***File 1: data-loading-and-cleaning***

1) Importing packages

2) Loading Data

3) Checking for missing values and dropping unrequired columns

=> Invoice data: There are multiple values in Job card number which are duplicate having multiple invoice number which is not possible.

Solution - To approach this problem we added numerical values and keep categorical as it is.

***File 2: feature-engineering***

Feature 1) Service Time: Invoice time - Job Card time

Feature 2) Place: There are multiple values in City, Locality which has spelling mistake.

Feature 3) Age of cars: based on Kilometers driven.

Feature 4) Day, Month, Year

Feature 5) Season: Summer, Monsoon, autumn, winter

***File 3: visualising-data***: Insights and questions answered based on data visualization

1) From where most of the Data has been collected?

2) Which are the most car types that comes for the service?

3) Which are the most Customer type that visits for the service?

4) What are the most type of orders received?

5) What are the Age of cars that usually comes for the service?

6) Which are the plants who have not received order for past 1 month?

7) Which location receives most orders?

8) What are the total orders of different car manufacturer types which comes at particular location?

9) What are the average revenue generated for particular car manufacturer and frequency of the orders for that manufacturer?

10) What is the average labour cost for different car make?

11) What is the Labour total for different location?

12) What is the average revenue generated for order types?

13) What is the most Order types at particular location?

14) What is the average service time at particular location?

15) What is the service time for particular order type and revenue generated from it?

16) What is the average revenue generated for different age of cars?

17) Which season receives most orders and revenue generated from it?

18) What are the order types for particular season?

19) What is average service time and average revenue year wise?

20) What is the Total revenue and Total orders received yearly?

21) Which are top 50 highest revenue generating plants and how many of the top revenue generating plants are non functional?

***File 4: productivity***

**Productivity of a Plant**: Total Amount generated per hour by per technician at particular plant

1. Calculate Number of unique Technicians present at the particular Plant over a year

2. Calculate the total amount of service time in hours over a year

3. Calculate the total amount of revenue generated at the respective plant over a year

4. Productivity = Total Revenue (in year)/ (Total Service Time (in year) \* Number of Technicians worked (in year))

**Note**:

- Large number of Technician names are missing and majority of null values is for year 2012 and 2016.

- Very Few values for Technician names for 2012 and 2016. So we will calculate productivity of plants only for 2013, 2014 and 2015

- We will calculate Productivity only for Running Repairs types because highest order type is Running Repairs and also as there is chance that different order type for example like Accidental may have more damage and more service time.

***File 5: customer-ltv***

**Customer Lifetime value**: Customer lifetime value is the metric that indicates the total revenue a business can reasonably expect from a single customer account.

- We are going to calculate CLTV based on Make, Model of a particular customer for year 2013-2015 for only Running type orders

- Formula:

LTV = (average order value \* purchase frequency)/churn rate

where,

Average order value = Money spent / Number of transactions

Purchase frequency = Total number of transactions / Total number of customers

Repeat rate = Number of customers that have come more than once/ Total customers

Churn rate = 1 - Repeat rate

- For this case, we have calculated LTV of a particular customer based on their car make and model.

So,

LTV = average\_order\_value (for a single customer) \* Purchase Frequency (for a single customer)

1) Finding CLTV of a customer based on Make and Model.

Note:

* Model has 1524 missing values. To handle this we tried to impute Model values based on Customer no. and car manufacturer of a customer is using.
* There is chances a single customer can have more than one car of different car manufacturer. Using that data we can impute remaining Model values of that customer.

Problem:

* While trying to impute Model values using above mentioned method groupby is taking too long to execute. To solve this we tried to find out how many customers are there who has visited more than once as we can impute Model of a particular customer based on previous visit data.
* We can see that there are 292 customers who has visited more than once and who has missing data in Model feature. Out of 292 customers only 37 customers has atleast one Model value which can be used to impute remaining Model values of that customer. But this will impute only 146 values out of 1524.

Workaround: Imputed Model values based on only Car manufacturer.

2) Forming Clusters based on CLTV

3) Applying models to predict CLTV