MAHINDRA FIRST CHOICE CAPSTONE

< Nishant Pandey>

PROBLEM DEFINITION

1. Identifying the ownership pattern of cars at state and city level to be able to understand the purchasing power of the customers

2. Identify the type of order of each state or city receives and any seasonality associated to it.

3. Classify the customers into Segments and device a classification model that would classify the customers into the designated segments.

DATA SET

The dataset provided had the following csv files:

- Customer_Data
- Final_invoice
- JTD

The dataset was not preprocessed and needed cleaning had to be able to bring out insights from it.

DATA PREPARATION

The steps performed for cleaning the dataset

- Removing columns with high percent of null values.
- Removing columns such as CGST, IGST etc which are not relevant to our analysis.
- Removing columns with high percent of junk values (such as CITY in invoice data table)

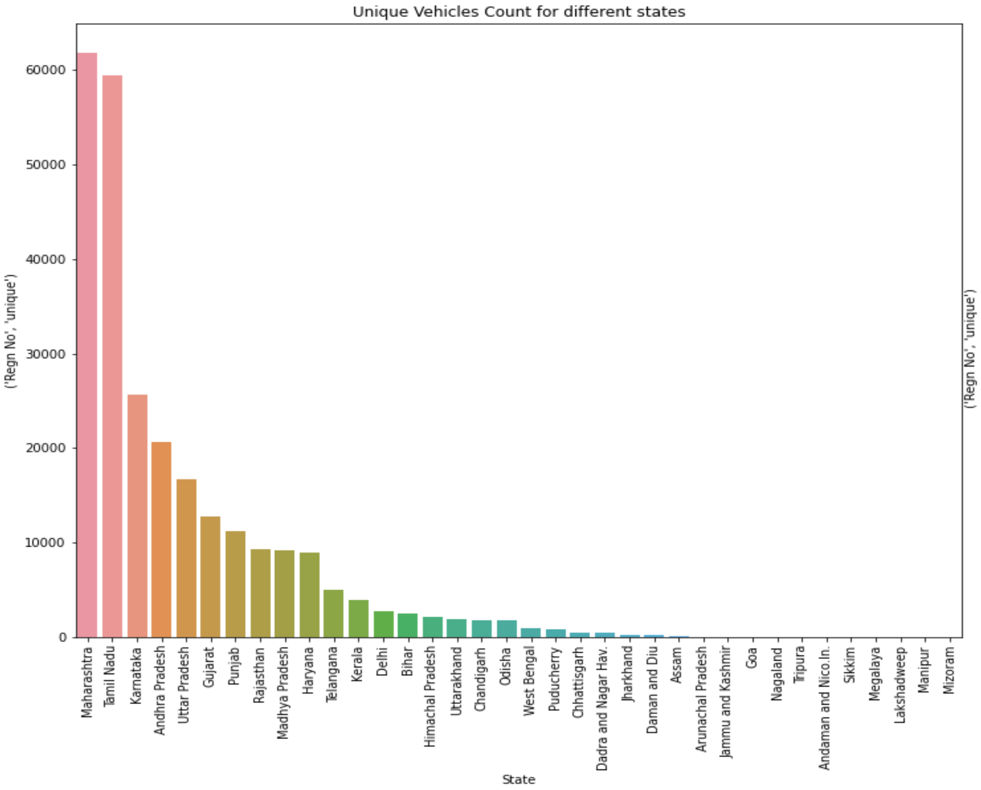
After Cleaning the dataset , imputation of missing values for the remaining columns was done using **SimpleImputer**.

FEATURE ENGINEERING

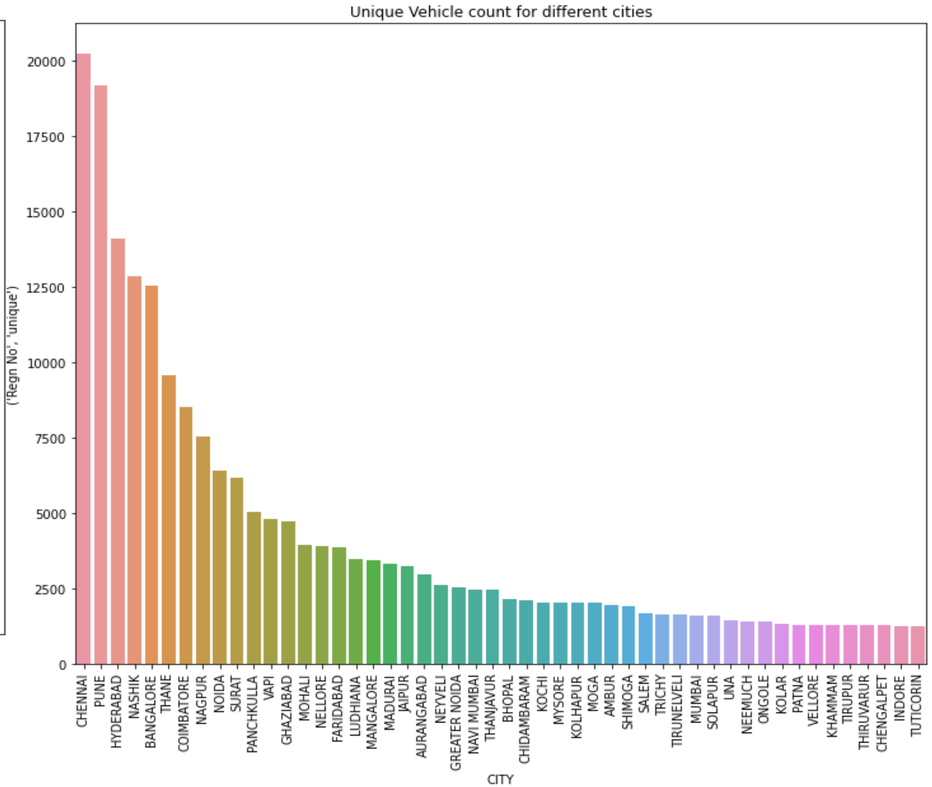
In Final Invoice table:

- Added a column **CITY** , copying the records of Plant Name1 and mapped all the different plant names to a set of cities near them , to make the City wise analysis easy
- Added a column **Time Taken to Service** that gives the time taken to process the service for all the records.
- Column(**Month Of Service** and **Year Of Service**) was added as a separate feature.

DEMOGRAPHIC DISTRIBUTION OF VEHICLES

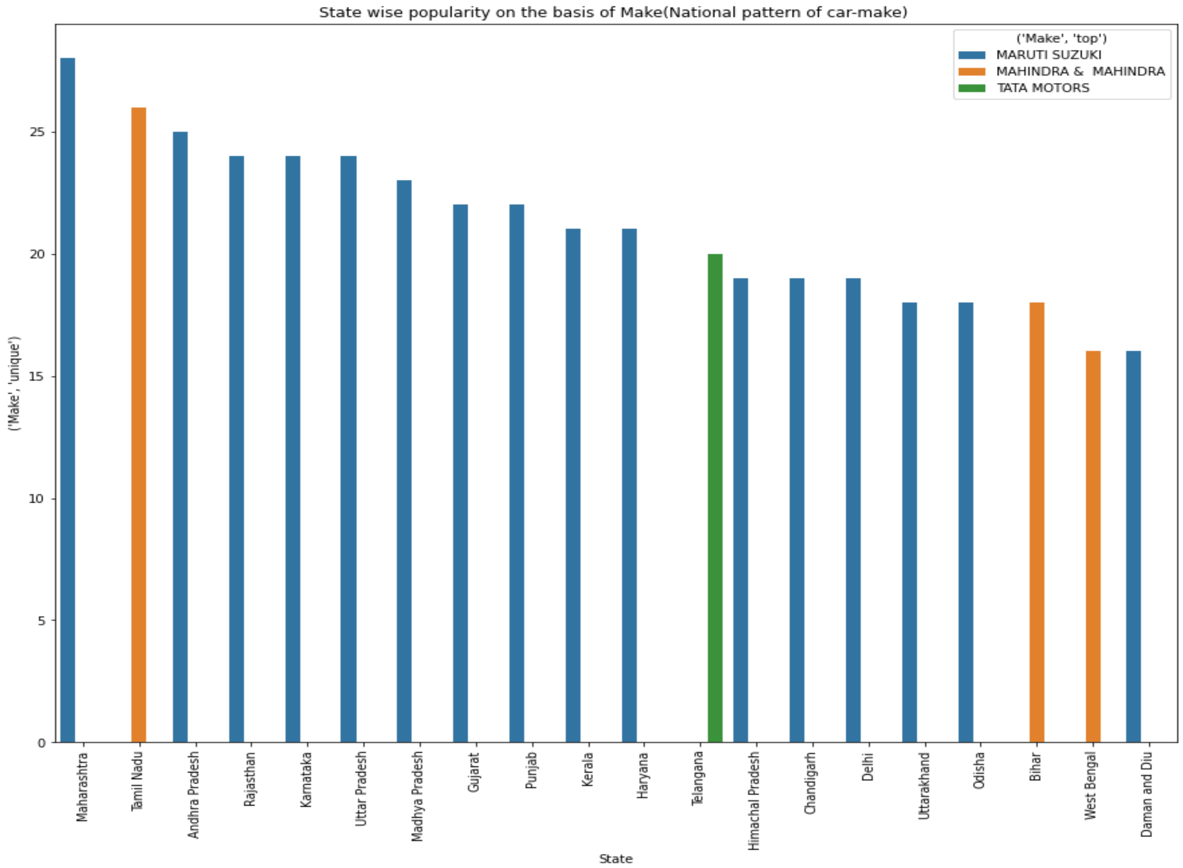


Maharashtra has largest number of vehicles



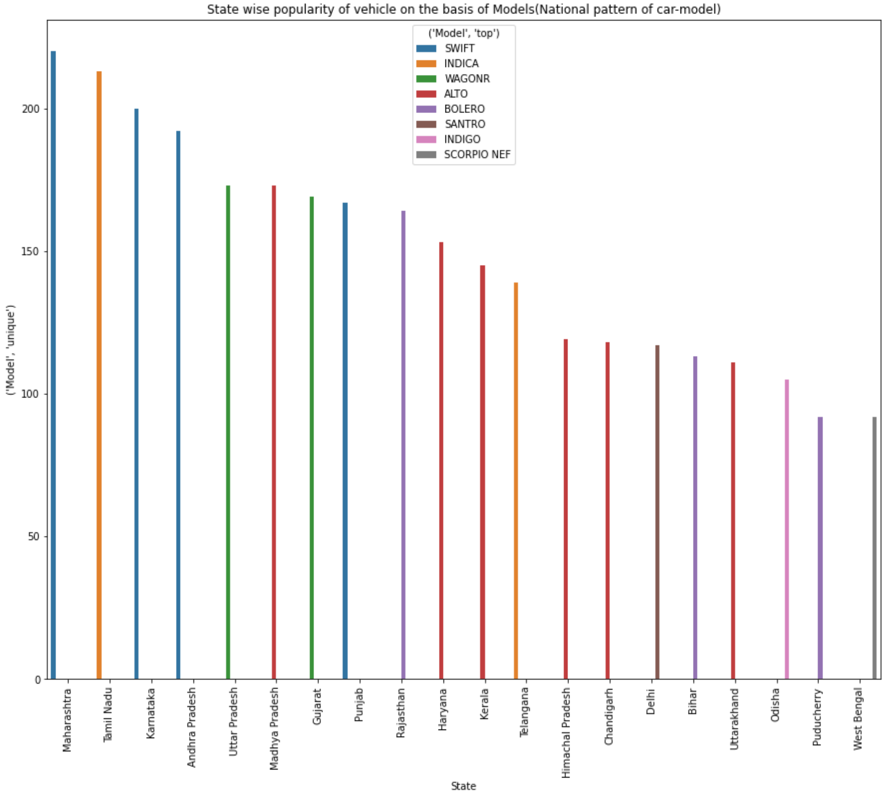
Chennai as a City has largest number of vehicles in the country. As per the data , Mumbai does not figure in the top list

