BDA Analytics

Yellow Taxi Data Analysis

Roshan Tushar

CB.EN.U4AIE19071

XX DATA ANALYST XX

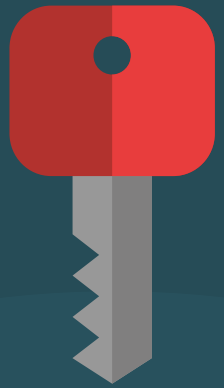
Asmitha U

CB.EN.U4AIE19065

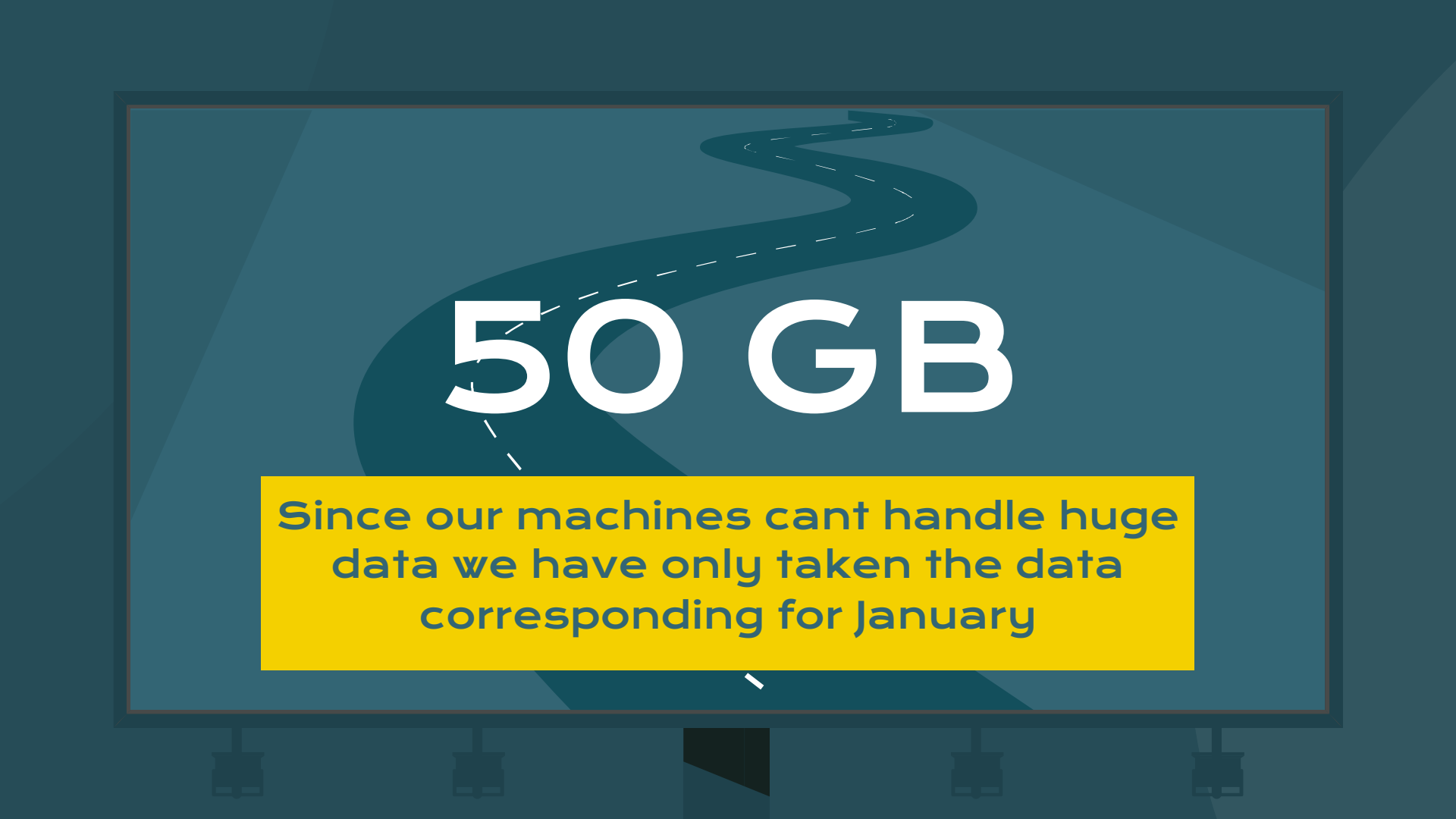
XX DATA ANALYST XX

01

Steps Taken



- ✓ Reading and cleaning data using Spark(Scala)
- ✓ Data Analysis using Spark(Scala)
- ✓ Linear Regression Models to estimate fare
- ✓ Linear Regression Models to estimate trip duration
- ✓ Visualizations using Jupyter Notebook


50 GB
Since our machines cant handle huge data we have only taken the data corresponding for January

The Data

File Format : CSV Format

File Size : 50 GB of data

14573157 rides | **2.99** GB after cleaning

Extracted features:

- **Average Speed**
- **Great Circle Distance**
- **Kilometer**
- **Taxi revenue**
- **Borough mapping from coordinates**

VendorID	A code indicating the TPEP provider that provided the record.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
Pickup_longitude	Longitude where the meter was engaged.
Pickup_latitude	Latitude where the meter was engaged.
RateCodeID	Taximeter rate code for which the fare is different for each rate code
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,”because the vehicle did not have a connection to the server.

Dropoff_longitude	Longitude where the meter was disengaged.
Dropoff_latitude	Latitude where the meter was disengaged.
Payment_type	A numeric code signifying how the passenger paid for the trip.
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Data Cleaning

- All rides with average speed above 120km/h
- Pickup and dropoff outside NYC boroughs
- Standard fare with distance smaller than great circle distance
- 0 passengers in the car
- Fare of 0 or less
- Trips longer than 24 hours
- Trips slowers than 1km/h on average

DATA ANALYSIS

Rate Code Analysis

Pickup Analysis

Drop-off Analysis

Pairs of boroughs analysis

Hack Licenses analysis

Medallions analysis

Day of week analysis

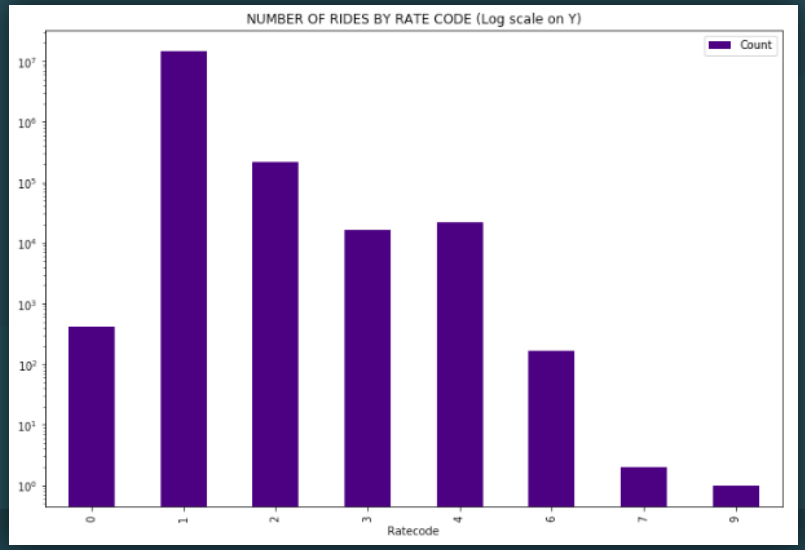
Hour of day analysis

Date analysis

Time-Zone analysis

Distance analysis

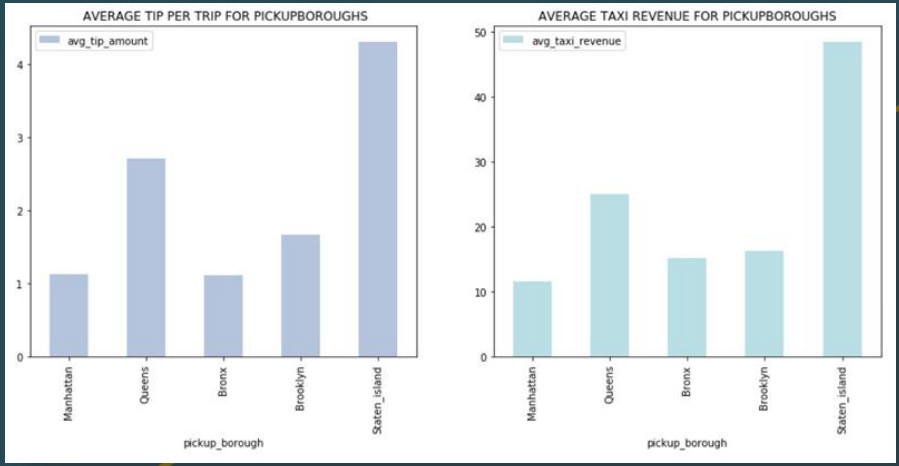
$ Rate Code Analysis

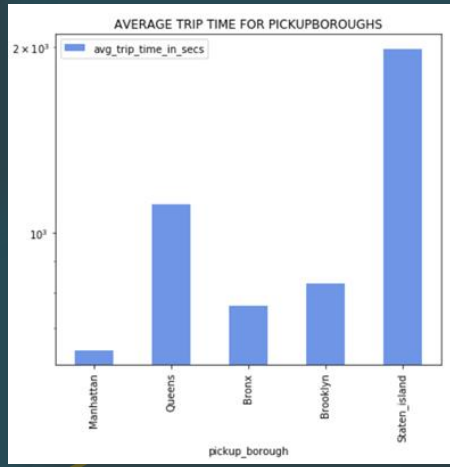


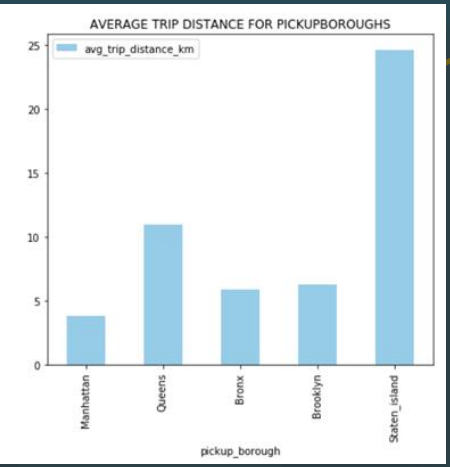
Ratecode	Count
0	419
1	14317811
2	217635
3	16007
4	21121
6	161
7	2
9	1

→ Rate Code 1 (Standard Rate) has the highest number of trips.

