THE TEAM

APOORVA JASTI

KELLY ZHANG

SATYA PACHIGOLLA

SEBASTIAN OSORIO

TUSHAR GUPTA

VICTOR NGUYEN

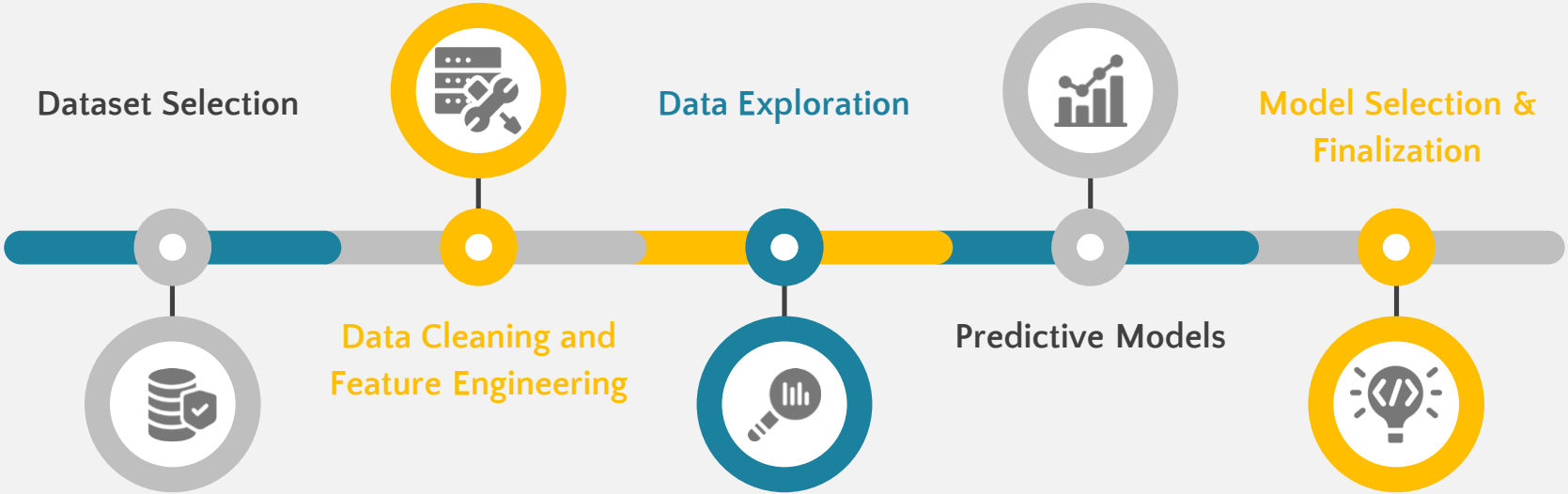
NYC YELLOW CABS

BUSINESS PROBLEM

Use of Machine Learning to improve the prediction of travel times in NYC by Yellow Cabs

Everyone wants to have an accurate estimate of time taken for a route. We hypothesize that techniques from machine learning, if carefully applied to such data, can improve the prediction of travel times

PROJECT FLOW



ORIGINAL PREDICTORS

Number of Trips
1,458,644

Number of Variables
12

Period of Data
Jan - June 2016

- Pick-up Datetime
- Pick-up Longitude
- Pick-up Latitude
- Drop-off Longitude
- Drop-off Latitude
- Drop-off Datetime
- Trip Duration
- Store and forward flag
- ID Vendor and ID
- Passenger Count