DEMOGRAPHIC DISTRIBUTION OF VEHICLES

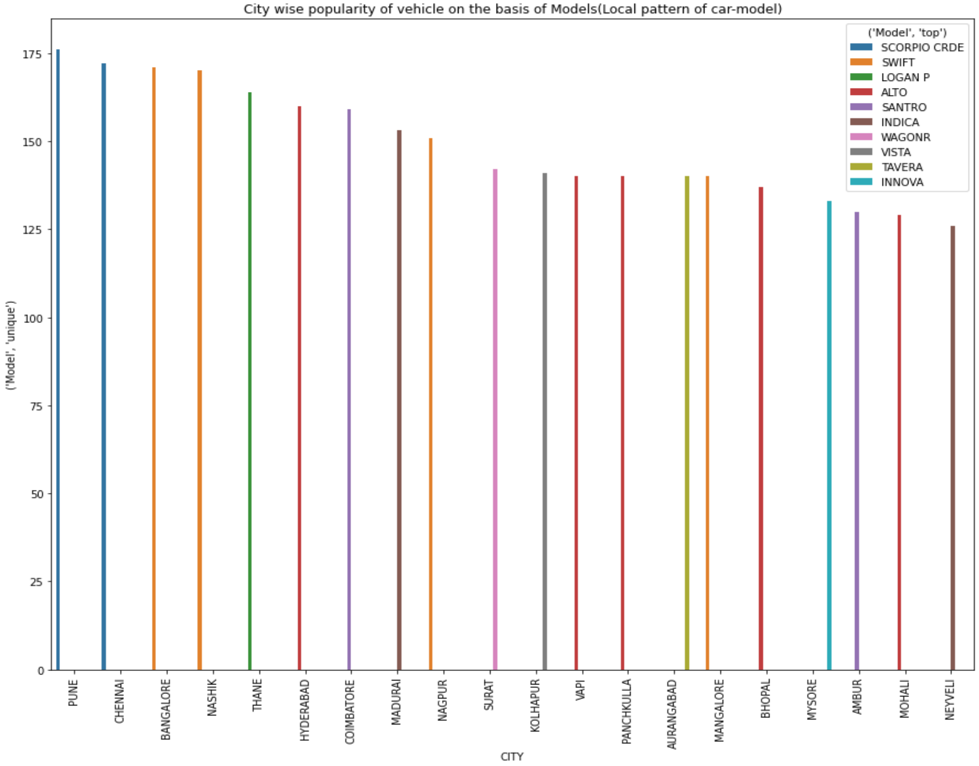


Maruti Suzuki dominates the market in most of the states. Higher expertise in Maruti is expected.

DEMOGRAPHIC DISTRIBUTION OF VEHICLES



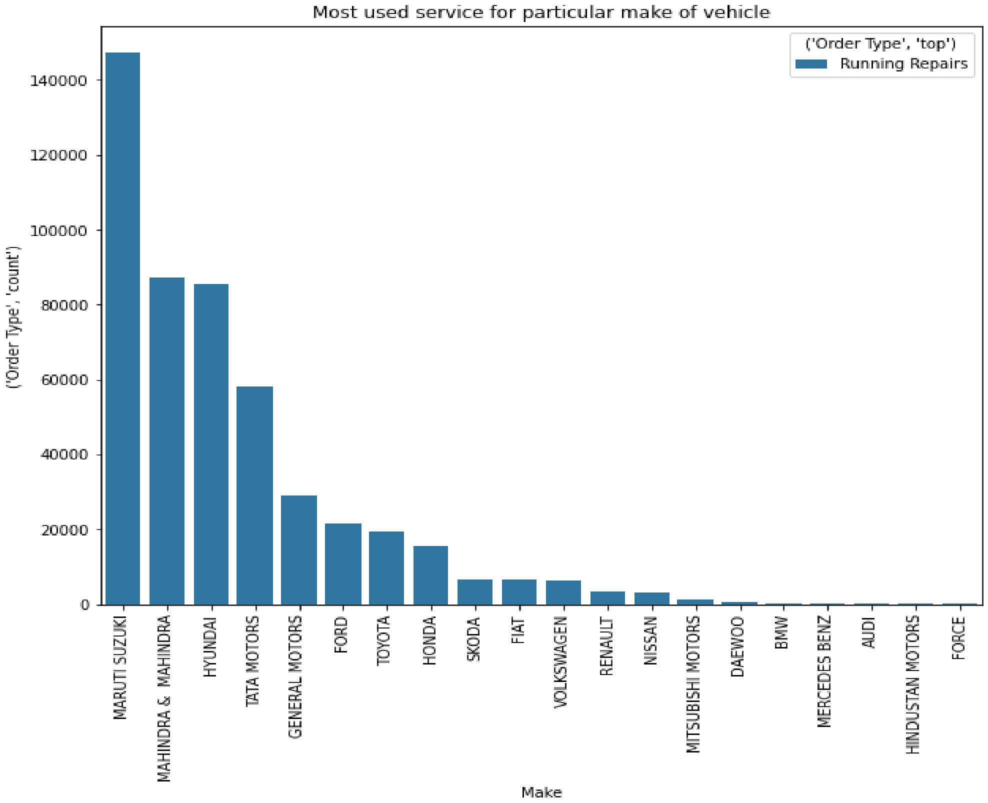
Swift is the most popular car in MH

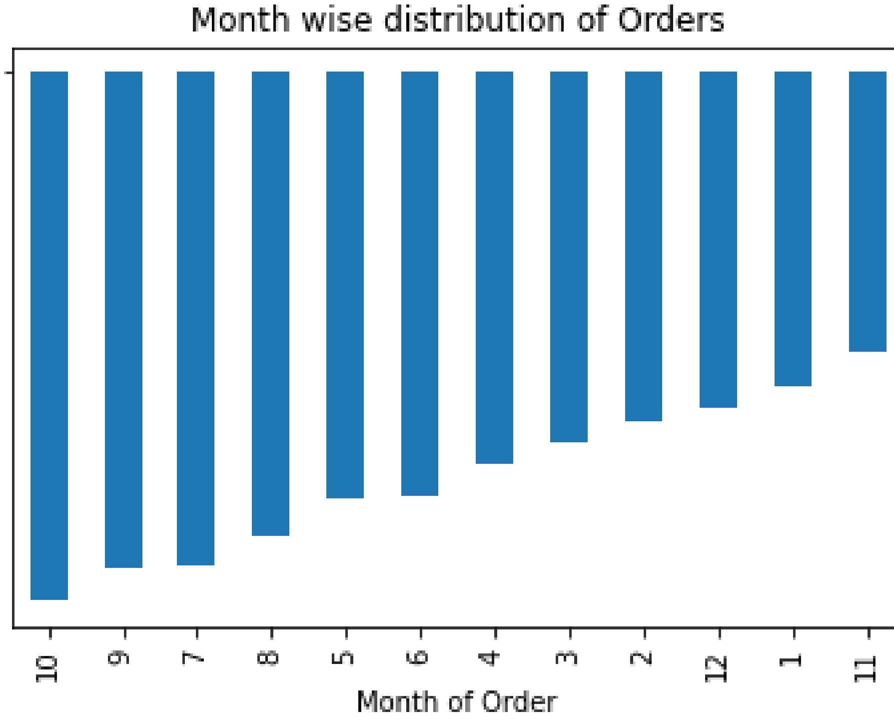


City wise pattern is quite different from state.