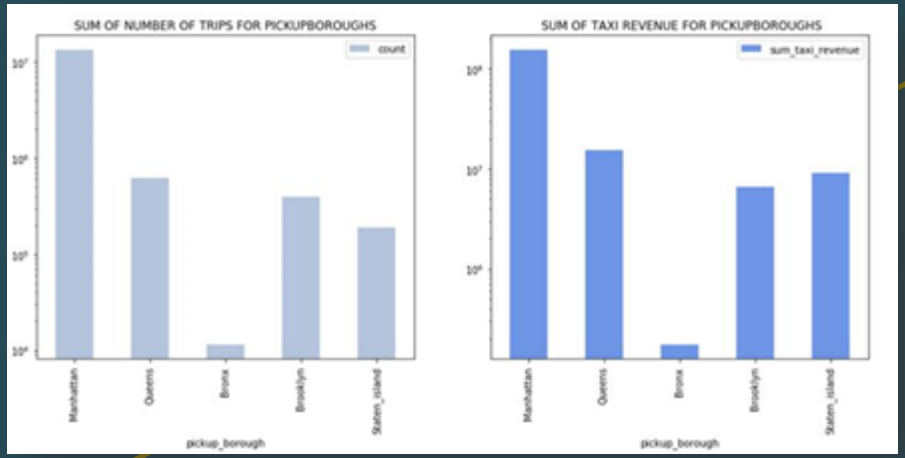
Pickup Analysis

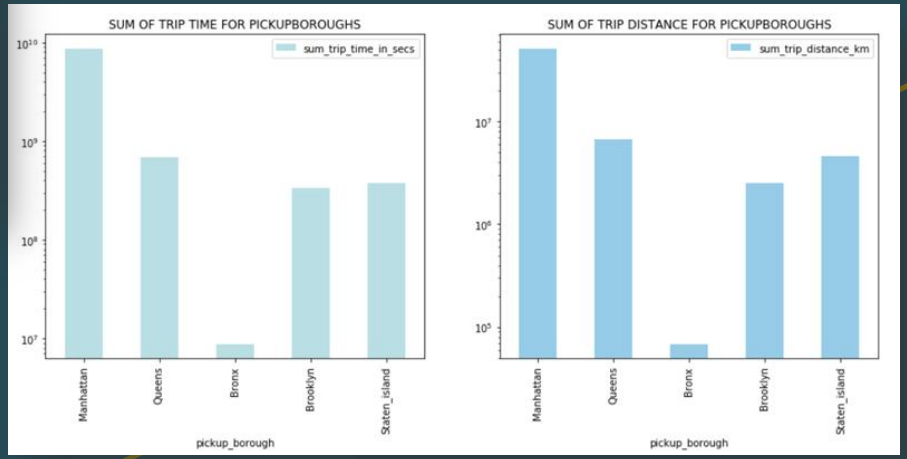




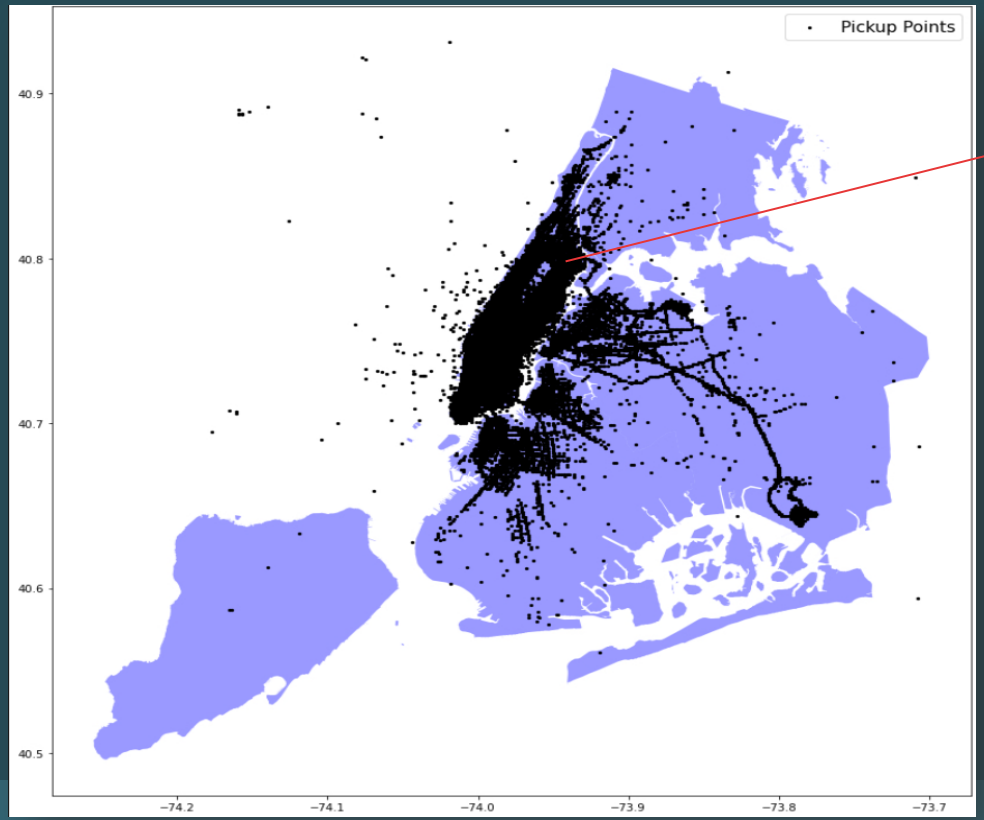


Pickup Analysis



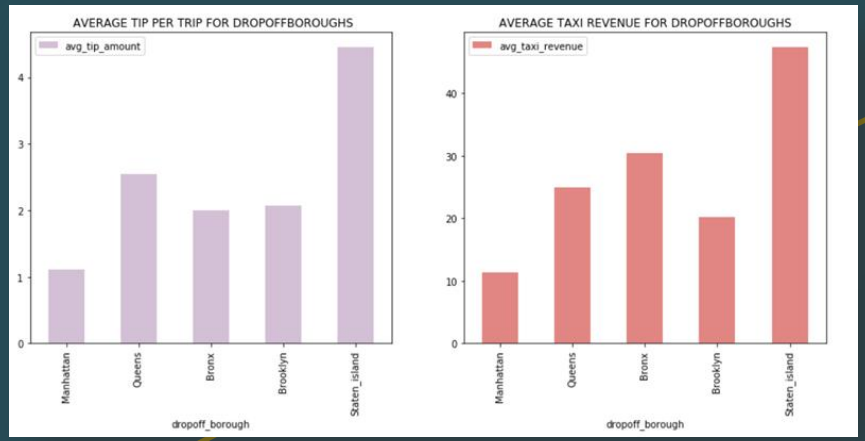


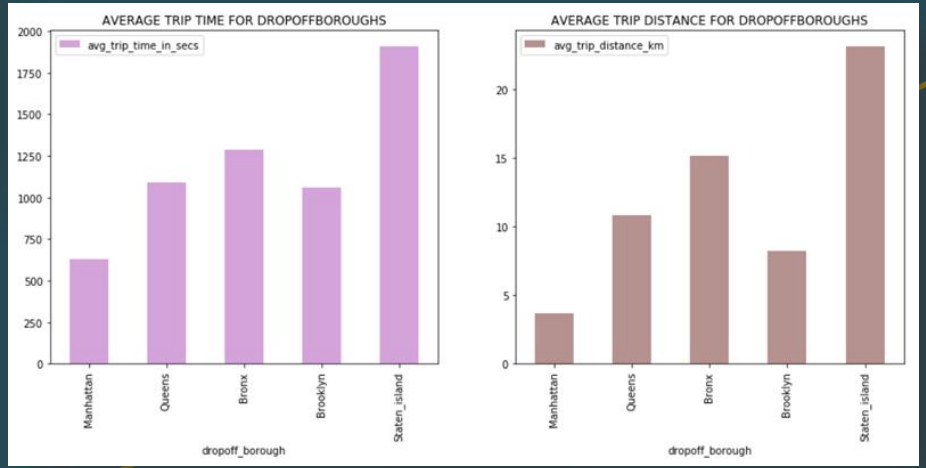
Top 10,000 pick up points



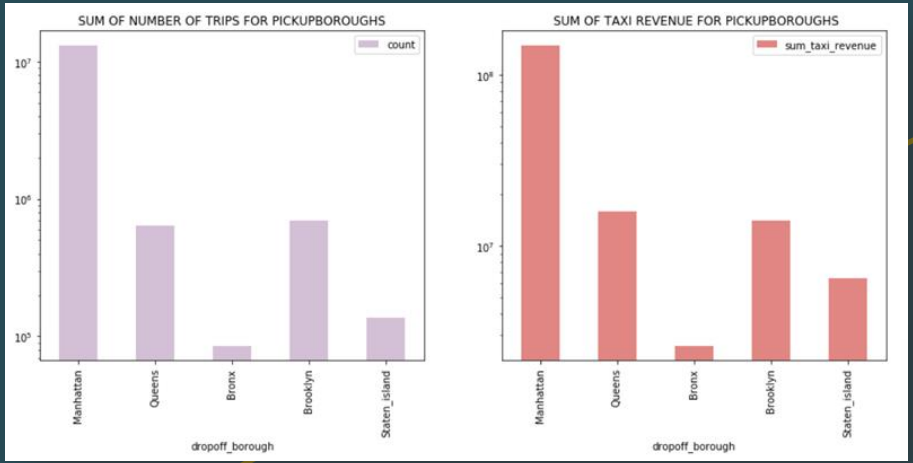
Manhattan

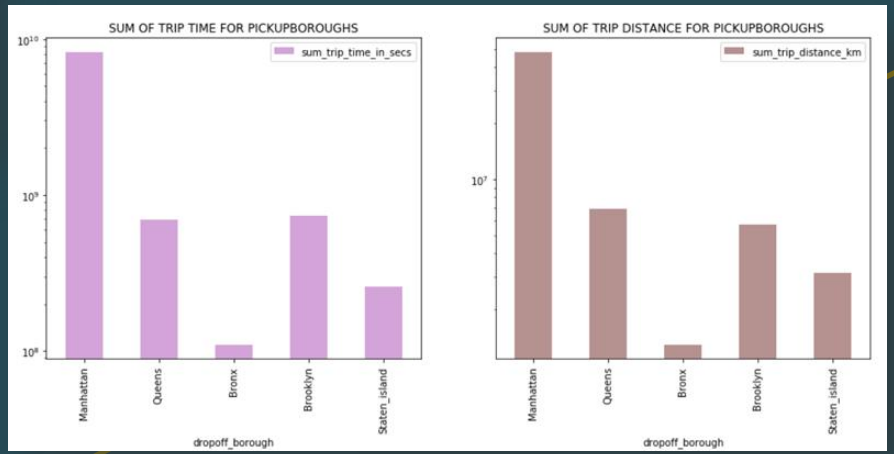
Drop-off Analysis



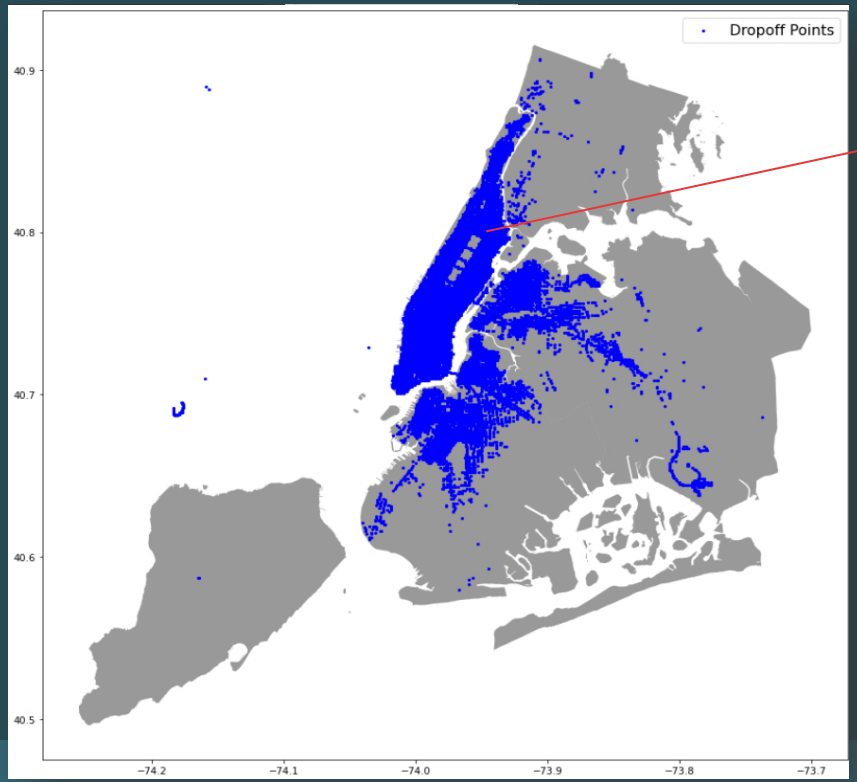


Drop-off Analysis



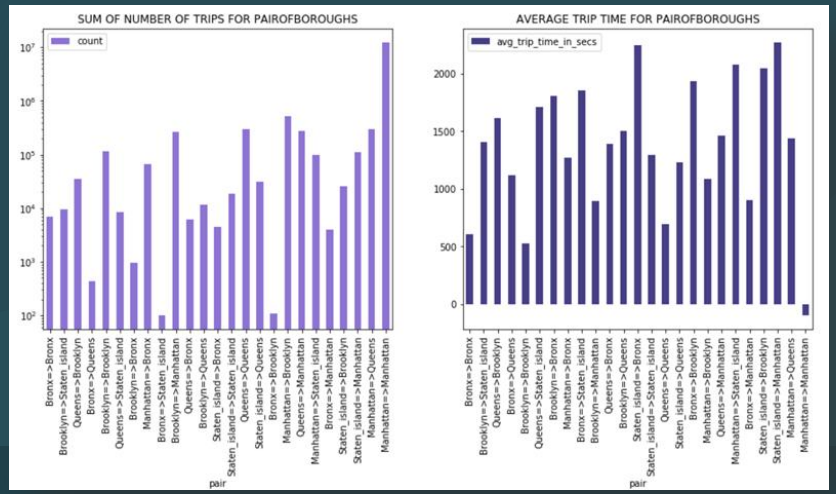


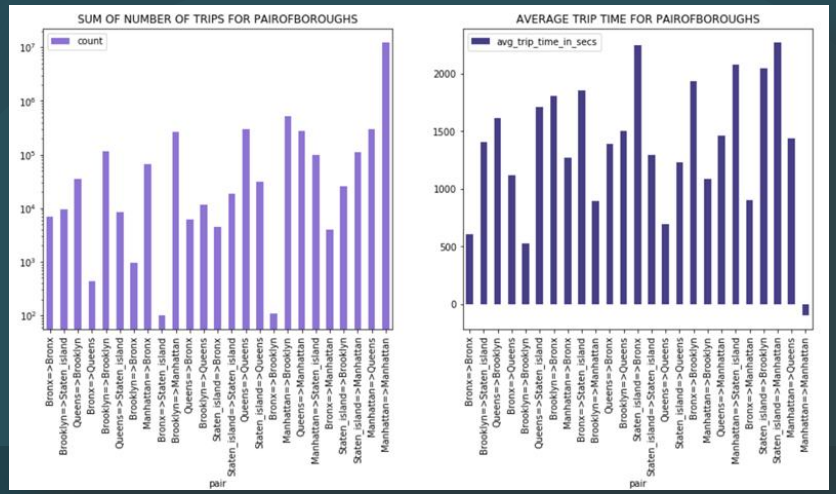
Top 10,000 drop off points



Manhattan

Pairs of boroughs analysis





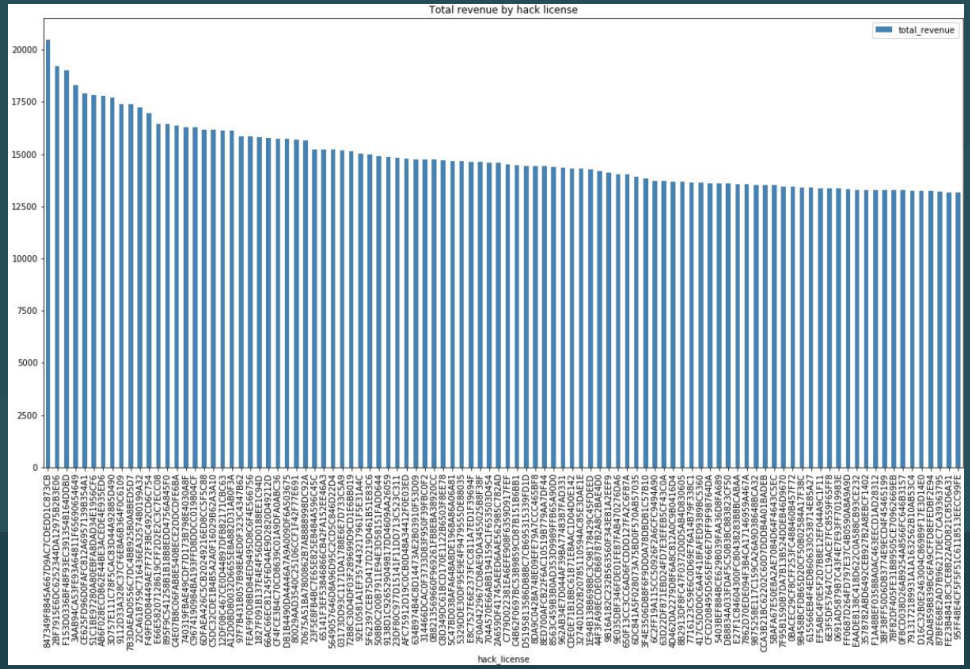
Hack Licenses analysis

Total Rides Per Month(Top-100)

hack_license	count
F49FD0D84449AE7F7...	1595
51C1BE97280A80EBF...	1593
CE625FD96D0FAFC81...	1539
847349F8845A667D9...	1538
3D757E111C78F5CAC...	1502
22CA618759C716436...	1481
3AAB94CA53FE93A64...	1470
E66E58207128619CF...	1406
03173DD93C1171DA1...	1377