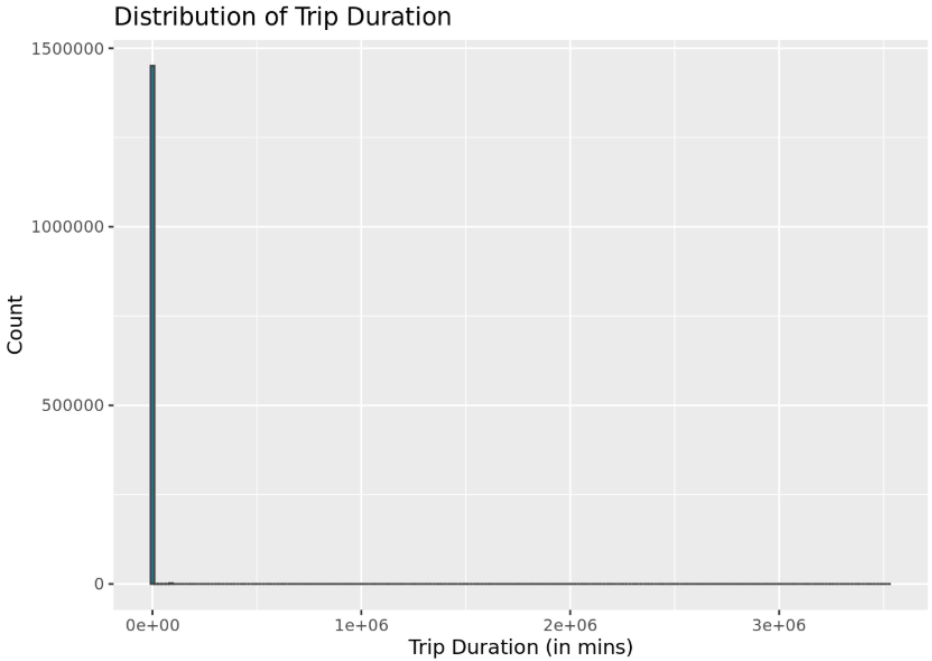
DATA SUMMARY

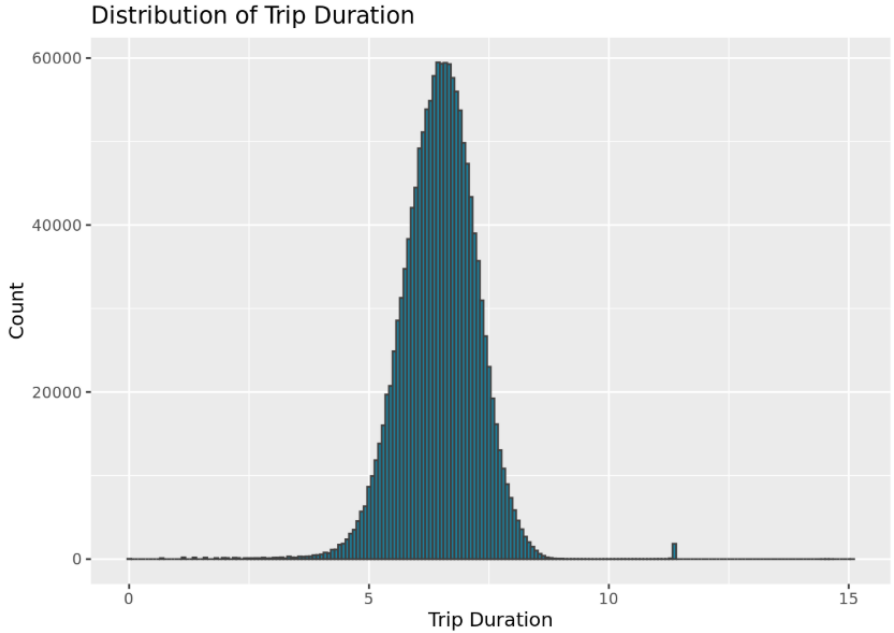
```
        id              vendor_id     passenger_count  trip_duration       day_of_week                time_of_day      temperature
id0000001:      1   Min.   :1.000   Min.   :0.000   Min.   :      1   Friday   :223533   afternoon/evening:358372   Min.   :-18.30
id0000003:      1   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:    397   Monday   :187418   earlyMorning      :155901   1st Qu.:  3.90
id0000005:      1   Median :2.000   Median :1.000   Median :    662   Saturday :220868   eveningRush       :264980   Median : 10.60
id0000008:      1   Mean   :1.535   Mean   :1.665   Mean   :    959   Sunday   :195366   lateNight         :171480   Mean   : 11.06
id0000009:      1   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:   1075   Thursday :218574   morningRush       :273449   3rd Qu.: 18.00
id0000011:      1   Max.   :2.000   Max.   :9.000   Max.   :3526282   Tuesday  :202749   night             :234462   Max.   : 32.20
(Other)  :1458638                                                     Wednesday:210136                             NA's   :12032
 pickup_month     pickup_hour      displacement        distance          dewpoint         visibility              conditions
Min.   :1.000   Min.   : 0.00   Min.   :   0.000   Min.   :   0.000   Min.   :-28.300   Min.   : 0.40   Clear        :730865
1st Qu.:2.000   1st Qu.: 9.00   1st Qu.:   1.232   1st Qu.:   1.679   1st Qu.: -6.100   1st Qu.:14.50   Overcast     :337086
Median :4.000   Median :14.00   Median :   2.094   Median :   2.753   Median :  1.100   Median :16.10   Mostly Cloudy: 92640
Mean   :3.517   Mean   :13.61   Mean   :   3.442   Mean   :   4.505   Mean   :  0.644   Mean   :14.69   Partly Cloudy: 77215
3rd Qu.:5.000   3rd Qu.:19.00   3rd Qu.:   3.875   3rd Qu.:   4.918   3rd Qu.:  8.300   3rd Qu.:16.10   Unknown      : 62959
Max.   :6.000   Max.   :23.00   Max.   :1240.510   Max.   :797.753   Max.   : 20.600   Max.   :16.10   Light Rain   : 49047
                                                   NA's   :3559      NA's   :12032     NA's   :56597   (Other)      :108832
```

