Taxi Trip Duration

Data Understanding

- **id** - a unique identifier for each trip
- **vendor_id** - a code indicating the provider associated with the trip record
- **pickup_datetime** - date and time when the meter was engaged
- **dropoff_datetime** - date and time when the meter was disengaged
- **passenger_count** - the number of passengers in the vehicle (driver entered value)
- **pickup_longitude** - the longitude where the meter was engaged
- **pickup_latitude** - the latitude where the meter was engaged
- **dropoff_longitude** - the longitude where the meter was disengaged
- **dropoff_latitude** - the latitude where the meter was disengaged
- **store_and_fwd_flag** - 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** - duration of the trip in seconds

Check Dataset

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1458644 entries, 0 to 1458643
Data columns (total 11 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   id                  1458644 non-null  object
 1   vendor_id           1458644 non-null  int64
 2   pickup_datetime     1458644 non-null  object
 3   dropoff_datetime    1458644 non-null  object
 4   passenger_count     1458644 non-null  int64
 5   pickup_longitude    1458644 non-null  float64
 6   pickup_latitude     1458644 non-null  float64
 7   dropoff_longitude   1458644 non-null  float64
 8   dropoff_latitude    1458644 non-null  float64
 9   store_and_fwd_flag  1458644 non-null  object
 10  trip_duration       1458644 non-null  int64
dtypes: float64(4), int64(3), object(4)
memory usage: 122.4+ MB
```

Check Dataset

3.1 Duplicates Value ¶

```
]:  1  data.duplicated().sum()
```

```
]: 0
```

3.2 Null Value

```
1  data.isnull().sum()
```

```
id                    0
vendor_id             0
pickup_datetime       0
dropoff_datetime      0
passenger_count       0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude     0
dropoff_latitude      0
store_and_fwd_flag    0
trip_duration         0
dtype: int64
```

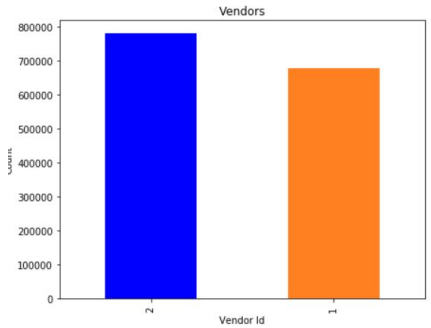
Exploratory Data Analysis

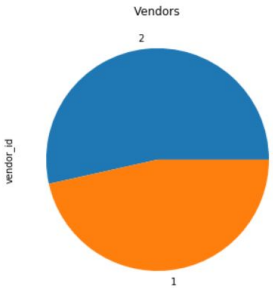
- id

  There are 1393253 Unique id's which represent each row in the data

- **vender Id**

  - Here we got to know that there are only 2 venders(1 and 2)
  - Both the venders share almost equal amount of trips, the difference is quite low between two venders
  - But Vendor 2 is evidently more famous among the population as per the above graphs.





Exploratory Data Analysis

- passengers

New York City Taxi Passenger Limit says:

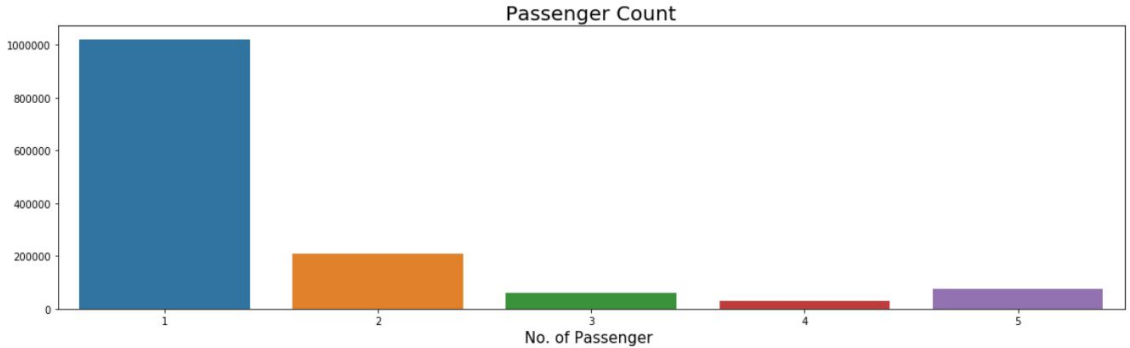
- A maximum of 4 passengers can ride in traditional cabs.
