NY Taxi Trip Duration

Submitted to

Assoc. Prof. Dr.Tiranee Achalakul
Asst Prof Dr.Santitham Prom-on

INTRODUCTION

Dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. Based on individual trip attributes, participants should predict the duration of each trip in the test set.

WHAT WE PREDICT

trip duration of New York city taxi based on given features.

DATASET

CONSISTS OF

- Train.csv which contains 1458644 trip records and 11 attributes.
- Test Data which contains 625134 trip records and 9 attributes

DATASET

TRAIN.CSV

(the training set contains 1458644 trip records and 11 attributes)

Attributes	Type	Decription
id	Character	a unique identifier for each trip
vendor_id	Integer	a code indicating the provider associated with the trip record
pickup_datetime	Character	date and time when the meter was engaged
dropoff_datetime	Character	date and time when the meter was disengaged
passenger_count	Integer	the number of passengers in the vehicle (driver entered value)
pickup_longtitude	Double	the longitude where the meter was engaged
pickup_latitude	Double	the latitude where the meter was engaged
dropoff_longitude	Double	the longitude where the meter was disengaged
dropoff_latitude	Double	the latitude where the meter was disengaged
store_and_fwd_flag	Character	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
trip_duration	Integer	duration of the trip in seconds

DATASET

TEST.CSV

(the testing set contains 625134 trip records and 9 attributes)

Attributes	Type	Decription
id	Character	a unique identifier for each trip
vendor_id	Integer	a code indicating the provider associated with the trip record
pickup_datetime	Character	date and time when the meter was engaged
passenger_count	Integer	the number of passengers in the vehicle (driver entered value)
pickup_longtitude	Double	the longitude where the meter was engaged
pickup_latitude	Double	the latitude where the meter was engaged
dropoff_longitude	Double	the longitude where the meter was disengaged
dropoff_latitude	Double	the latitude where the meter was disengaged
store_and_fwd_flag	Character	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip

METHOD

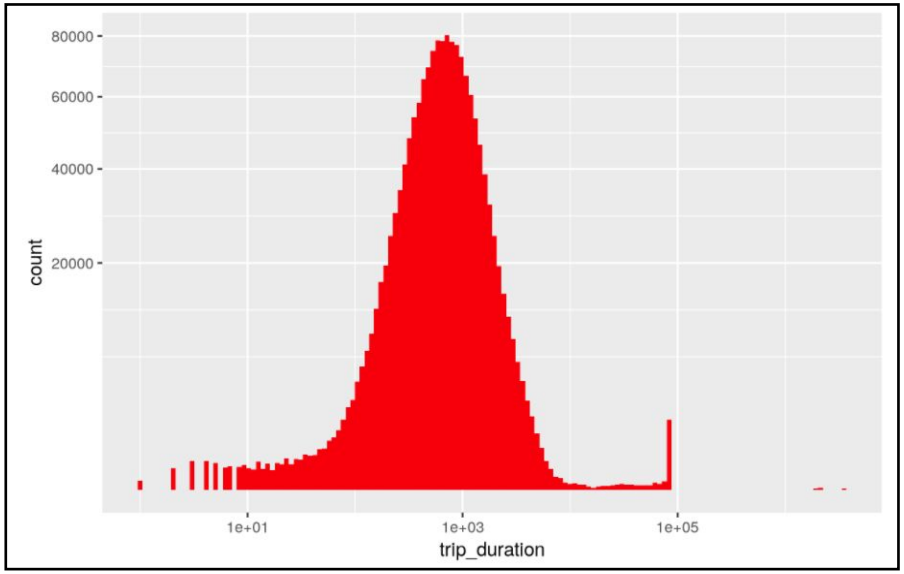
1.EDA Exploration Data Analysis

- Individual visualization
- Feature relations

EDA : Individual visualization

(relationship between the number of trips and each attribute)

A. Trip Duration

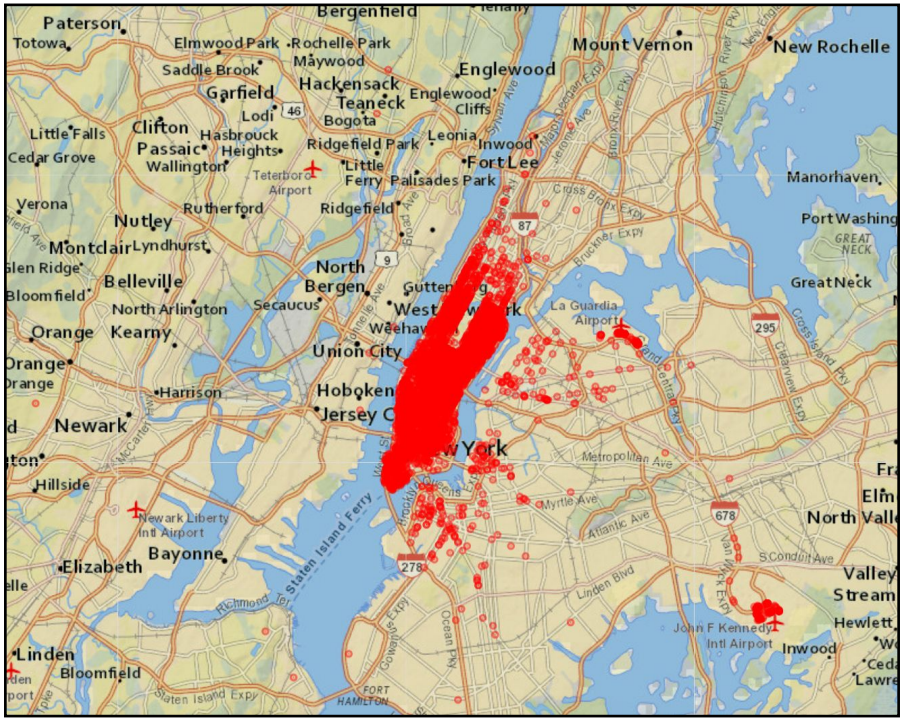



```
# A tibble: 10 x 11
   trip_duration     pickup_datetime    dropoff_datetime        id vendor_id passenger_count pickup_longitude
           <int>              <dttm>              <dttm>     <chr>     <fctr>          <fctr>            <dbl>
 1       3526282 2016-02-13 22:46:52 2016-03-25 18:18:14 id0053347         1               1       -73.78391
 2       2227612 2016-01-05 06:14:15 2016-01-31 01:01:07 id1325766         1               1       -73.98379
 3       2049578 2016-02-13 22:38:00 2016-03-08 15:57:38 id0369307         1               2       -73.92168
 4       1939736 2016-01-05 00:19:42 2016-01-27 11:08:38 id1864733         1               1       -73.78965
 5         86392 2016-02-15 23:18:06 2016-02-16 23:17:58 id1942836         2               2       -73.79453
 6         86391 2016-05-31 13:00:39 2016-06-01 13:00:30 id0593332         2               1       -73.78195
 7         86390 2016-05-06 00:00:10 2016-05-07 00:00:00 id0953667         2               1       -73.99601
 8         86387 2016-06-30 16:37:52 2016-07-01 16:37:39 id2837671         2               1       -73.99228
 9         86385 2016-06-23 16:01:45 2016-06-24 16:01:30 id1358458         2               1       -73.78209
10         86379 2016-05-17 22:22:56 2016-05-18 22:22:35 id2589925         2               4       -74.00611
# ... with 4 more variables: pickup_latitude <dbl>, dropoff_longitude <dbl>, dropoff_latitude <dbl>,
#   store_and_fwd_flag <chr>
```