DATA CLEANING – Heatmap

We can see some of the trips are on water.

DATA CLEANING – Trip Duration

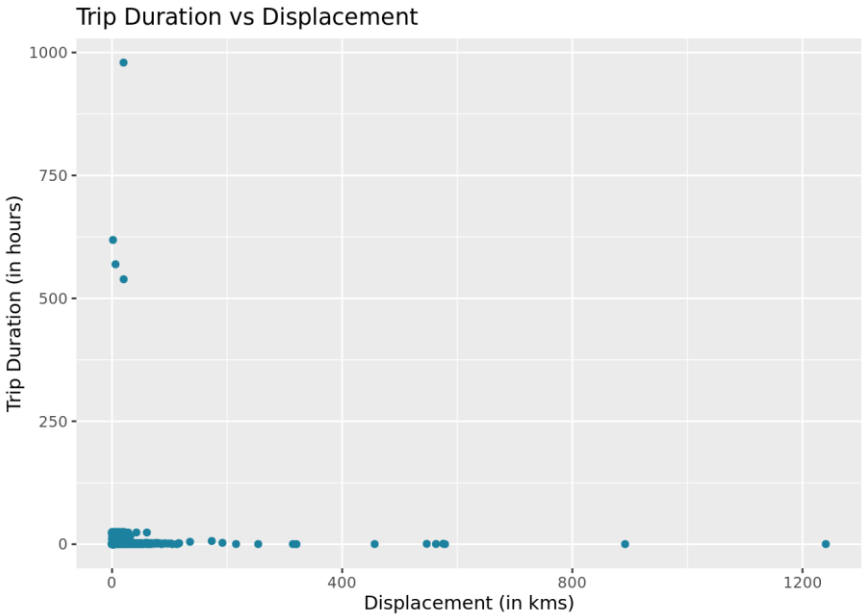




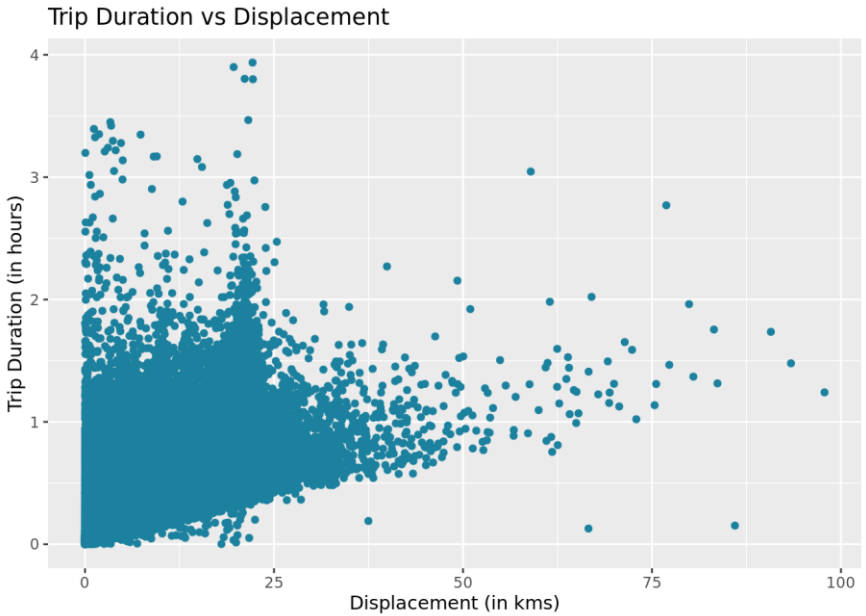
The distribution of trip duration is unreadable because of the extreme outliers.

