SUBJECT: ENGINEERING OF BIG DATA SYSTEMS

FINAL PROJECT REPORT

PRESENTED BY: KRISH PRAVIN JAIN NU ID: 001881885

TOPIC: New York City Taxi and Limousine Commission (TLC) Trip Record Data

DESCRIPTION:

NYC Taxi & Limousine Commission data is available for yellow, green and FHV taxi data set, this data set is freely available for analysis. This taxi records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. I have taken the dataset of Yellow Taxi to perform the analysis.

LINK TO THE DATASET:

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

TECHNOLOGY USED:

- Hadoop Map Reduce
- Pig
- Mahout
- Hive
- Amazon Elastic Map Reduce
- Tableau for data visualization

DICTIONARY:

VendorID	1= Creative Mobile Technologies, LLC 2= VeriFone Inc
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.
Trip_distance	The elapsed trip distance in miles reported by the taximeter
PULocationID	Pickup location
DOLocationID	Drop location
RateCodeID	1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride

Store_and_fwd_flag	Y= store and forward trip N= not a store and forward trip
Payment_type	1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown

	6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Surcharges tax etc
MTA_tax	Tax
Improvement_surcharge	The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers

ANALYSIS PERFORMED:

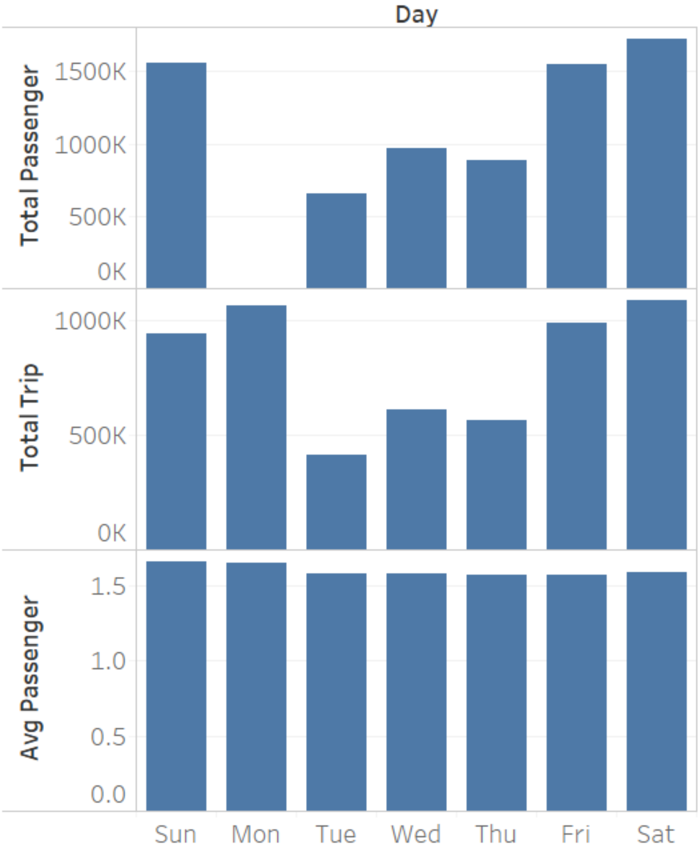
- **Calculate the Average Number of Passenger travelling by Taxi month wise and day wise**
- **Calculate the Average Trip Distance of taxi vendor day wise**
- **Analyze the Fare collected by taxi per day of week for past months**
- **Analyze the day wise Surcharge collected**
- **Analyze the total Fare distribution**
- **Find out the distinct pickup location**
- **Find the top 10 busy pickup location**
- **Analyze the density of taxi during Holiday vs Normal routine day**
- **Analyze the mode of payment passenger prefer the most to pay the fare**
- **Calculate the average speed of taxi to see how the traffic affects the speed for each hour of the day**
- **Analyze the density of taxi during Morning Noon Evening and Night**
- **Analyze the trend of Taxi for weekday vs weekend**
- **Generate bins to classify the trips according to the rate card and find the top 10 rides on basis of the total fare**
- **Give prediction/recommendation about distance and total fare**
- **Give prediction/recommendation about distance and total tip**
- **Calculate the total number of round trips using hive**
- **Calculate the total payment percentage using hive**
- **Calculate the total payment month wise using hive**
- **Find out the top 10 popular drop location using pig**
- **Find max trip for popular pickup and drop location using pig**

ANALYSIS DESCRIPTION:

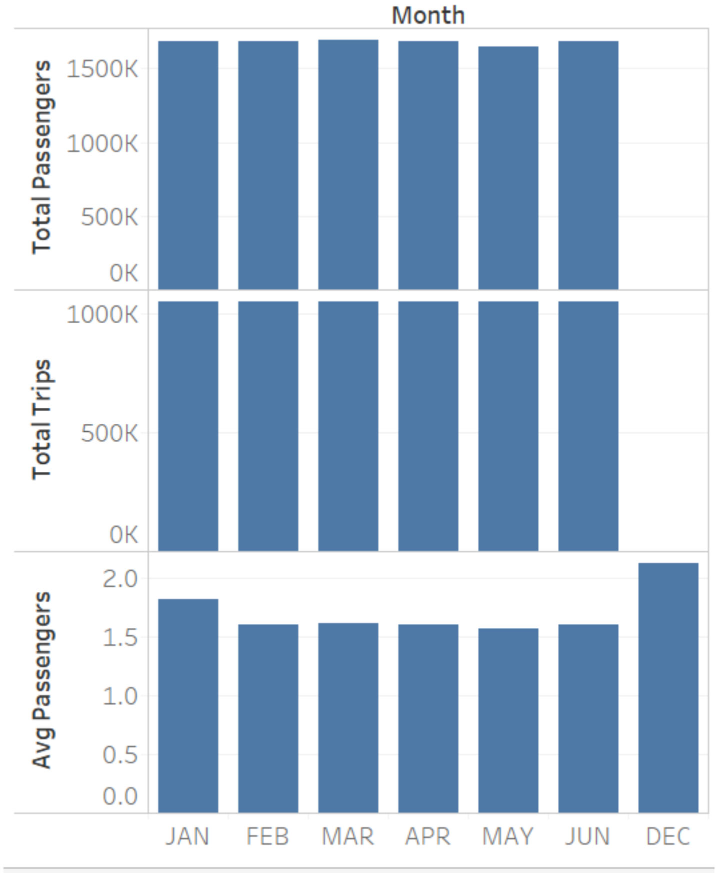
1) Calculate the Average Number of Passenger travelling by Taxi month wise and day wise

- The analysis performed to identify the average number of passengers travelling by Taxi. The analysis was performed to observe the trend of passengers in each month as well as for each day. The main aim of this analysis was to identify the density of passengers travelling on each day.
- I have implemented the Map Reduce job using Hadoop and have taken the advantage of Combiner implementation by making the average an associative and commutative operation by multiplying the count to the average and add it to the running sum.