EDA : Individual visualization

(relationship between the number of trips and each attribute)

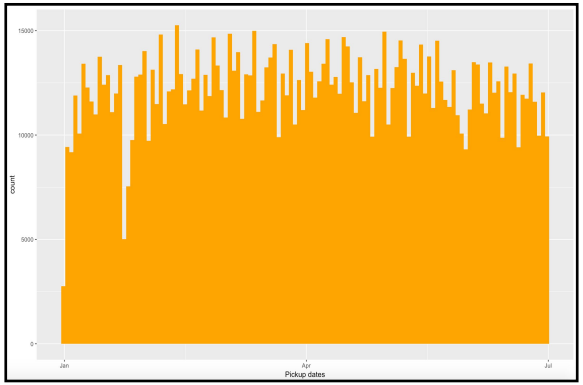
B. pickup_longtitude, pickup_latitude, dropoff_longtitude, dropoff_latitude



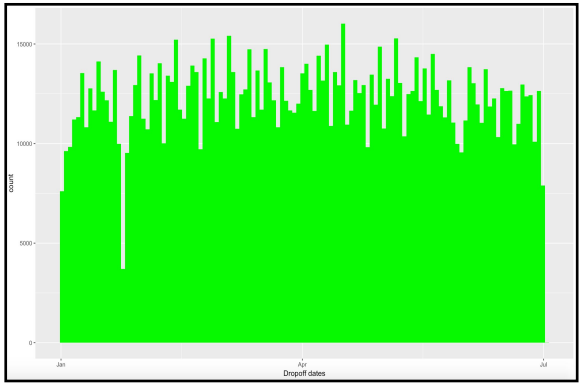
EDA : Individual visualization

(relationship between the number of trips and each attribute)

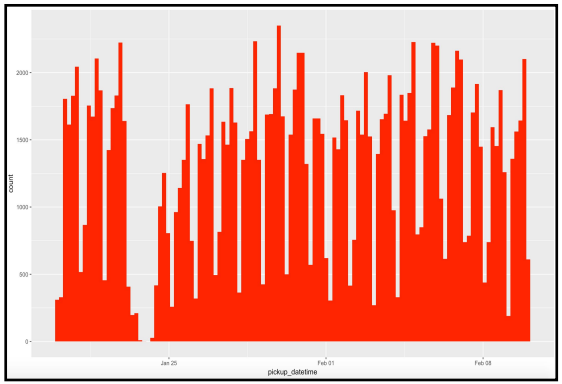
C. pickup and dropoff month



graph between the number of trips
& pickup date from January to July 2016



graph between the number of trips
& dropoff date from January to July 2016

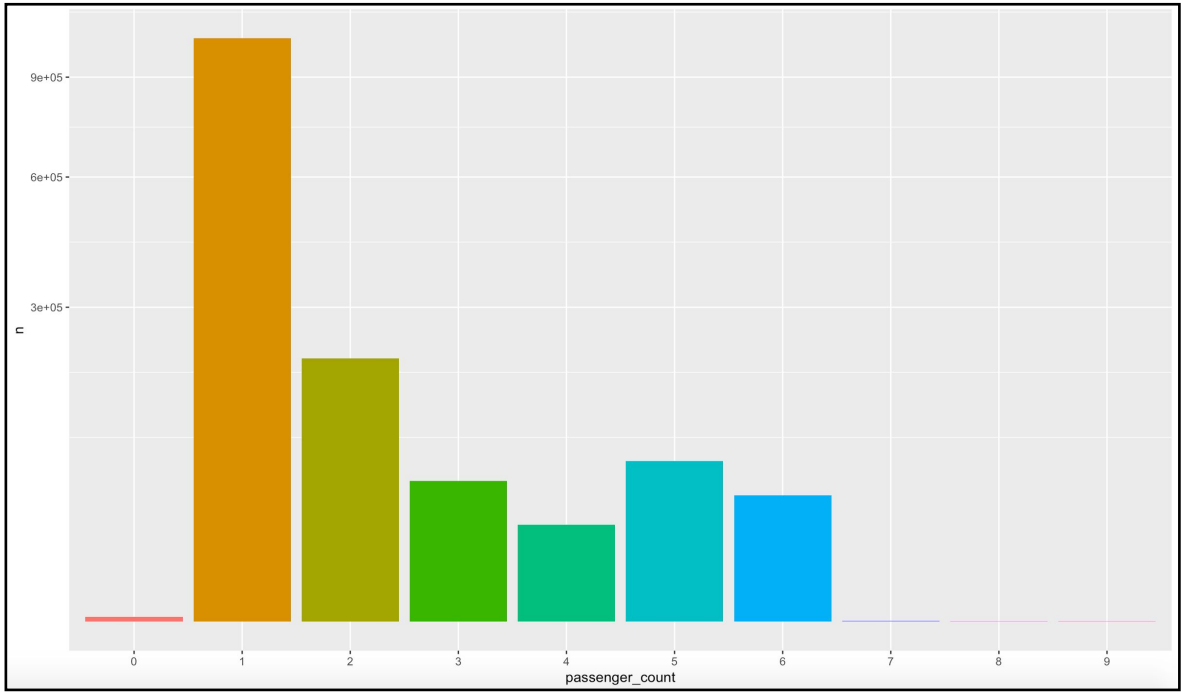


graph between the number of trips
& dropoff date from January to February 2016

EDA : Individual visualization

(relationship between the number of trips and each attribute)