We removed the trip duration that was causing the extreme skewness and now we have a normal distribution

DATA CLEANING – Trip duration vs

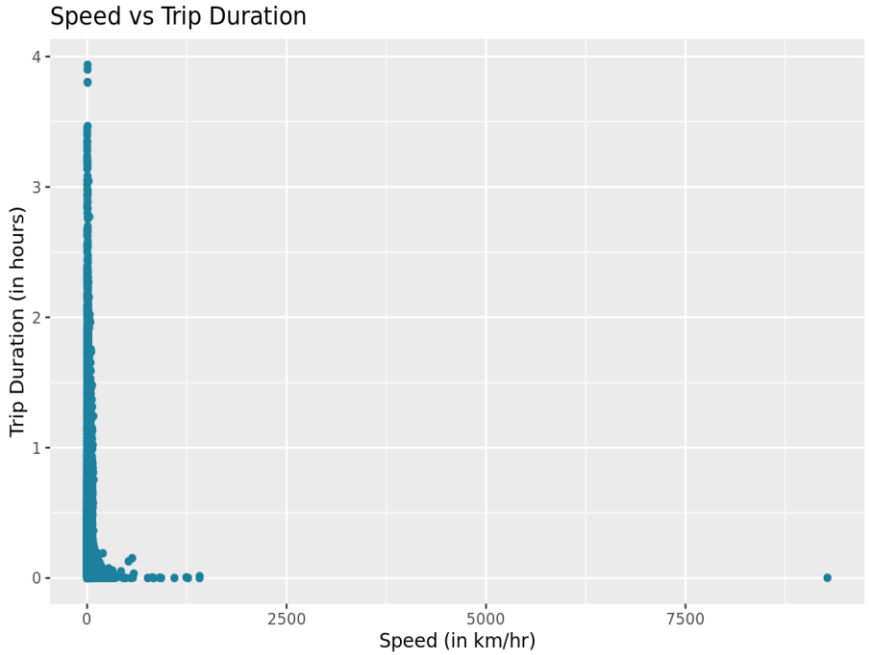


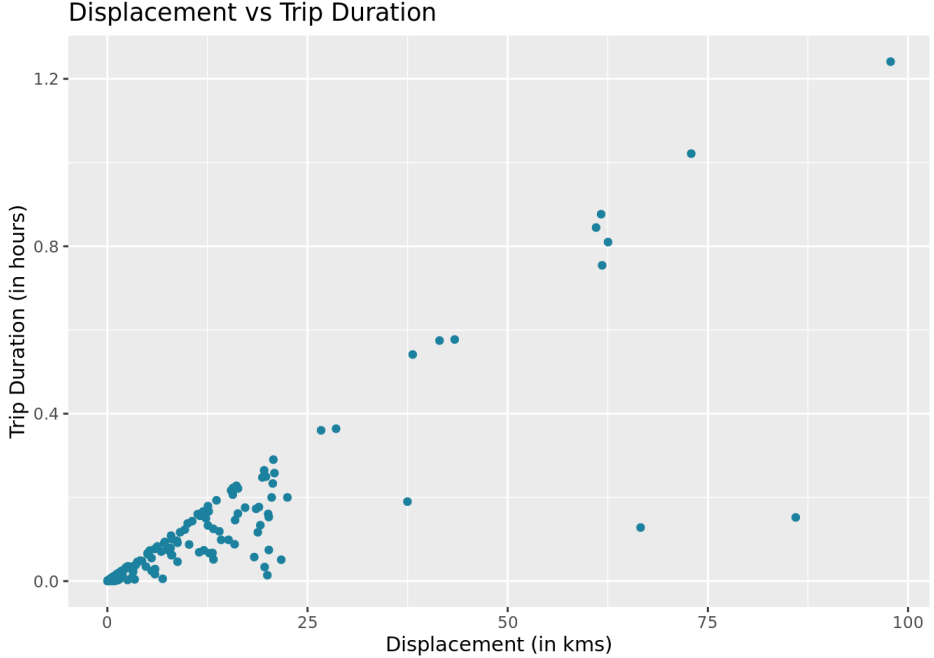
We can clearly see the outliers here. In some of the trips, trip duration is more than 500 hours, which is impossible while in some of the trips the displacement is extremely large.



