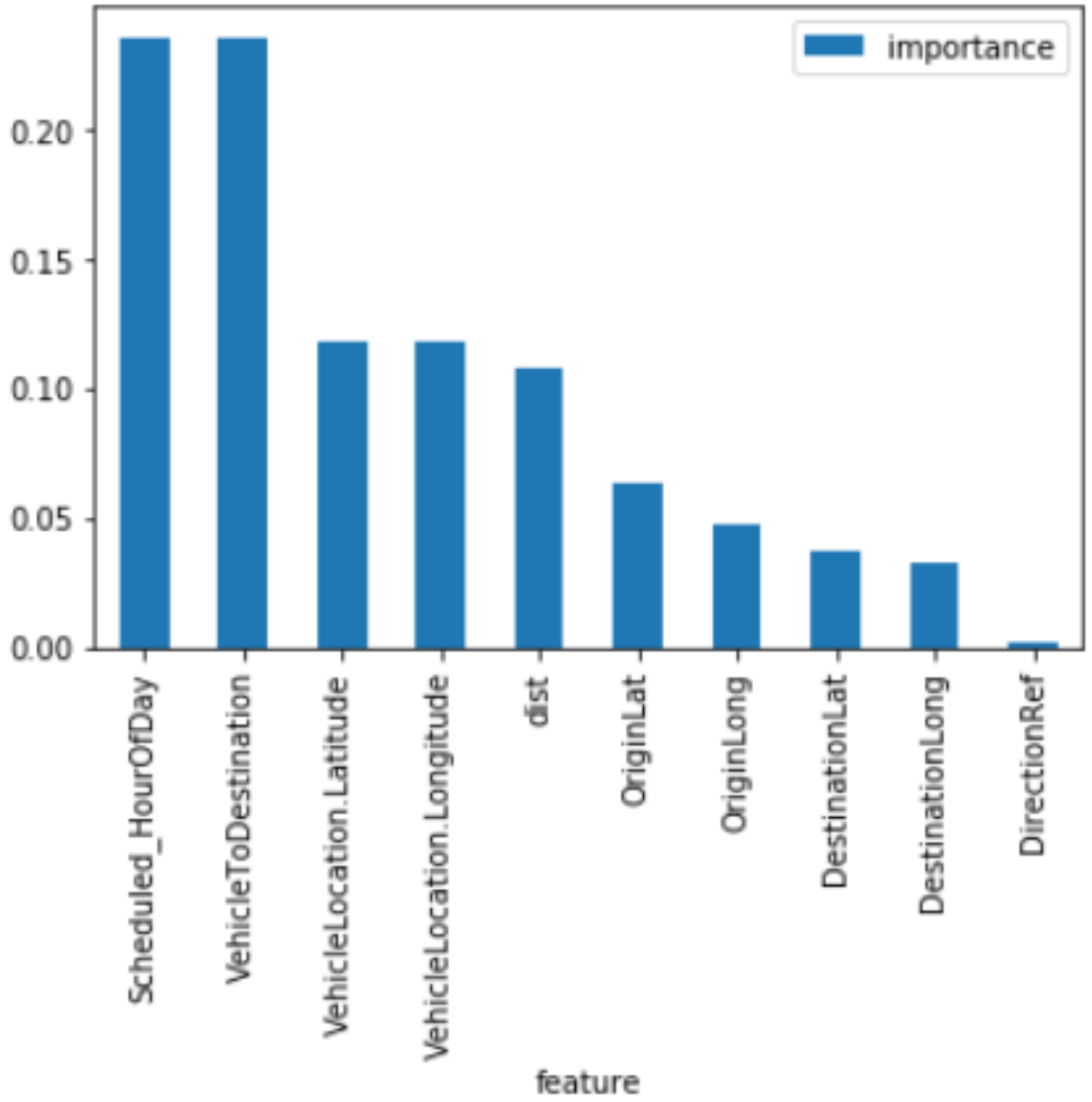
```
[0.6843015  0.68445179 0.68255261 0.68207665 0.68197645 0.68434369
 0.68439379 0.68311623 0.68106212 0.68747495]
Average cross validated score from our Random Forest Model is: 0.6835749780729266
```