AvgPassDay

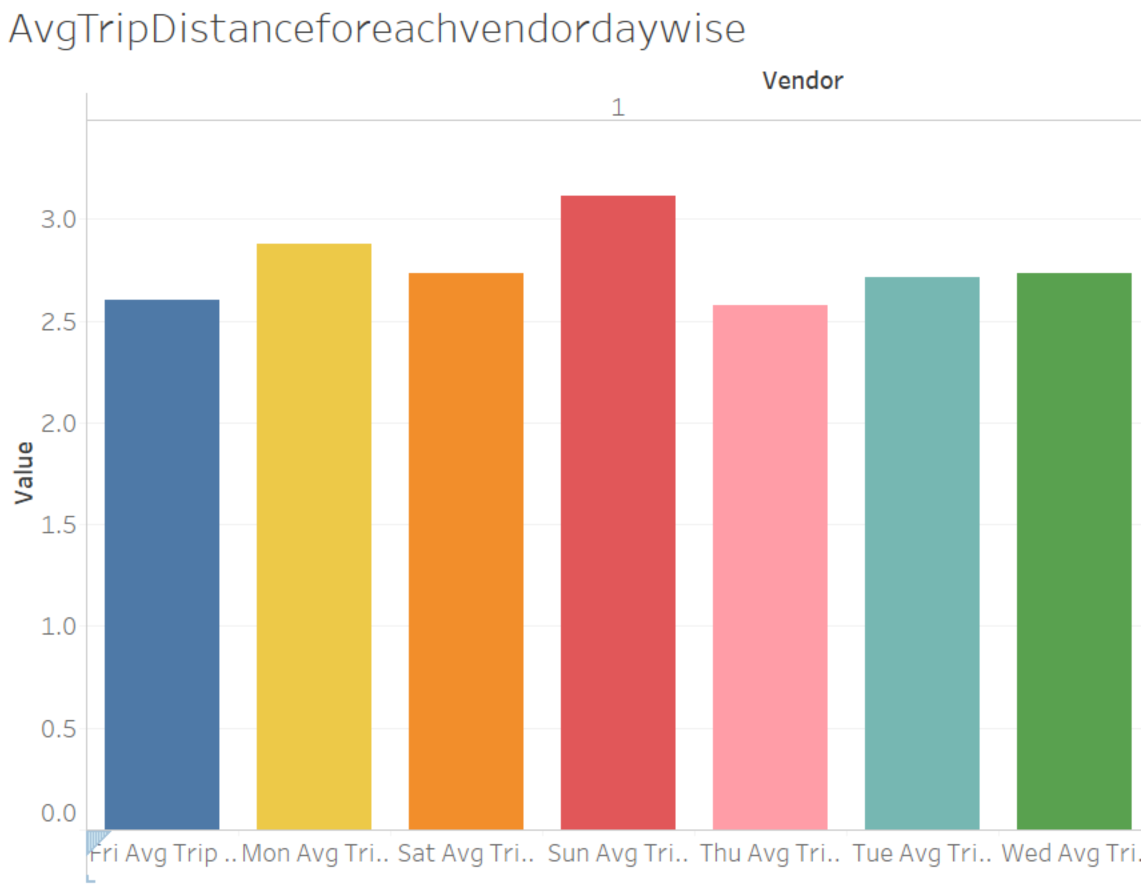


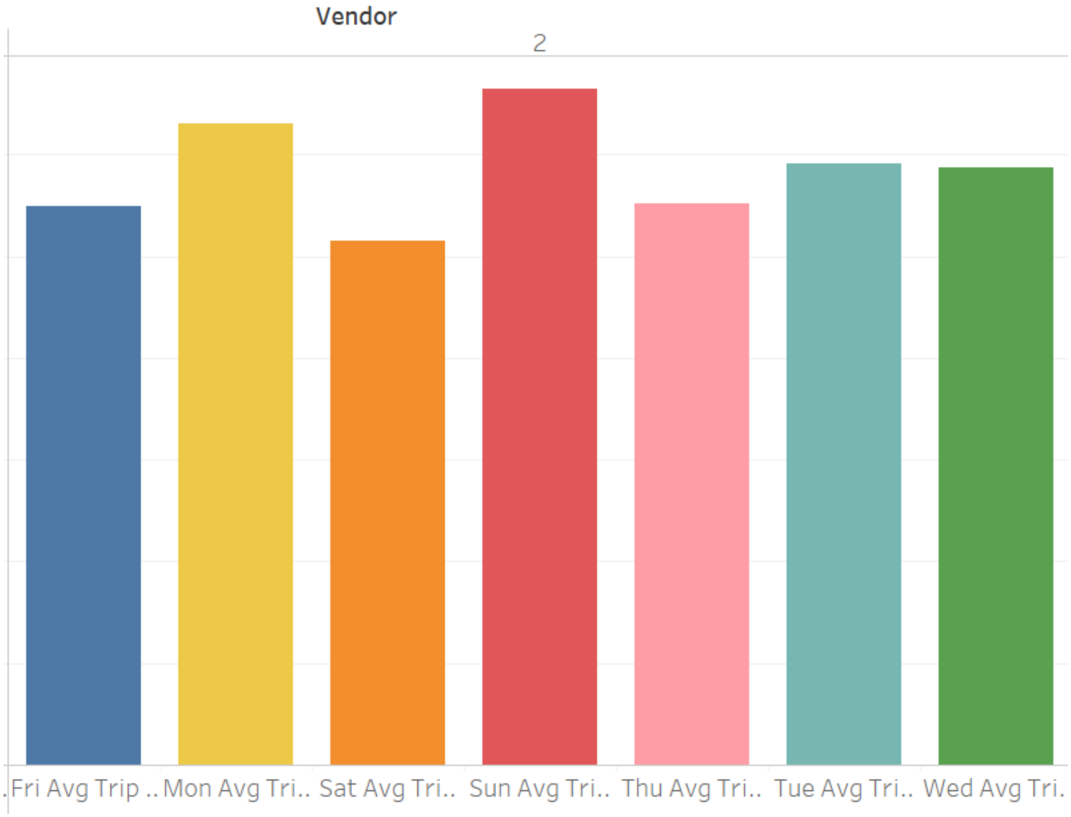
AvgPassMonth



2) <u>Calculate the Average Trip Distance of taxi vendor day wise</u>

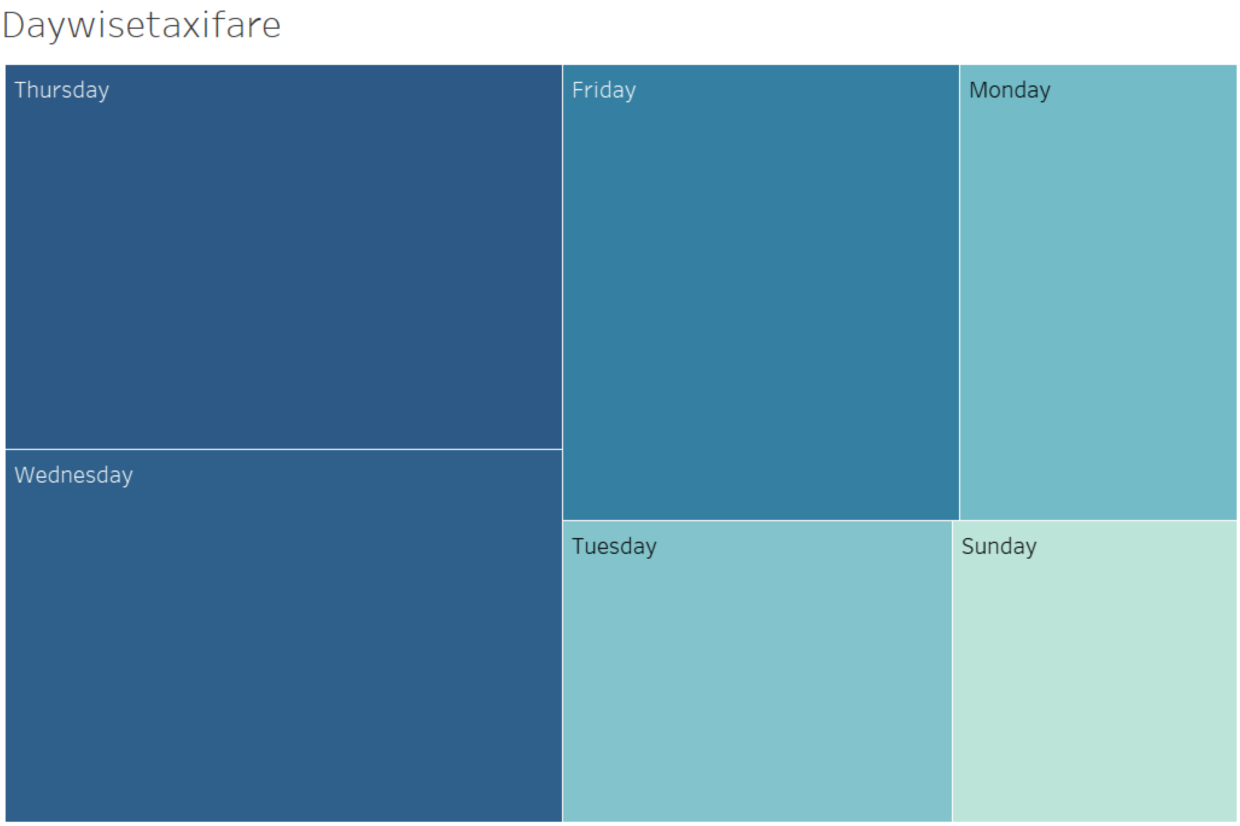
- The analysis is performed to give the Taxi vendor a brief overview of the demand of taxi people have and the distance they cover on each day. This will help them to predict the days where they have to increase the amount of taxi and gain more profit
- I implemented the Map Reduce Job using Hadoop