We removed the trips which are longer than 4 hrs or have more than 100 kms distance.

DATA CLEANING – Speed



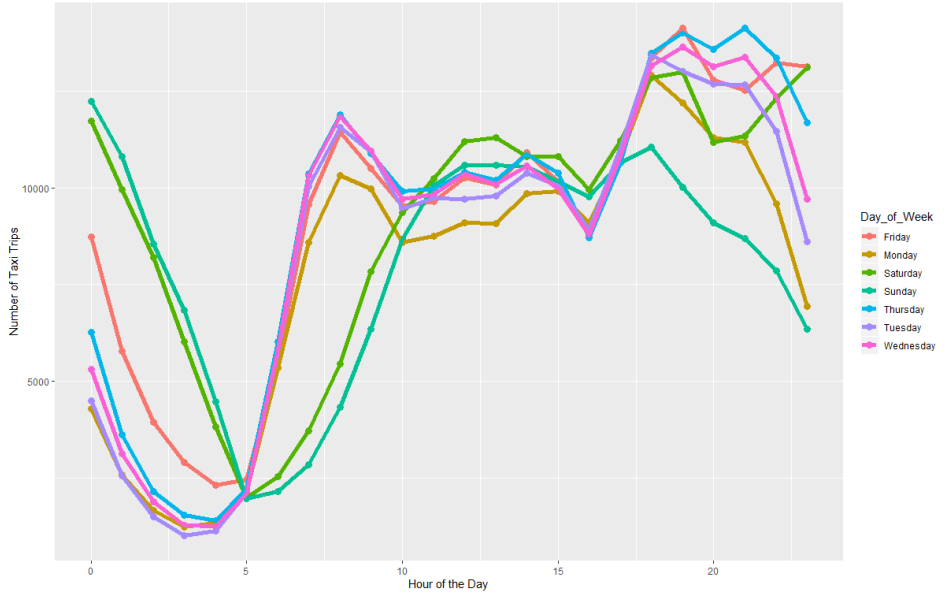


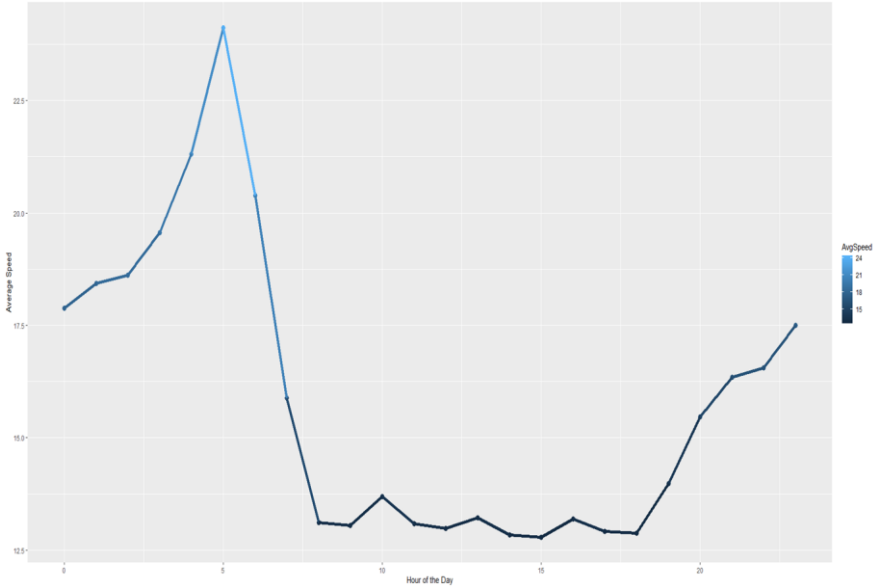
We see some cabs with unusually high speeds. And one particular cab beat the speed of light!