C9674190984BA193F...	1369

only showing top 10 rows

Hack Licenses analysis



hack_license	total_revenue
847349F8845A667D9...	20347.429999999997
2BF7915E6DC625234...	19177.07
F153D0336BF48F93E...	18999.239999999998
3AAB94CA53FE93A64...	18292.200000000004
CE625FD96D0FAFC81...	17895.870000000006
AB6F028ECDB62E44B...	17750.590000000004
3D757E111C78F5CAC...	17705.790000000005
51C1BE97280A80EBF...	17669.099999999995
9112D33A328C37CF6...	17360.76
7B3DAEAD0556C7DC4...	17276.56

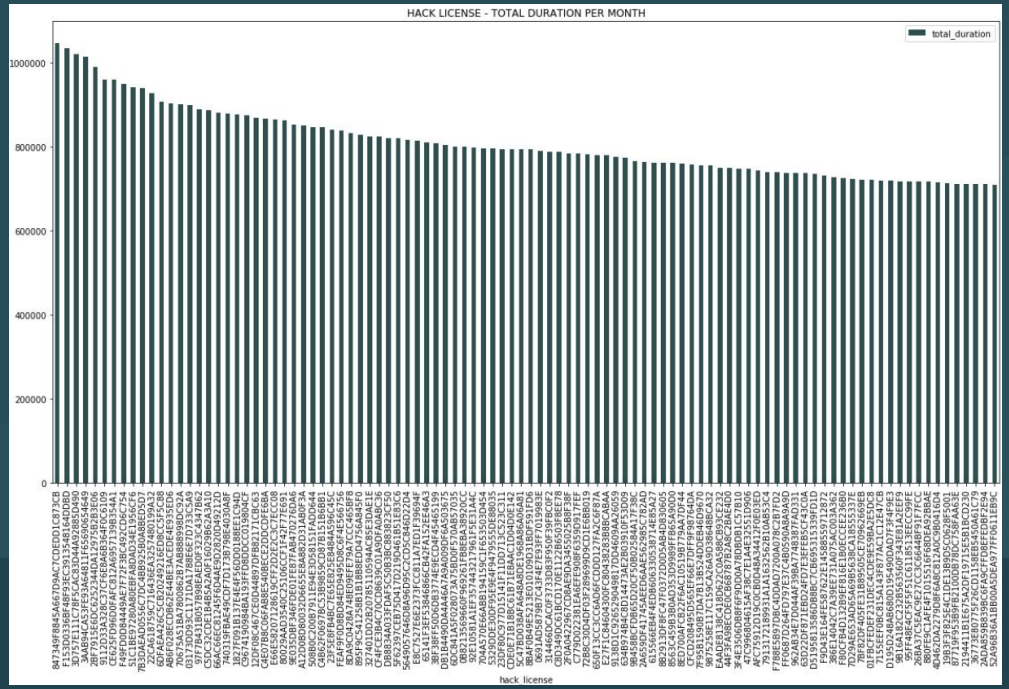
only showing top 10 rows

Hack Licenses analysis

hack_license	total_tips
2BF7915E6DC625234...	1879.5699999999997
847349F8845A667D9...	1836.9299999999998
AB6F028ECDB62E44B...	1829.5900000000001
3AAB94CA53FE93A64...	1791.1999999999998
9112D33A328C37CF6...	1729.7600000000002
7B3DAEAD0556C7DC4...	1717.0599999999997
F153D0336BF48F93E...	1706.7400000000005
A12D0BD80032D6655...	1702.069999999999
895F9C541258B1B18...	1688.64
51C1BE97280A80EBF...	1651.099999999999

