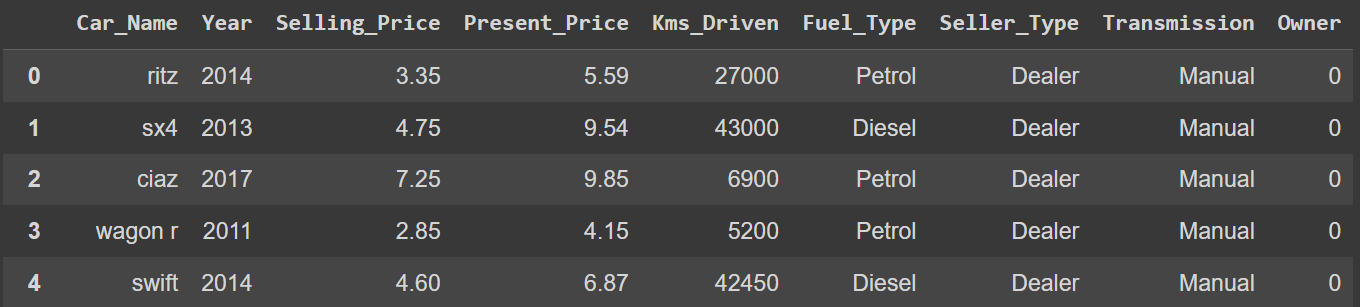
**LINEAR REGRESSION PROJECT – CAR PRICE PREDICTION**

**The main objective of this report** is to predict vehicle prices using different regression machine learning models on second-hand car data set.

**The data set** has the following 8 features:

* Car\_Name
* Year
* Selling\_Price
* Present\_Price
* Kms\_Driven
* Fuel\_Type
* Seller\_Type
* Transmission
* Owner

Here is an example of the first 5 rows of the data:

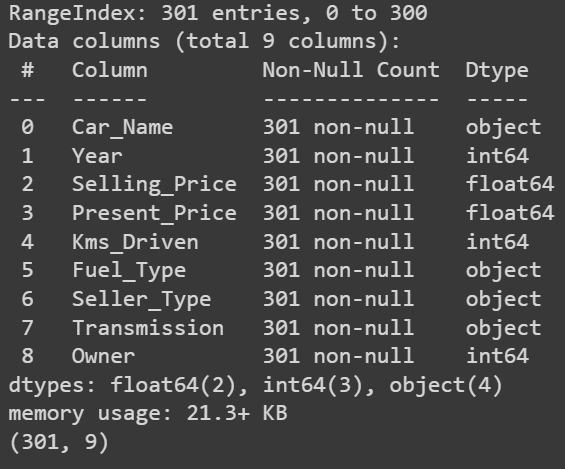


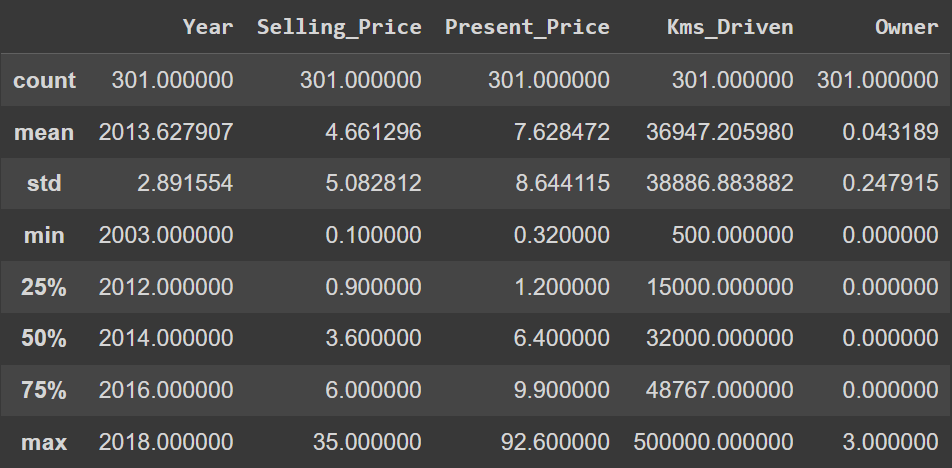
The length of the data set is 301 without any null or missing values. Considering that I will be predicting Selling\_Price, I have 301 rows and 8 features to perform regression. Therefore this dataset seems a bit challenging for regression. However, I wanted to use this dataset to see how well I can perform the regression.

