Second hand car price Prediction

using

Supervised Learning Regression

Presented by
Gopi Selvaraj

Problem Definition

- The used car market in India is a dynamic and ever-changing landscape. Prices can fluctuate wildly based on a variety of factors including the make and model of the car, its mileage, its condition and the current market conditions. As a result, it can be difficult for sellers to accurately price their cars.

- This dataset contains information about used cars. This data can be used for a lot of purposes such as Used Car Price Prediction using different Machine Learning Techniques.

- The goal of the project is to predict the price of a second-hand car that will be listed in future in the E-Commerce website using statistical modelling and machine learning techniques.

Data dictionary

S.No	Field name	Description	Data type
1	car_name	Car's Full name, which includes brand and specific model name	Object
2	brand	Exact brand Name of the particular car	Object
3	model	Exact model name of the car of a particular brand.	Object
4	vehicle_age	The count of years since car was bought.	Int64
5	km_driven	Number of kilometers the car has been driven	Int64
6	seller_type	Type of seller that is selling the used car	Object
7	fuel_type	Fuel used in the used car, which was put up on sale	Object
8	transmission_type	Transmission used in the used car, which was put on sale	Object
9	mileage	It is the number of kilometers the car runs per litre.	float64
10	engine	Engine capacity in cc(cubic centimeters).	int64
11	max_power	Max power the car produces in BHP	float64
12	seats	Total number of seats in car	int64
13	selling_price	The sale price which was put up on the website	Object

Shape and distribution of the target variable

- The dataset has **15411** observation & **13** variables.
- The complexity is in finding the solution to the problem based on the chosen techniques.
- Since the target variable is ‘Numerical', we will be building a Supervised Learning Regression model.

- **Python Version:**

  3.11.7 | packaged by Anaconda, Inc.

Dropping Variables

- None of the variable have been dropped from the data frame.

Excluded Variables

- These features are not directly correlated with the price of car and they can actually introduce noise into the model. For example, two cars with the same features but different brands may have different prices.

- This is because brand reputation and perceived quality can play a role in determining the price of a car. By dropping these variables, we can create a model that is more accurate and reliable.

car_name	brand
Model	

Duplicate Features

- These features are not directly correlated with the price of car and they can actually introduce noise into the model. For example, two cars with the same features but different brands may have different prices.

- This is because brand reputation and perceived quality can play a role in determining the price of a car. By dropping these variables, we can create a model that is more accurate and reliable.

Duplicate Observations:

- There were 167 duplicate rows which have been removed.

Missing Value Imputation

- There were to direct missing values in any columns.

- But there were two columns which had indirect missing values:

- ‘**seats**’:

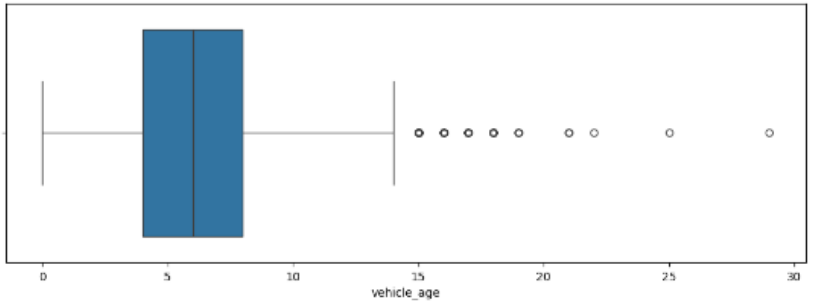
- There were 2 indirect missing values for ‘Honda City’ & ‘Nissan Kicks’ which as ‘0’ as the seat value.
- We have imputed the missing values according to the car’s make since it doesn’t make sense to have a car with ‘0’ seats.

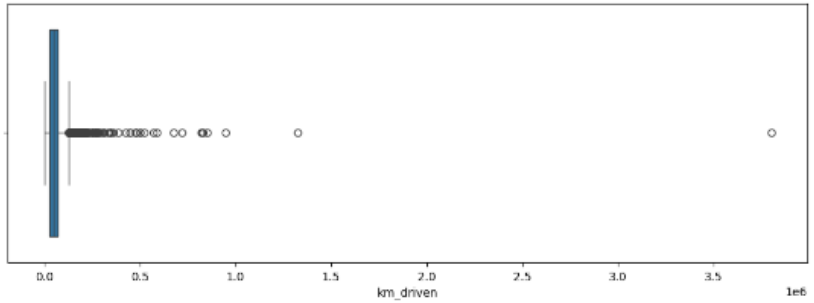
- ‘**vehicle_age**’:

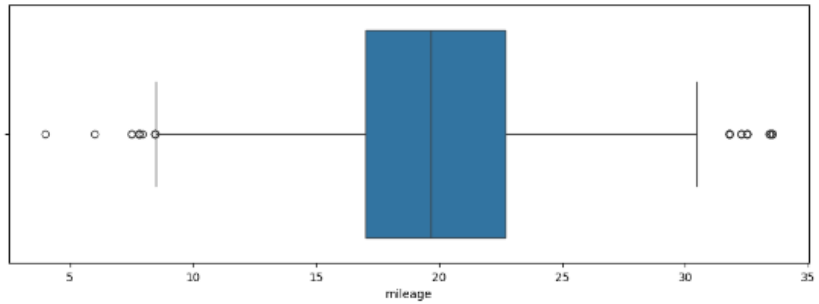
- There were 4 indirect missing values for 4 rows in this column. After closer inspection we found that these are cars were new cars hence had ‘0’ years as the column value which seems appropriate given the ‘kilometers_driven’.
- Hence, we didn’t need to treat them as missing values.

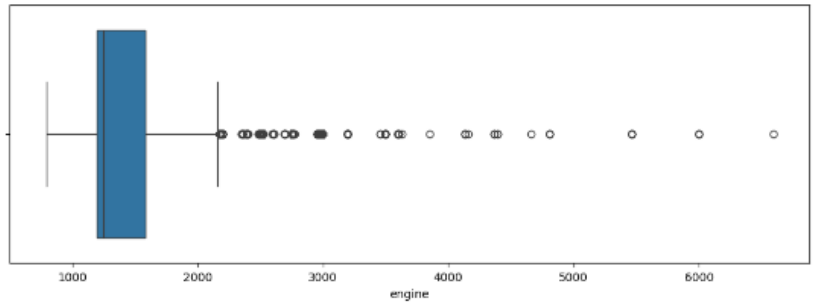
Exploratory Data Analysis **Univariate analysis - Numeric variables**

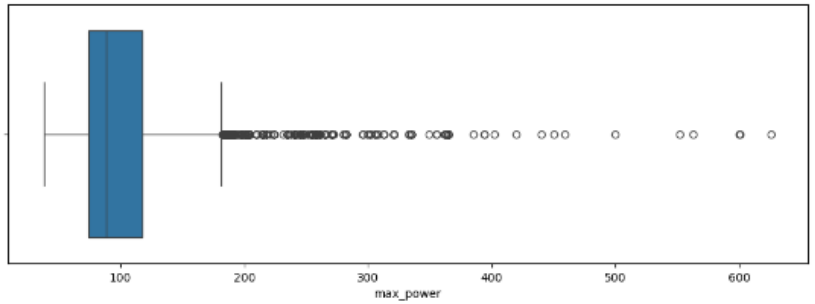
BOX PLOT: (Detection of IQR, Minimum, Maximum, 25%, 50%, 75% & outliers in a numeric column)