Focused inventory management

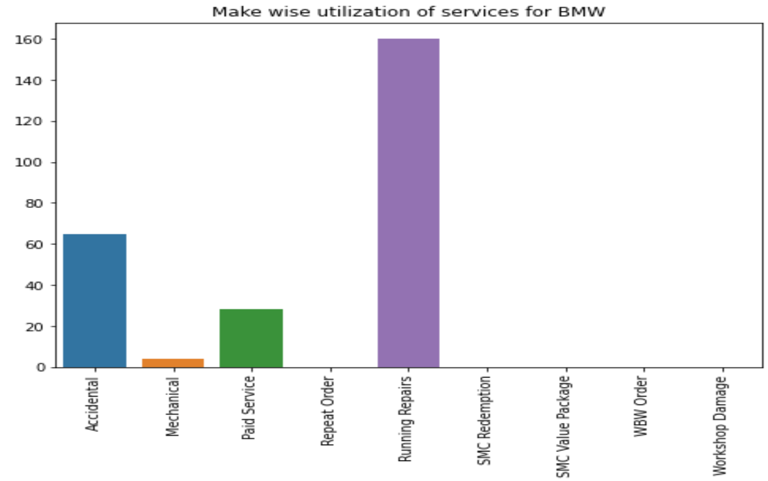
ORDER TYPE

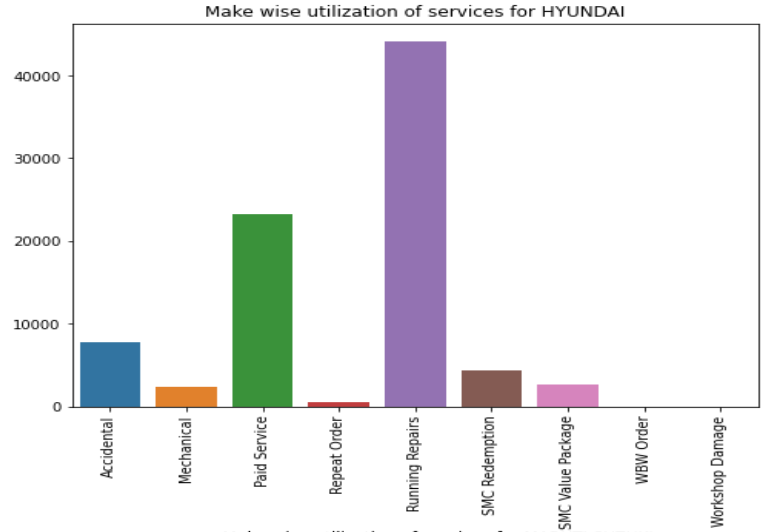


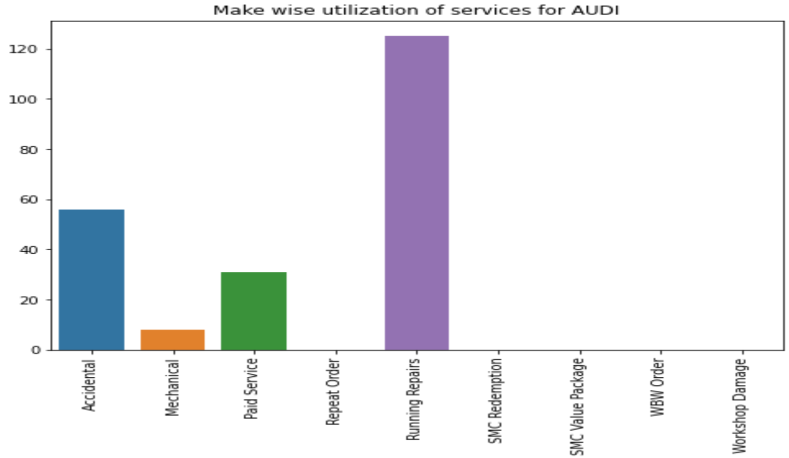


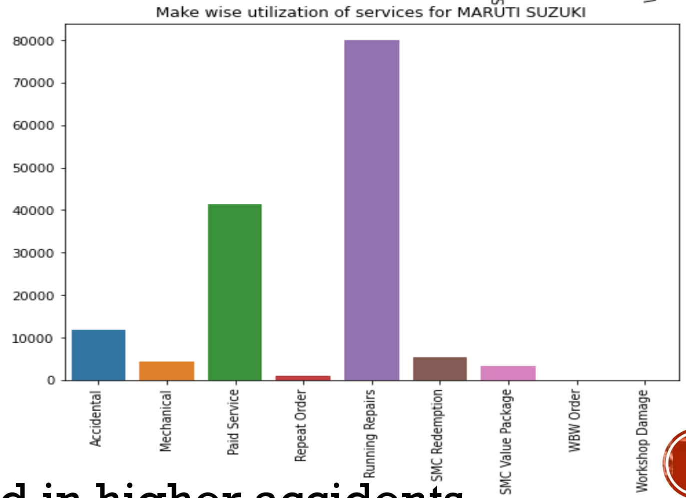
Running repairs is the most used service for all the make of vehicles

Surge in orders is seen in months that are after the Monsoon season.



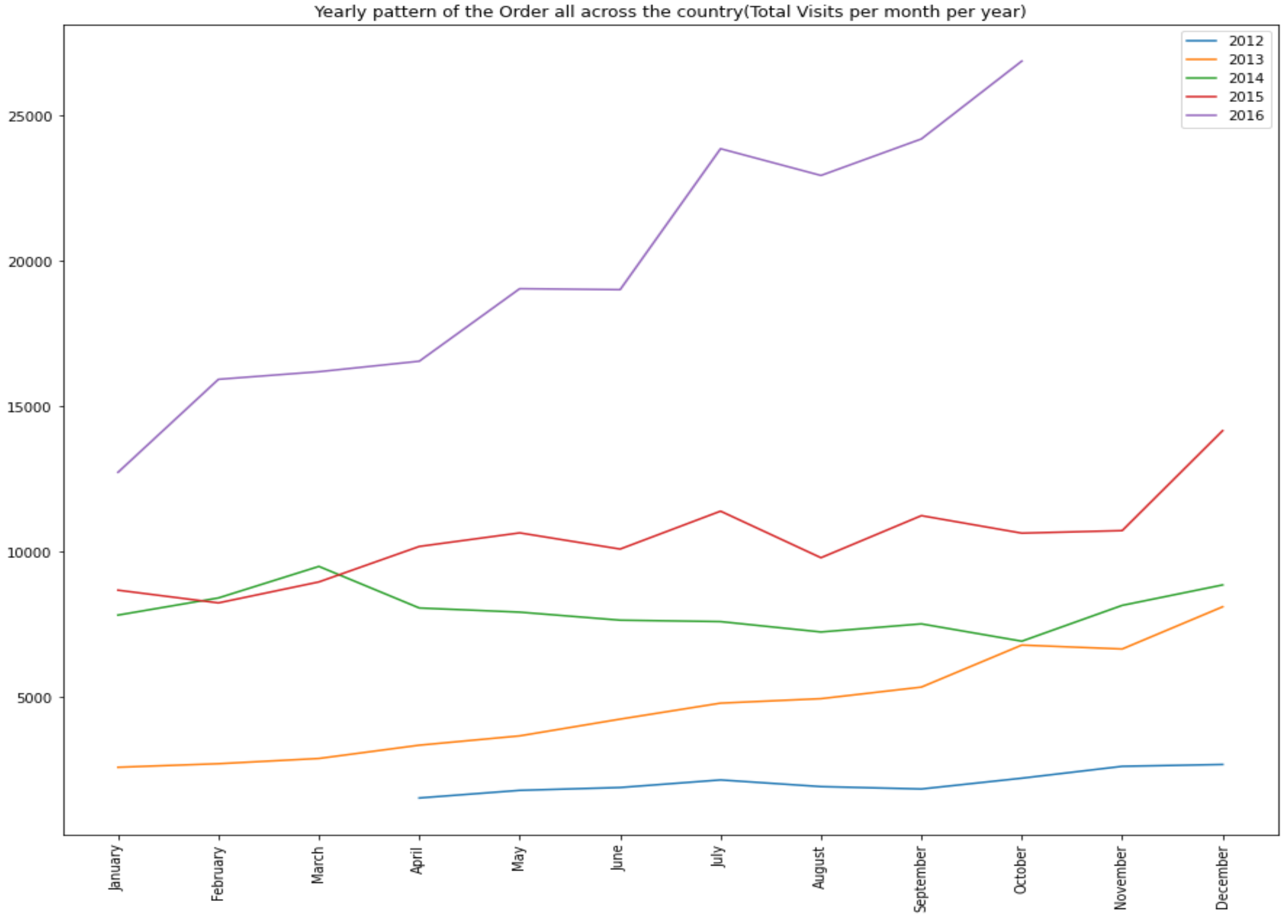






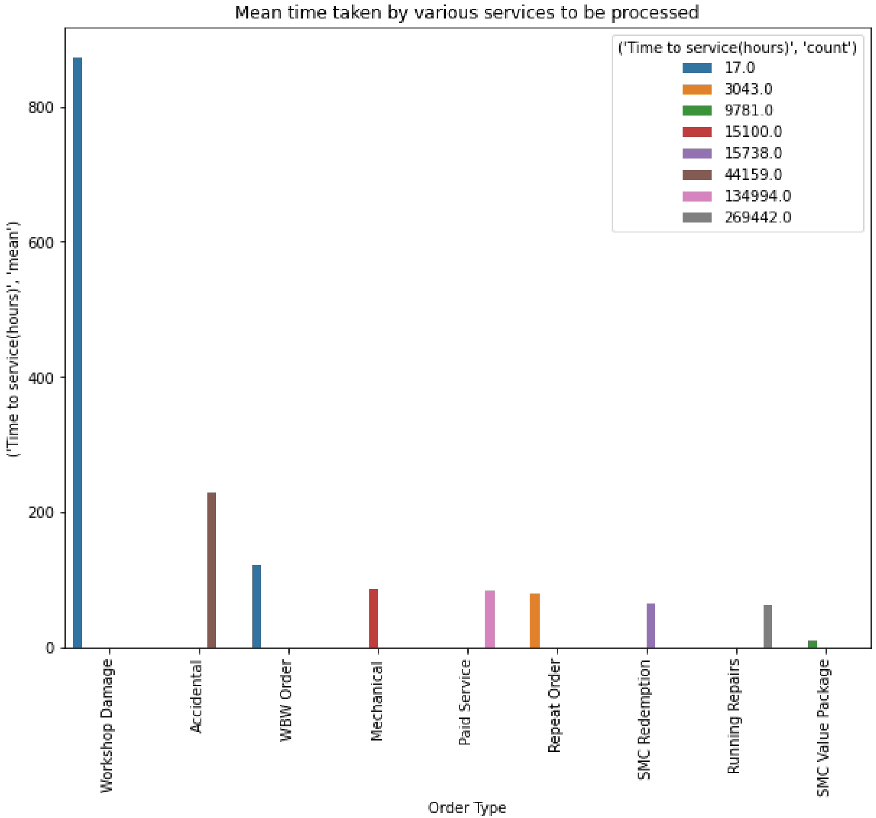
High end cars are involved in higher accidents