In the rest of the report, I will perform EDA and FE, apply different regularization techniques, summarize the key findings, and make recommendations on the analysis that I performed. Therefore, information regarding all the necessary steps that should be undertaken to perform this regression project will be presented in the next sections of the report.

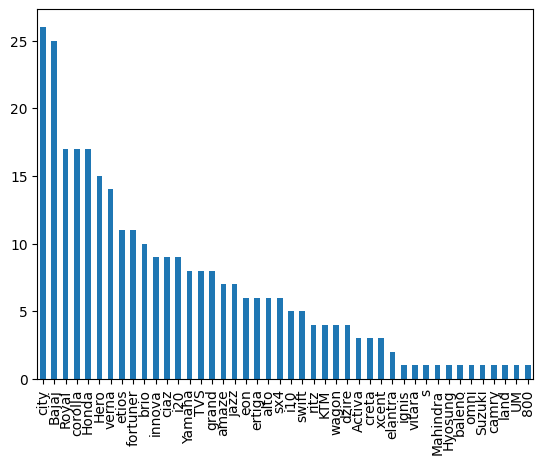
1. **PRELIMINARY EDA AND FEATURE ENGINEERING**

* Data info:



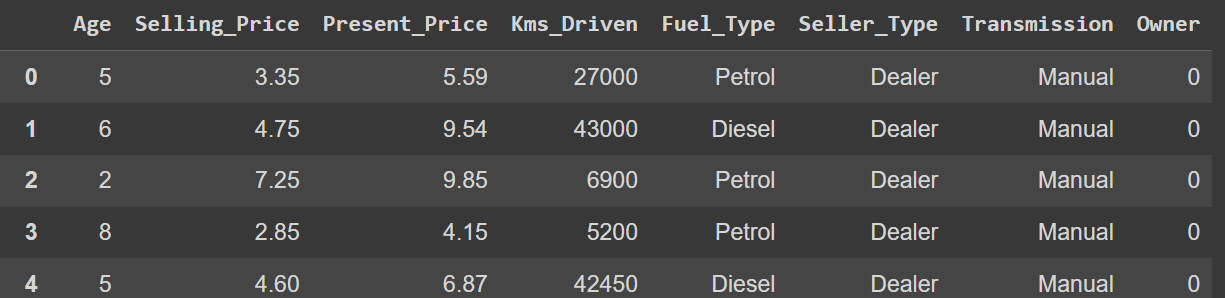


* The data set has an average of approximately 2013 model cars, driven approximately 37K kms on average. Most of the cars had no other owners.
* The figure below shows the count of different car names. Since (i) the data set has 301 rows, (ii) there are 98 unique car names, and (iii) almost half of them has less than 5 counts in the data set, I decided to drop the car name count from the dataset since I think it will not be a useful feature for the regression. Including car names can cause too many fluctuations in the dataset, making the regression perform worse.