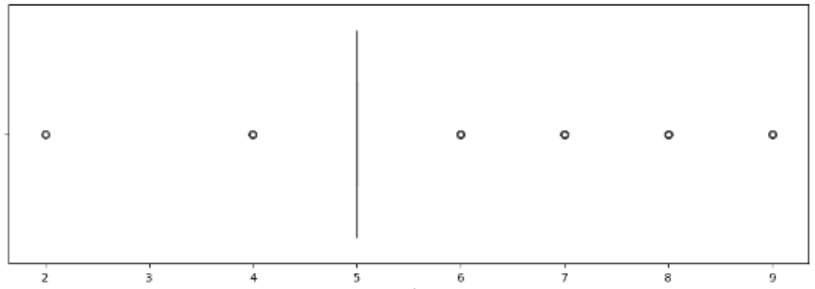


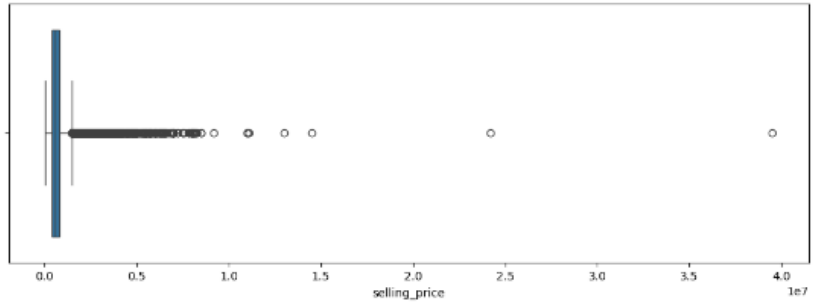




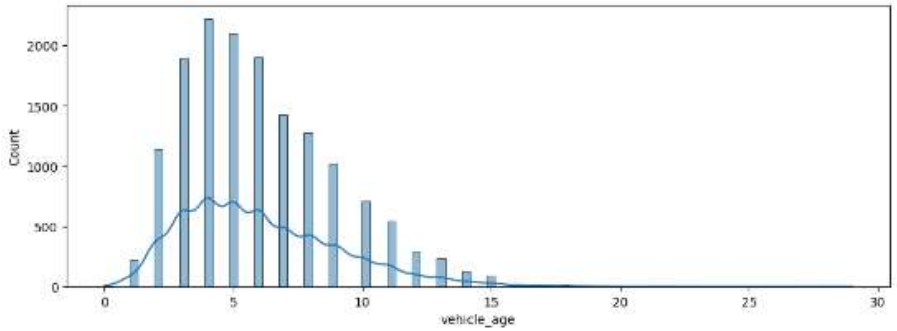


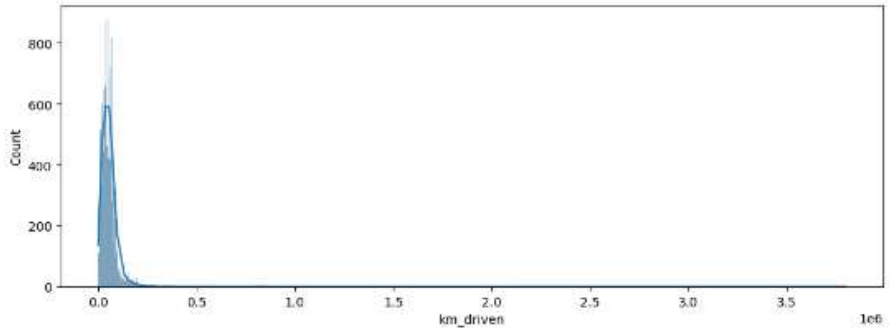


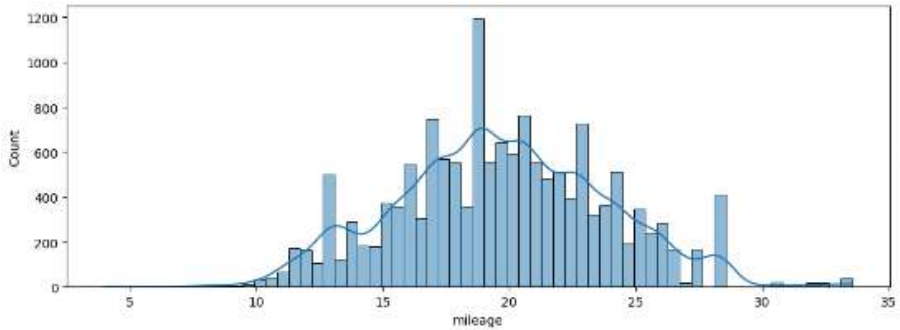


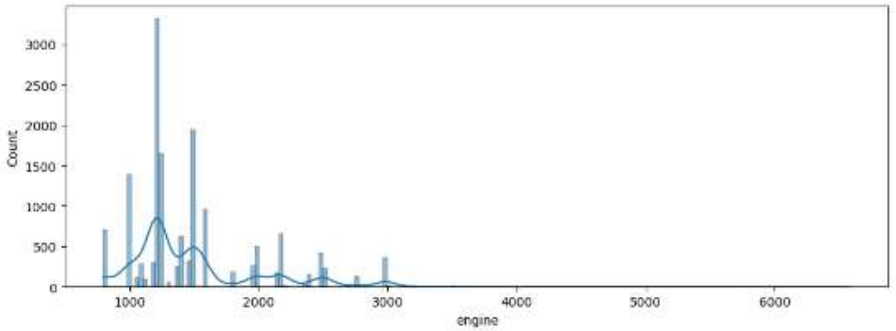


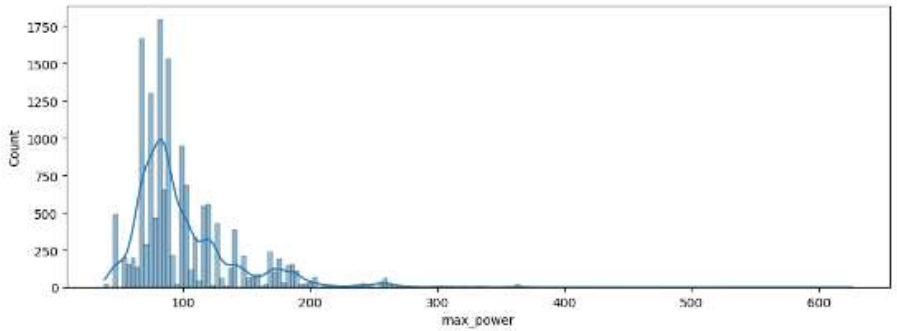
HISTOGRAM: (Detect skewness; visualize the distribution of the data; detect the position of mean median & mode)

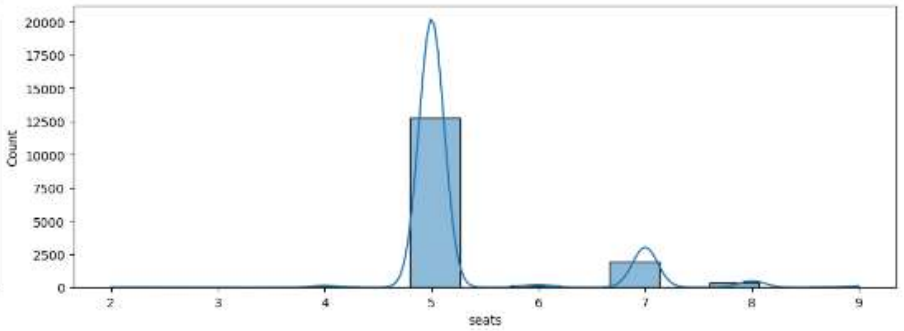


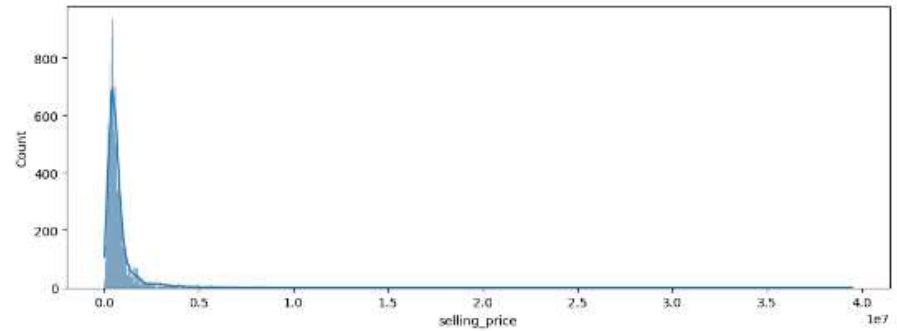












Inference:

'vehicle_age' :
- Distribution of data : High positive skewness(0.83), indicating presence of few outliers in the higher scale.
- Most of the vehicle age lie between 4 and 8 years.
- We can observe a maximum outlier value of 29 years.

‘km_driven’ :
- Distribution of data : High positive skewness(28.2), indicating presence of outliers in the higher scale & further notice huge no. of outliers.
- Most of the vehicles have been driven around 30000 and 70000 kilometers indicating mgiht have been sparsely used.
- We can observe a maximum outlier value of 3800000 years indicating there are also cars on sale which have been used a lot.

‘mileage’ :
- Distribution of data : Near normal distribution(0.11).
- Most of the vehicles give a mileage between 17 and 22.7, indicating most of the cars have acceptable mileage.
- We can observe a maximum outlier value of 33.54 mileage, which is a very good mileage for a second had car.

'engine' :
- Distribution of data : High positive skewness(1.6), indicating presence of outliers in the higher scale & we can notice huge no. of outliers.
- Most of the vehicles have engine capacity between 1197 and 1582, indicating they might be small hatchbacks.
- We can observe a maximum outlier value of 6592, which might be a luxury or sports car.

'max_power'
- Distribution of data : High positive skewness(2.4) indicating, presence of outliers in the higher scale & we can notice huge no. of outliers.
- Most of the vehicles have maximum power between 74 and 117.3, indicating most of the cars might be small hatchbacks or sedans.
- We can observe a maximum outlier value of 626, which might be a luxury or sports car.

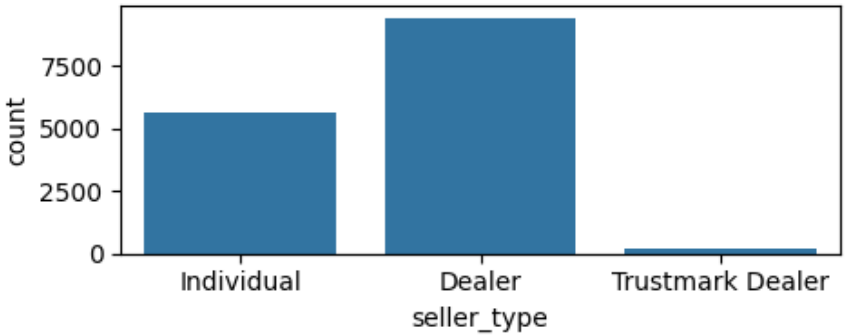
'seats'
- Distribution of data : High positive skewness(2), indicating presence of outliers in higher scale.
- Most of the cars seems to have 5 seats followed by 7 seats & 6 seats.
- There are even cars with 9 seats which may be mini vans & cars with 2 seats which may be luxury or sports cars.