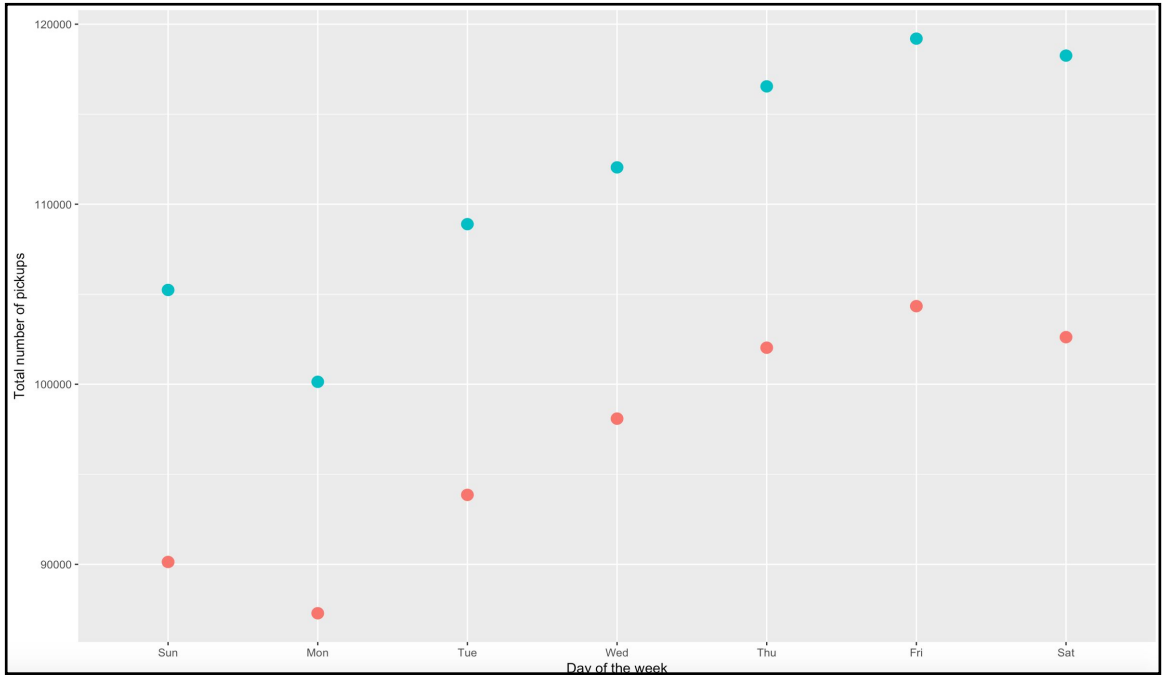
D. passenger counts



EDA : Individual visualization

(relationship between the number of trips and each attribute)

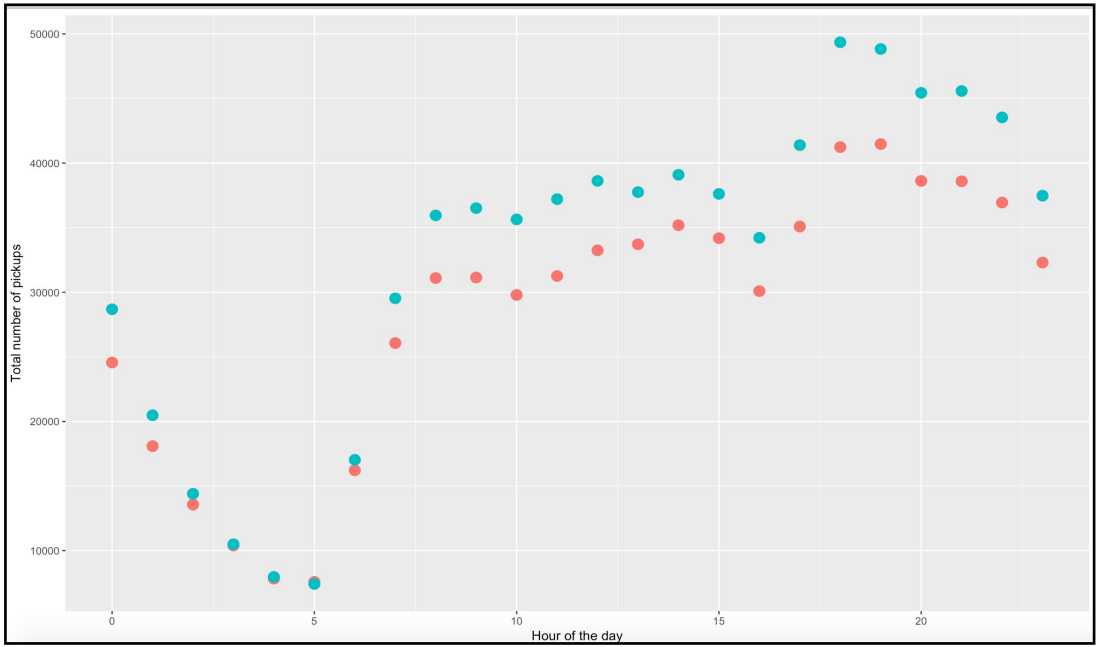
E. Weekday between two vendors



EDA : Individual visualization

(relationship between the number of trips and each attribute)