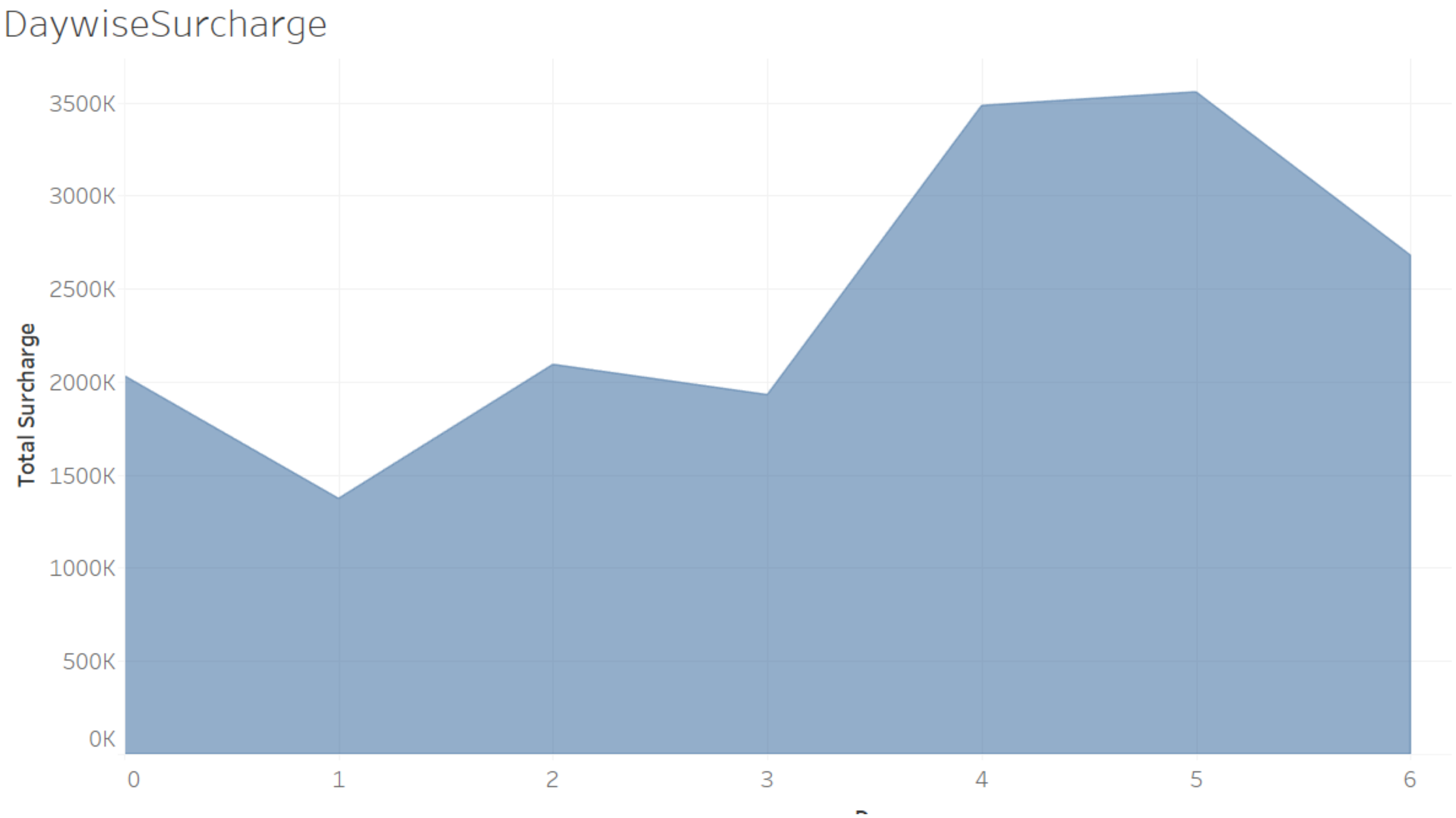
3) Analyze the Fare collected by taxi per day of week for past months

- This analysis is performed using Map Reduce framework in order to analyze the days which is most busiest and profitable to the vendor.
- The total fare collected on weekdays is more than weekends. The weekday obviously sees a lot of passenger travelling in taxi to their respective location.



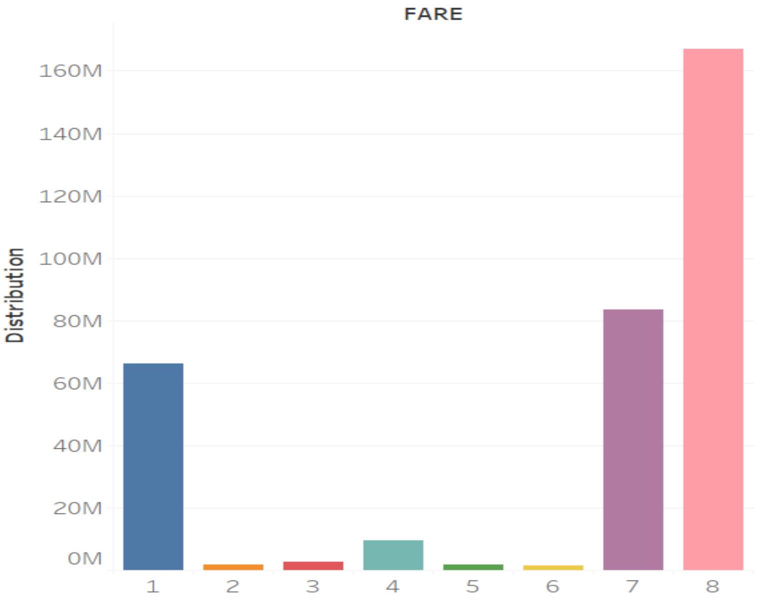
4) **Analyze the day wise Surcharge collected**

- To perform this analysis, I used Partitioning Pattern and divided the data into 7 partitions one for each day of the week respectively. After partitioning, the total surcharge for the entire dataset on that particular day of the week has been calculated from the respective partitions. I have made use of Custom writable class, Partitioner class as well as Identity Mapper and reducer along with Mapper and Reducer. I have tried implementing chaining of map reduce jobs here.
- The surcharge is more on weekends than on weekdays