YEARLY PATTERN OF SERVICE

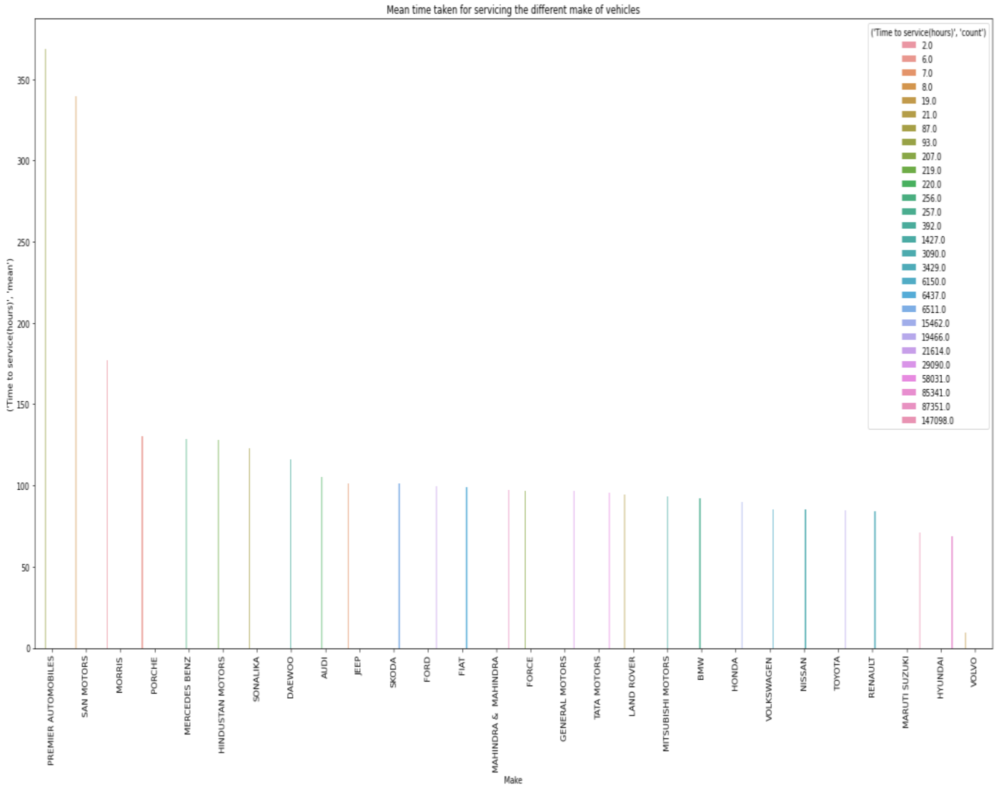


Cyclical pattern of peak after monsoon can be observed

SERVICE TIME



Running Repairs takes least mean time and has most orders



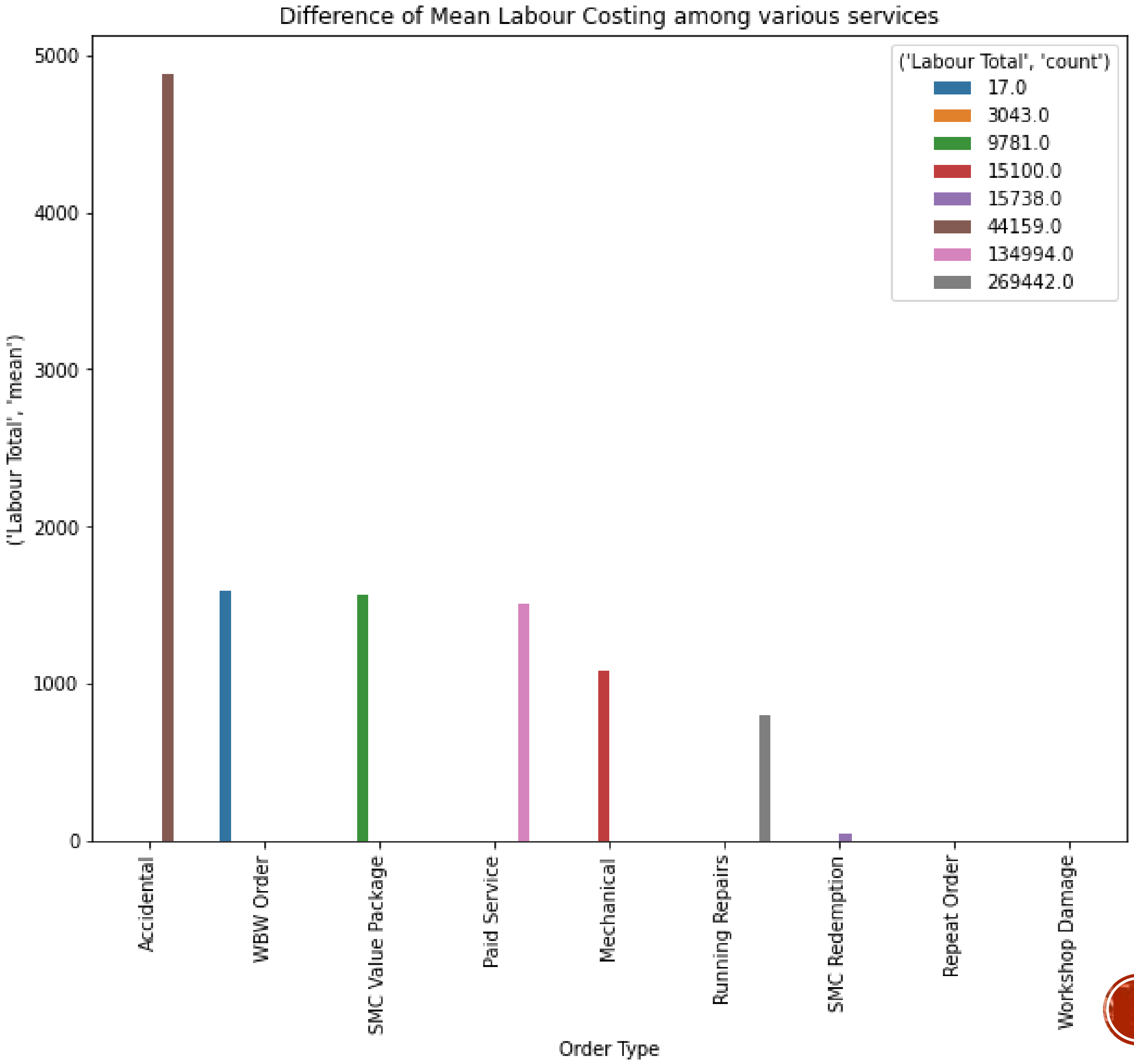
Mean time for different make of vehicle.

Maruti lies at the bottom , surprisingly Premier Auto vehicles take highest service time.

LABOUR COSTING

Mean Labour costing of various services

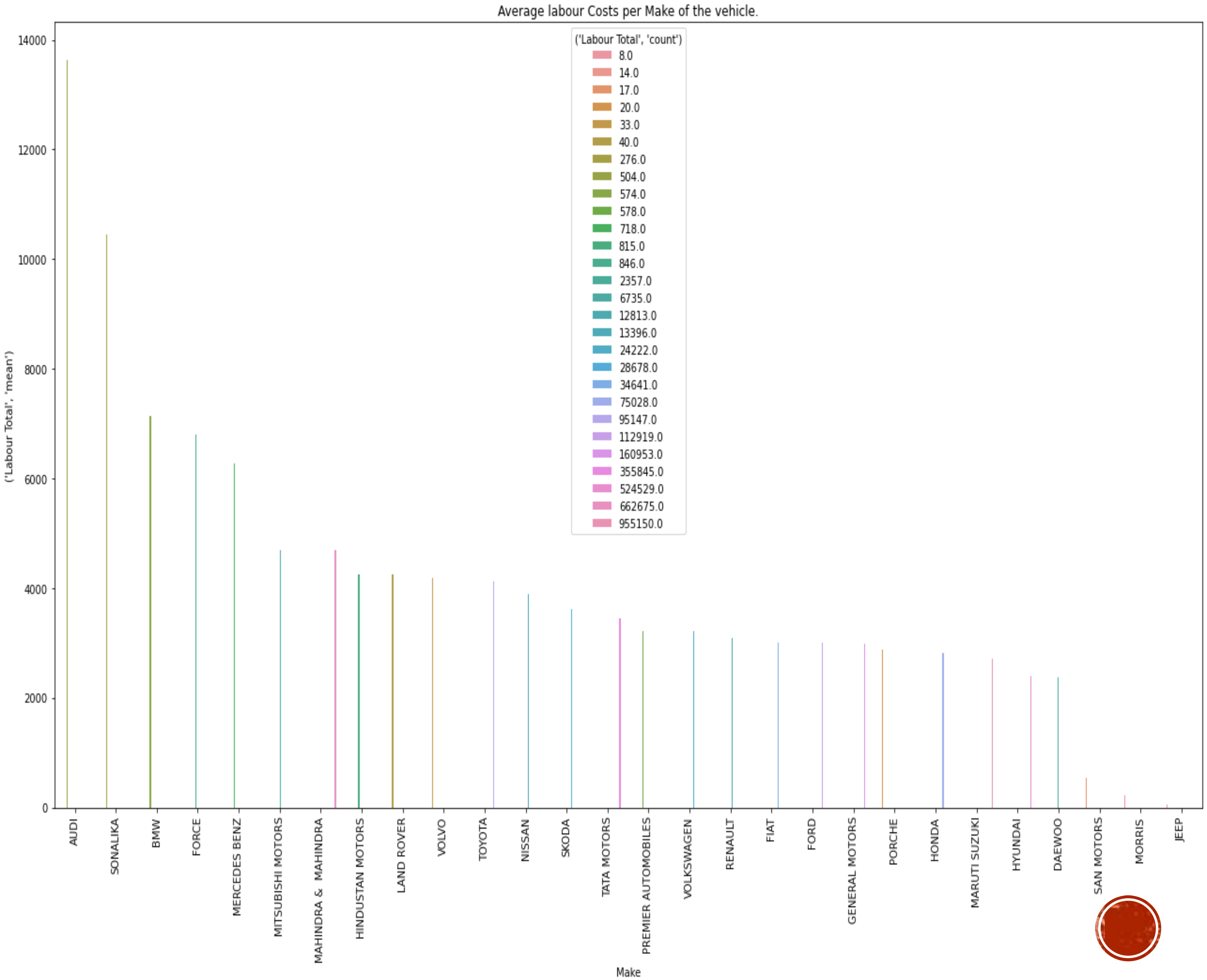
Labour Cost for Accidental Services far exceeds other types