It is safe to assume that average trip speed of over 100 km/hr is unlikely within New York. We removed these outliers

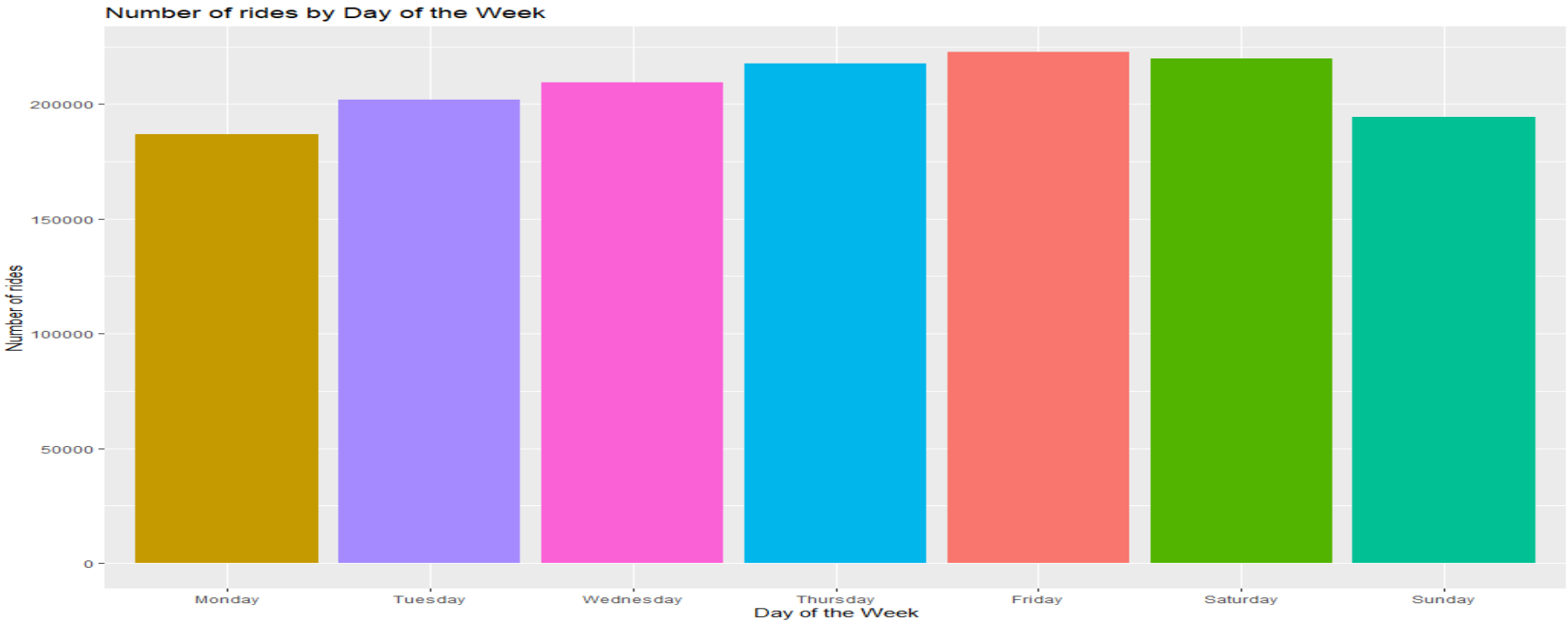
DATA Exploration – by Hour of the Day

During the week, ride-requests are at their highest during the morning and evening commutes. More rides are requested after work than before work.