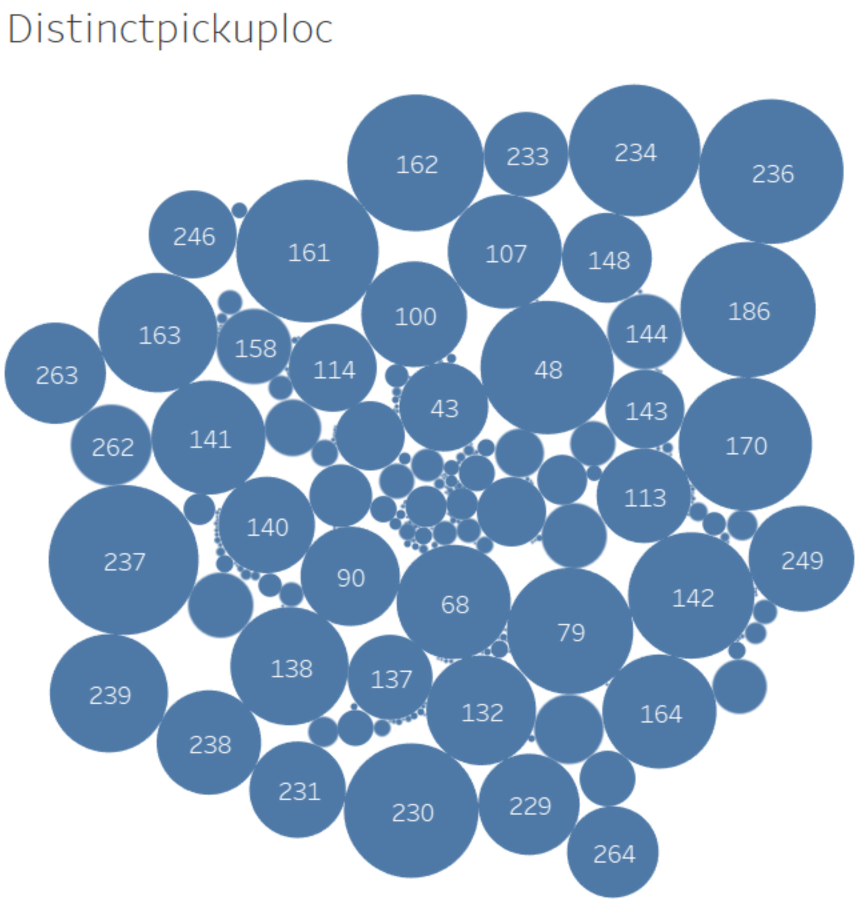
5) **Analyze the total Fare distribution**

- This analysis is done to analyze the total amount collected at the end of the ride. There are various taxes and surcharges which add up to the ride amount and tip to sum up to total amount. This analysis gives an overview about the fare distribution. I have implemented it using Map Reduce Hadoop Job.



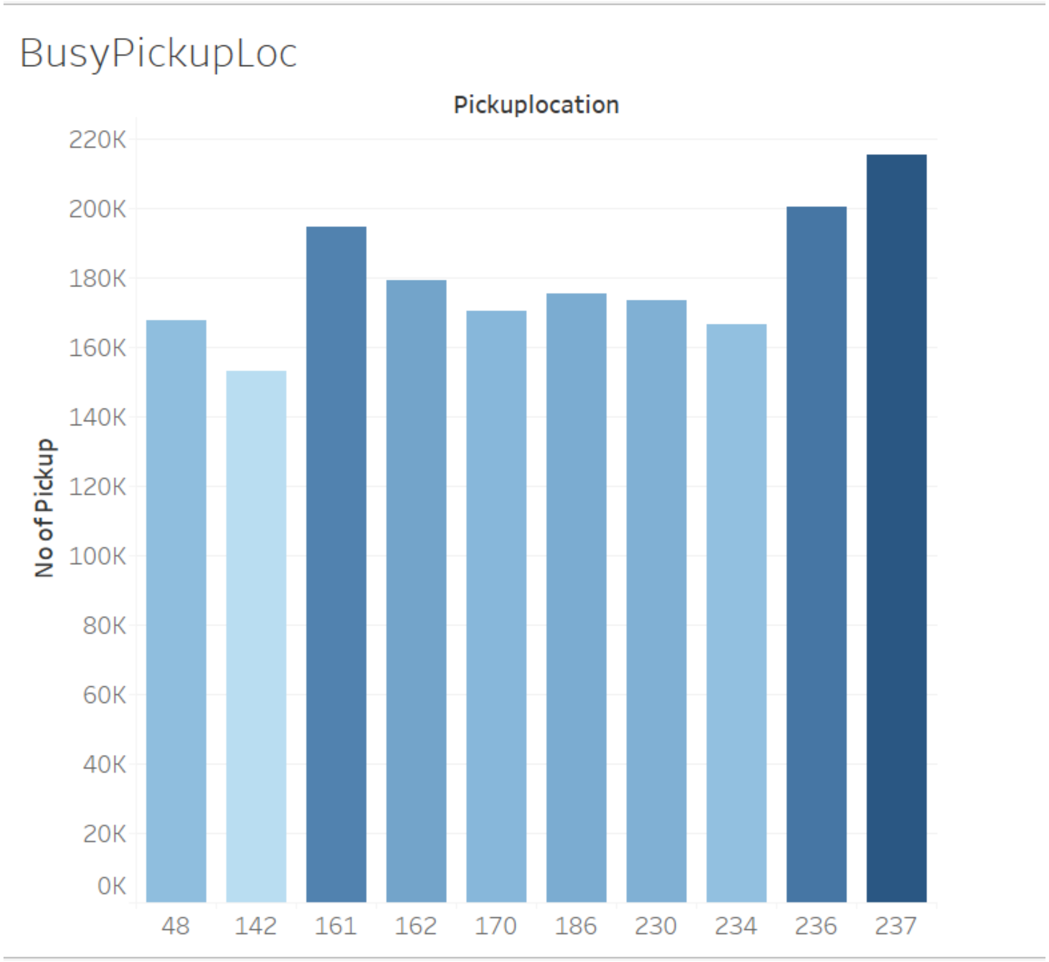
6) Find out the distinct pickup location

- It is always important to find the distinct locations where the taxi is providing service. It helps the vendor to know which other locations the vendor can start the taxi service.
- I have implemented this analysis using the Distinct Pattern of map reduce framework.



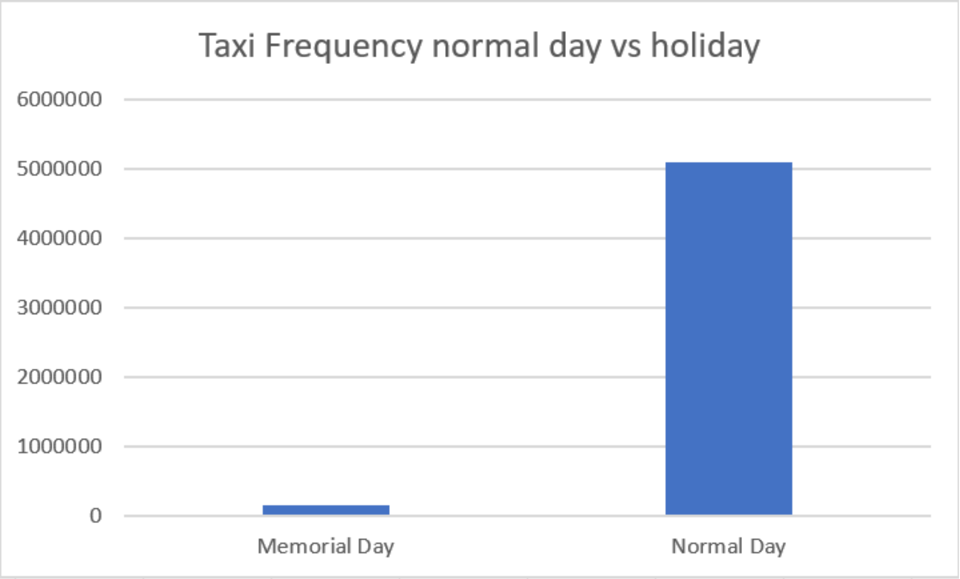
7) Find the top 10 busy pickup location

- I have implemented this analysis by performing the map reduce job to find the number of rides which starts from every location and than have used the TOP 10 Pattern to give me the top most pickup location.
- This analysis gives the brief overview of the busy pickup location which is useful in order to increase the number of taxi services there to fulfil the demand.