LABOUR COSTING

Mean Labour costing per Make of the vehicle

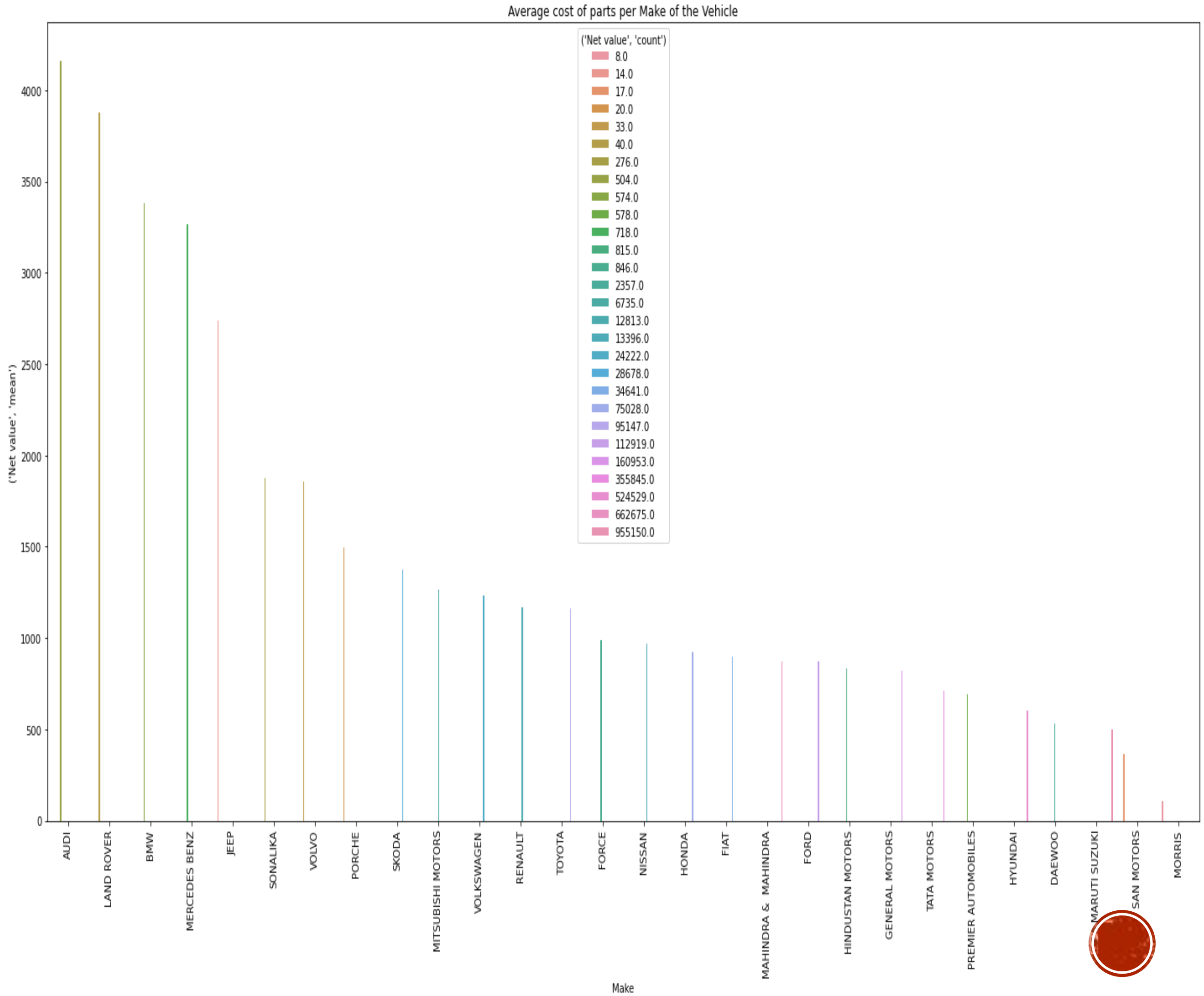
High end cars also entails higher revenue in a single visit



COST OF PARTS

Average Cost of Parts per make of the vehicle

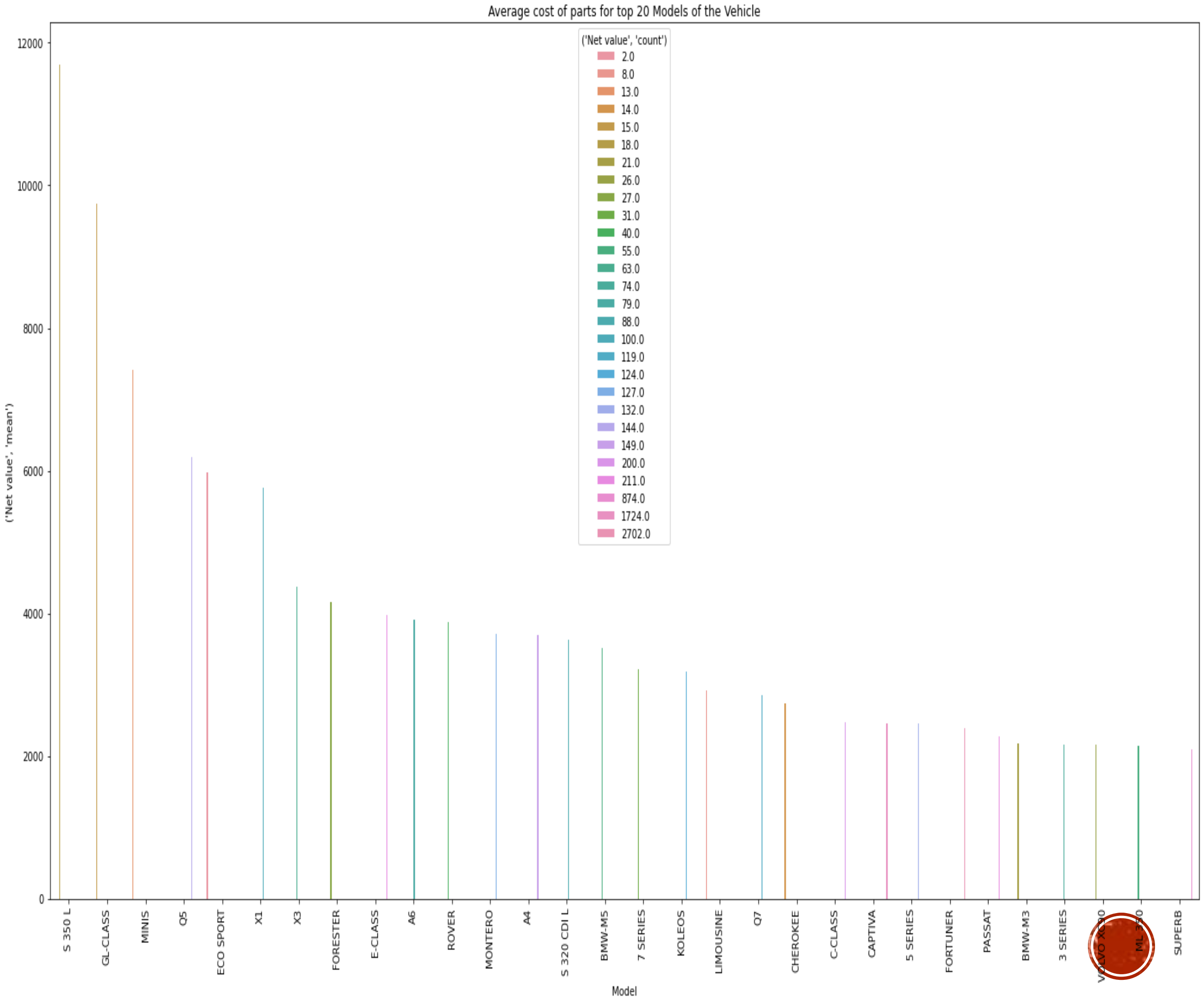
High end cars have High Cost of parts while Maruti features at the bottom



COST OF PARTS

Average Cost of Parts per model of the vehicle

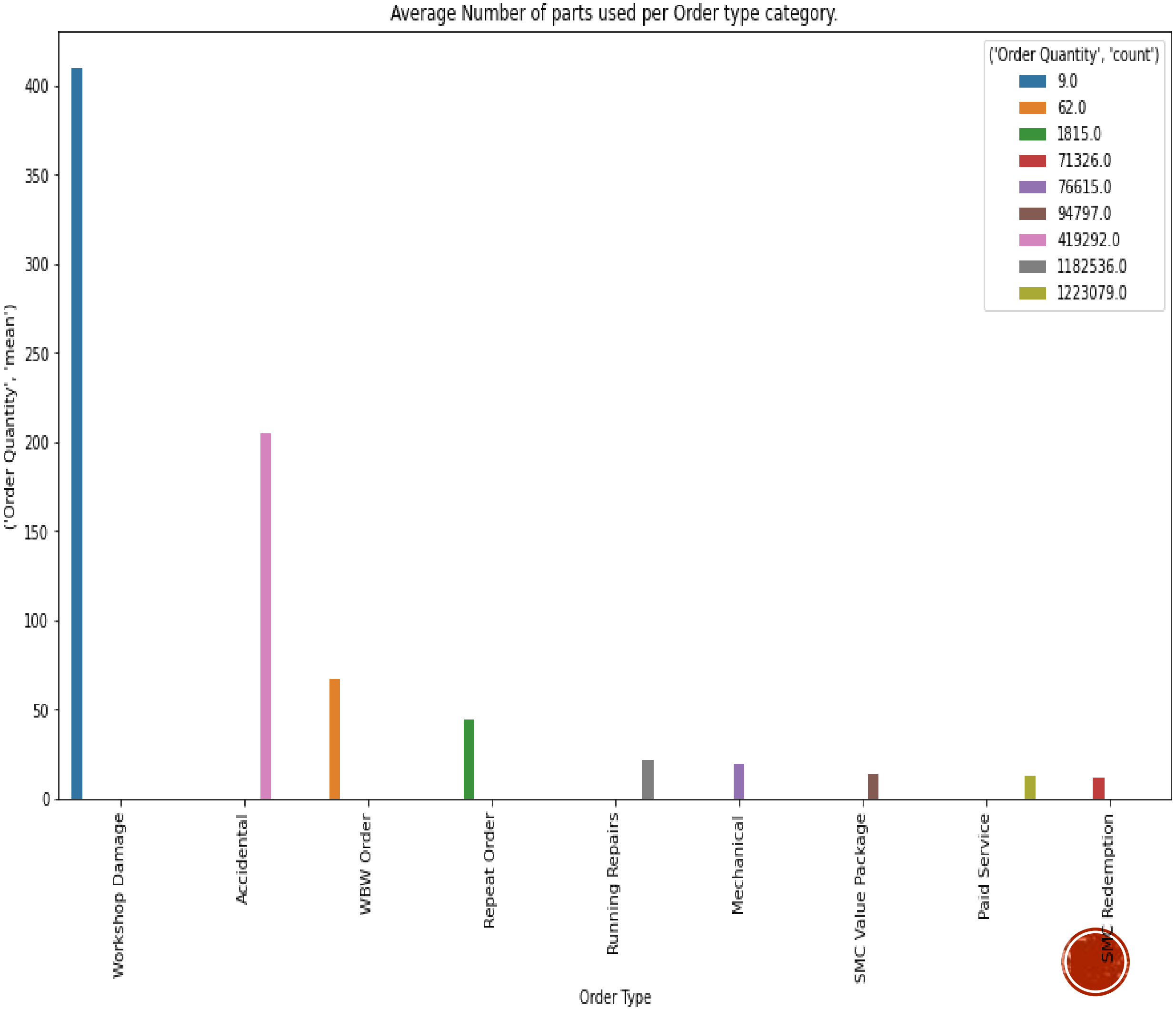
Not a single model of Maruti/Tata/Mahindra appears in the top 20