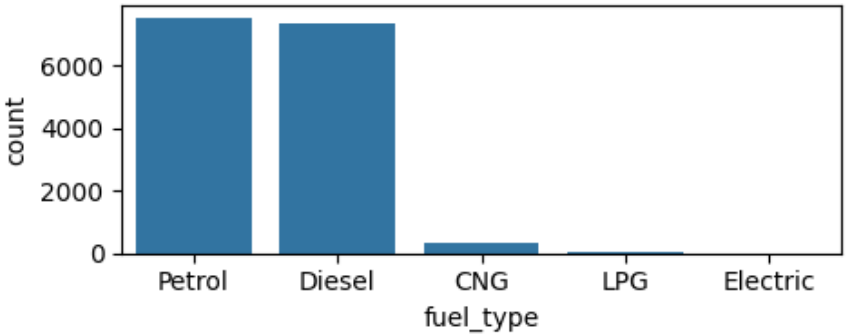
'selling_price'
- Distribution of data : High positive skewness(10.1) indicating, presence of outliers in higher scale & we can notice huge no. of outliers.
- Most of the vehicles have maximum power between 385000 and 825000, indicating they might be small hatchbacks.
- We can observe a maximum outlier value of 39500000, which might be a luxury or sports car.
- Target variable doesn't follow normal distribution.

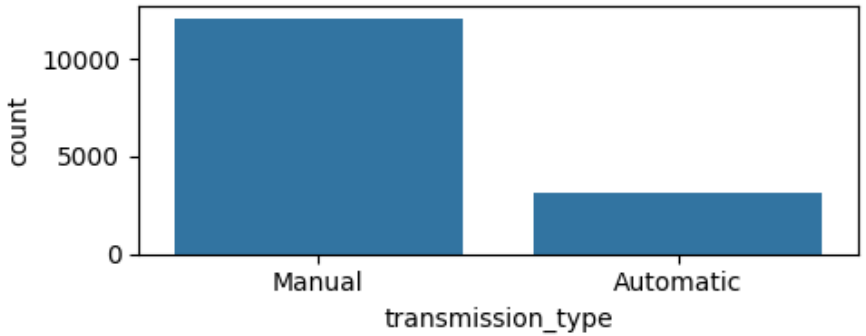
Univariate analysis - Categorical variables

COUNT PLOT:

(Visualize each of the subclasses total count in the form of a bar chart)

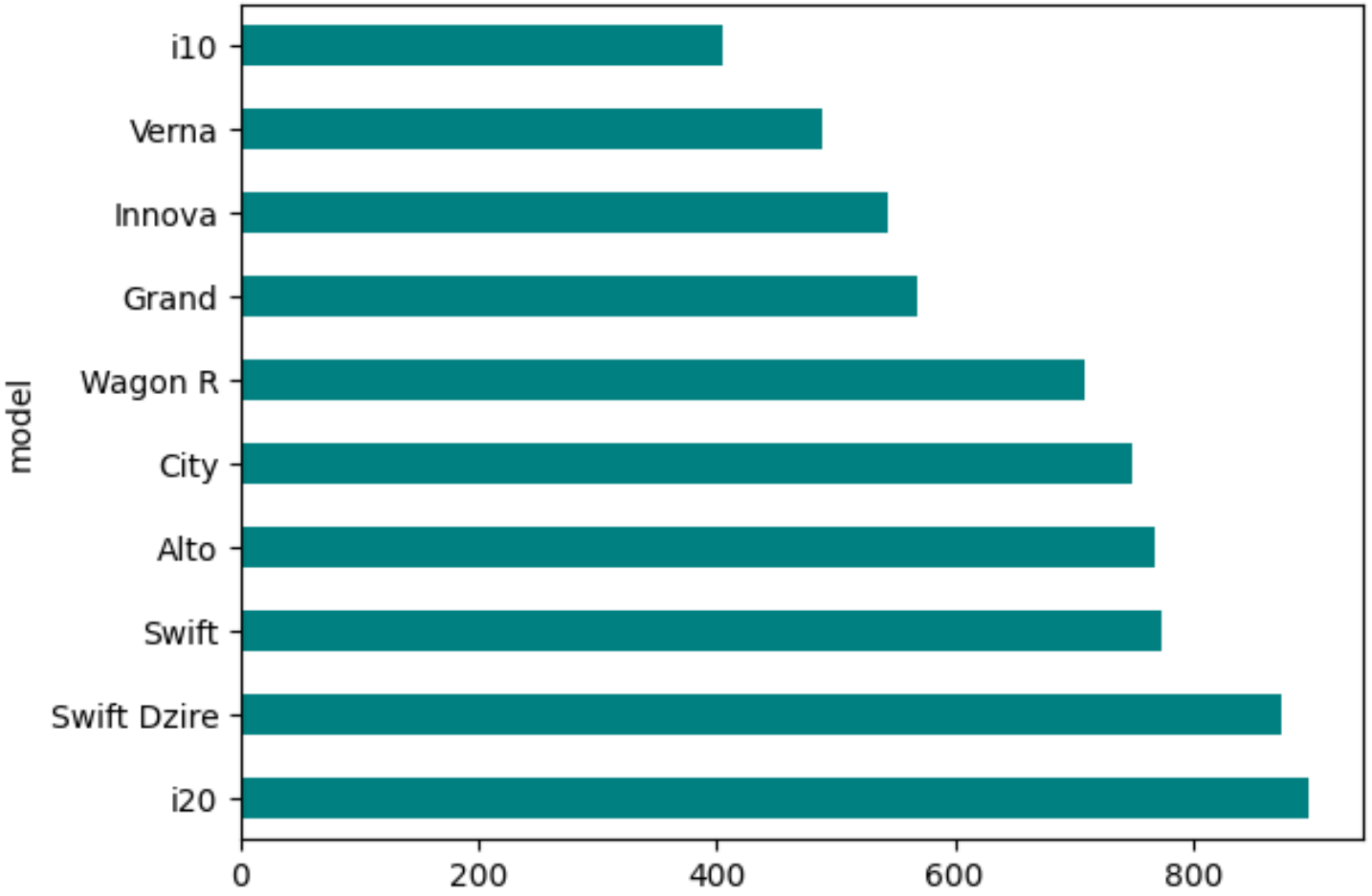






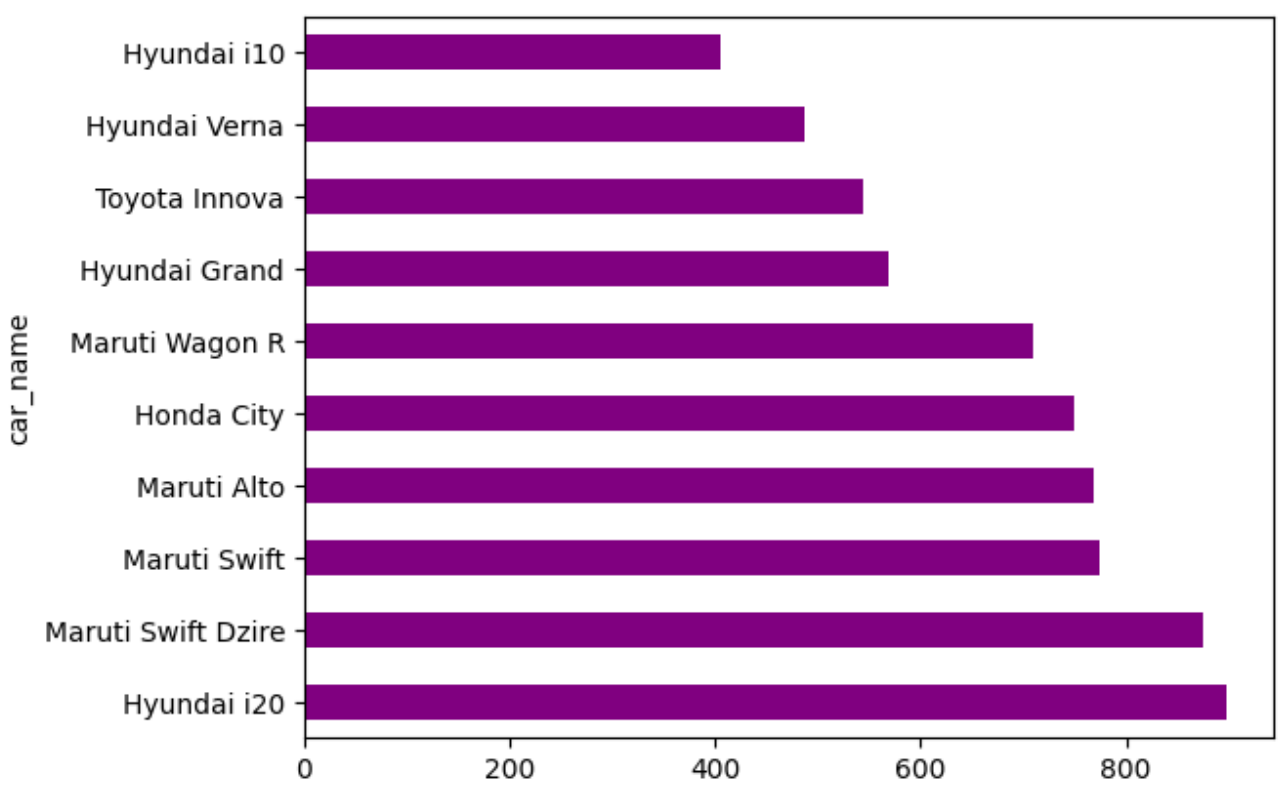
HORIZONTAL BAR PLOT:

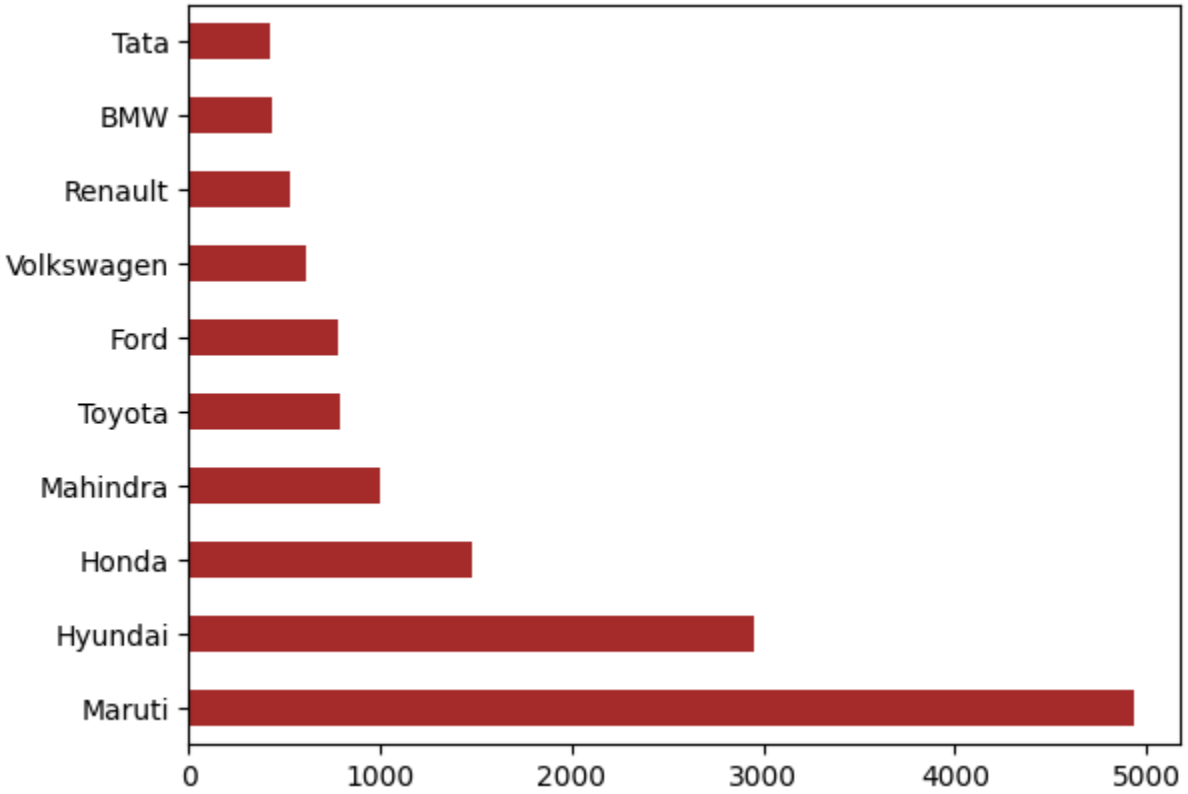
(Visualize each of the subclasses total count in the form of a horizontal bar graph)



HORIZONTAL BAR PLOT:

(Visualize each of the subclasses total count in the form of a horizontal bar graph)





Inference:

'seller_type'

- Most of the cars are being sold from 'Dealer' (9459) followed by 'Individual' sellers (5612).
- There are fewer no. of cars(173) sold from 'Trustmark dealer'.

'fuel_type'

- Most of the cars are being sold from 'Petrol' cars (7555), closely followed by 'Diesel' cars (7342) and followed by 'CNG' cars (299).
- There are very few no. of 'Electric' cars(4) & 'LPG' cars(44) on sale.

'transmission_type'

- Most of the cars on sale have 'Manual' transmission (12094).
- There are fewer no. of cars that have 'Automatic' transmission (3150).

'brand'

- Most of the cars on sale are from the brand 'Maruti' (4933) which are smaller cars like 'Maruti Swift Dzire' & 'Maruti Swift' followed by 'Hyundai'.
- There are fewer luxury brands on sale like 'Maserati', 'Ferrari', 'Mercedes-AMG', 'Rolls-Royce' ,'Force' & 'Bentley'.

'car_name’

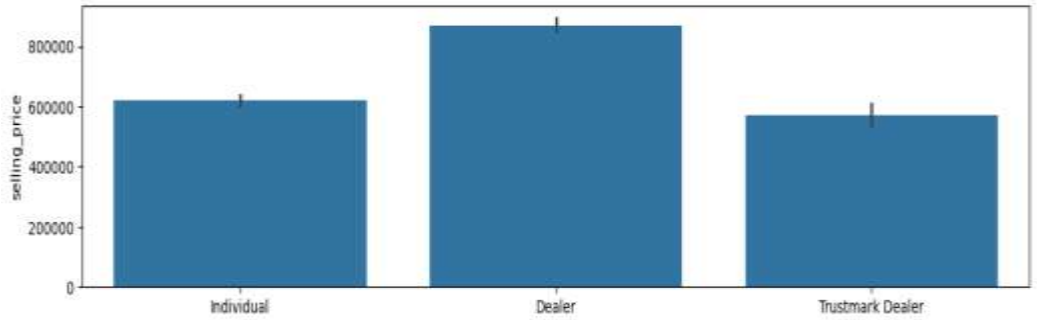
- Most of the cars on sale are 'Hyundai i20' (898), 'Maruti Swift Dzire' (875), 'Maruti Swift' (774) & 'Maruti Alto’ (768) & 'Honda City' (750).
- There are also luxury cars on sale like 'Mercedes-AMG C' , 'Ferrari GTC4Lusso', 'Force Gurkha' but fewer in numbers.
- There are fewer luxury cars on sale compared to smaller cars like hatchbacks

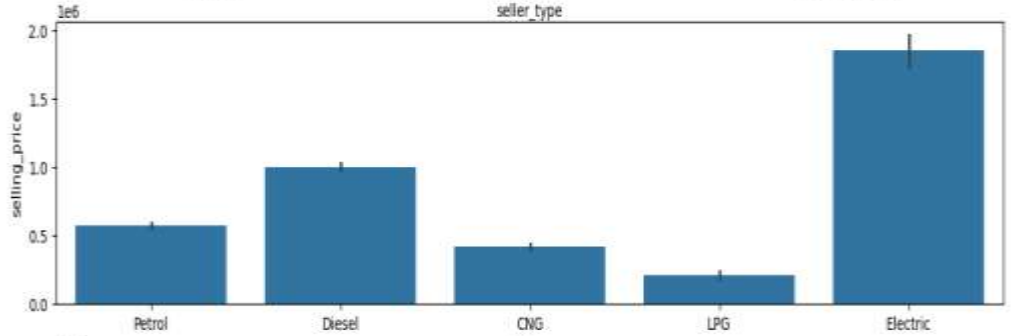
- 'model’

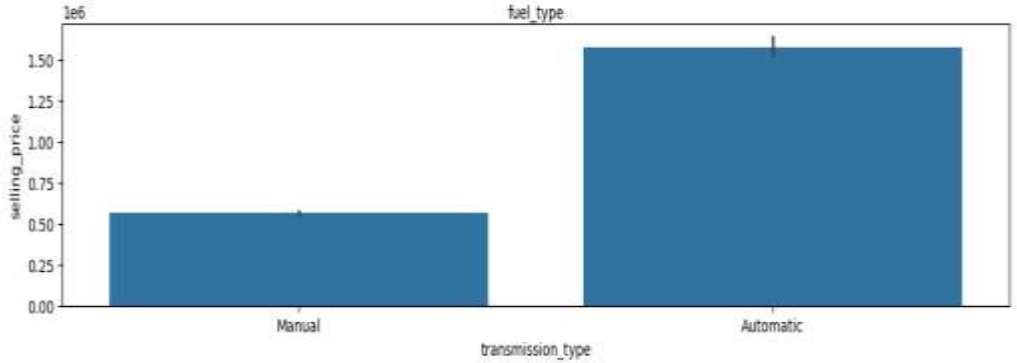
- Most of the cars on sale are 'i20' (898) closely followed by 'Swift Dzire' (875) & 'Swift' (774).
- There are few luxury cars on sale while most of the cars are smaller sedans or hatchbacks.