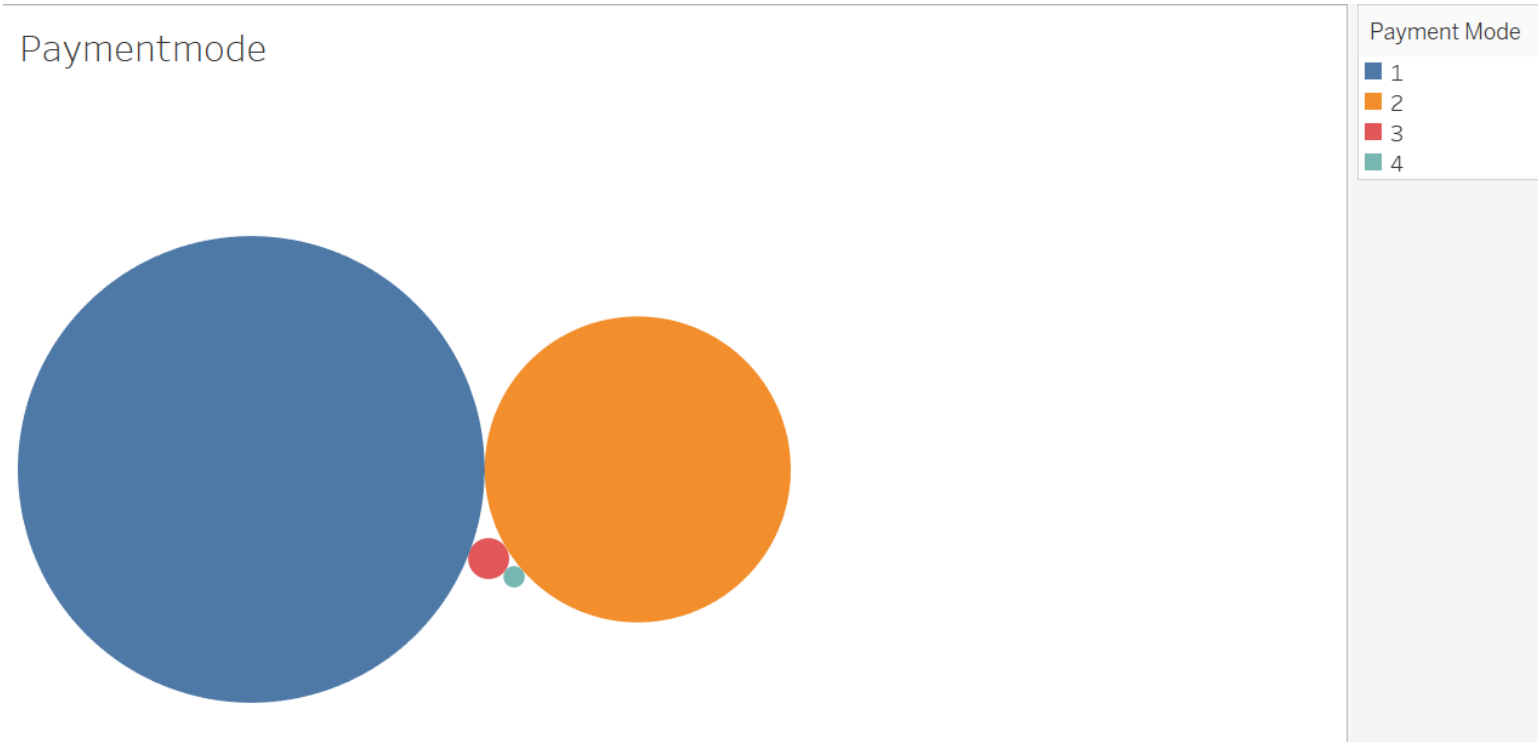
8) **Analyze the density of taxi during Holiday vs Normal routine day**

- This analysis gives an insight of the taxi frequency on a normal day vs on a holiday. I have implemented this analysis using Hadoop Map Reduce framework.



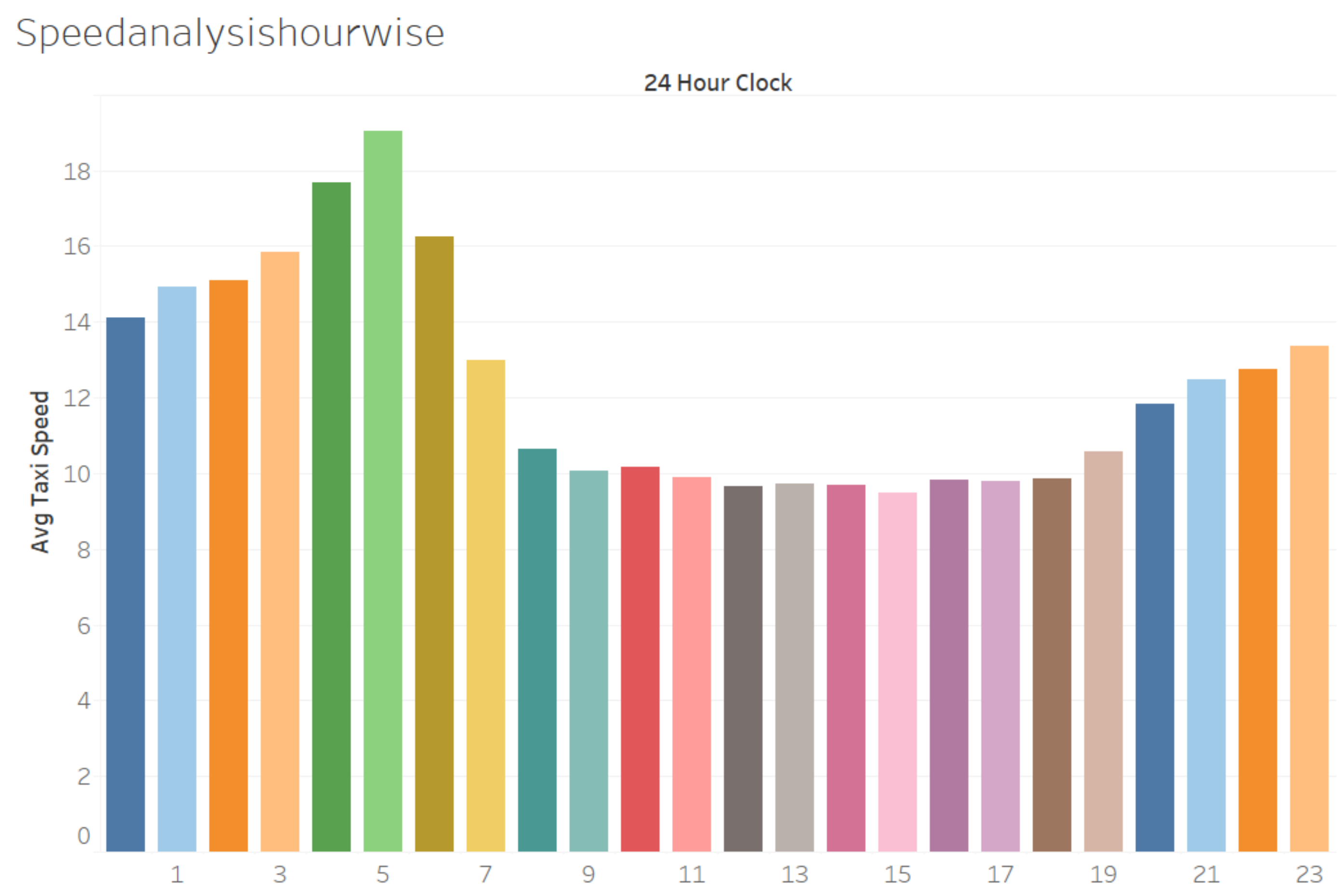
9) Analyze the mode of payment passenger prefer the most to pay the fare