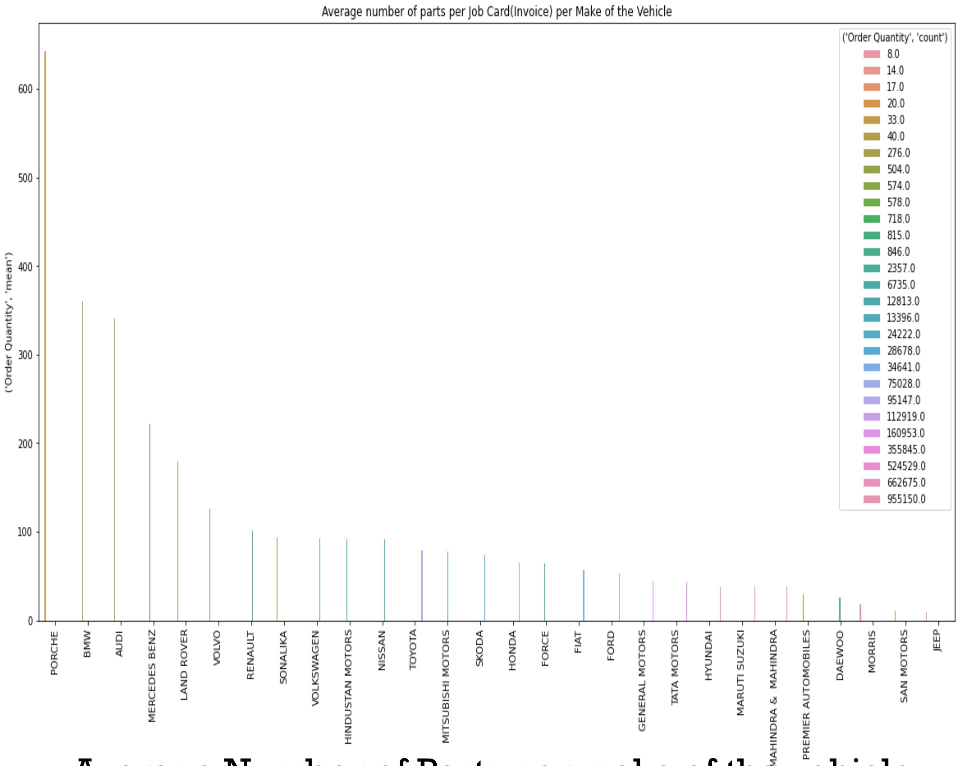
NUMBER OF PARTS

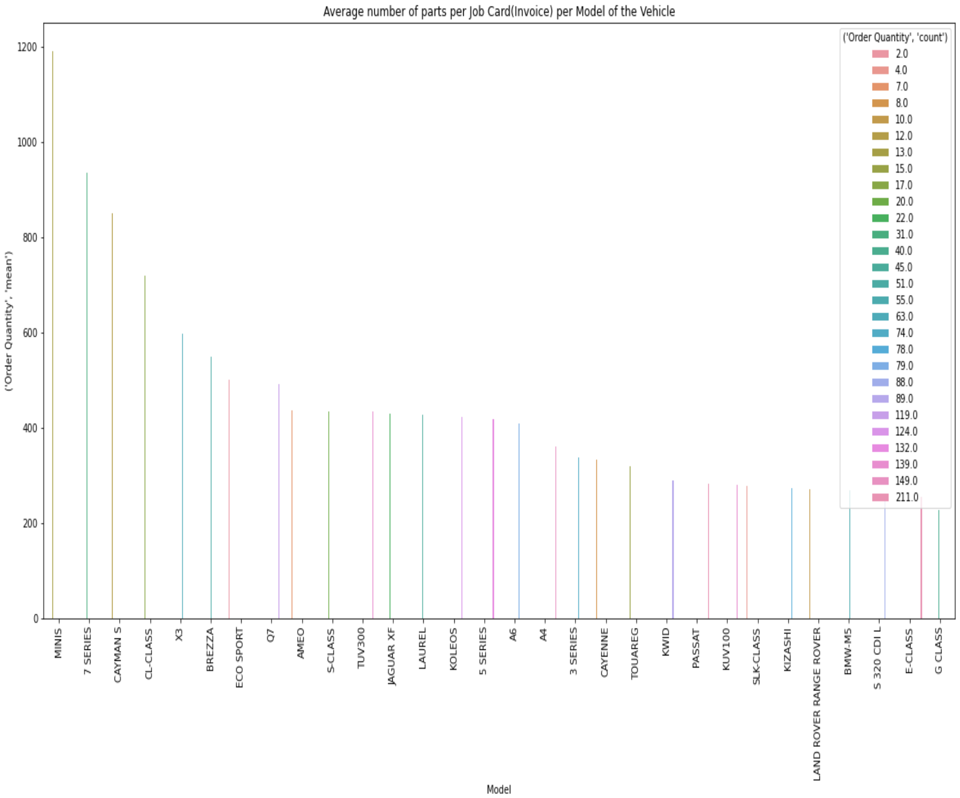
Average Number of Parts per Order Type

Running Repairs requires very less parts to be replaced.



NUMBER OF PARTS



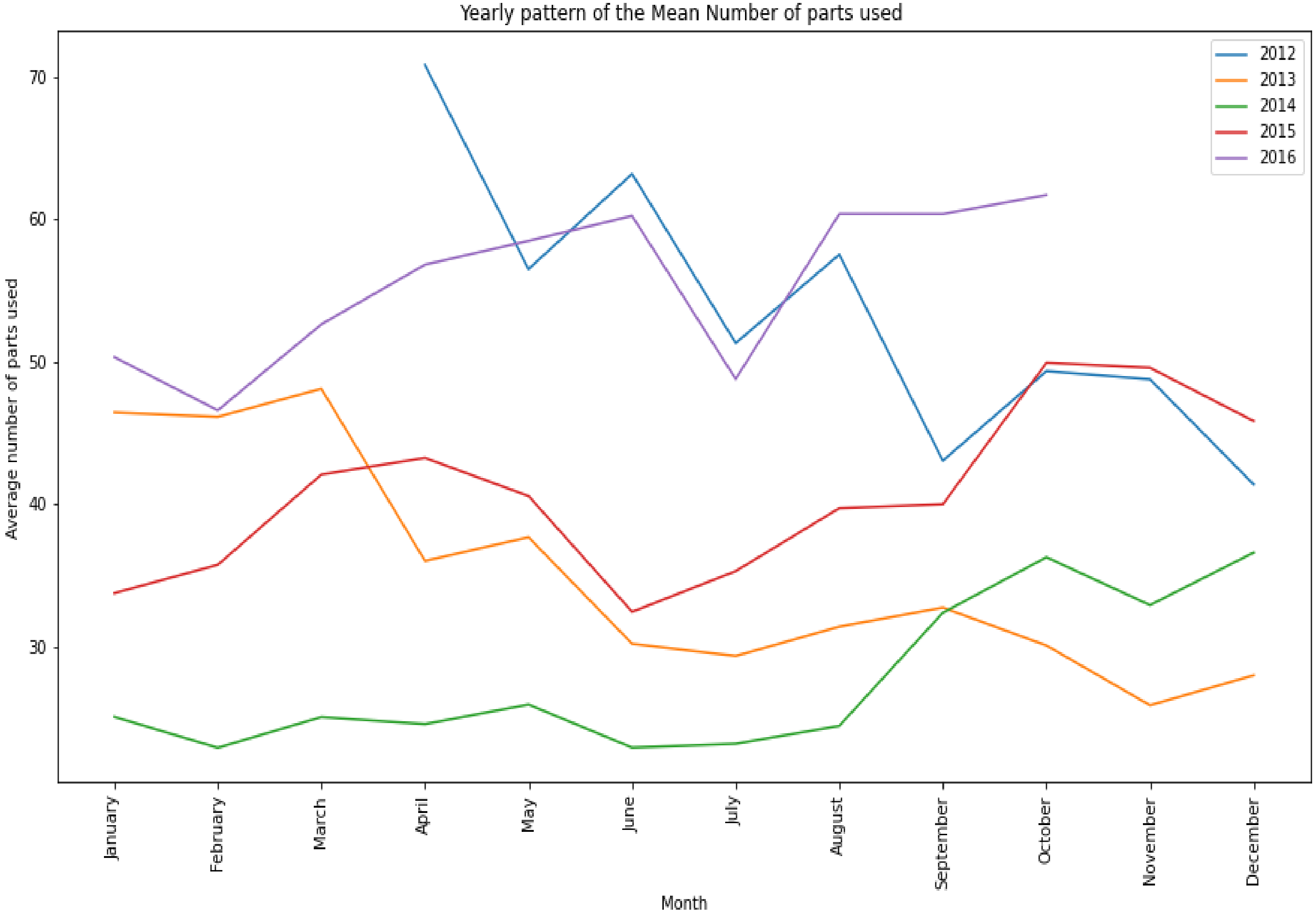


Average Number of Parts per make of the vehicle

Expensive vehicles require more parts for servicing, but the frequency of order is low.

Maruti BREZZA is a surprise entry in the list

NUMBER OF PARTS



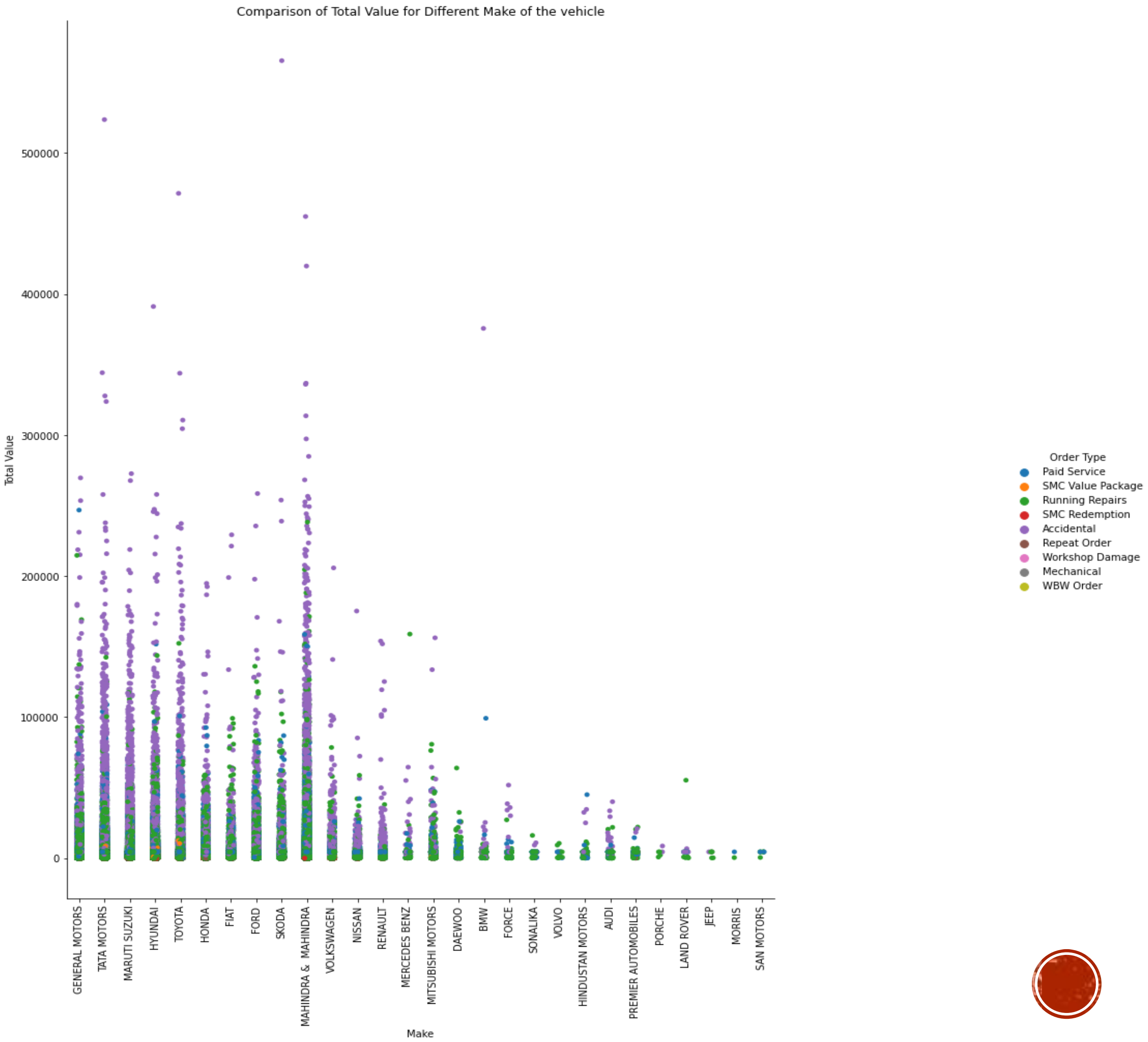
Peak after Monsoon months is observed.