- A child under 7 is allowed to sit on a passenger's lap in the rear seat in addition to the passenger limit.

So, in total we can assume that maximum 5 passenger can board the new york taxi i.e. 4 adult + 1 minor

```
1    1033540
2     210318
5      78088
3      59896
6      48333
4      28404
0         60
7          3
9          1
8          1
Name: passenger_count, dtype: int64
```

Exploratory Data Analysis

- passengers



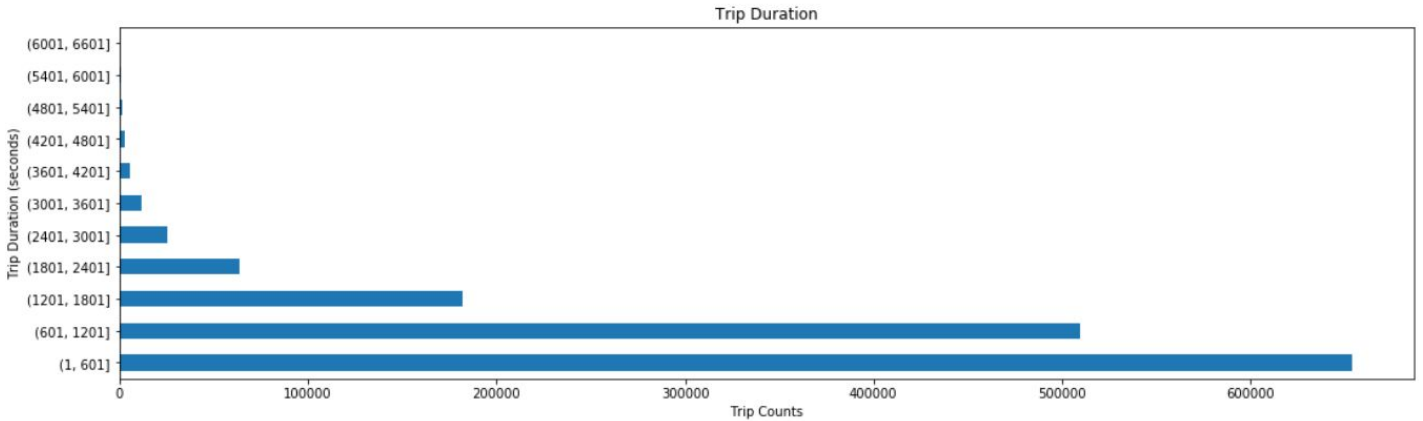
- There are some trips with 0 passenger count.
- Few trips consisted of even 6, 7, 8 or 9 passengers. Clear outliers and pointers to data inconsistency
- Most of trip consist of passenger either 1 or 2.

Exploratory Data Analysis

- ## trip duration
  - Some trip durations are over 100000 seconds which are clear outliers and should be removed.
  - There are some durations with as low as 1 second. which points towards trips with 0 km distance.
  - Major trip durations took between 10-20 mins to complete.
  - Mean and mode are not same which shows that trip duration distribution is skewed towards right
  - These trips ran for more than 20 days, which seems unlikely by the distance travelled.
  - All the trips are taken by vendor 1 which points us to the fact that this vendor might allows much longer trip for outstations.
  - All these trips are either taken on Tuesday's in 1st month or Saturday's in 2nd month. There might be some relation with the weekday, pickup location, month and the passenger.
  - But they fail our purpose of correct prediction and bring inconsistencies in the algorithm calculation.