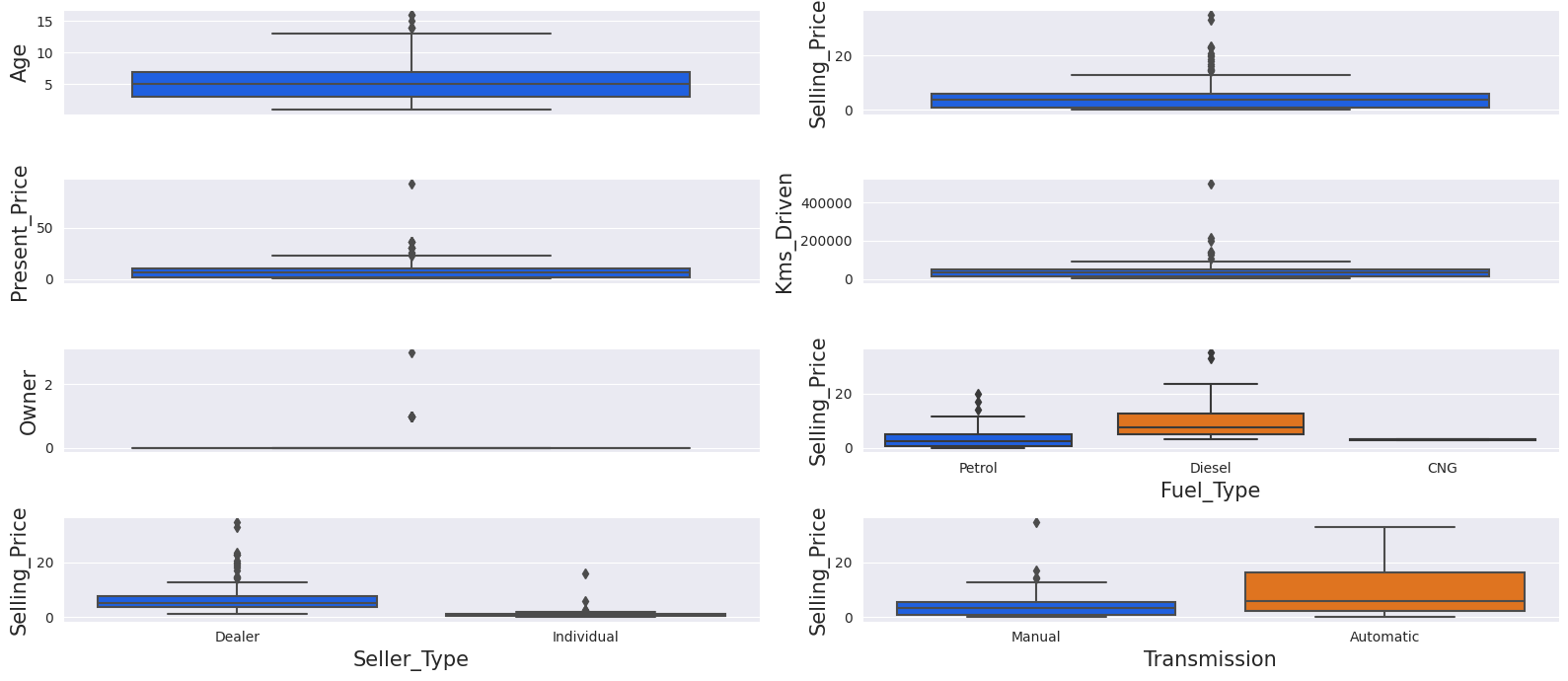


1. **FEATURE ENGINEERING AND SECONDARY EDA**

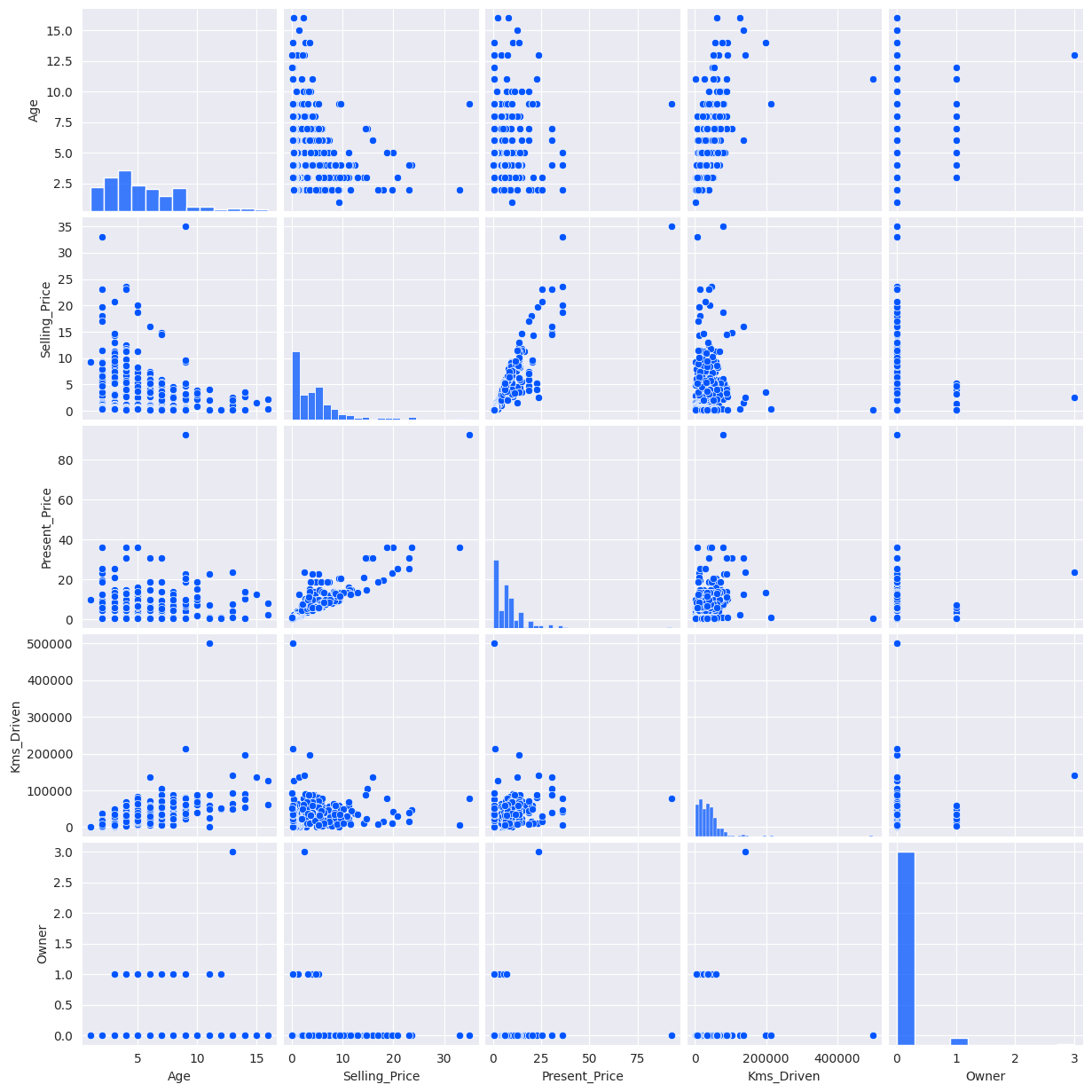
* After the preliminary analysis and dropping the car names, the model of the cars is converted to age since it will make the regression model more meaningful. For this conversion, the latest model in the data set was considered as age “1”. The modified dataset had now 8 columns:



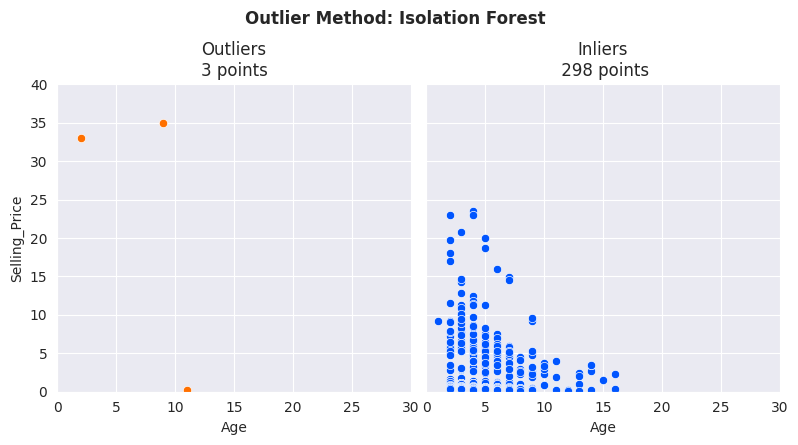
* Now performing data cleaning: no missing or fully duplicated rows. Therefore, moving to outlier detection and removal analysis. Using IQR method, the graphs below are plotted. The graphs show that Selling\_Price, Present\_Price and Kms\_Driven features seem to have some outliers.