Help in timing the inventory

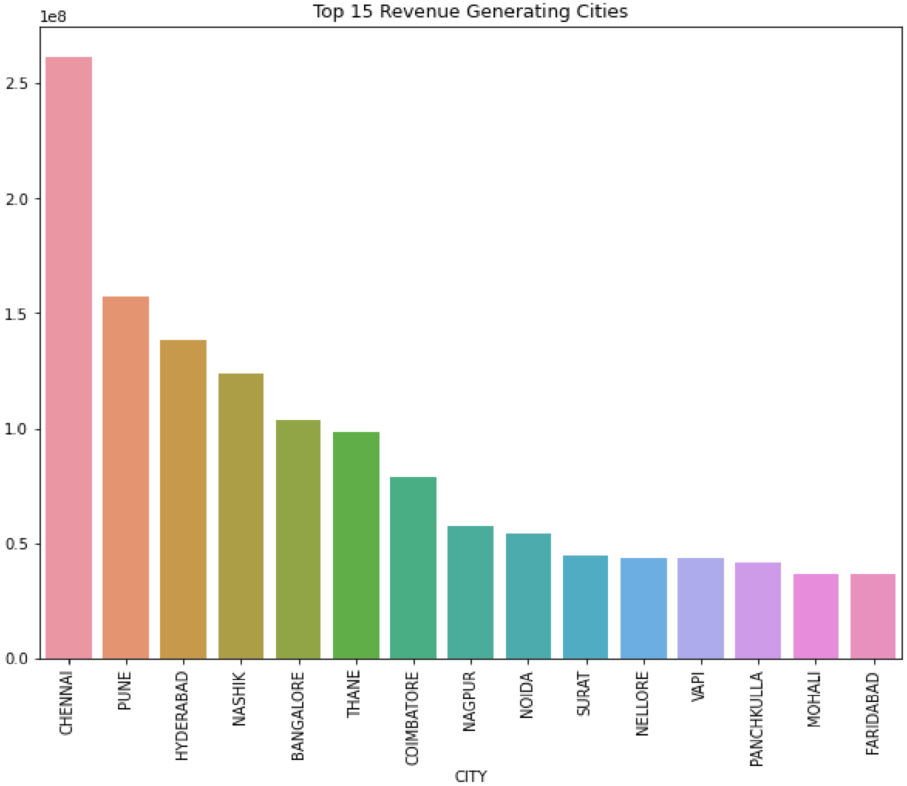
TOTAL REVENUE

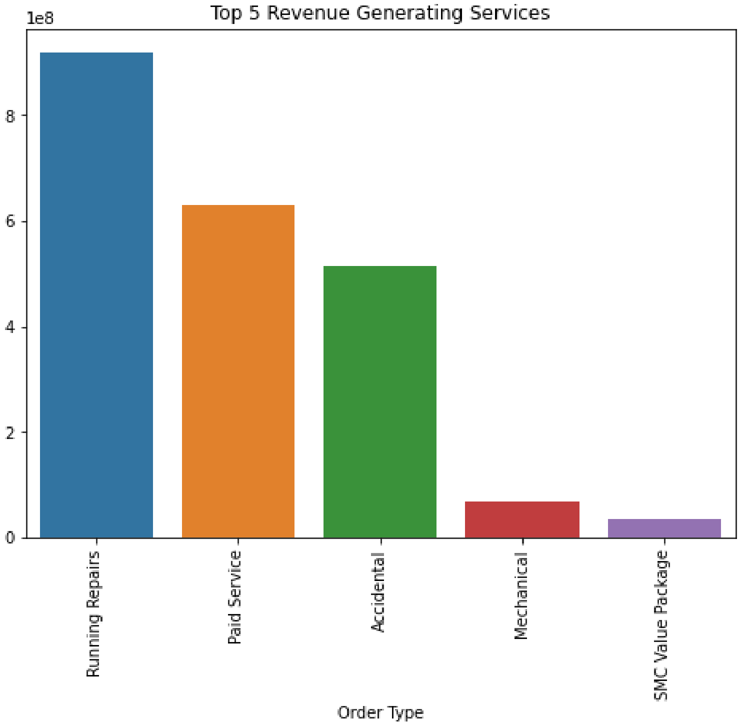
Comparison of Total Order Value of different make of vehicle

Accidental service accounts for maximum revenue for any Make type



TOTAL REVENUE





Chennai tops the list as it's having most number of vehicles.

SEGMENTATION

Why Segmentation : Focussed marketing Strategies for different segments

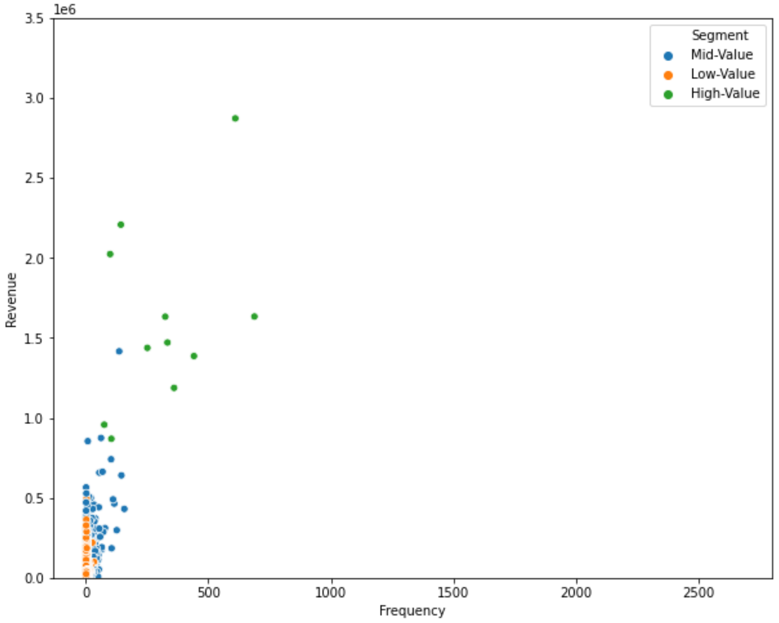
Segmentation done on the basis of **RFM**:

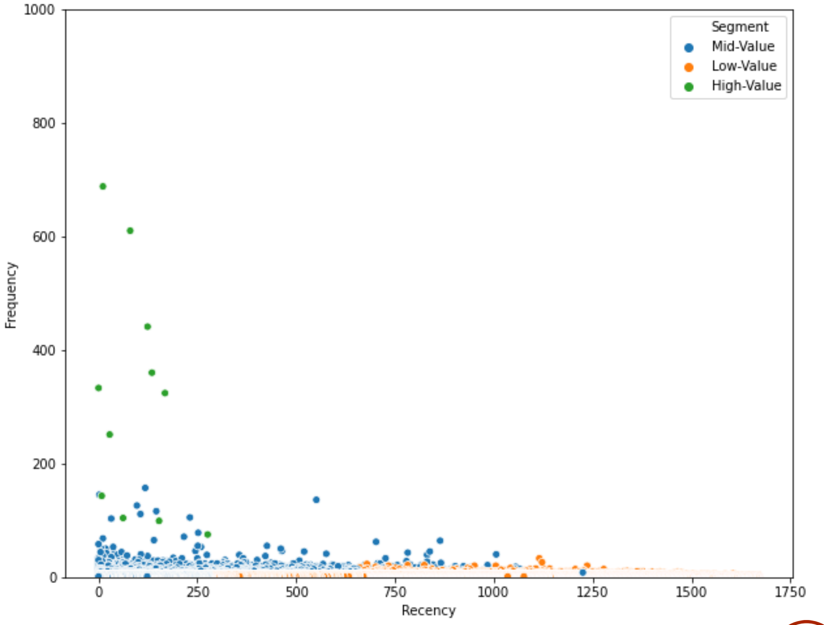
Recency: How recently the customer made the purchase?
Frequency: How frequently the customer made the purchase?
Monetary Value : What is monetary value of the purchase made by the customer