- The nyc taxi has the following mode of payments: 1= Credit card 2= Cash 3= No charge 4= Dispute
- This analysis is implemented to analyze the most frequent mode of payment the passengers use to pay for their ride.
- Passengers pay highest by Credit card followed by Cash. There have been very few incidents where the ride was not charged or there was dispute regarding the fare.



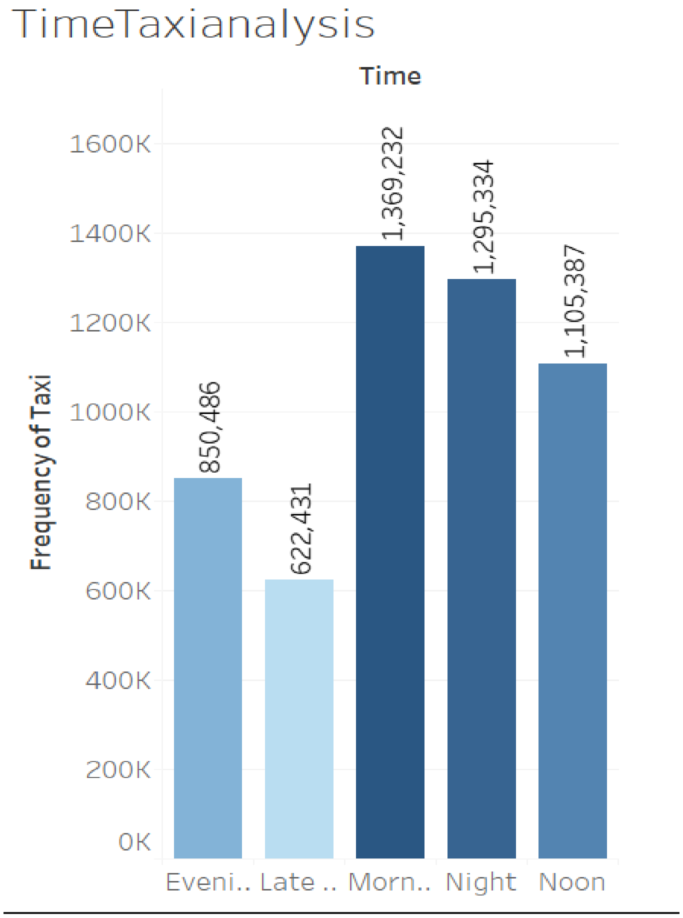
10) Calculate the average speed of taxi to see how the traffic affects the speed for each hour of the day

- This analysis is an interesting analysis which is helpful to view at what average speed the taxi runs during each hour of the day. This will give us an idea about the traffic during the day and night time we can say cause of the speed the taxi runs to some extent.



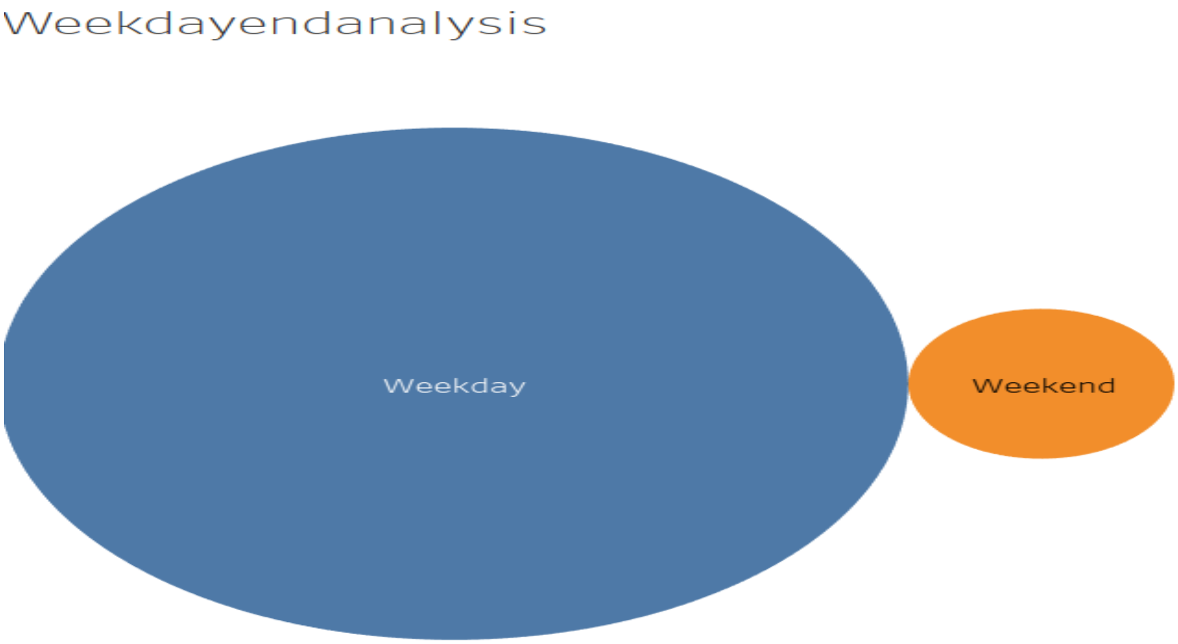
11) Analyze the density of taxi during Morning Noon Evening and Night

- This analysis is done to see the demand of taxi during each time of the day. I have implemented the logic by categorizing the time into 5 segments namely morning, noon, evening, night and late night.
- We can see that the demand of taxi is more during office pickup and drop hours in comparison to late night.



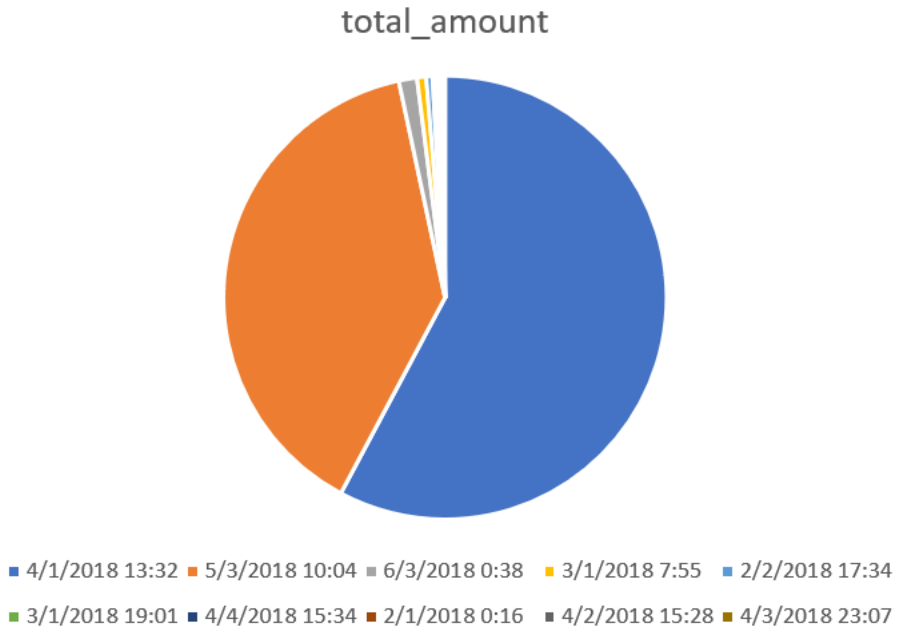
12) <u>Analyze the trend of Taxi for weekday vs weekend</u>

- This analysis gives an insight about the demand of taxi during weekdays and weekends.
- We can see that there is a huge demand of taxi during weekdays in comparison with weekends.