Exploratory Data Analysis

- trip duration



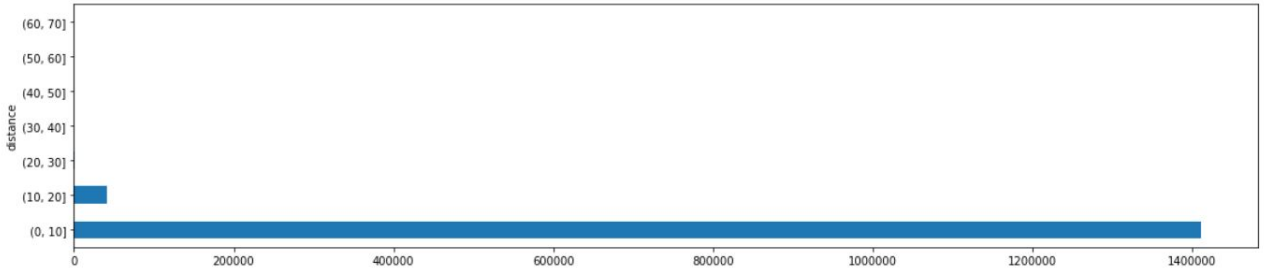
- We can observe that most of the trips took 0 - 30 mins to complete i.e. approx 1800 secs.

Exploratory Data Analysis

- **distance**
  - There some trips with over 60 miles distance.
  - Some of the trips distance value is 0 miles.
  - mean distance travelled is approx 2.1 miles.
  - Around 6K trip record with distance equal to 0. Below are some possible explanation for such records.
    - Customer changed mind and cancelled the journey just after accepting it.
    - Software didn't recorded dropoff location properly due to which dropoff location is the same as the pickup location.
    - Issue with GPS tracker while the journey is being finished.
    - Driver cancelled the trip just after accepting it due to some reason. So the trip couldn't start
    - Or some other issue with the software

Exploratory Data Analysis

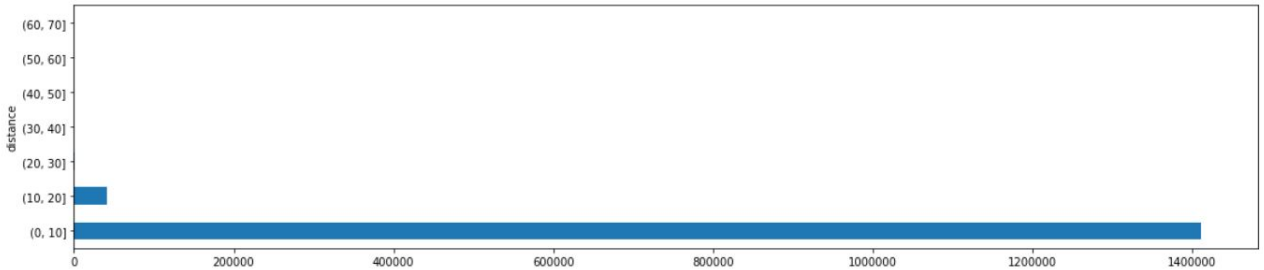
- **distance**



From the above observation it is evident that most of the rides are completed between 1-10 miles with some of the rides with distances between 10-30 miles. Other slabs bar are not visible because the number of trips are very less as compared to these slabs

Exploratory Data Analysis

- **distance**



  - From the above observation it is evident that most of the rides are completed between 1-10 miles with some of the rides with distances between 10-30 miles. Other slabs bar are not visible because the number of trips are very less as compared to these slabs
  - According to the dustribution of trip distances and the fact that it takes about 30 miles to drive across the whole New York City, we decided to use 30 as the number to split the trips into short or long distance trips.
    - Short Trips: 1458545 records in total.
    - Long Trips: 99 records in total.