F. Hour of the day between two vendors



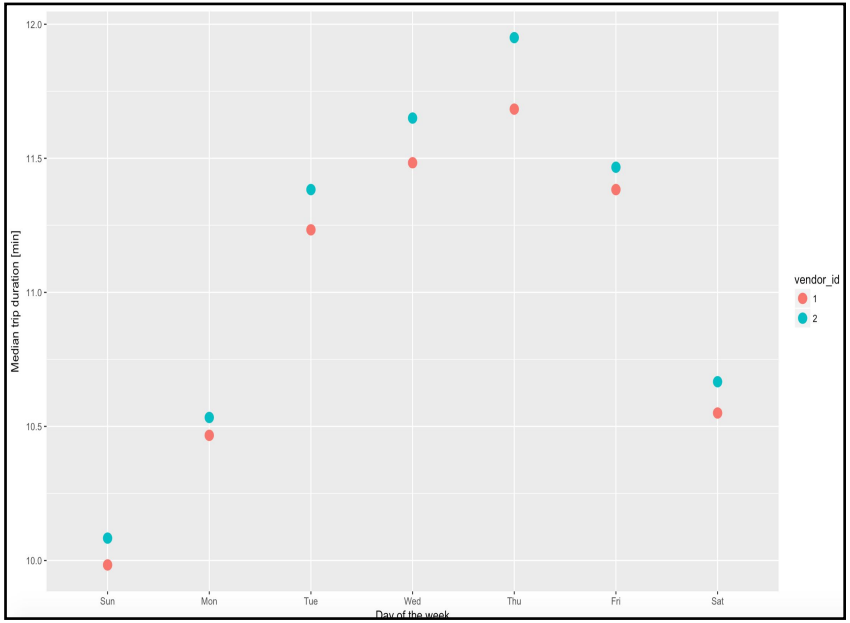
METHOD

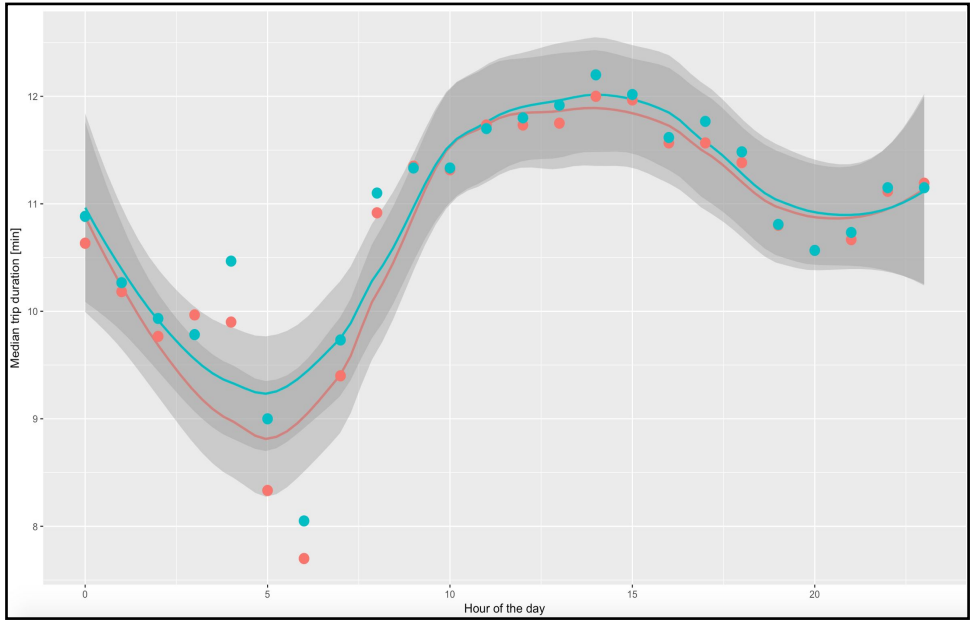
2.FEATURE RELATION

EDA : Feature relations

(relationship between the target attributes;trip_duration and each attribute)

A. pickup weekday and hour of the day

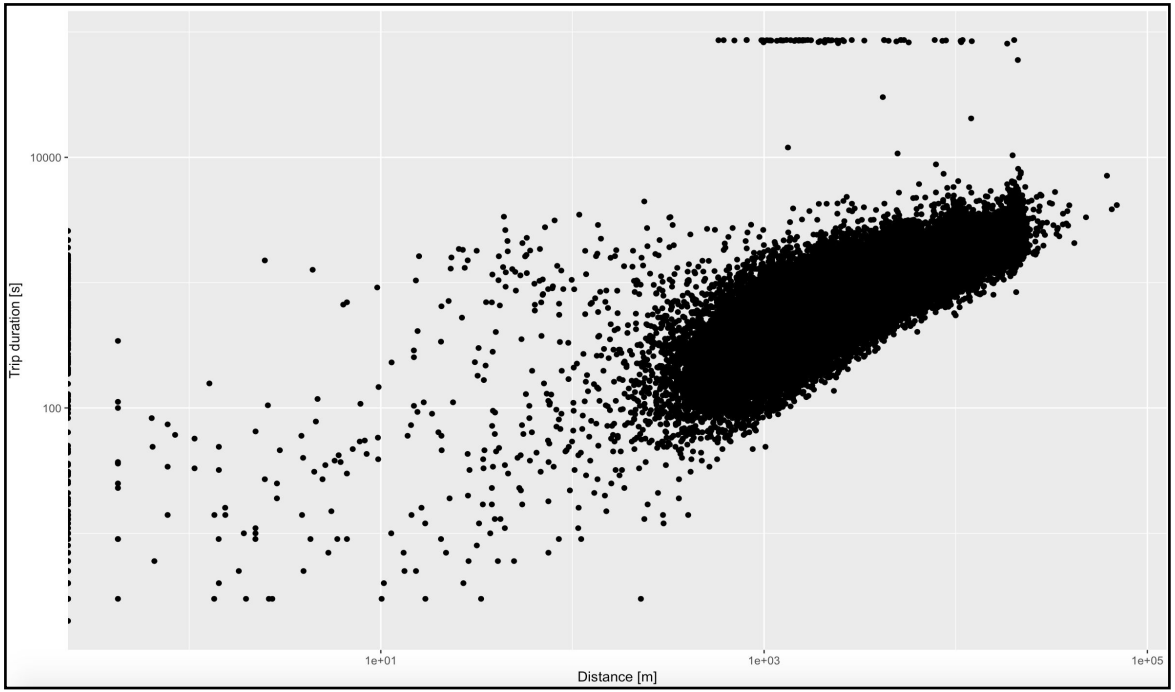




EDA : Feature relations

(relationship between the target attributes;trip_duration and each attribute)