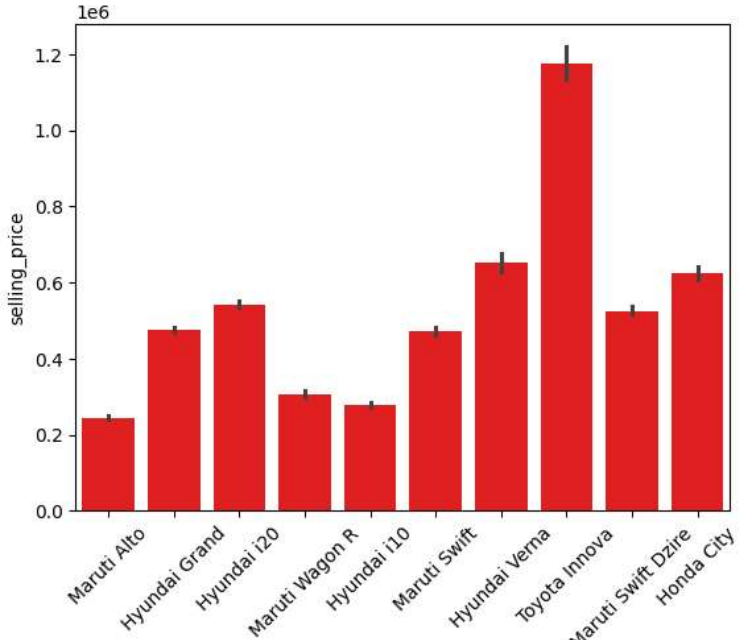
Bivariate analysis – Categorical vs Numerical

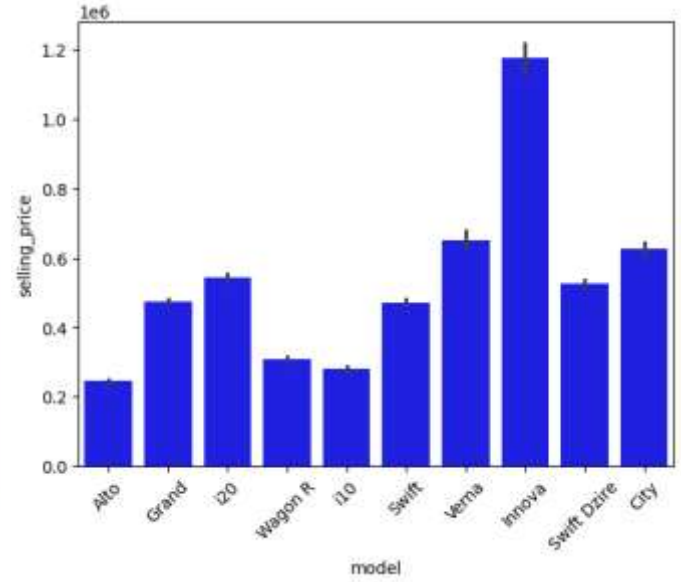
BAR PLOT: [Seaborn bar plot gives the mean of each subclass (of the target variable) in regards to the numeric variable]

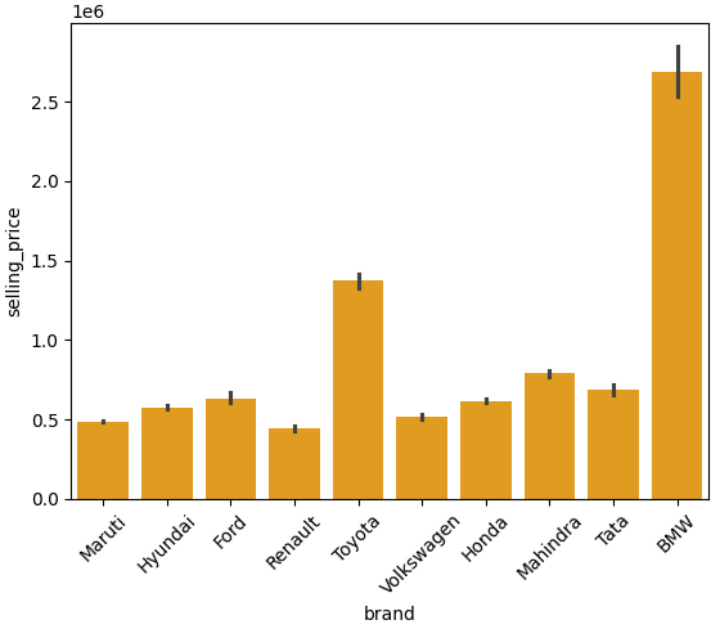












Inference:

'seller_type' VS 'selling_price'
- On average, most of the 'Dealer'ship cars have higher price followed by 'Individual' sellers.
- Cars from 'Trustmark Dealer' have lower selling price on average.

'fuel_type' VS 'selling_price'
- On average, 'Electric' cars have higher selling price compared to other type of cars considering the aim of EVs is higher upfront cost followed by higher savings in the longer run whihc is followed by 'Diesel' type cars.
- Lowest selling price is for 'LPG' fueled cars on average which might be to tempt buyers with lower price into this category given that most states lack good infrastructures for LPG fuel stations.

'transmission_type' VS 'selling_price'
- On average, Automatic cars have higher selling price which is expected given the additional technology involved in these cars.
- Manual cars have lower selling price on average.

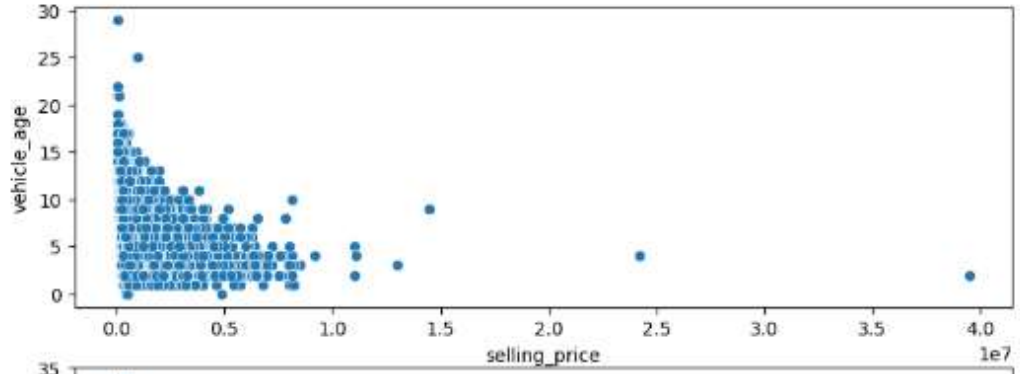
'car_name', 'brand' & 'model' VS 'selling_price'
- On average, most of the hatchbacks & smaller sedans (i10, i20, Swift, Alto, Swift Dzire) have smaller selling price.
- On average, hatchback & smaller sedans brands like Maruti, Hyundai, Tata have smaller selling price.
- There are also others brands like Toyota (Innova car) & BMW on sale which has considerably higher price.