Applied **K-Means Clustering** to the RFM to identify different clusters :-

- High Value
- Mid Value
- Low Value

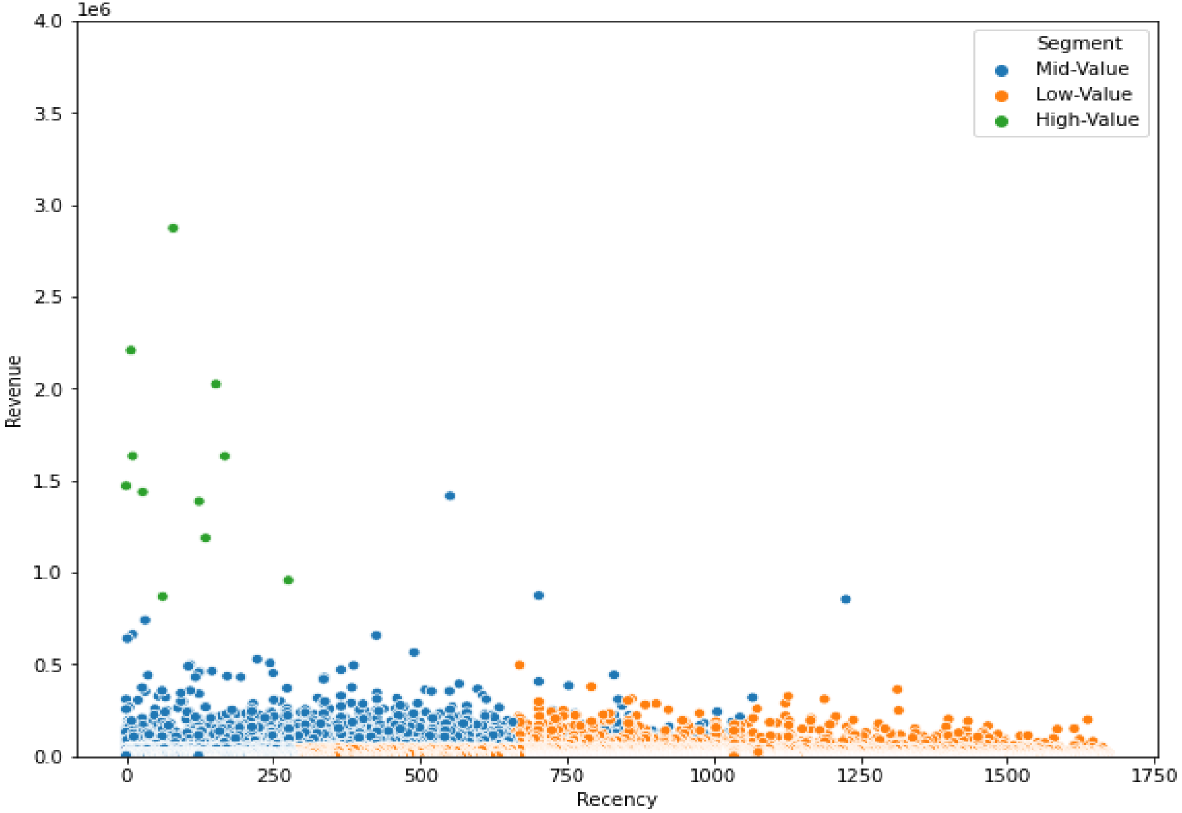
SEGMENTATION





Clustering on Customer No.

SEGMENTATION

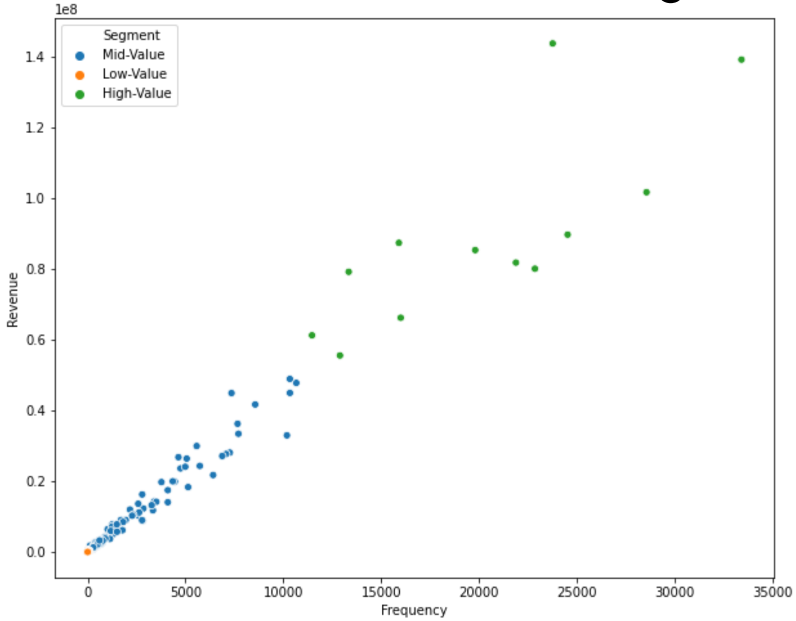


Clustering on Customer No.

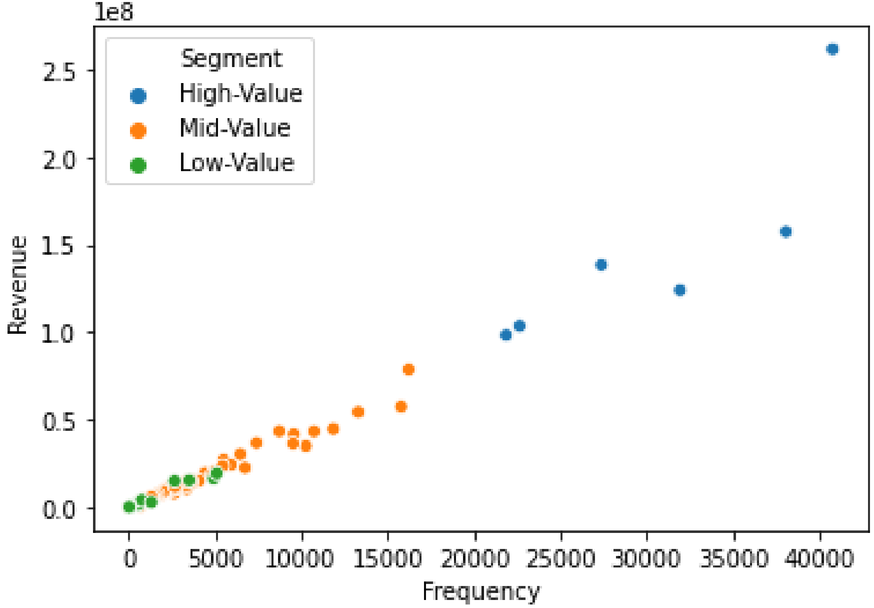
High value customers have High Revenue and low Recency associated with them

SEGMENTATION

Model wise clustering



City wise clustering



Strong correlation between Revenue and Frequency is observed.

LIFETIME VALUE PREDICTION

Lifetime Value = Total Gross Revenue from the customer

Time frame selected for this was 6 months

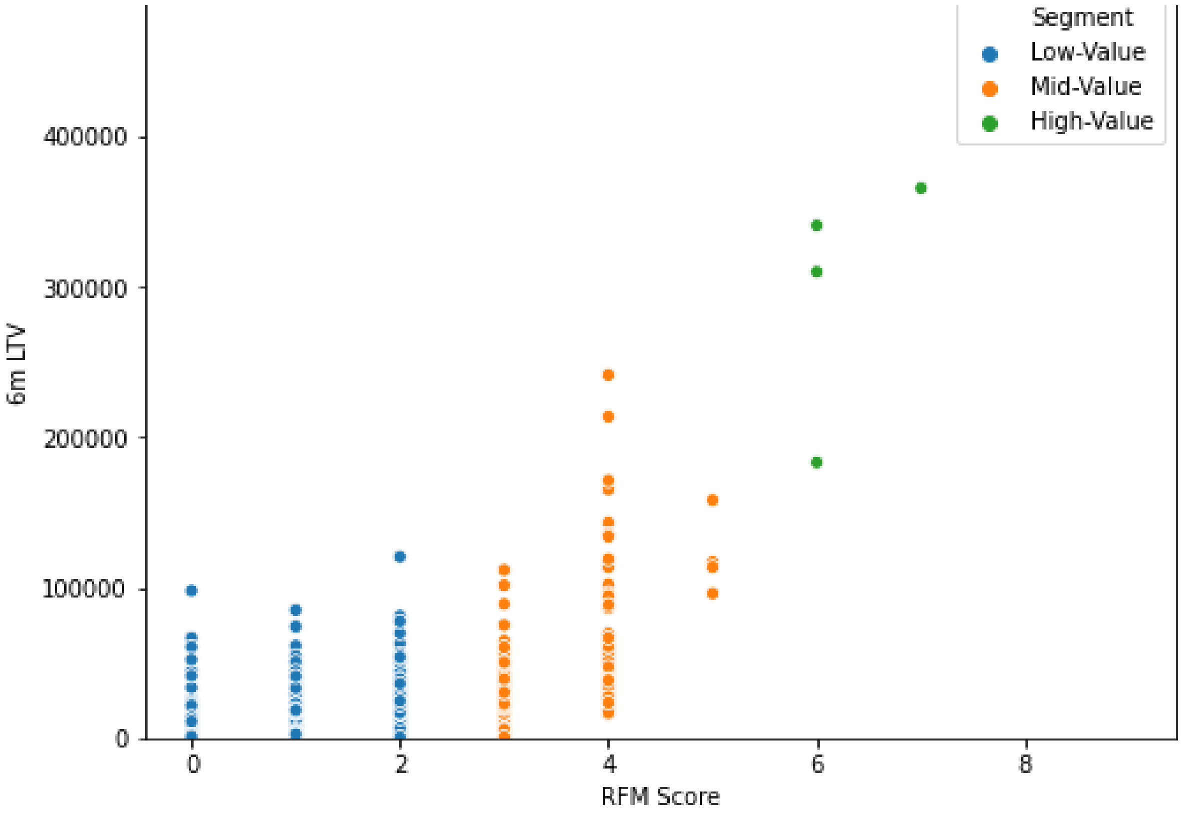
Overall Score is calculated on basis of RFM

OverallScore	Recency	Frequency	Revenue
0	80.168798	1.078005	3.238091e+03
1	59.865005	1.120457	3.986359e+03
2	36.186047	1.236279	5.313862e+03
3	11.584270	1.354869	5.740340e+03
4	9.081633	2.394558	3.466430e+04
5	6.750000	10.250000	1.124622e+05
6	3.000000	32.666667	2.392879e+05
7	2.000000	94.000000	3.254139e+05
9	2.000000	679.000000	3.388183e+06

LTV

LTVCluster	count	mean	std	min	25%	50%	75%	max
0	3051.0	2904.853471	2329.117176	0.00	935.155	2338.53	4345.970	8886.43
1	773.0	14836.203014	4741.064791	8890.60	10824.910	13643.75	18254.480	26859.16
2	179.0	39038.852514	9464.070311	26997.76	30837.100	36529.51	45004.265	62691.12

- K-means clustering was used to decide the LTV Cluster.
- Cluster 2 is the best wit h average of 38K , whereas 0 is the worst with 2.8K



LIFETIME VALUE

High RFM score entails High LTV

CLASSIFICATION MODEL

- XG Boost Classifier was to used predict the **LTV Segment** of the Customer.
- Accuracy of XGB classifier on train set is 93%
- Accuracy of the XGB classifier on test set is 92%
- Accuracy is better than the baseline accuracy of 76%

```
0    0.762178
1    0.193105
2    0.044716
Name: LTVCluster, dtype: float64
```

CONFUSION MATRIX

	precision	recall	f1-score	support
0	0.92	1.00	0.96	1531
1	0.92	0.70	0.80	371
2	0.94	0.61	0.74	100
accuracy			0.92	2002
macro avg	0.93	0.77	0.83	2002
weighted avg	0.92	0.92	0.92	2002

CEO DECK

We can start taking actions with this segmentation.The main strategies should be:

- High Value: Improve Retention
- Mid Value: Improve Retention + Increase Frequency
- Low Value: Increase Frequency

THANK YOU