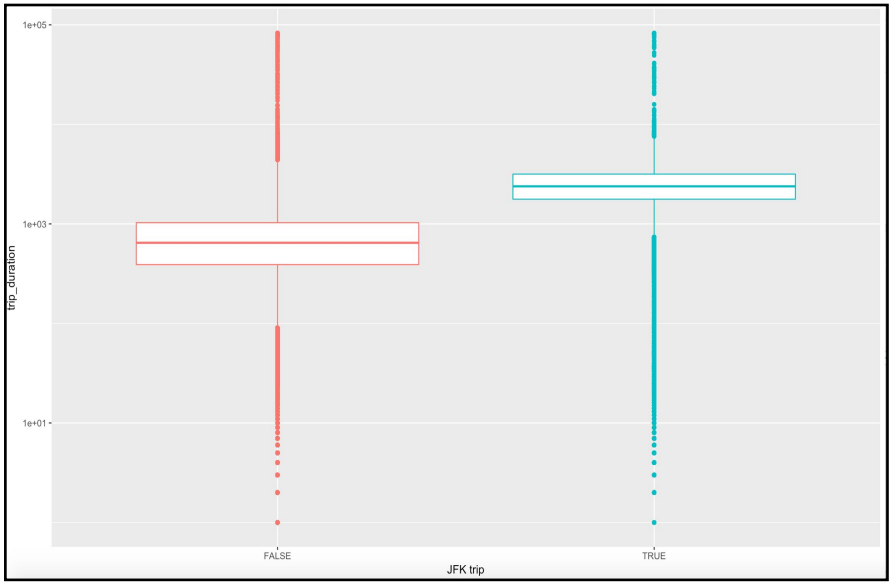
A. Direct distance of the trip

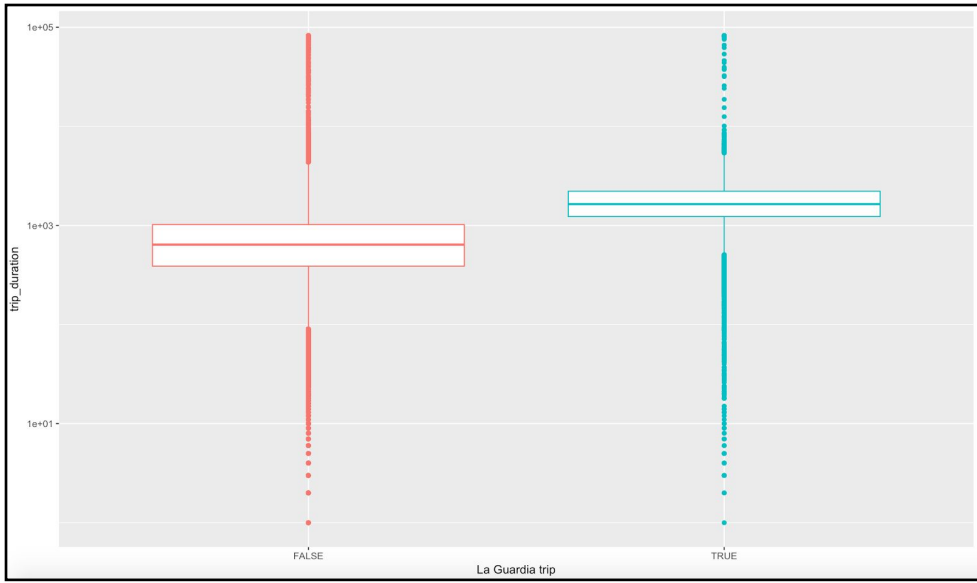


EDA : Feature relations

(relationship between the target attributes;trip_duration and each attribute)

C. Airport distance





METHOD: 3.FEATURE ENGINEERING

No	Attributes	Decription
12.	distance<dbl>	The shortest distance computed from the direct distance between the two points, and compare with our trip_durations.
13.	bearing<dbl>	It is the direction that tells about tstart out for instance in the direction of North-West or South-East.
14.	jfk_dist_pick<dbl>	A number of trips began or ended at either of the two NYC airports. The pick up point is JFK airport.
15.	jfk_dist_drop<dbl>	A number of trips began or ended at either of the two NYC airports. The destination drop off point is JFK airport.
16.	lg_dist_pick<dbl>	A number of trips began or ended at either of the two NYC airports. The pick up point is La Guardia airport
17.	lg_dist_drop<dbl>	A number of trips began or ended at either of the two NYC airports. The destination drop off point is LGA airport.
18.	date<date>	The date time data that will display on Month/Day/Year
19.	work<lgl>	Delays rate during work hours, its result will display on True or False.
20.	jfk_trip<lgl>	It indicating journeys to JFK airport or not created by define a pickup or dropoff distance of less than 2 km from the JFK airport
21.	lg_trip<lgl>	It indicating journeys to La Guardia airport or not created by define a pickup or dropoff distance of less than 2 km from the LGA airport

METHOD: 3.FEATURE ENGINEERING

22.	pickup_weekday<int>	The number of pickup day in one week which covers a peak in raw trip_duration distribution.
23.	pickup_month<int>	The number of pickup month on the highest pickup day which covers with pickup_wday (which day in the week).
24.	pickup_hour<int>	The number of pickup hour on the highest pickup day which covers with pickup_wday (which day in the week).
25.	night_trip<lgl>	The feature checks which trip is on night trip or not. (Result will display on True or False)
26.	rush_hour<lgl>	The feature checks which trip is on rush hour or not. (Result will display on True or False)
27.	weekday<lgl>	The feature checks which trip is on weekend day or not. (Result will display on True or False)

METHOD:

2.DATA CLEANING

DATA CLEANING

Formatting Criteria:

```
train <- train %>%
  filter(trip_duration < 22*3600,
         dist > 0 | (near(dist, 0) & trip_duration < 60),
         jfk_dist_pick < 3e5 & jfk_dist_drop < 3e5,
         trip_duration > 10,
         speed < 100)
```