ROC CURVE



AUC Score of 75%

FEATURE IMPORTANCE

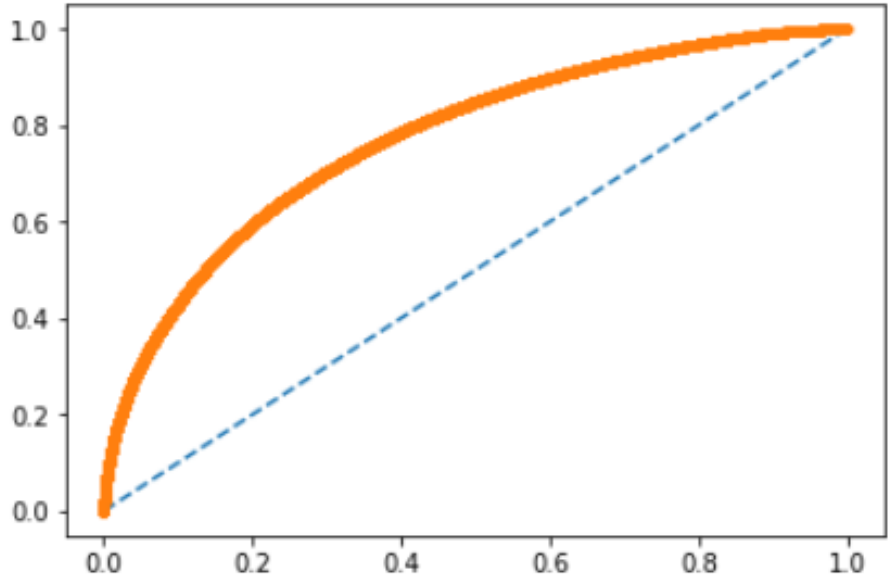


RANDOM FOREST USING PCA

- 200 trees created using full features.
- Seven components explained 95% of variance.
- Max depth set at 16.
- Using “gini”.
- Max features set at 4
- Average Run time of 24 minutes and 12 seconds
- Accuracy score of 68%
- 10-fold cross validation score of 68%

```
[0.68258561 0.67904111 0.67917084 0.67871994 0.68151303 0.68191383
 0.68232715 0.68041082 0.67954659 0.68280311]
Average cross validated score from our Random Forest Model is: 0.680803202337894
```

ROC CURVE

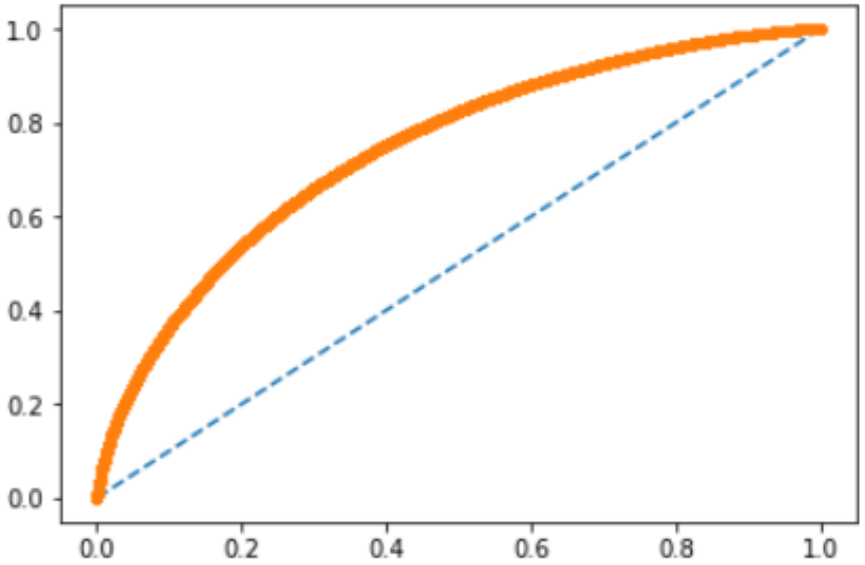


AUC Score of 77%

GRADIENT BOOSTING WITH PCA

- 200 trees created using full features.
- Max depth set at 16.
- Learning rate set at .02
- Subsample parameter set at .7
- Average Run time of 2 hrs and 57 minutes
- Accuracy score of 68%