13) <u>Generate bins to classify the trips according to the rate card and find the top 10 rides on basis of the total fare</u>

- To implement this analysis, I used the Binning Data organization technique to classify the entire dataset into bins according to various rate card. I further used the Top 10 design pattern to display the Top most ride on basis of total fare amount from all rate card category.



14) **Give prediction/recommendation about distance and total fare**

- I have implemented this analysis using the Mahout Machine learning framework to give recommendation/prediction about what will be the estimated total fare for a particular distance. It will be helpful for the passenger to have a rough estimate about the total fare for their trip.

OUTPUT:

```
For Yellow Taxi When trip distance is 4 meters the estimated fair is 3.2(Tolls are not included)
For Yellow Taxi When trip distance is 4 meters the estimated fair is 3.1494684(Tolls are not included)
For Yellow Taxi When trip distance is 4 meters the estimated fair is 3.1335955(Tolls are not included)
```

15) **Give prediction/recommendation about distance and total tip**

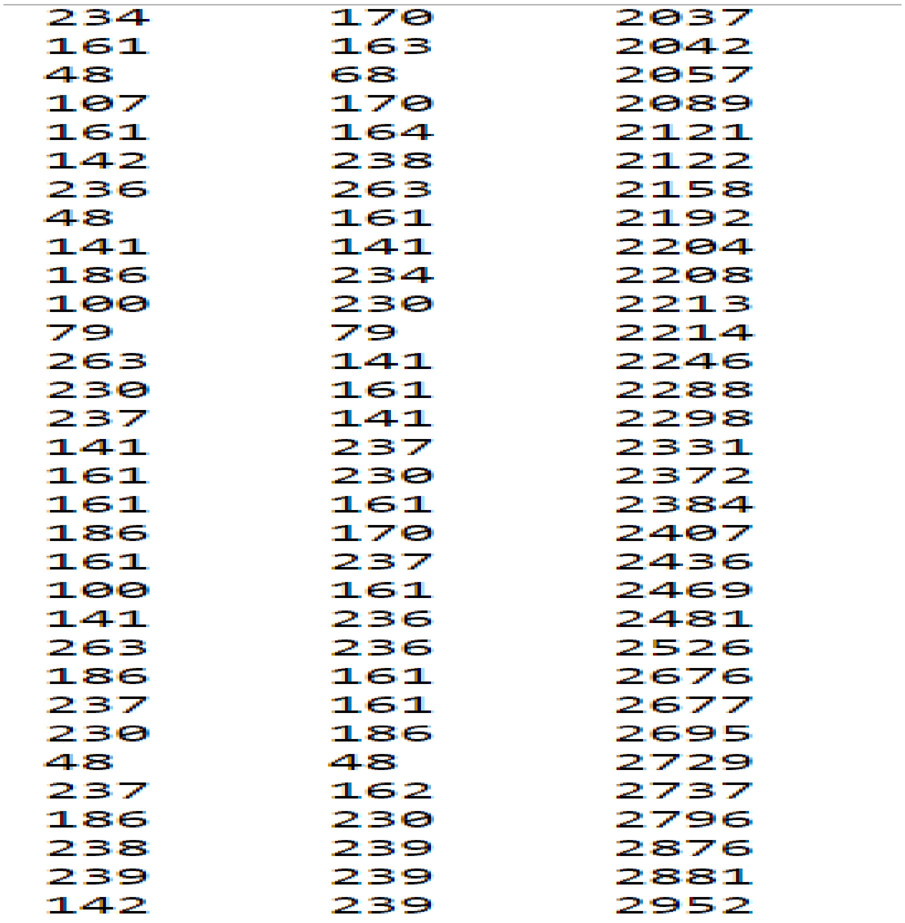
- I have implemented this analysis using the Mahout Machine learning framework to give recommendation/prediction about what will be the estimated total tip for a particular distance ride. It will be helpful for the passenger to have a rough estimate about the total tip for their trip and the driver will also know about the estimated tip he can expect for the trip based on the total distance of the ride.

OUTPUT:

```
For Yellow Taxi When trip distance is 30 meters the popular tip is around 2.0
```

16) **Calculate the total number of round trips using hive**

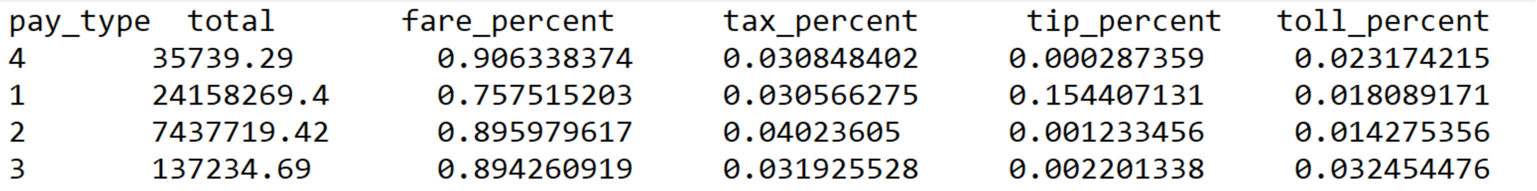
- This analysis is done to count the total round trip from Pickup location to Drop location.
- I have implemented this analysis using HIVE query and Amazon Web Service Elastic Map Reduce technique
- I stored my data in table and used hive query to retrieve the result



```
234      170      2037
161      163      2042
48       68       2057
107      170      2089
161      164      2121
142      238      2122
236      263      2158
48       161      2192
141      141      2204
186      234      2208
100      230      2213
79       79       2214
263      141      2246
230      161      2288
237      141      2298
141      237      2331
161      230      2372
161      161      2384
186      170      2407
161      237      2436
100      161      2469
141      236      2481
263      236      2526
186      161      2676
237      161      2677
230      186      2695
48       48       2729
237      162      2737
186      230      2796
238      239      2876
239      239      2881
142      239      2952
```