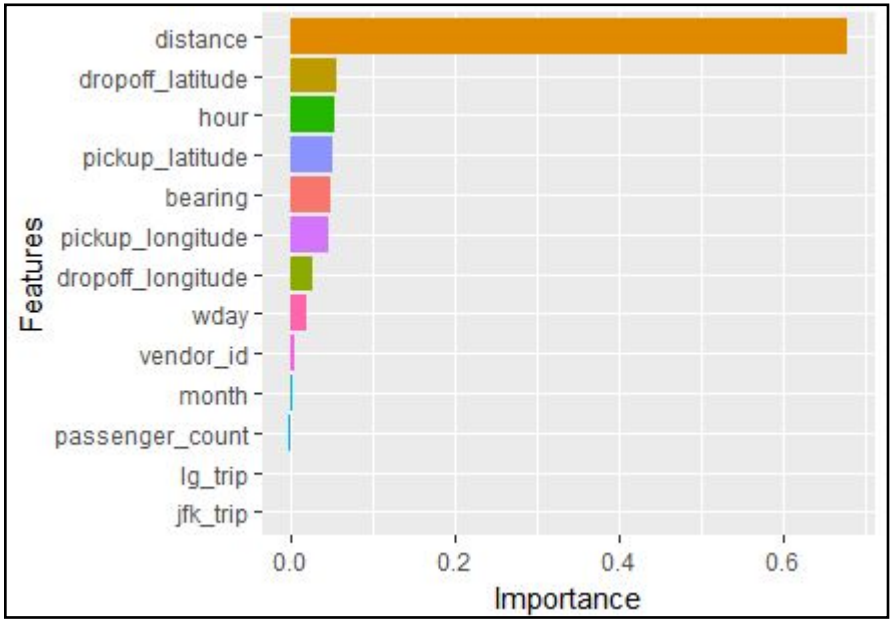
- trip duration longer than 22 hours and less than 10 seconds
- trip distance equal zero which trip duration is more than a minutes
- long distance location which pickup or dropoff locations more than 300 km away from NYC (JFK airport)
- travel speed more than 100 km/hr

METHOD:

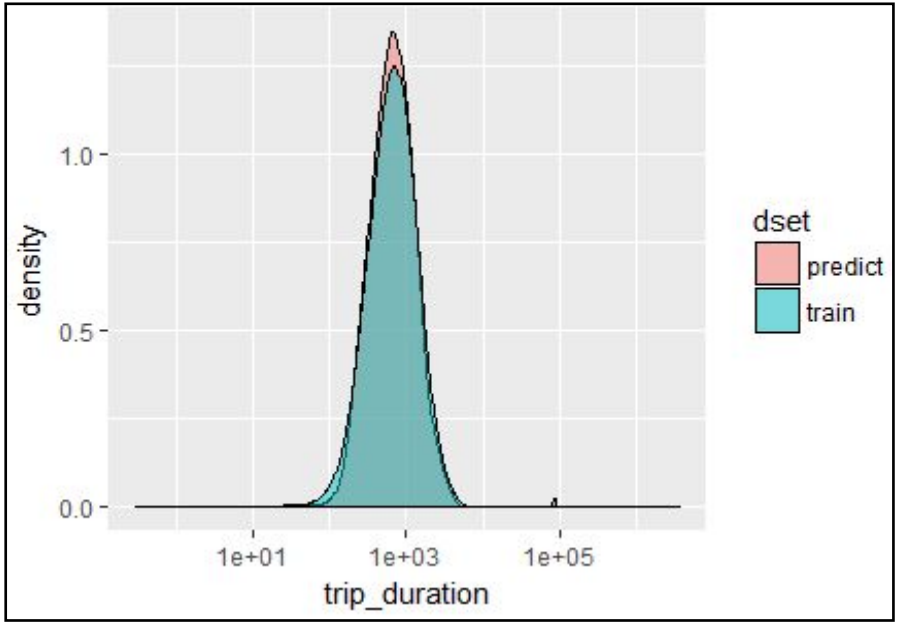
TRAINING MODEL (RESULT)

MODEL1: XGBoost

Feature importance



Prediction result



MODEL1: XGBoost

Result

RMSLE FROM TRAINING

```
[251]   train-rmse:0.386086     valid-rmse:0.385573
[256]   train-rmse:0.385781     valid-rmse:0.385295
[261]   train-rmse:0.385537     valid-rmse:0.385029
[266]   train-rmse:0.385273     valid-rmse:0.384692
[271]   train-rmse:0.384980     valid-rmse:0.384401
[276]   train-rmse:0.384662     valid-rmse:0.384020
[281]   train-rmse:0.384428     valid-rmse:0.383766
[286]   train-rmse:0.384052     valid-rmse:0.383398
[291]   train-rmse:0.383666     valid-rmse:0.383092
[296]   train-rmse:0.383397     valid-rmse:0.382835
[300]   train-rmse:0.383108     valid-rmse:0.382591
```

RMSLE FROM CROSS VALIDATION

```
[990]   train-rmse:0.375571+0.000594    test-rmse:0.397627+0.002093
[991]   train-rmse:0.375555+0.000596    test-rmse:0.397627+0.002091
[992]   train-rmse:0.375535+0.000601    test-rmse:0.397634+0.002086
[993]   train-rmse:0.375519+0.000594    test-rmse:0.397629+0.002086
[994]   train-rmse:0.375502+0.000591    test-rmse:0.397638+0.002082
[995]   train-rmse:0.375483+0.000589    test-rmse:0.397636+0.002082
[996]   train-rmse:0.375462+0.000594    test-rmse:0.397623+0.002070
[997]   train-rmse:0.375445+0.000595    test-rmse:0.397617+0.002072
[998]   train-rmse:0.375422+0.000592    test-rmse:0.397604+0.002074
[999]   train-rmse:0.375401+0.000594    test-rmse:0.397602+0.002071
[1000]  train-rmse:0.375382+0.000591    test-rmse:0.397606+0.002079
```