ROC CURVE



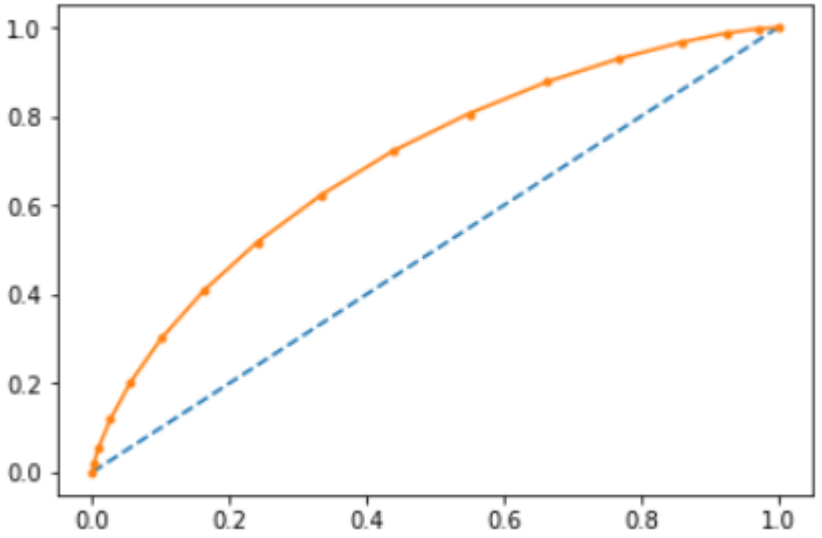
AUC Score of 75%

K NEAREST NEIGHBORS

- Initialized using 15 neighbors
- “uniform” weight setting.
- Average run time of 11.4 seconds
- Accuracy score of 65%
- 5-fold cross validation score of 65%

```
[0.64595884 0.64572713 0.64475827 0.64651177 0.64449524]
Average cross validated score from our KNN model is: 0.6454902512876356
```

ROC CURVE



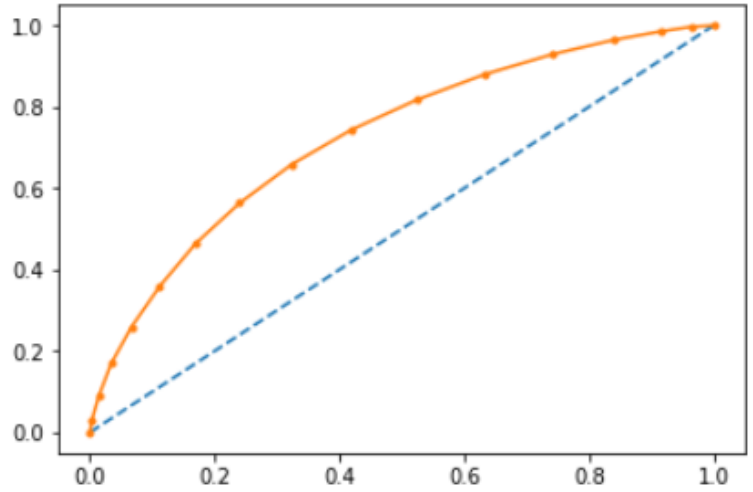
AUC Score of 70%

K NEAREST NEIGHBORS (KNN) WITH PCA

- Initialized using 15 neighbors
- “uniform” weight setting.
- Average run time of 5.9 seconds
- Accuracy score of 67%
- 5-fold cross validation score of 67%

```
[0.66665623 0.66637442 0.66568763 0.66816759 0.66534319]
Average cross validated score from our KNN model is: 0.6664458090414618
```