only showing top 10 rows

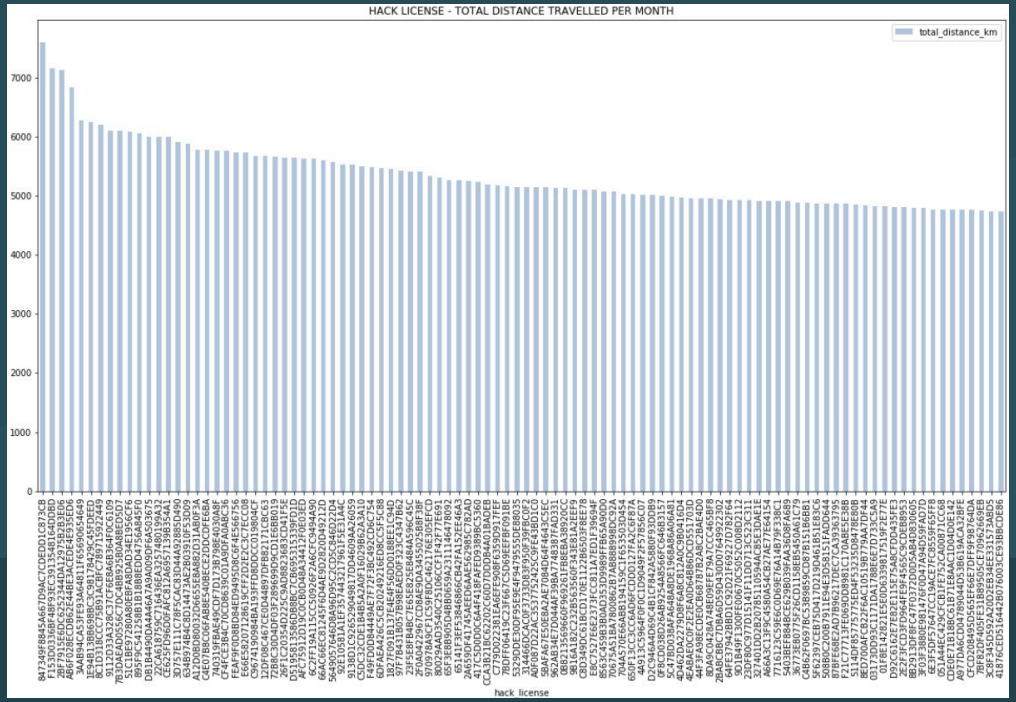
Hack Licenses analysis



hack_license	total_duration
847349F8845A667D9...	1046631
F153D0336BF48F93E...	1035060
3D757E111C78F5CAC...	1021726
3AAB94CA53FE93A64...	1014820
2BF7915E6DC625234...	989522
9112D33A328C37CF6...	960840
CE625FD96D0FAFC81...	960037
F49FD0D84449AE7F7...	950580
51C1BE97280A80EBF...	940800
7B3DAEAD0556C7DC4...	938820

only showing top 10 rows

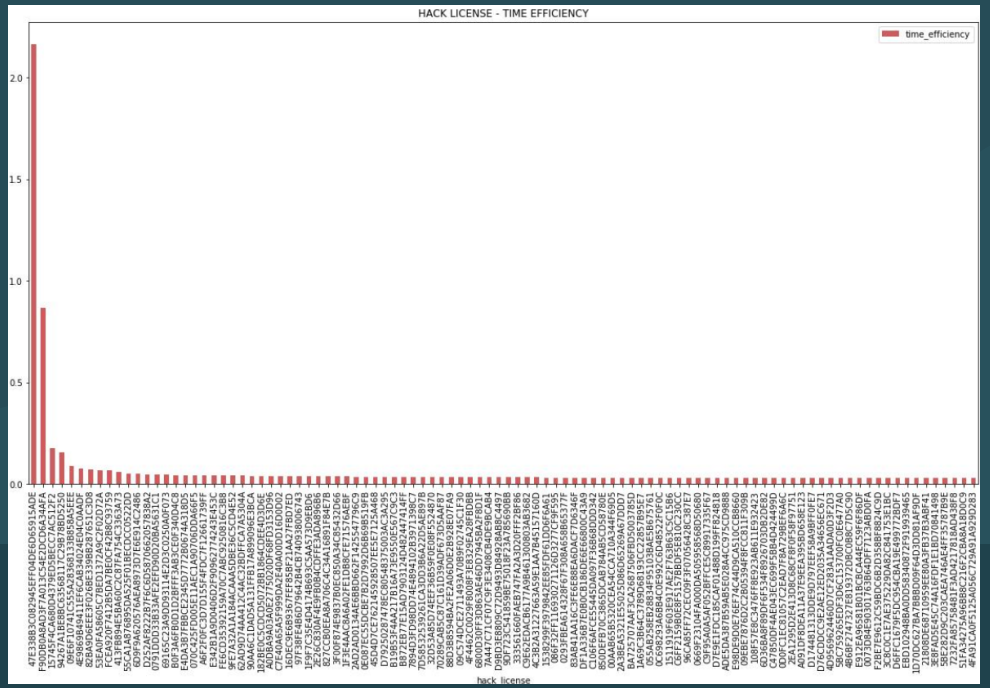
Hack Licenses analysis



hack_license	total_distance_km
847349F8845A667D9...	7588.3277812
F153D0336BF48F93E...	7154.3692502
2BF7915E6DC625234...	7118.4326880000035
AB6F028ECDB62E44B...	6828.622740799999
3AAB94CA53FE93A64...	6266.609026
1E94B13BB698BC3C9...	6239.089312
8C1D318CCC75B91AC...	6196.763670000003
9112D33A328C37CF6...	6096.405227599998
7B3DAEAD0556C7DC4...	6083.723628400002
51C1BE97280A80EBF...	6083.031612200001

only showing top 10 rows

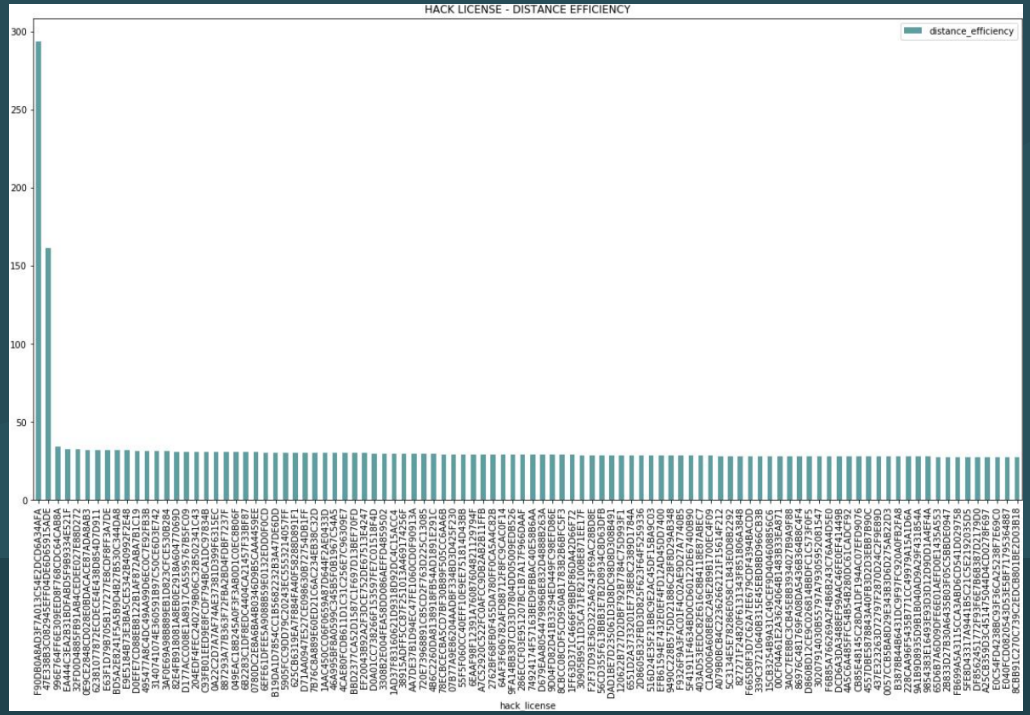
Hack Licenses analysis



hack_license	time_efficiency
09E6F10741C555A28...	0.09085213032581453
82BA9D6EEE3F026BE...	0.07142857142857142
FCEA920F7412B5DA7...	0.06553398058252427
413FB894E5BA60C2C...	0.05976095617529881
6D9F9A620576AEA89...	0.04950495049504951
D252AF8222B7F6C6D...	0.046825396825396826
69165C23A9DD93114...	0.045454545454545456
E4DA3B7FBBCE2345D...	0.0432900432900432
0791DDD33B33AFE2E...	0.043178037279148374
070C425FD005E11AE...	0.043010752688172046