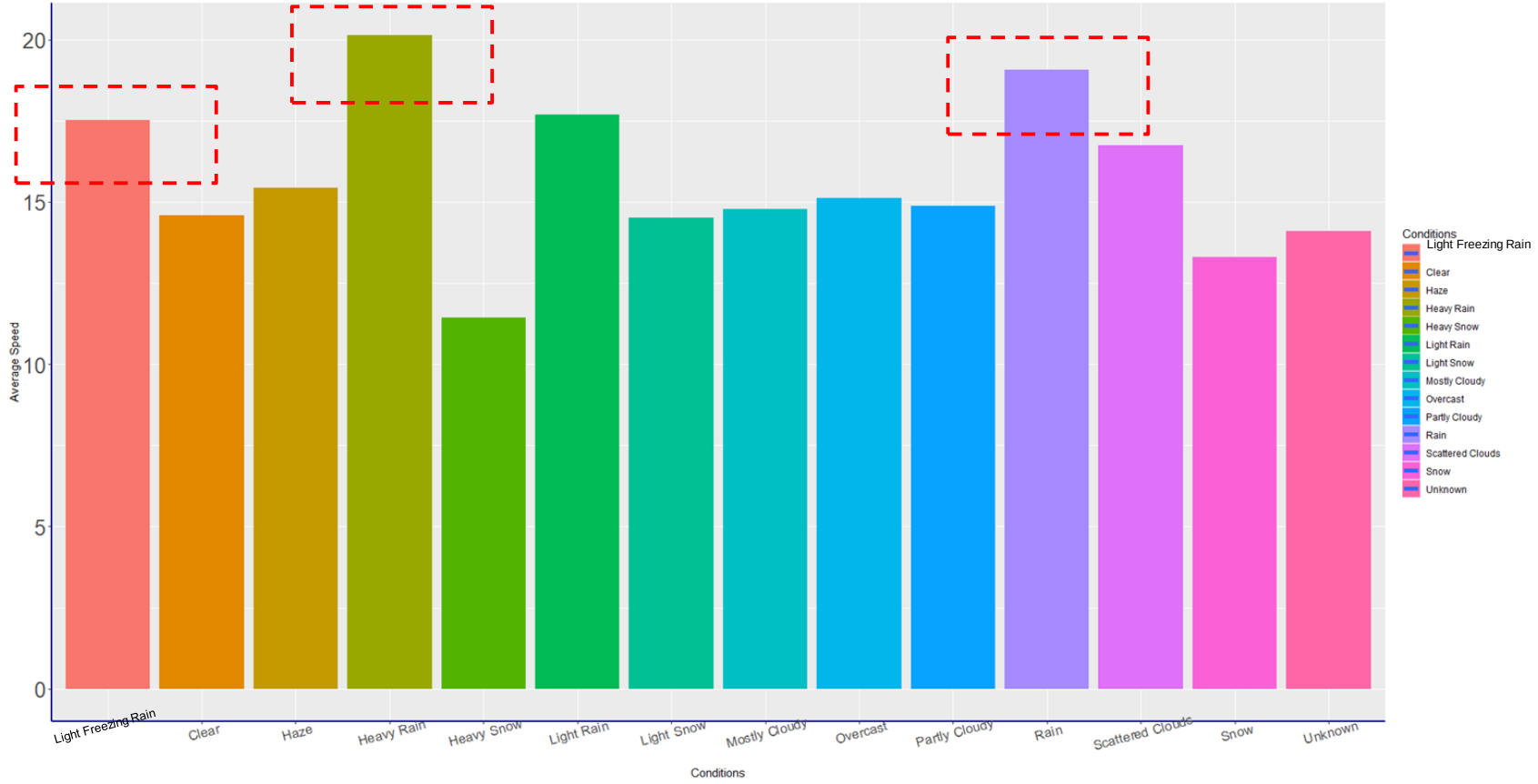


DATA Exploration- By day of week



More rides are requested as the work week progresses

DATA Exploration- Speed vs. Weather Conditions



ADDITIONAL FEATURES

Number of Trips
1,458,644 ➡ 1,450,435

Number of Variables
18

Period of Data
Jan - June 2016

- Pickup Month
- Time of the day
- Condition
- Temperature
- Visibility
- Day of the week
- Pickup Hour
- Distance /Displacement
- Dewpoint
- Degree

DATA MODELLING (1)

OLS	Did 10-fold cross validation. Considered all variables	RMSE (in secs) 380	R-SQUARE 63.7%
RIDGE	Did 10-fold cross validation. Bestlam = 0.51	RMSE (in secs) 399.3	R- SQUARE 63.5%
LASSO	Did 10-fold cross validation. Bestlam = 0.51	RMSE (in secs) 398,7	R- SQUARE 63.5%
DECISION TREE	Depth of tree = 2 Displacement, Distance, Pickup Hour	RMSE (in secs) 402	R- SQUARE 62.8%

DATA MODELLING (2)