Exploratory Data Analysis

- ## speed

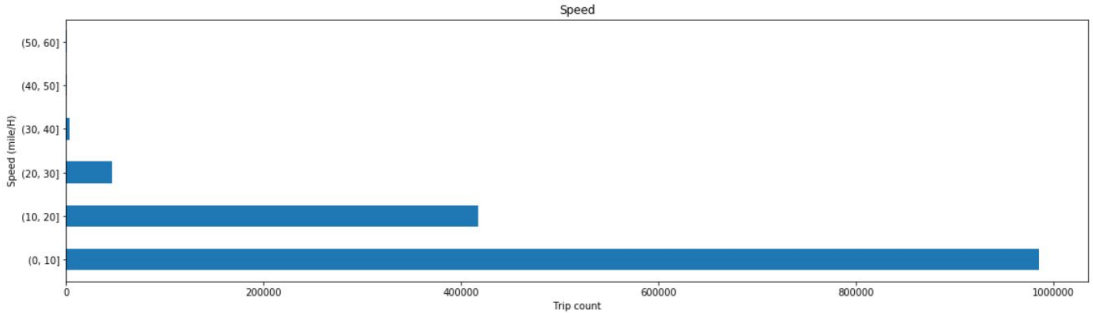
  Speed is a function of distance and time. Let's visualize speed in different trips.

  - Maximum speed limit in NYC is as follows:
    - 25 mph in urban area
    - 65 mph on controlled state highways
  - Many trips were done at a speed of over 125 mile/h. Going SuperSonic..!!

    remove and focus on the trips which were done at less than 65 mile/h as per the speed limits

Exploratory Data Analysis

- speed

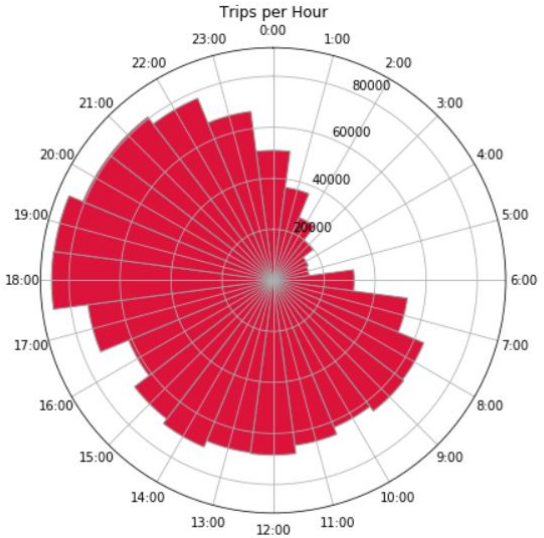


- Trips over 15 miles/h are being considered as outliers but we cannot ignore them because they are well under the highest speed limit of 65 mile/h on state controlled highways.
- Mostly trips are done at a speed range of 6-12 miles/h with an average speed of around 8 miles/h.

Exploratory Data Analysis

- Trip Per Hour

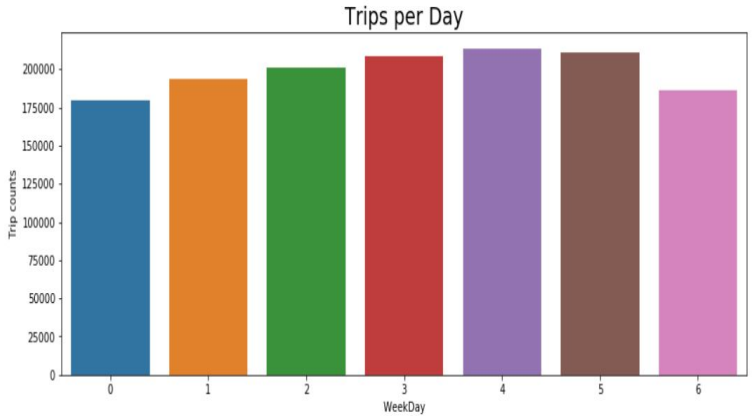
  - It's inline with the general trend of taxi pickups which starts increasing from 6AM in the morning and then declines from late evening i.e. around 8 PM. There is no unusual behavior here.
  - The number of pickup is maximum at 6-7 pm.