only showing top 10 rows

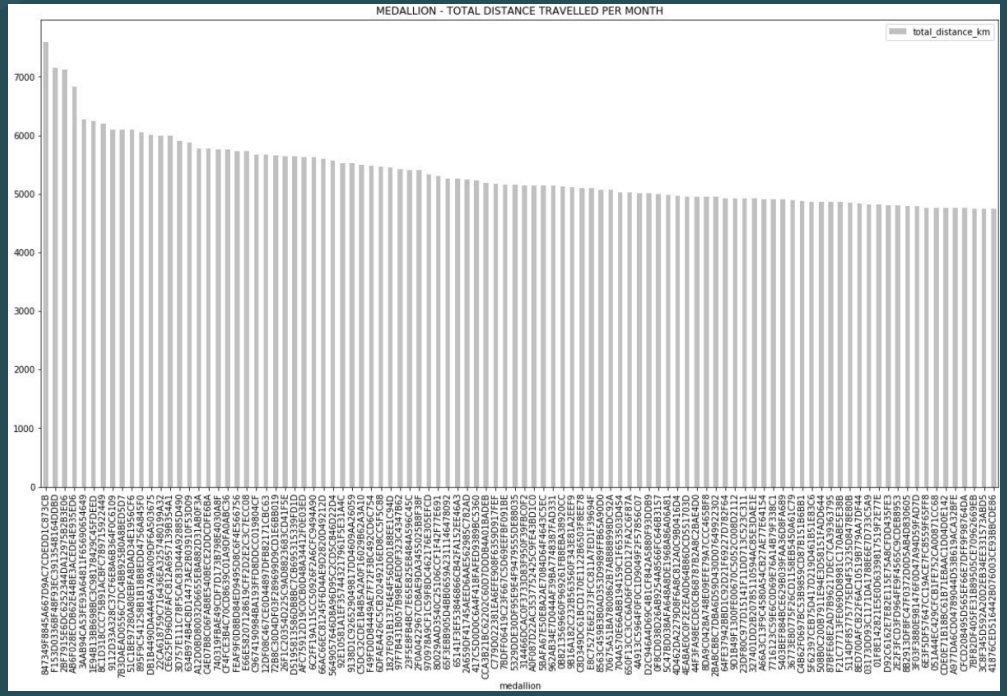
Hack Licenses analysis



hack_license	distance_efficiency
996E764FF6E309EED...	34.01429461096695
EA444C3EA2B31BDFA...	32.628750726173465
32FD0D4885FB91AB4...	32.345246218235553
0E9CE2848CD028E0B...	32.09983953113987
62381077872ECDECE...	32.02161099929729
BD5A2E8241F5A5B9D...	31.65636697175748
E9E5184B73E3BA5CE...	31.60685114691983
E63F1D79B705B1772...	31.452324967114894
D0EFE7CD88EB8122B...	31.35584803142254
495477A8C4DC49AA9...	31.297727707663775

only showing top 10 rows

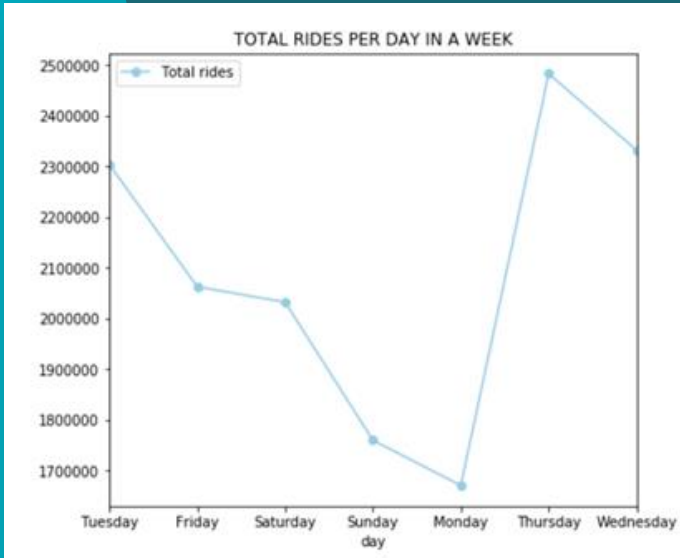
Medallions

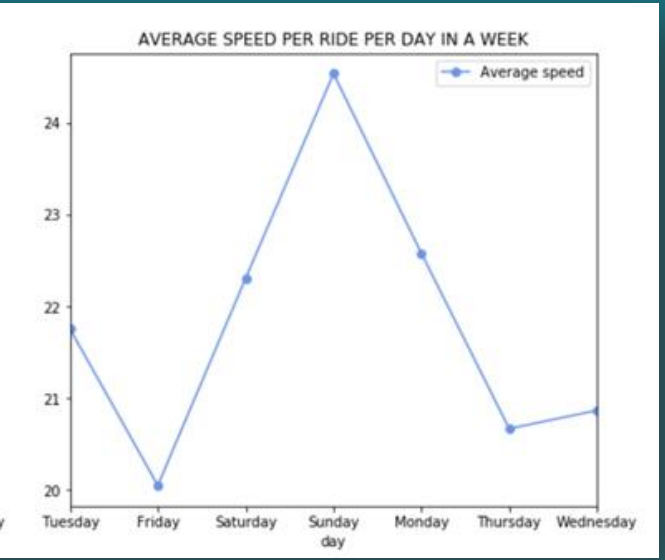


medallion	total_distance_km
847349F8845A667D9...	7588.3277812
F153D0336BF48F93E...	7154.3692502
2BF7915E6DC625234...	7118.4326880000035
AB6F028ECDB62E44B...	6828.622740799999
3AAB94CA53FE93A64...	6266.609026
1E94B13BB698BC3C9...	6239.089312
8C1D318CCC75B91AC...	6196.763670000003
9112D33A328C37CF6...	6096.405227599998
7B3DAEAD0556C7DC4...	6083.723628400002
51C1BE97280A80EBF...	6083.031612200001

only showing top 10 rows

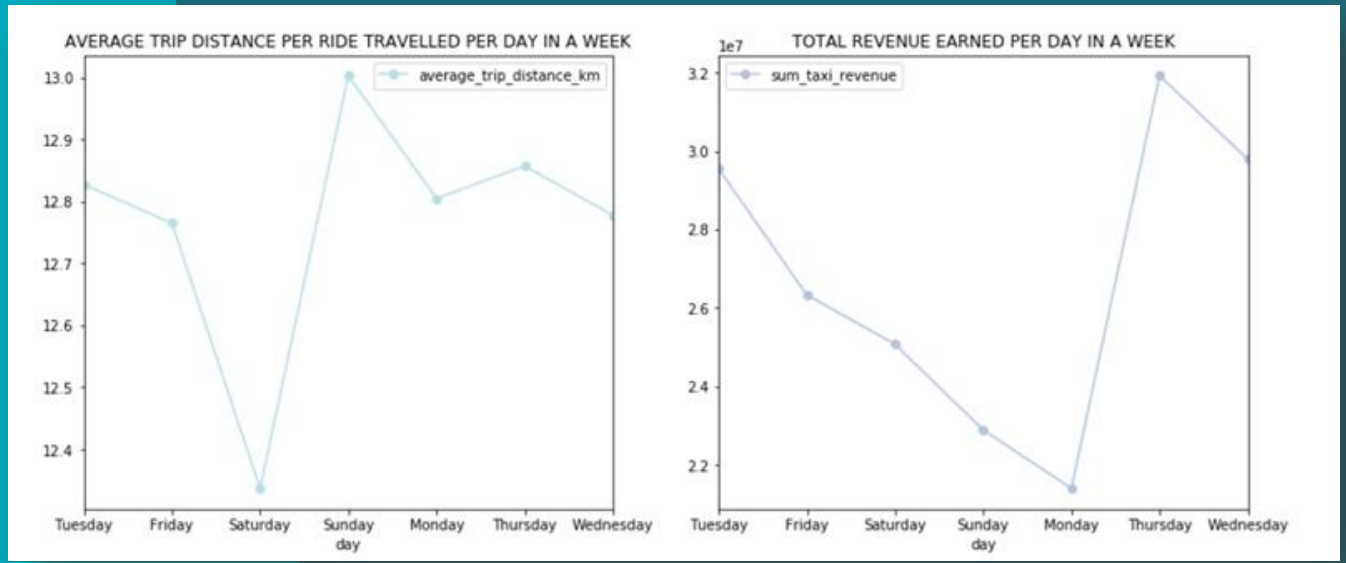
Day of week analysis





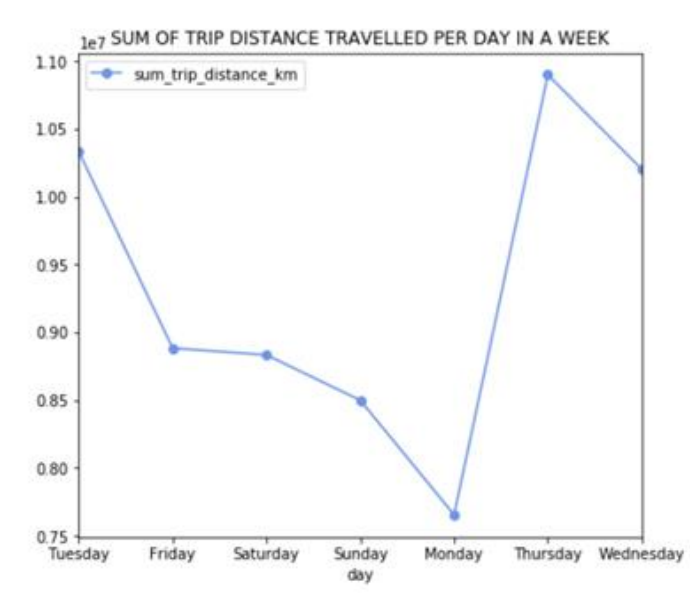
- Thursday has the most number of trips.
- Less traffic in during Sundays.
- Monday has the least number of trips.
- Friday is the most busiest day.