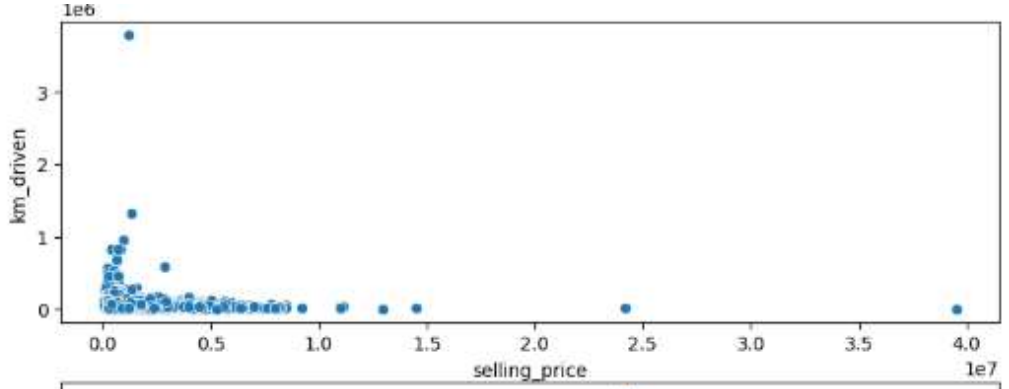
Bivariate analysis

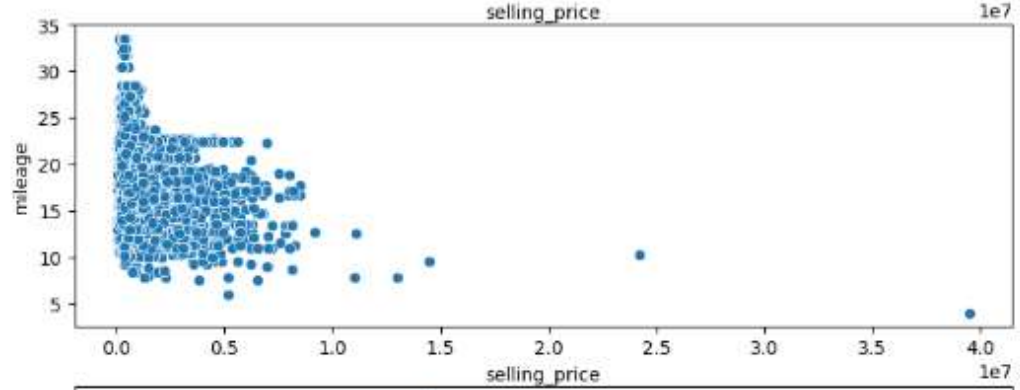
Numerical VS Numerical

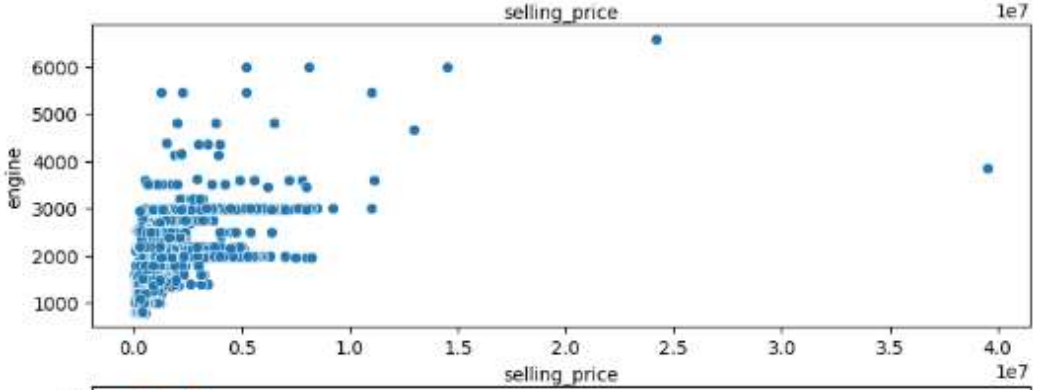
SCATTERPLOT

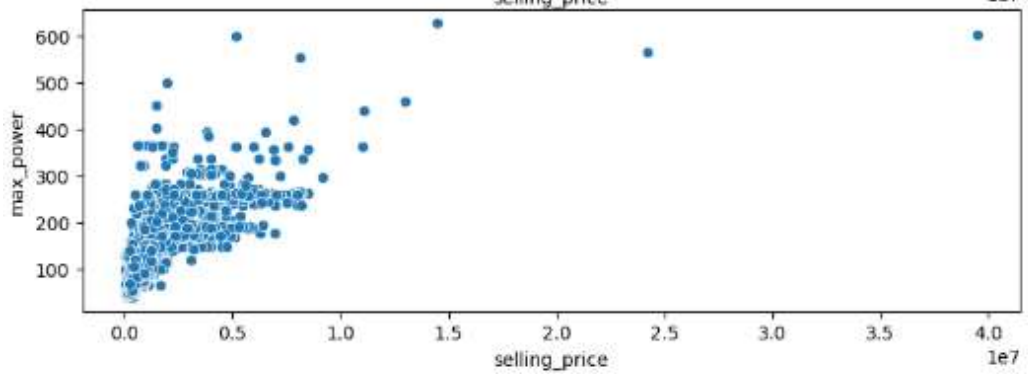
(Understand the relationship between two numerical variables)

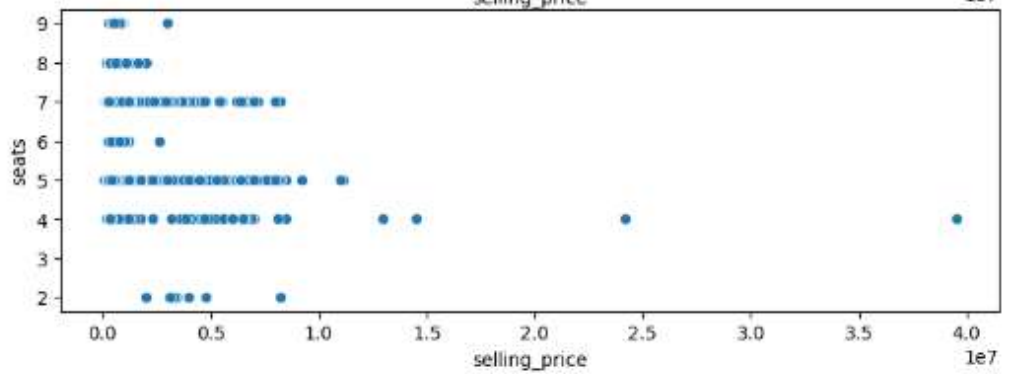






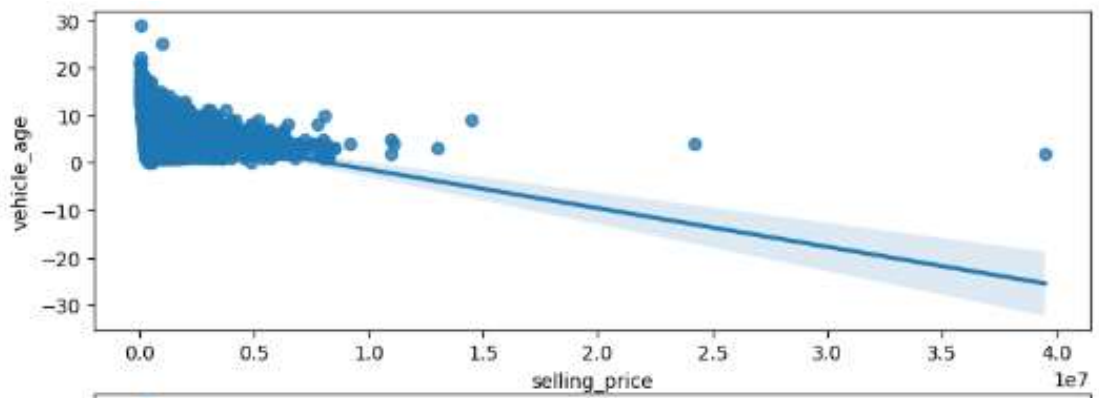


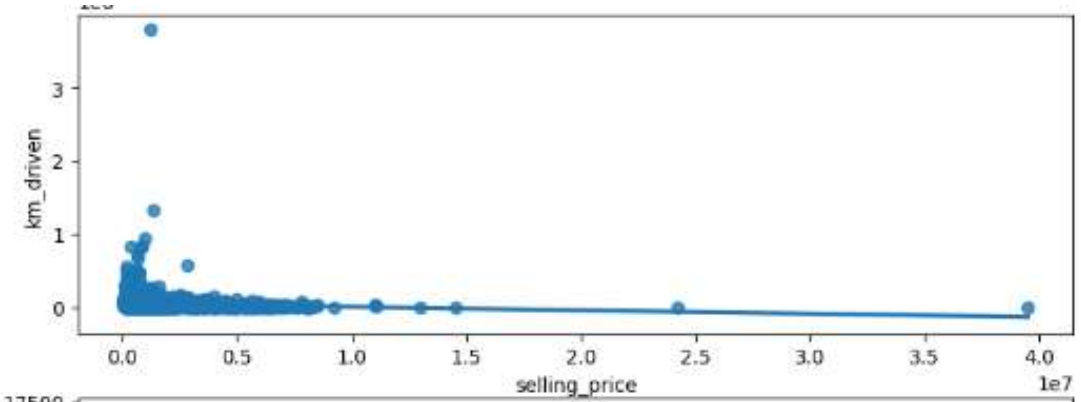


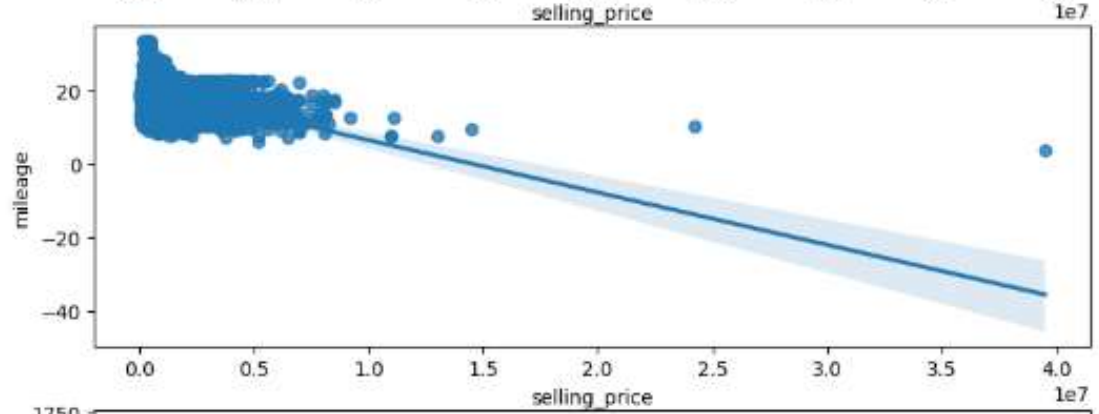


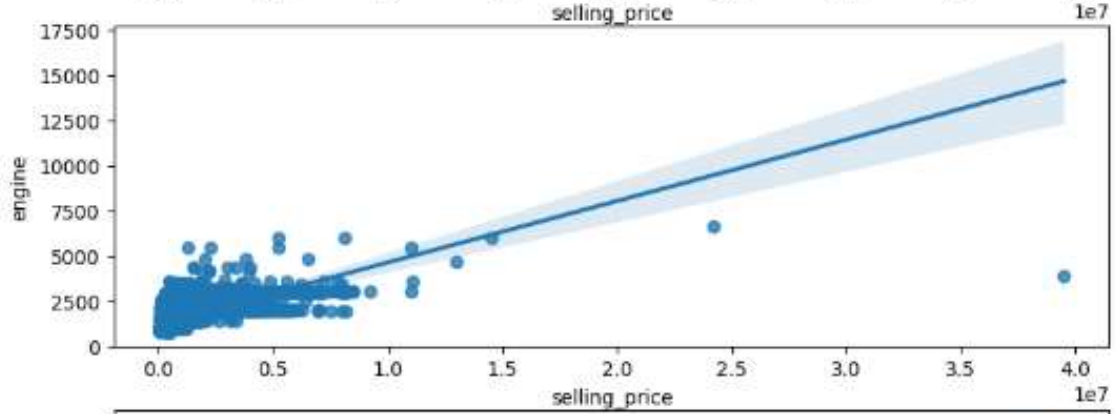
REGRESSION PLOT

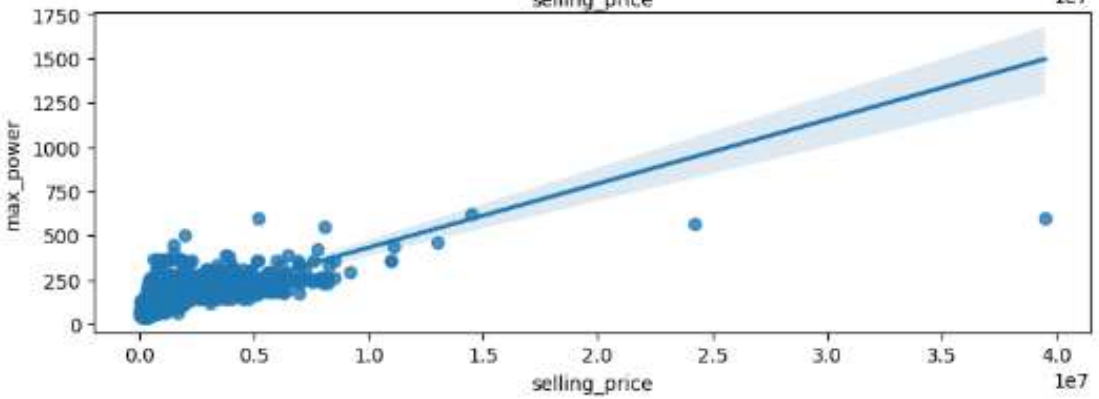
(Understand the relationship
between two numerical variables)