Exploratory Data Analysis

- **Total trips per weekday**

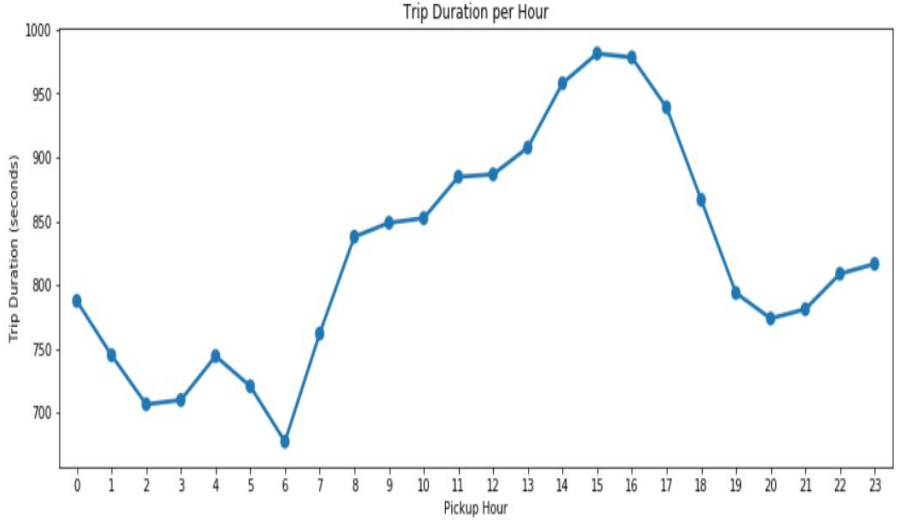
- Here we can see an increasing trend of taxi pickups starting from Monday till Friday. The trend starts declining from saturday till monday which is normal where some office going people likes to stay at home for rest on the weekends.



Exploratory Data Analysis

- **Trip Duration per hour**

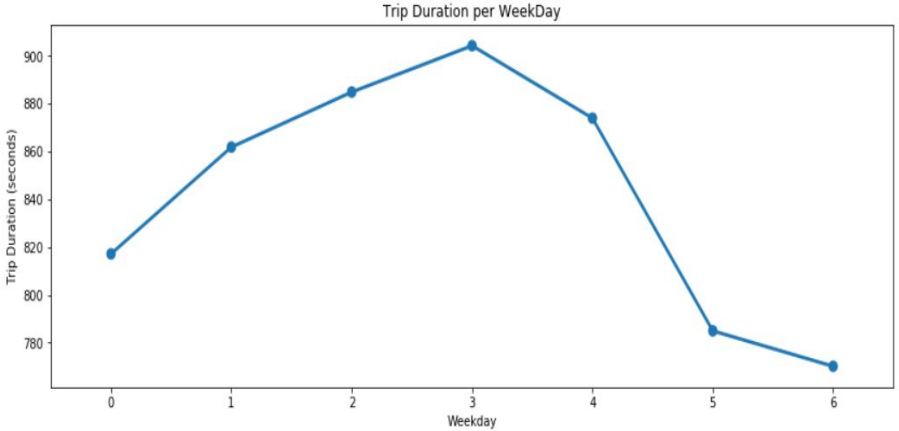
- We need to aggregate the total trip duration to plot it agaist the month. The aggregation measure can be anything like sum, mean, median or mode for the duration. Since we already did the outlier analysis, so we can take the mean to visualize the pattern which should not result in the bias of the general trend.
- Average trip duration is lowest at 6 AM when there is minimal traffic on the roads.
- Average trip duration is generally highest around 3 PM during the busy streets.
- Trip duration on an average is similar during early morning hours i.e. before 6 AM & late evening hours i.e. after 6 PM.



