RISHIKA RAVICHANDRAN

DATA SCIENCE INTERN

SAPIENCE EDU CONNECT PVT. LTD.

WEEK 1

**ANALYSIS REPORT**

1. **Data Cleaning and Preprocessing**

* The dataset contains 8128 rows and 12 columns.
* Features in the dataset : name, year, selling\_price, km\_driven, fuel, seller\_type, transmission, owner, mileage(km/ltr/kg), engine, seats, and max\_power.
* Features containing missing values : mileage(km/ltr/kg), engine, seats, and max\_power
* Missing values are replaced. The attribute, ‘max\_power’ was of mixed data type, so the missing values were first replaced by mode, converted into numeric, and then replaced with their median.
* Categorical features were fuel, seller\_type, transmission and owner. They were label-encoded.
* A new feature named ‘car\_age’ was derived from the given feature ‘year’ to use for visualization.

1. **Data Visualization**

* A moderate positive correlation is observed between ‘year’ and ‘selling\_price’ indicating newer cars tend to have higher selling prices.
* ‘engine’ and ‘max\_power’ also show a moderate positive correlation with ‘selling price’ indicating that the cars with larger engines and higher power (maybe luxury and high-performance cars) also have higher selling prices.
* The distribution of selling prices is highly skewed to the right. We can also infer that most of the car prices are concentrated in the lower range with a few expensive outliers
* Car prices decrease with increase in car age.
* Most of the car prices are low with few expensive outliers
* We can observe that the car prices remain stable with increasing kilometres but then after a certain value it started decreasing.
* A weak negative correlation between mileage and selling price indicating higher mileage leads to lower prices, but again exceptions exist.
* Diesel cars tend to have higher selling prices compared to petrol, LPG, CNG.
* Automatic cars have higher selling prices than manual cars.
* First owners have higher selling prices followed by second owners.

1. **Machine Learning feasibility:**

* The target variable we set here is ‘selling\_price’ which is a continuous numerical value. So, this is a **Regression Problem**. Our objective is to predict the car prices based on the given features like mileage, engine, max\_power fuel and transmission. Therefore, different regression models like Linear Regression, Logistic Regression, etc can be used.
* The given dataset contains labelled data, that is, we have input features and these features are used to train the model to make accurate predictions. So, this is a **Supervised Learning Problem.**