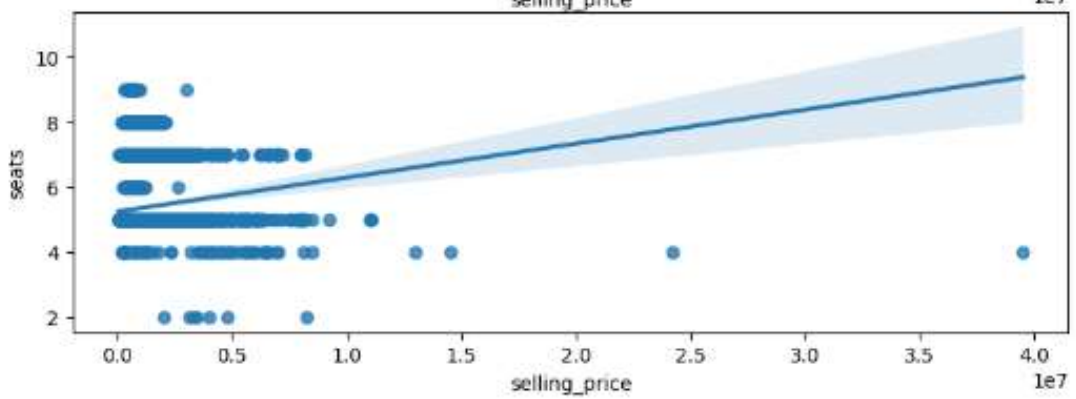






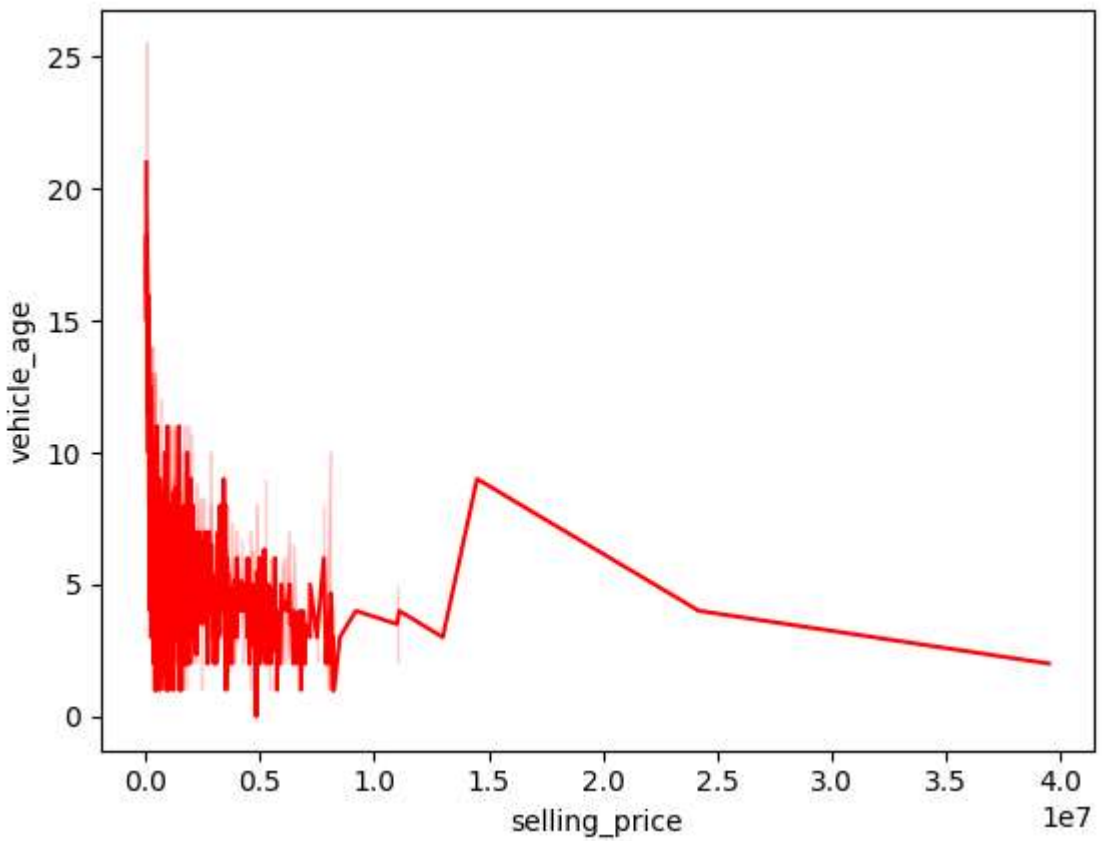






LINE PLOT

(Relationship between timeseries variables)



Inference:

'vehicle_age' & 'selling_price'-

- We can see as the vehicle age decreases the selling price tends to increase but on a smaller magnitude.
- Sparsely used cars have higher selling prices.
- Cars that have been used a lot have lower selling price.

Inference:

'km_driven' & 'selling_price':

- We can't observe any direct relation between the no. of kilometers driven and the selling price of a car.

'mileage' & 'selling_price':

- We can see as the mileage decreases the selling price increases.
- Since most of the cars are hatchbacks we can see they might provide good mileage hence selling price of the hatchbacks might be low.
- There are few luxury cars which give lower mileage but have high selling price.

'engine' & 'selling_price':

- We can see observe significant positive correlation because as the engine capacity increases the selling price also increases.
- Hatchback and sedans have smaller price compared to luxury cars which have a higher selling price.

'max_power' & 'selling_price':

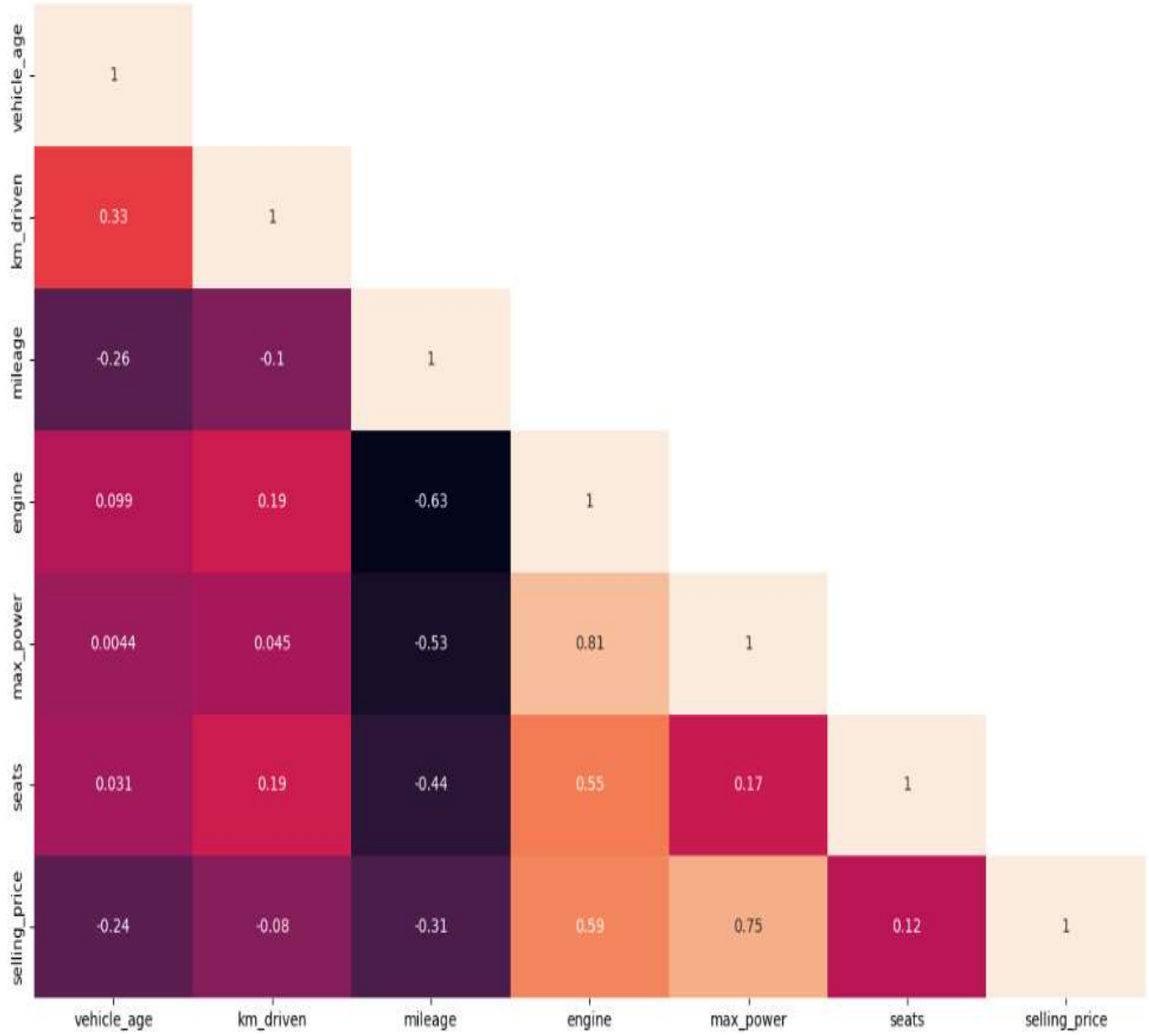
- We can see observe significant positive correlation because as maximum power increases the selling price also increases. Hatchback and sedans (will have lower power) have smaller price compared to luxury cars (will have higher power) which have a higher selling price.