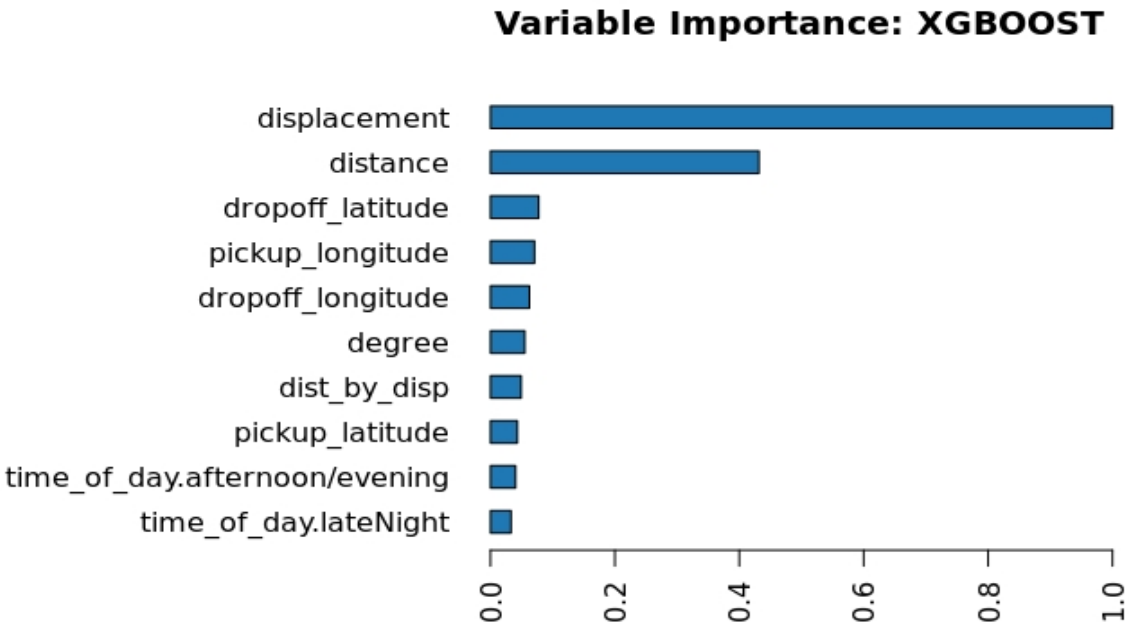
BAGGING	ntrees = 500 Displacement, Distance, Pickup Hour	RMSE 329	R-SQUARE 74%
RANDOM FOREST	ntrees = 500 , mtry = p/3 Displacement, Distance, Pickup longitude	RMSE (in secs) 294.1	R- SQUARE 80.4%
GRADIENT BOOSTING	ntrees = 1000, shrinkage = 0.3, interaction depth = 6 Displacement, Distance, Pickup Hour	RMSE 308	R- SQUARE 77.6%
XG BOOST	ntrees = 3000, learning rate = 0.1,max depth = 8 subsample = 0.6 , colsample_by tree = 0.5	RMSE (in secs) 280.8	R- SQUARE 81.9%

XGBOOST (GRID SEARCH)

ETA	Depth	N-Trees	RMSE
0.1	4	500	307.6928
0.1	4	1000	301.5163
0.1	4	1500	298.7062
0.1	6	500	294.9437
0.1	6	1000	290.7747
0.1	6	1500	288.8377
0.1	8	500	287.7832
0.1	8	1000	284.5928
0.1	8	1500	282.857

ETA	Depth	N-Trees	RMSE
0.3	4	500	301.9141
0.3	4	1000	298.7645
0.3	4	1500	297.2909
0.3	6	500	292.8685
0.3	6	1000	291.5232
0.3	6	1500	291.3419
0.3	8	500	292.0629
0.3	8	1000	293.1799
0.3	8	1500	294.8082

XGBOOST (VARIABLE IMPORTANCE)



SOURCES

- Dataset https://www.kaggle.com/c/nyc-taxi-trip-duration/data
- Weather https://www.kaggle.com/mathijsen/new-york-city-taxi-trip-hourly-weather-data
- Driving distance data using geo-coordinates from Google Maps Distance Matrix API
- Background photo: laurapuig4 @ Pixabay
- Slides base by 24Slides
- Some icons by Flaticon