KAGGLE SCORE - RMSLE FROM TEST SET

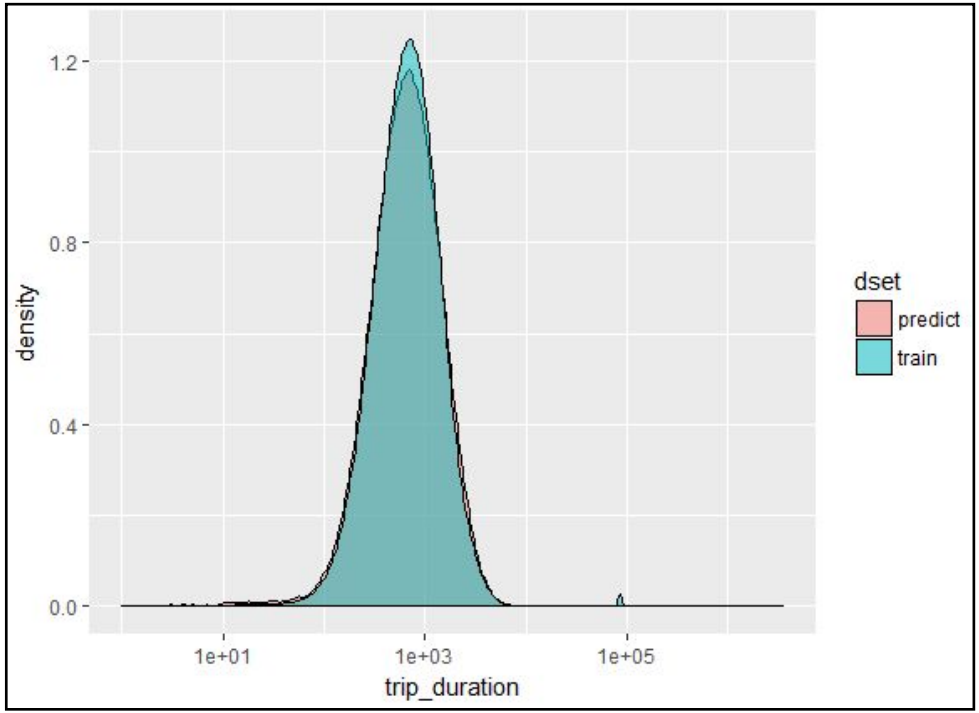
Submission and Description	Private Score	Public Score
submit2.csv a few seconds ago by May Junejuly xgboost with eta 0.1	0.39993	0.40219

Model2: Decision tree

Feature importance

```
Feature ranking Decision tree:
1. feature distance (0.679895)
2. feature bearing (0.063712)
3. feature dropoff_latitude (0.051363)
4. feature pickup_hour (0.051241)
5. feature dropoff_longitude (0.034659)
6. feature pickup_longitude (0.033078)
7. feature pickup_latitude (0.030436)
8. feature pickup_weekday (0.016866)
9. feature pickup_month (0.013026)
10. feature weekday (0.008122)
11. feature passenger_count (0.006291)
12. feature rush_hour (0.004465)
13. feature vendor_id (0.003336)
14. feature work (0.001758)
15. feature night_trip (0.001347)
16. feature lg_trip (0.000317)
17. feature jfk_trip (0.000088)
```

Prediction result



Model3: Decision tree

RMSLE FROM TRAINING

```
With model: DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
Train RMSLE: 0.000276448745885
Val. RMSLE: 0.482543207943
```

KAGGLE SCORE - RMSLE FROM TEST SET

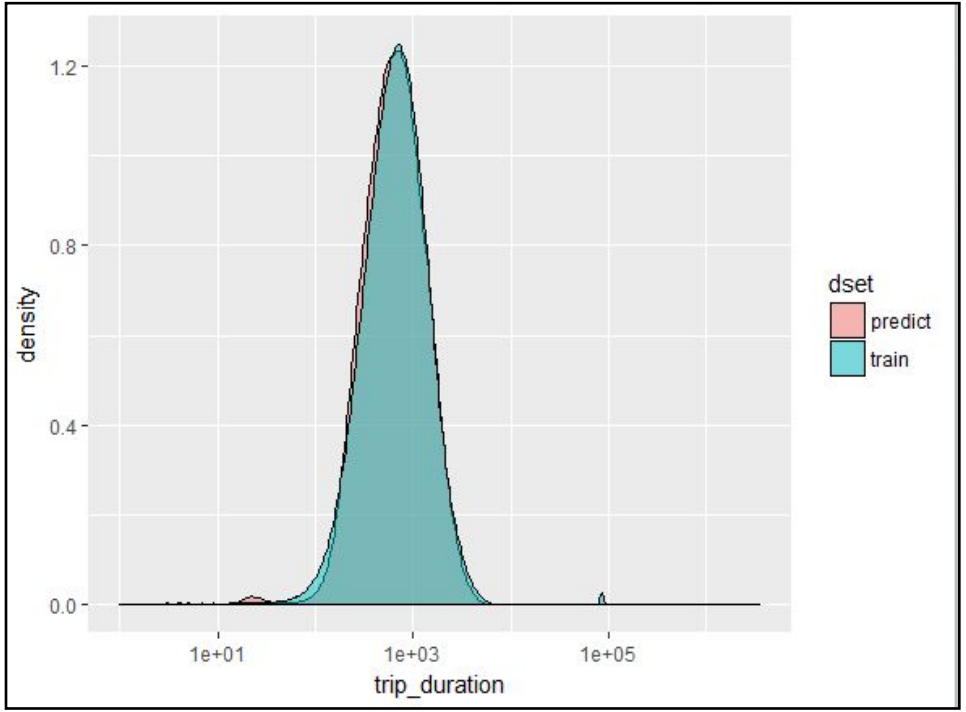
submission-dt.csv a day ago by May Junejuly using decision tree	0.63833	0.63711

Model3: Random Forest

Feature importance

```
Feature ranking Random Forest:
1. feature distance (0.680009)
2. feature bearing (0.064019)
3. feature dropoff_latitude (0.051785)
4. feature pickup_hour (0.049699)
5. feature dropoff_longitude (0.034452)
6. feature pickup_longitude (0.033023)
7. feature pickup_latitude (0.030092)
8. feature pickup_weekday (0.017583)
9. feature pickup_month (0.013074)
10. feature weekday (0.007332)
11. feature passenger_count (0.006287)
12. feature rush_hour (0.005450)
13. feature vendor_id (0.003300)
14. feature work (0.001923)
15. feature night_trip (0.001576)
16. feature lg_trip (0.000282)
17. feature jfk_trip (0.000114)
```