'seats' & 'selling_price':

- The number of seats might play a minor role in determining the selling price of a car as we can see for most cars as the seats increase the selling price also increases

Multivariate analysis

Correlation matrix : (Visualize the correlation between the numeric variables)



Inference:

We can observe high positive correlation between:

- 'selling_price' & 'max_power' (0.75)
- 'selling_price' & 'engine' (0.59)
- 'seats' & 'engine' (0.55)
- 'max_power' & 'engine' (0.81)

We can observe weak positive correlation between:

- 'selling_price' & 'seats' (0.12)
- 'seats' & 'max_power' (0.17)
- 'km_driven' & 'engine' (0.19)
- 'km_driven' & 'seats' (0.19)
- 'km_driven' & 'vehicle_age' (0.33)

We can observe strong positive correlation between:

- 'max_power' & 'mileage' (-0.53)
- 'engine' & 'mileage' (-0.63)

We can observe weak negative correlation between:

- 'selling_price' & 'vehicle_age' (-0.24)
- 'selling_price' & 'mileage' (-0.31)
- 'seats' & 'mileage' (-0.44)
- 'mileage' & 'vehicle_age' (-0.26)

We can observe the presence of Multi-collinearity in few of the independent variables.

- 'engine' & 'seats' [0.55]
- 'engine' & 'max_power' [0.81]
- 'km_driven' & 'vehicle_age' [0.33]
- 'engine' & 'mileage' [-0.63]
- 'max_power' & 'mileage' [-0.53]

Statistical test of significance:

Since the data has failed to meet the assumptions

(i) Data normality (Jarque Bera test)

(ii) Data has equal variance (Levene test)

we have opted for Non-parametric test to try to understand if any relationship exists between the independent variables and dependent variable.

Numeric VS Numeric – Spearman R (Non-Parametric)

Categorical VS Numeric – Kruskal Wallis test (Non-Parametric)

- Observing the pvalue of the Numeric & Categorical variables there seems to be a relationship between most of the 'x' variables and the 'y' variable.

Data preprocessing

Scaling the data

- Scaling is performed in the dataset for the numeric variables after train test split as we can hide the mean standard deviation of training data from the test data.
- Scaling has been performed as part of one of Yeo-Johnson transformation (which performs both scaling & transformation)

Outlier treatment:

- The popular methods of handling the outliers are Trimming/removing the outlier based on IQR or z-Score or capping them.
- But, we would like to consider the outliers in this dataset as a pattern and prefer not to treat them but to handle them differently since it is important for the model to get trained based on some extreme values in order to predict in an efficient way consider their nature in the E-Commerce sector.

Transformation technique:

- We prefer transformation of the numeric variables (using Yeo-Johnson transformation) so that we can try to convert them to near normal distribution.

Encoding the Categorical Variables:

- The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model. Since most machine learning models only accept numerical variables, pre-processing the categorical variables becomes a necessary step. Converting these categorical variables to numbers such that the model is able to understand and extract valuable information is known as Encoding. There are various encoding techniques available like Dummies, One Hot Encoder, Label Encoder, Ordinal Encoder etc., We have used Label Encoder for encoding our categorical variables considering the no. of categorical variables and we have also seen there is not a significant improvement in model performance when dummy encoding was utilized.

- We have utilized **Label encoding technique**. In this case, retaining the order is important. Hence encoding should reflect the sequence. In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person likewise.

- The final processed data which we would be using in building various Regression Models & Non-Parametric Models like Decision Tree and Random Forest. We, with the help of the built models we would infer on the significance and effects of each independent variable on our target variable for predicting the patterns and rate of successful conversion to give some insightful ideas for effective marketing.

Model Building & Evaluation

- The Modeling is the core of any machine learning project. This step is responsible for the results that should satisfy or help satisfy the project goal.

- Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data.

- Our problem statement come under the Classification thus we have decided to use various models namely

    1. Linear Regression Model
    2. OLS (Ordinary Least Squares) Model
    3. Decision Tree model
    4. Random Forest model

- **Evaluation metrics**: R2 & RMSE

Hyper tuning the model

Model: Decision Tree Classifier

GridSearchCV (run for below parameters)

Criterion	**squared_error, friedman_mse, absolute_error, poisson**
Max Depth	30, 35, 40
Min Samples Split	50, 60, 70, 80, 90
Min Samples Leaf	10, 20, 30, 40

Best Parameters

Criterion	**squared_error**
Max Depth	30
Min Samples Split	10
Min Samples Leaf	60

Model: Random Forest Classifier

GridSearchCV (run for below parameters)

Criterion	**squared_error, friedman_mse, absolute_error, poisson**
Max Depth	30, 35, 40
Min Samples Split	50, 60, 70, 80, 90
Min Samples Leaf	10, 20, 30, 40

Best Parameters

Criterion	**squared_error**
Max Depth	30
Min Samples Split	10
Min Samples Leaf	50

Model performance – Test & Training scores

	Model	R2 Score	RMSE
0	Linear Regression [Base model] - Test	0.640675	466942.068266
1	Linear Regression [Base model] - Train	0.622916	577109.425480
2	Linear Regression [Scaled data] - Test	0.641636	466317.192875
3	Linear Regression [Scaled data] - Train	0.622916	577109.425480
4	Linear Regression [Scaled data] - Test	0.534126	531683.529732
5	Linear Regression [Scaled data] - Train	0.443828	700879.537796
6	Linear Regression [RFE columns] - Test	0.638889	468100.608902
7	Linear Regression [RFE columns] - Train	0.622118	577719.592026
8	Linear Regression [RFE & scaled] - Test	0.637924	468725.685926
9	Linear Regression [RFE & scaled] - Train	0.622118	577719.592026
10	Linear Regression [RFE & transformed] - Test	0.501424	550028.281176
11	Linear Regression [RFE & transformed] - Train	0.415874	718277.027065
12	Decision Tree - Test	0.826161	324782.878399
13	Decision Tree - Train	0.999595	18909.843483
14	Decision Tree (Tuned) - Test	0.791917	355334.313095
15	Decision Tree (Tuned) - Train	0.819902	398834.694215
16	Random Forest - Test	0.884105	265186.680192
17	Random Forest - Train	0.978694	137179.563863
18	Random Forest (Hyper parameter tuned) - Test	0.904958	240147.240156
19	Random Forest (Hyper parameter tuned) - Train	0.807299	412554.184033

Best models that have been achieved without SMOTE

1. Random Forest (Hyper parameter tuned)
2. Random Forest
3. Decision tree

Conclusion:

- The linear regression model that we built didn't perform well because the data didn't satisfy the conditions of a linear regression model which are:

- **Presence of strongly influential outliers:**
  Most of the cars on sale are smaller hatchbacks and smaller sedans but there are also (outliers) luxury cars on sale
- **Presence of strong correlation between numeric 'x' variables & 'y' variable**:
  Out of the 6 numeric variables only 2 have strong correlation to the 'y' variable rest of the variables lack a strong correlation
- **Residuals follow normality:**
  Based on the pvalue from the Jarque-Bera test we can infer the residuals do not follow normality.
- **Data has linear relation with 'y' variable:**
  Based on the pvalue from Rainbow test we can see data lacks significant relationship between the 'y' variable.
- **Absence of Multi-Collinearity:**
  Based on the Correlation matrix, Condition number & VIF (Variance Inflation Factor) we can observe presence of Multi-Collinearity between 'x' variables.
- **Absence of Auto-Correlation**:
  Based on the pvalue from Durbin Watson test we can confirm the presence of relationship between immediate succeeding residuals (Auto-Correlation)
- **Absence of Heteroscedasticity:**
  Based on the pvalue from Breusch pagan test we can confirm the presence of Heteroscedasticity in the data.

- Since the model doesn’t satisfy the requirements of a Parametric model (model that has assumptions on the data) like Linear regression model we have tried to build Non-parametric model (model that has no assumptions on the data) like decision tree which had good prediction power compared to the Linear regression model we ran into a issue of overfitting as expected in a Decision tree but we ran a Grid Search CV to find the best hyper parameters and no. of cross validations to minimize the effect of overfitting by pruning the decision trees.

- Instead of depending on a single decision tree we also tried and built multiple trees via Random forest which also faced overfitting which we have reduced by finding the appropriate Hyper Parameters to prune the individuals trees that were built.

- Hence the model performance was significantly better on Non-parametric models than Parametric models given the nature of the data at hand which is evident based on the Model performance metrics like R2 and RMSE scores.