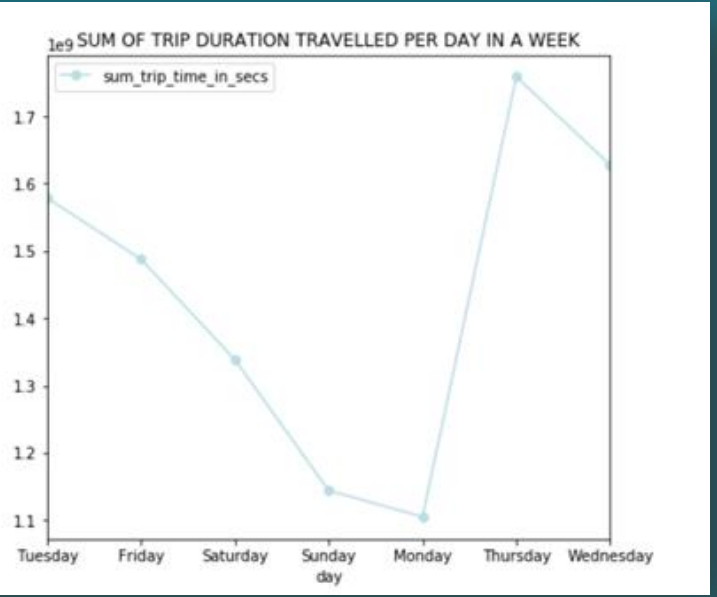
Day of week analysis



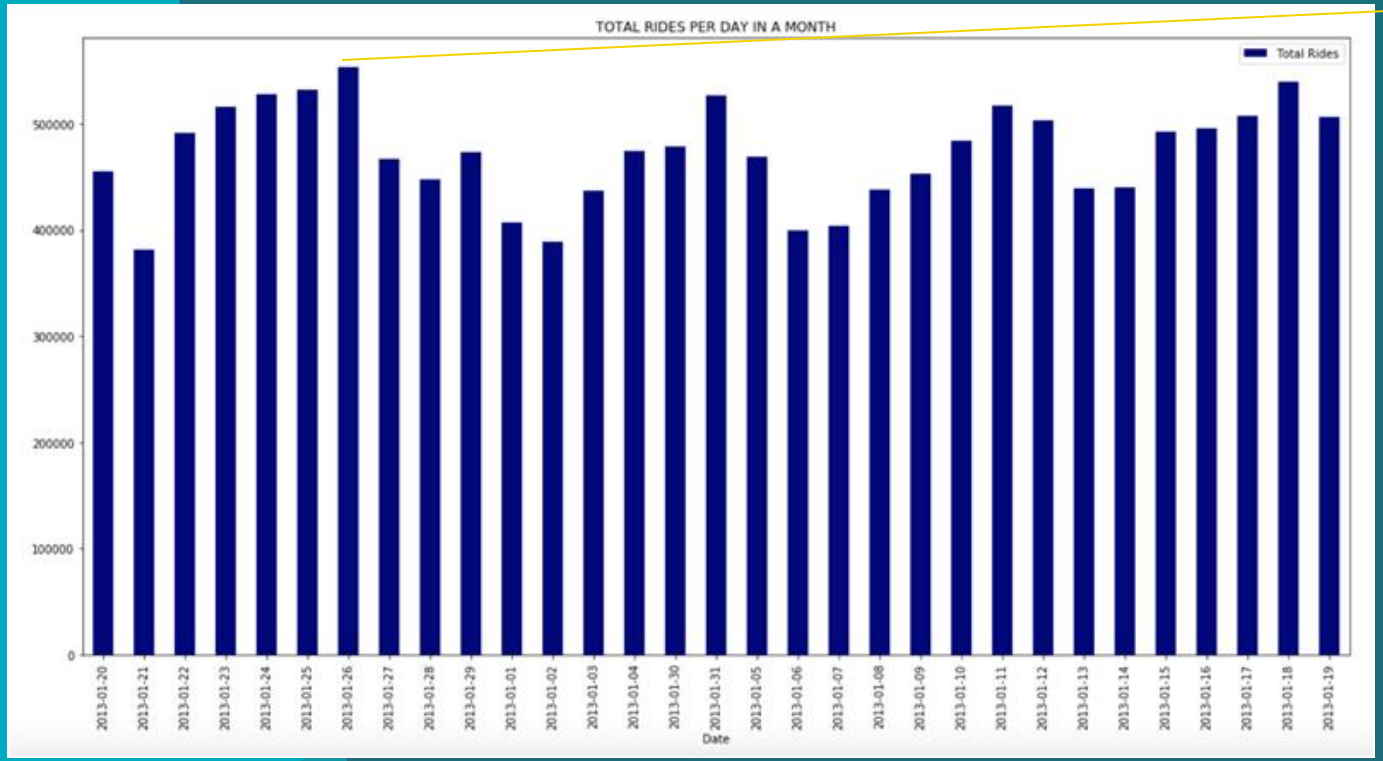
- → Highest revenue on Thursdays.
- → Lowest revenue on Mondays.
- → People have travelled longer on Sundays.

Day of week analysis



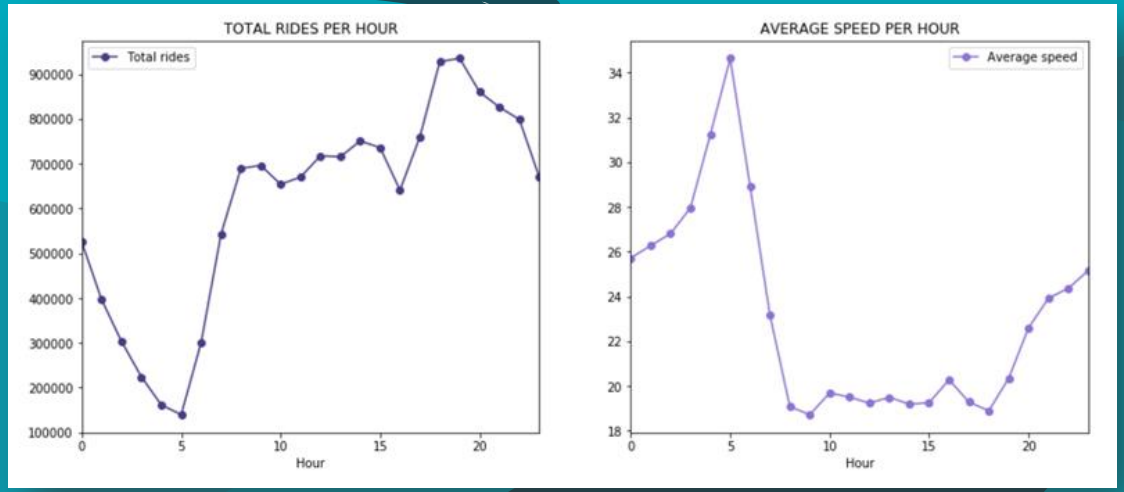


Date analysis



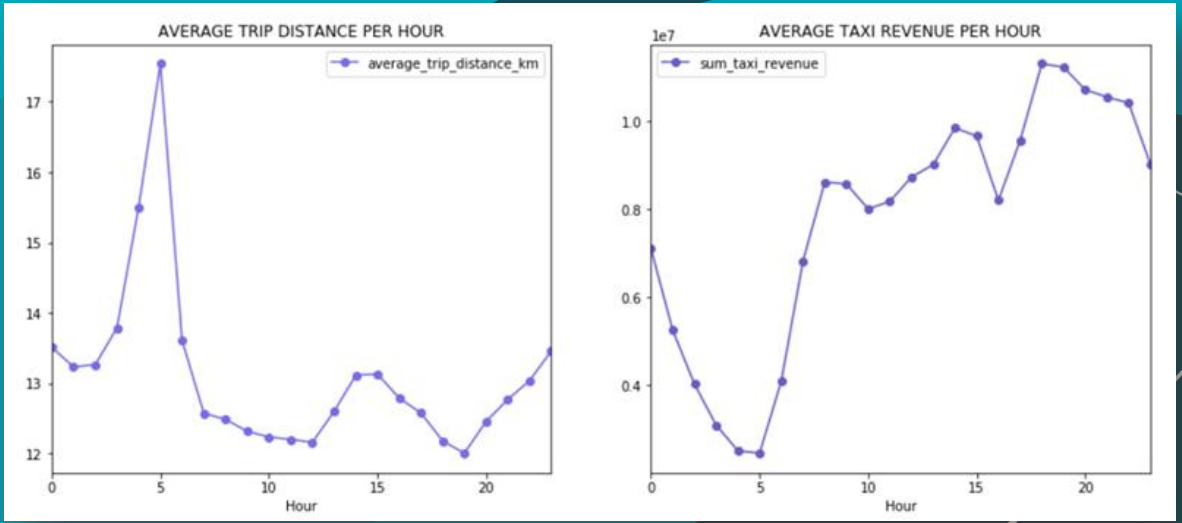
26-01-2013 has the most number of rides

Hour of day analysis



➔There is a drastic increase in the number of rides from 5:00 to 10:00.
➔Since 5:00 has the least number of rides , the traffic is less. This can be inferred from the average speed per hour.
➔The Average speed is almost 20km/h from 8:00 to 18:00.

Hour of day analysis



Time-Zone analysis

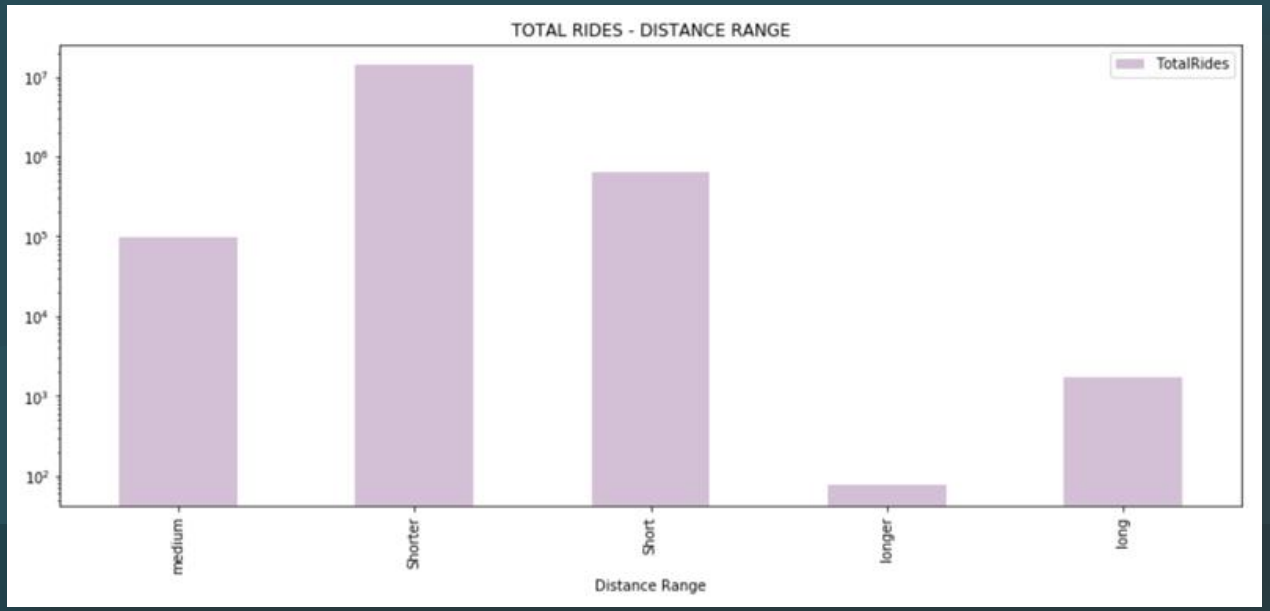
Evening slot has the most number of trips.