Exploratory Data Analysis

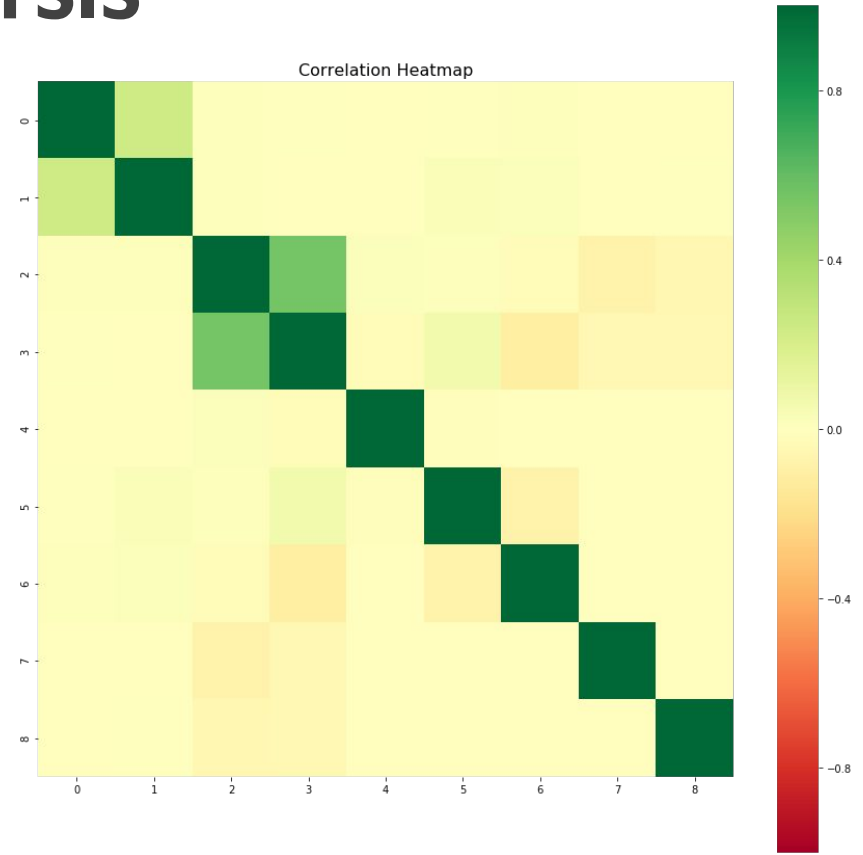
- **Trip duration per WeekDay**

- We can see that trip duration is almost equally distributed across the week on a scale of 0-1000 minutes with minimal difference in the duration times. Also, it is observed that trip duration on thursday is longest among all days.

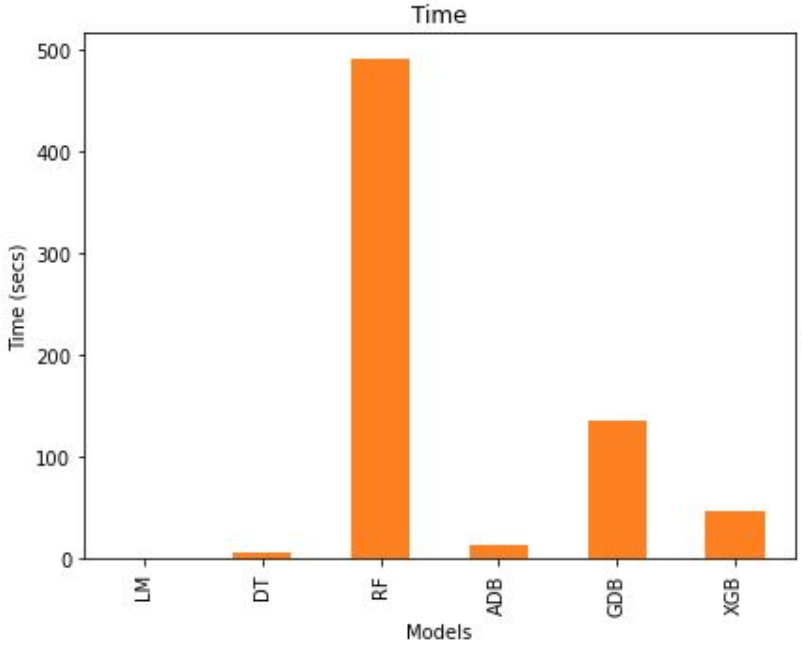


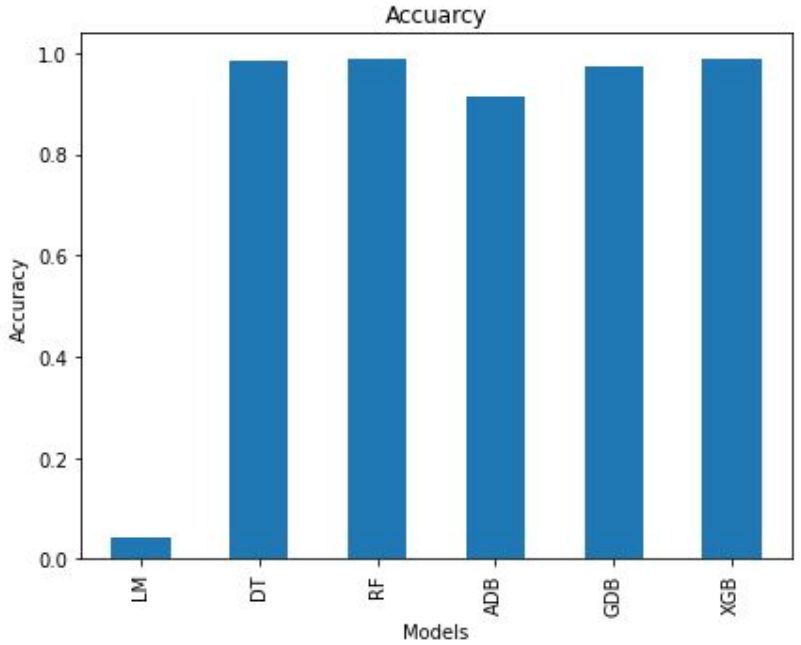
CORELATION ANALYSIS

- Some combinations of features shows slight correlation.
- But most of the features shows no correlation
- There is no negative correlation