17) Calculate the total payment percentage using hive

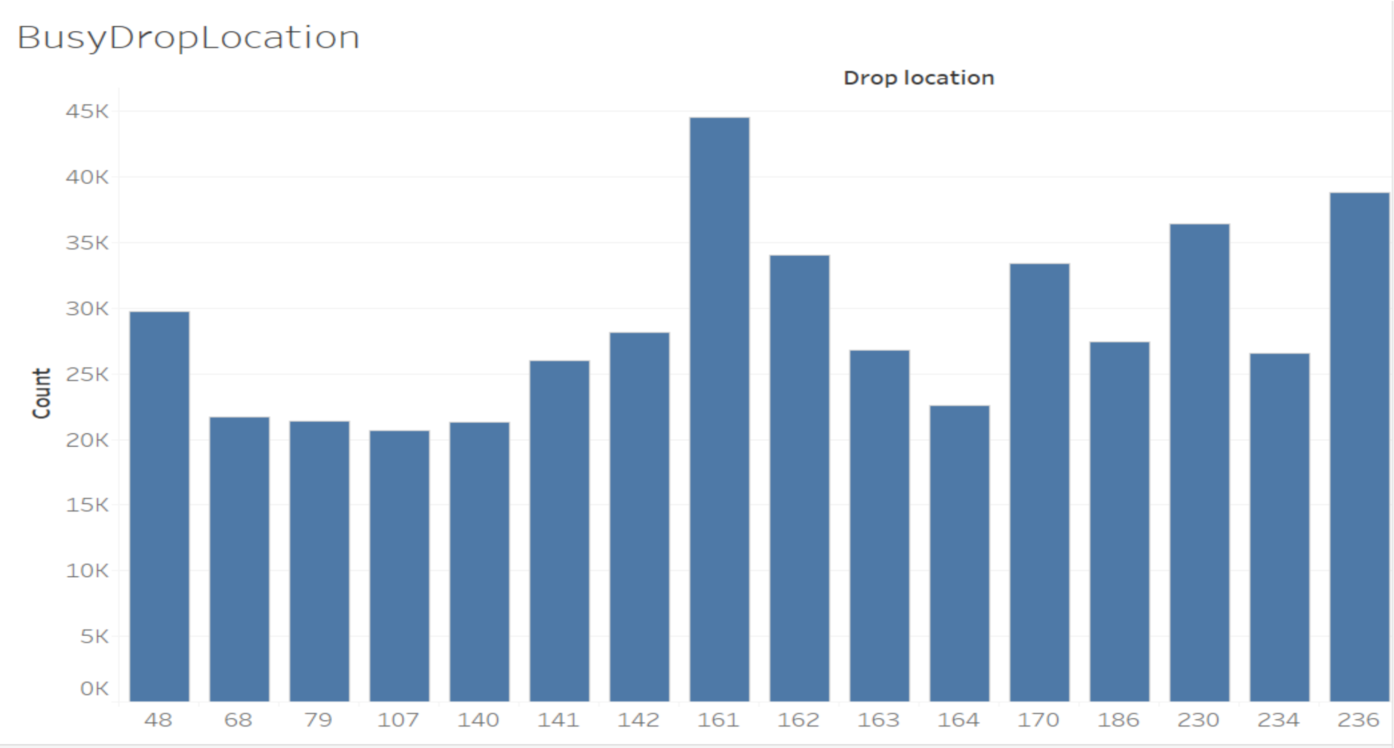
- This analysis is done to count the total round trip from Pickup location to Drop location.
- I have implemented this analysis using HIVE query and Amazon Web Service Elastic Map Reduce technique
- I stored my data in table and used hive query to retrieve the result



```
pay_type  total         fare_percent       tax_percent        tip_percent      toll_percent
4        35739.29         0.906338374      0.030848402      0.000287359      0.023174215
1        24158269.4       0.757515203      0.030566275      0.154407131      0.018089171
2        7437719.42       0.895979617      0.04023605       0.001233456      0.014275356
3        137234.69        0.894260919      0.031925528      0.002201338      0.032454476
```

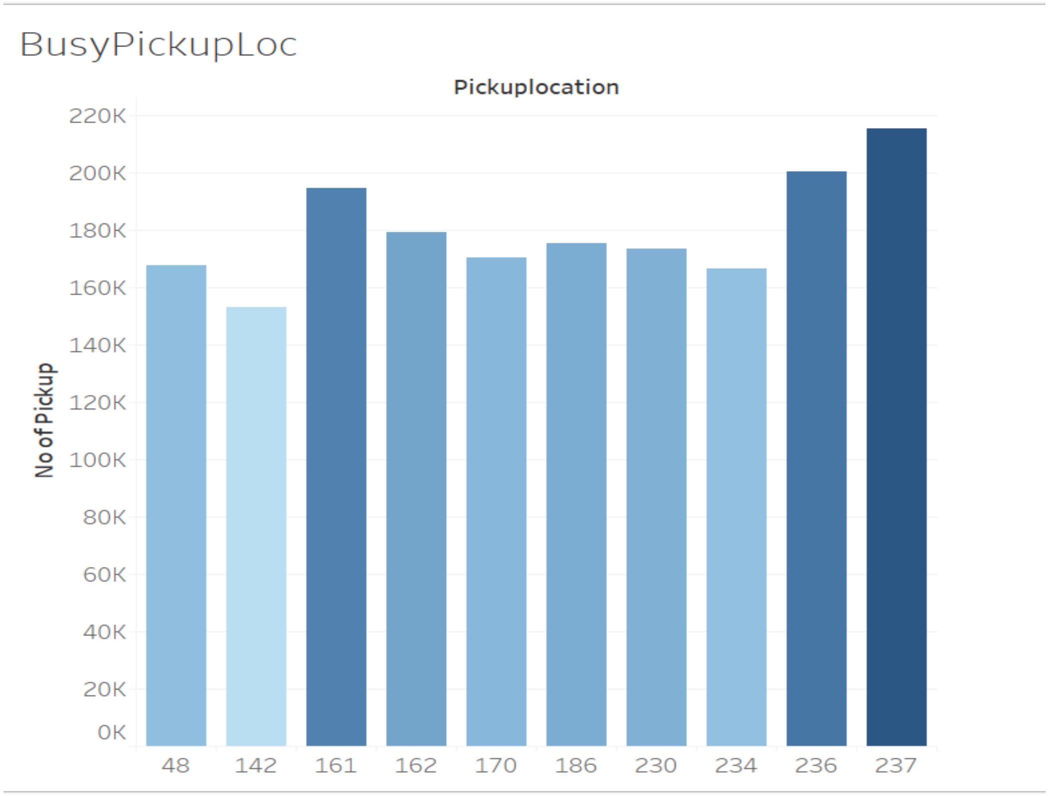
18) Find out the top 10 popular drop location using pig

- This analysis is done to find the most popular drop location using Pig query



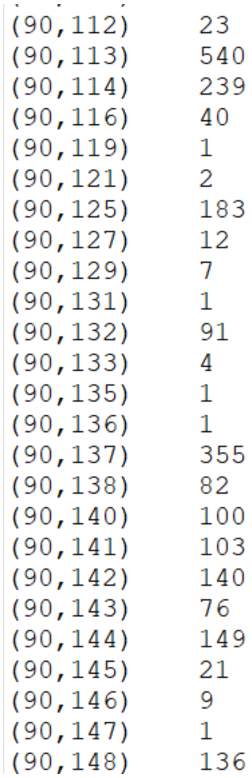
19) Find out the top 10 popular pickup location using pig

- This analysis is done to find the most popular pickup location using Pig query



20) Find max trip for popular pickup and drop location using pig

- This analysis is done to find the max number of round trips for pickup and drop location using Pig query



```
(90,112)     23
(90,113)     540
(90,114)     239
(90,116)     40
(90,119)     1
(90,121)     2
(90,125)     183
(90,127)     12
(90,129)     7
(90,131)     1
(90,132)     91
(90,133)     4
(90,135)     1
(90,136)     1
(90,137)     355
(90,138)     82
(90,140)     100
(90,141)     103
(90,142)     140
(90,143)     76
(90,144)     149
(90,145)     21
(90,146)     9
(90,147)     1
(90,148)     136
```

CONCLUSION:

The entire nyc yellow taxi dataset was implemented successfully using various Big Data Techniques to analyze and produce an overall big picture about the nyc taxi data set.