Late Night slot has the highest average speed.

Evening slot has the most revenue.

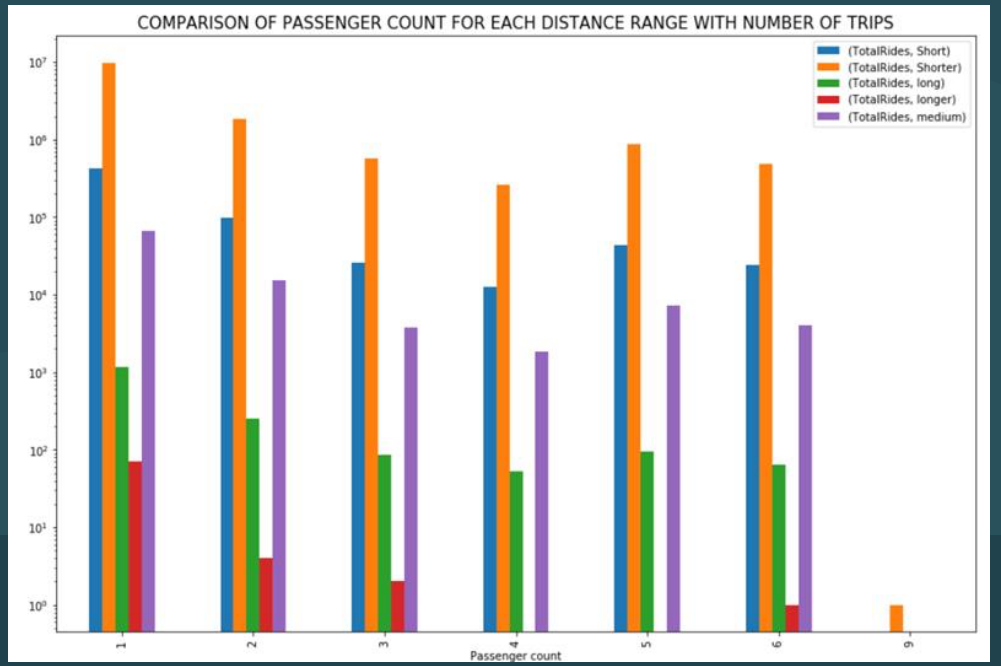
Late Night slot has the longest travelled trips.

Distance analysis



There are more number of rides for shorter trips(<15)

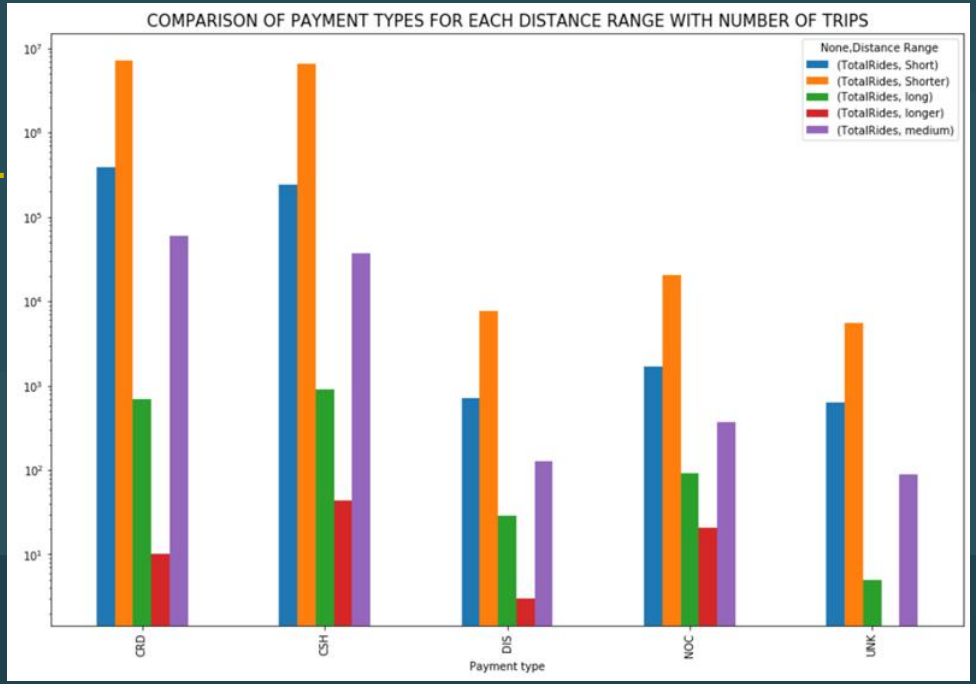
Distance analysis



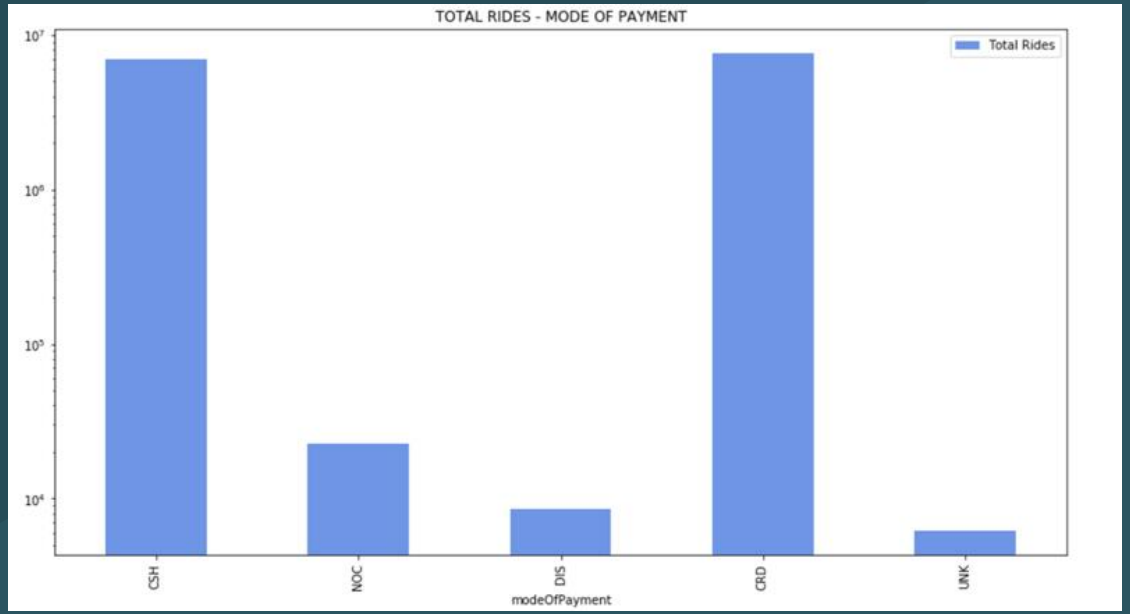
Distance analysis

Card is most used payment method.

Cash being the second.



Mode of payment

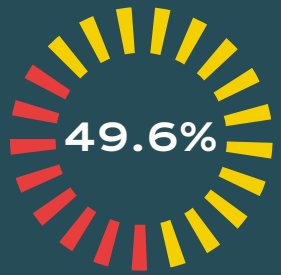


Card is most used payment method.

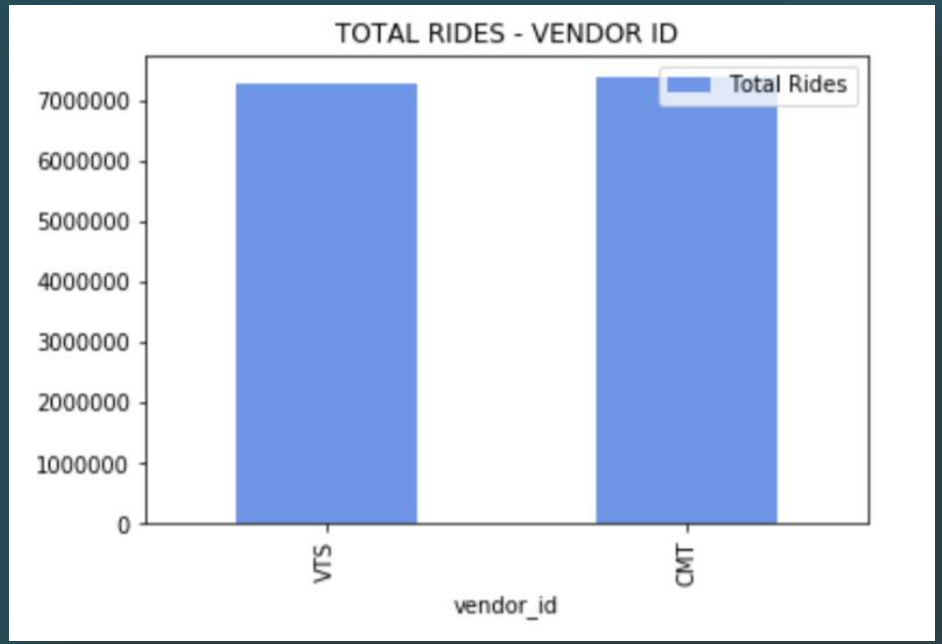
Cash being the second.