MODELING


Time
Time (secs)
500
400
300
200
100
0
LM
DT
RF
ADB
GDB
XGB
Models

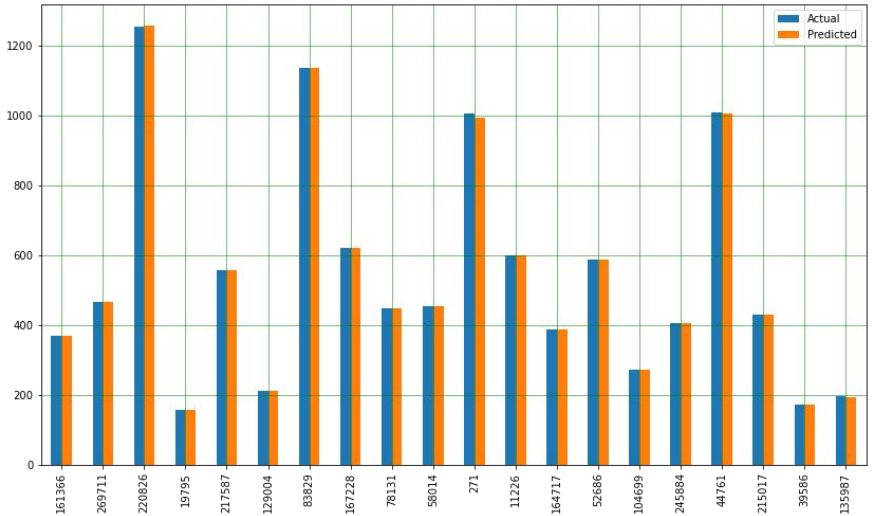

Accuarcy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
LM
DT
RF
ADB
GDB
XGB
Models

BEST MODEL

- DECISION TREE

```
1 predictions
```

	Actual	Predicted
0	1040	1037.0
1	830	831.0
2	614	615.0
3	867	867.0
4	4967	4950.0
...	...	...
291724	1303	1301.0
291725	1351	1357.0
291726	857	854.0
291727	535	535.0
291728	1530	1532.0




```
1 dt_score = r2_score(y_test, trips)
2 print(dt_score)
```

```
0.985774051284175
```

THANK YOU