Prediction result



Model3: Random Forest

RMSLE FROM TRAINING

```
With model: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=50, n_jobs=1,
           oob_score=False, random_state=None, verbose=0, warm_start=False)
Train RMSLE: 0.128683034987
Val. RMSLE: 0.340619719201
```

KAGGLE SCORE - RMSLE FROM TEST SET

Submission and Description	Private Score	Public Score
submission-rf.csv a few seconds ago by May Junejuly Using random forest with n_estimators=50	0.52711	0.52752

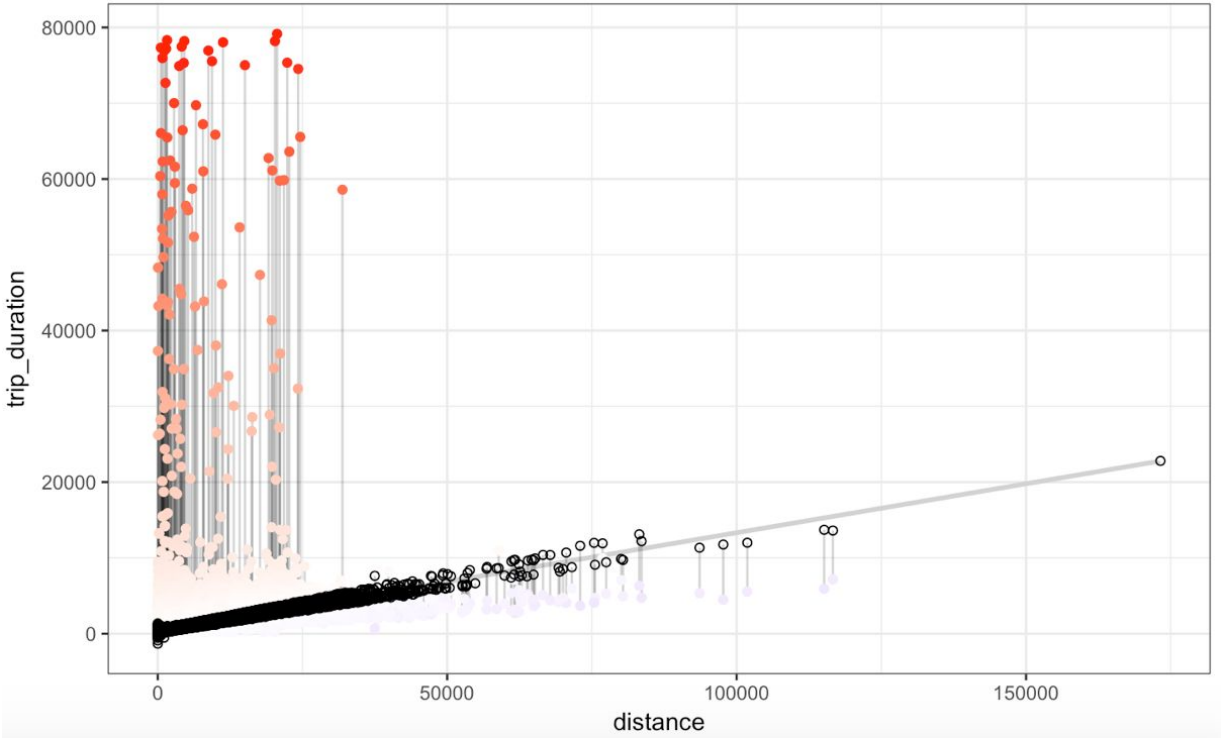
Model4: Mulriple Linear Regression

LINEAR REGRESSION EQUATION

Trip_duration = 3.338e+04+(1.291e-01*distance)+(7.700e+02*pickup_latitude)
-(1.581e+03*dropoff_latitude)+(4.053+00*pickup_hour)

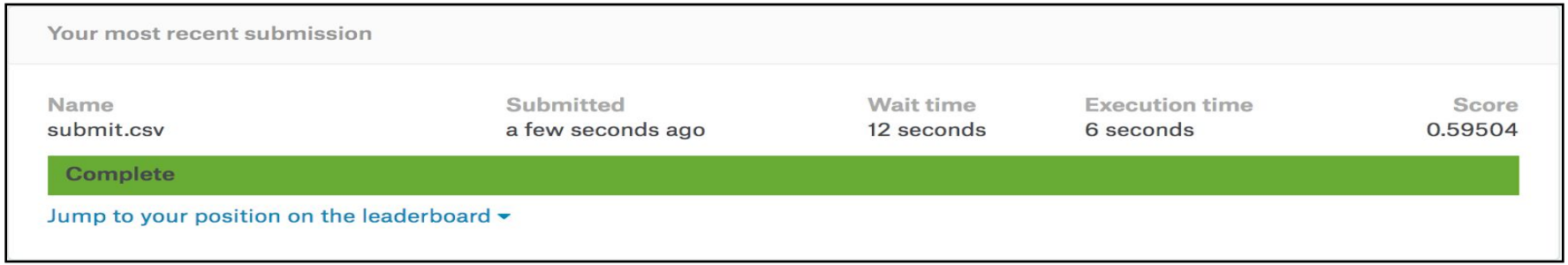
```
Residual standard error: 660.7 on 1048570 degrees of freedom
Multiple R-squared:  0.3772,    Adjusted R-squared:  0.3772
F-statistic: 1.588e+05 on 4 and 1048570 DF,  p-value: < 2.2e-16
```

Model4: Multiple Linear Regression



Model4: Multiple Linear Regression

KAGGLE SCORE - RMSLE FROM TEST SET



CONCLUSION:

CONCLUSION

	XGBoost	Decision tree	Random Forest	Linear Regression
Kaggle Score (RMLSE)	0.39993	0.63833	0.53662	0.59504

XGBoost has the best performance to predict the trip duration

THANK YOU

FOR YOUR ATTENTION