Vendor ID

CMT



VTS




MACHINE
LEARNING

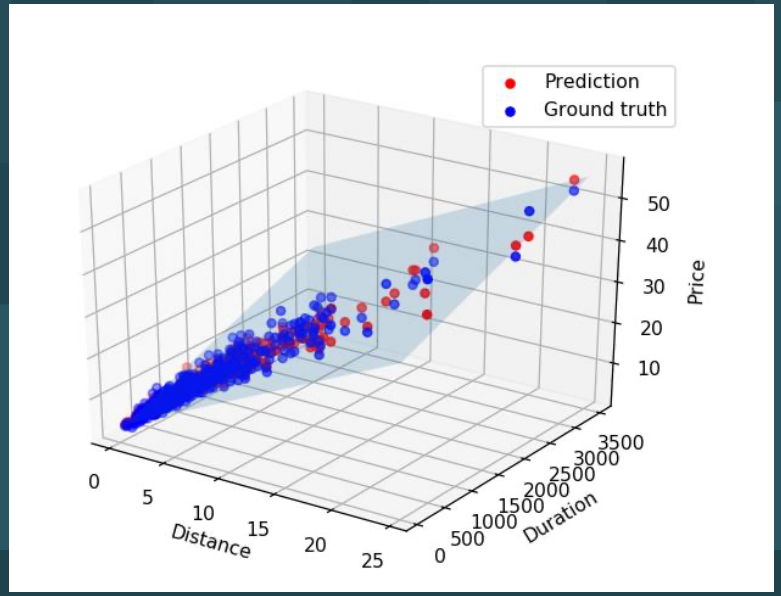
MACHINE LEARNING-ESTIMATING FARES

- ✓ **Trained one model for each rate code from 1 to 4.**
- ✓ **Linear Regression Model**
- ✓ **Features are distance and duration**
- ✓ **Aim to estimate plan**

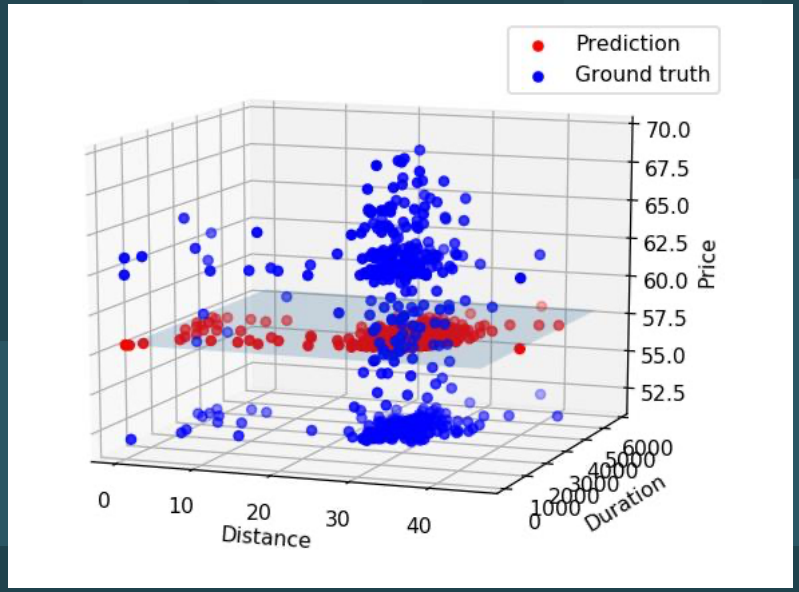
Rate Code	MSE	Formula
1	2.0045058905671587	[0.005894266881238054,1.3119317716666068] + 2.765520293236782
2	5.646270696955837	[-1.3583280798446978E-5,0.0] + 57.99429532382103
3	8.806923079695014	[0.005805615555620736,1.3208252274956496] + 26.08966744542558
4	8.234342104755951	[0.0024098358183701736,2.2886797170897886] + 1.1938434430034144

Scatter Plot Predicted vs Actual

Rate Code 1

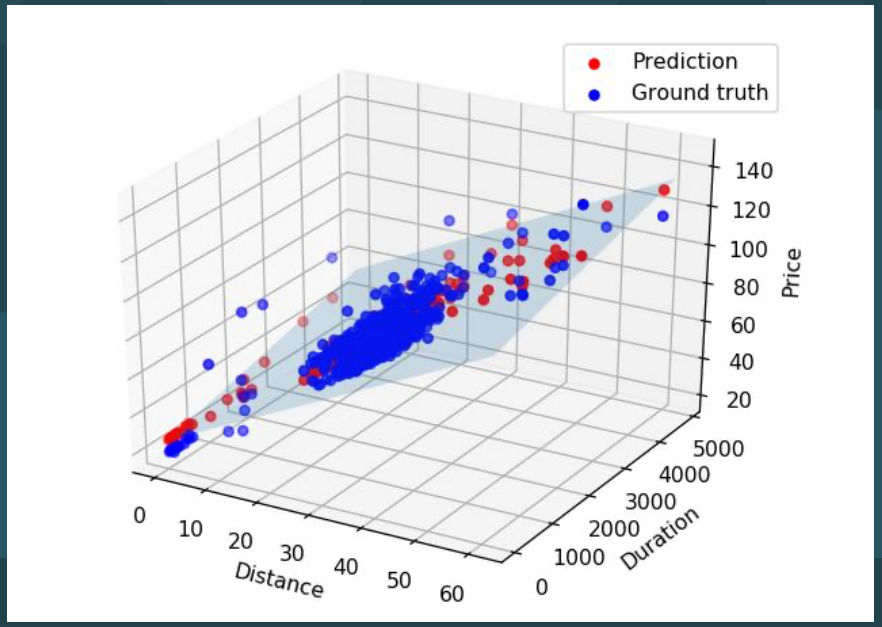


Rate Code 2

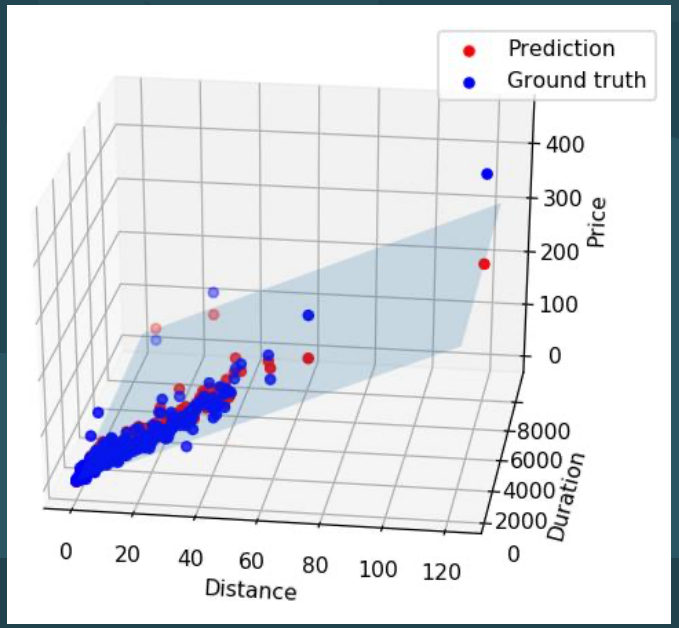


Scatter Plot Predicted vs Actual

Rate Code 3



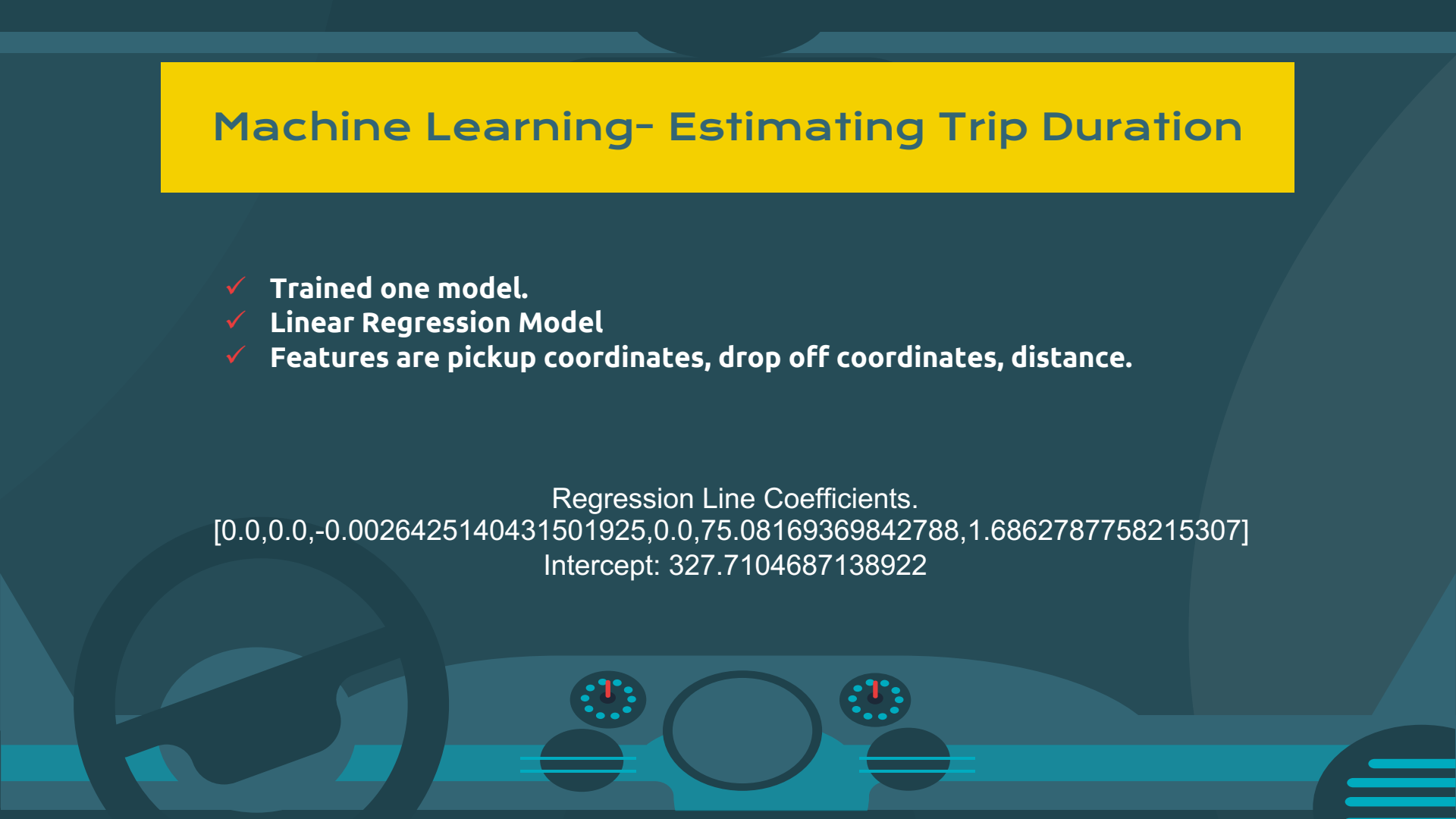
Rate Code 4



Machine Learning- Estimating Trip Duration

- ✓ **Trained one model.**
- ✓ **Linear Regression Model**
- ✓ **Features are pickup coordinates, drop off coordinates, distance.**

Regression Line Coefficients.
[0.0,0.0,-0.0026425140431501925,0.0,75.08169369842788,1.6862787758215307]
Intercept: 327.7104687138922




Thank You