ROC CURVE



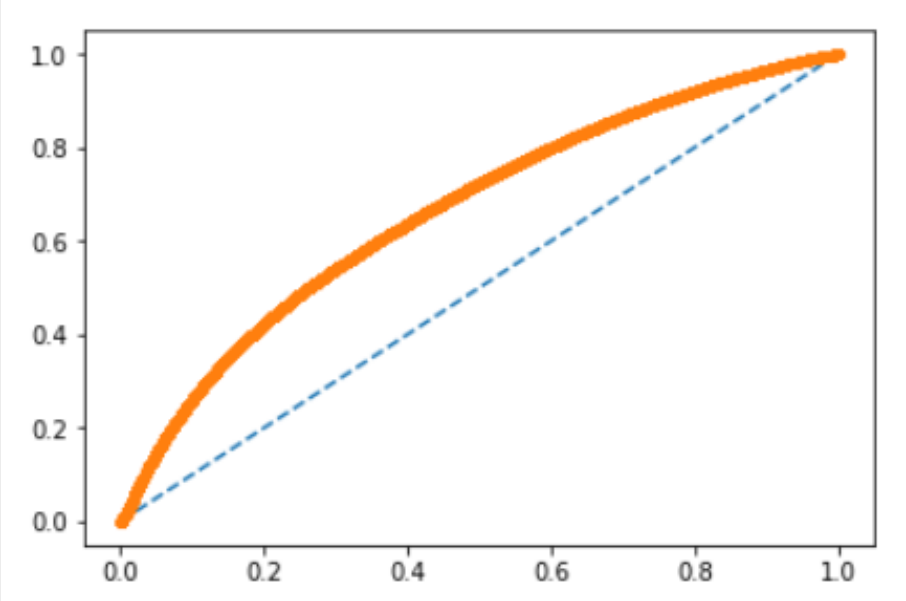
AUC Score of 73%

LOGISTIC REGRESSION

- C=1e9
- Accuracy score of 61%
- Runtime of 21.3 seconds
- CV score of 62%
- AUC Score of 66%
- Accuracy score of 61% with PCA.

```
[0.61056837 0.61363693 0.61419088 0.61162325 0.60996994 0.61272545
 0.60971944 0.61172345 0.61262525 0.6134519 ]
Average cross validated score from our log regression model is: 0.6120234864596557
```

ROC CURVE



AUC SCORE OF 66%

CLASSIFICATION REPORT

DECISION TREE

- AUC score of 72%.
- Recall rating of 67% for both classes.
- Precision rating of 66% for class 0 and 67% for class 1.
- F1 score of 67% for both classes.

```
Confusion Matrix:
 [[66166 33095]
 [33335 67005]]
```

RANDOM FOREST

- AUC score of 75%.
- Recall rating of 68% for both classes.
- Precision rating of 68% for both classes.
- F1 score of 68% for both classes.

```
Confusion Matrix:
 [[68309 31850]
 [31773 67669]]
```

RANDOM FOREST PCA

- AUC score of 77%.
- Recall rating of 50% for class 0 and 49% for class 1.
- Precision rating of 50% for both classes.
- F1 score of 50% for both classes.

```
Confusion Matrix:
 [[50330 49610]
 [50533 49128]]
```

GRADIENT BOOSTING PCA

- AUC score of 75%.
- Recall rating of 69% for class 0 and 67% for class 1.
- Precision rating of 68% for both classes.
- F1 score of 68% for both classes.

```
Confusion Matrix:
[[68974 31193]
 [32618 66816]]
```

CLASSIFICATION REPORT

KNN

- AUC score of 70%.
- Recall rating of 67% for class 0 and 62% for class 1.
- Precision rating of 64% for class 0 and 65% for class 1.
- F1 score of 65% for class 0 and 64% for class 1.

```
Confusion Matrix:
 [[66095 33166]
 [37760 62580]]
```

KNN WITH PCA

- AUC score of 73%.
- Recall rating of 68% for class 0 and 66% for class 1.
- Precision rating of 67% for both classes.
- F1 score of 67% for class 0 and 66% for class 1.

```
Confusion Matrix:
 [[67682 32485]
 [33883 65551]]
```

Logistic Regression

- AUC score of 66%.
- Recall rating of 59% for class 0 and 63% for class 1.
- Precision rating of 62% for class 0 and 61% for class 1.
- F1 score of 60% for class 0 and 62% for class 1.

```
Confusion Matrix:
 [[59251 40540]
 [37087 62723]]
```

CONCLUSION

Random Forest is the best option.

Additional insights still needed.