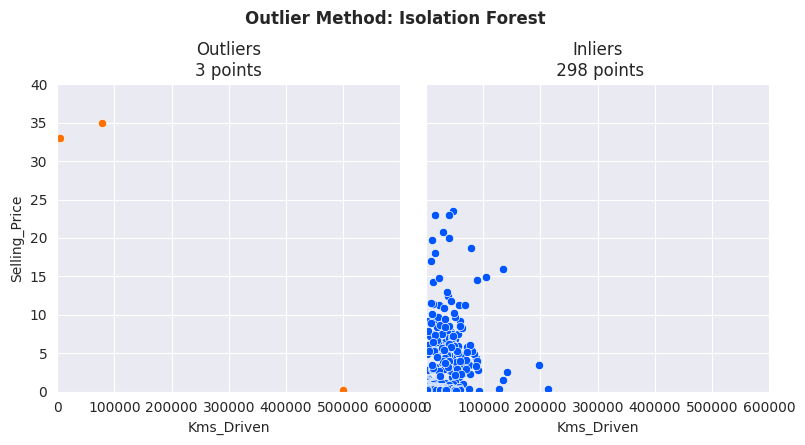


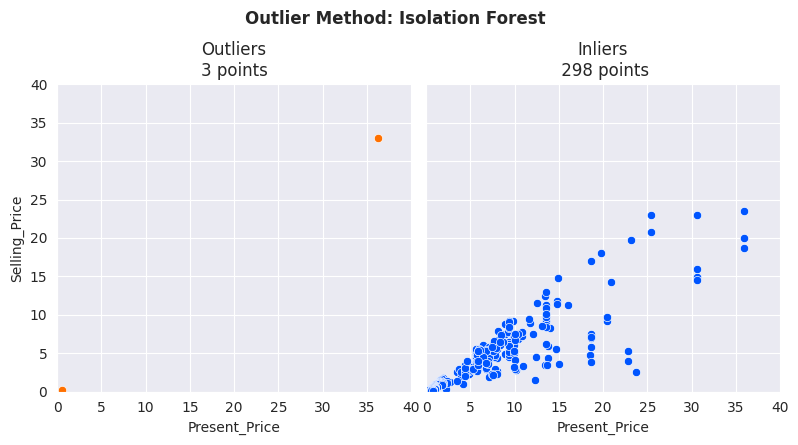
* Checking the outliers also using pair plots to see if there are any additional outliers that are visible when using multi-variate plots. It looks like the same 3 features are still the ones with outliers.



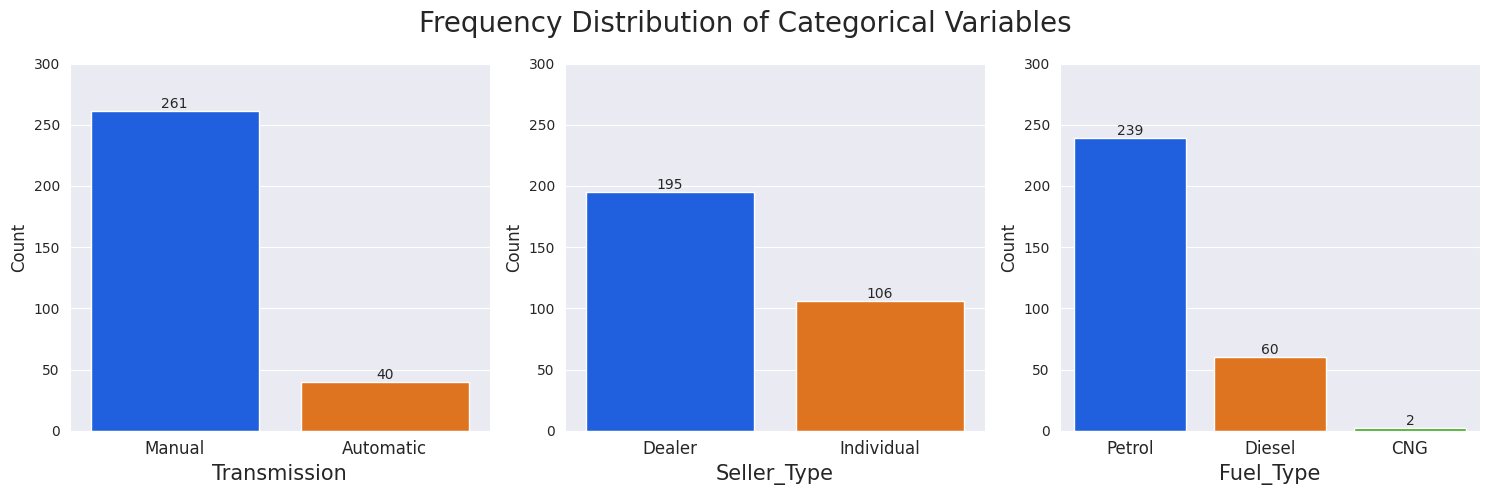
* The outliers are also checked with the Isolation Forest method to look deeper into outliers using the Selling\_Price, Present\_Price and Kms\_Driven. However, Age is also used in making the pair plots using Isolation Forest. The parameters required in the computation of isolation forest, especially the contamination ratio is selected based on the initial observations using IQR method and pair plots presented above.



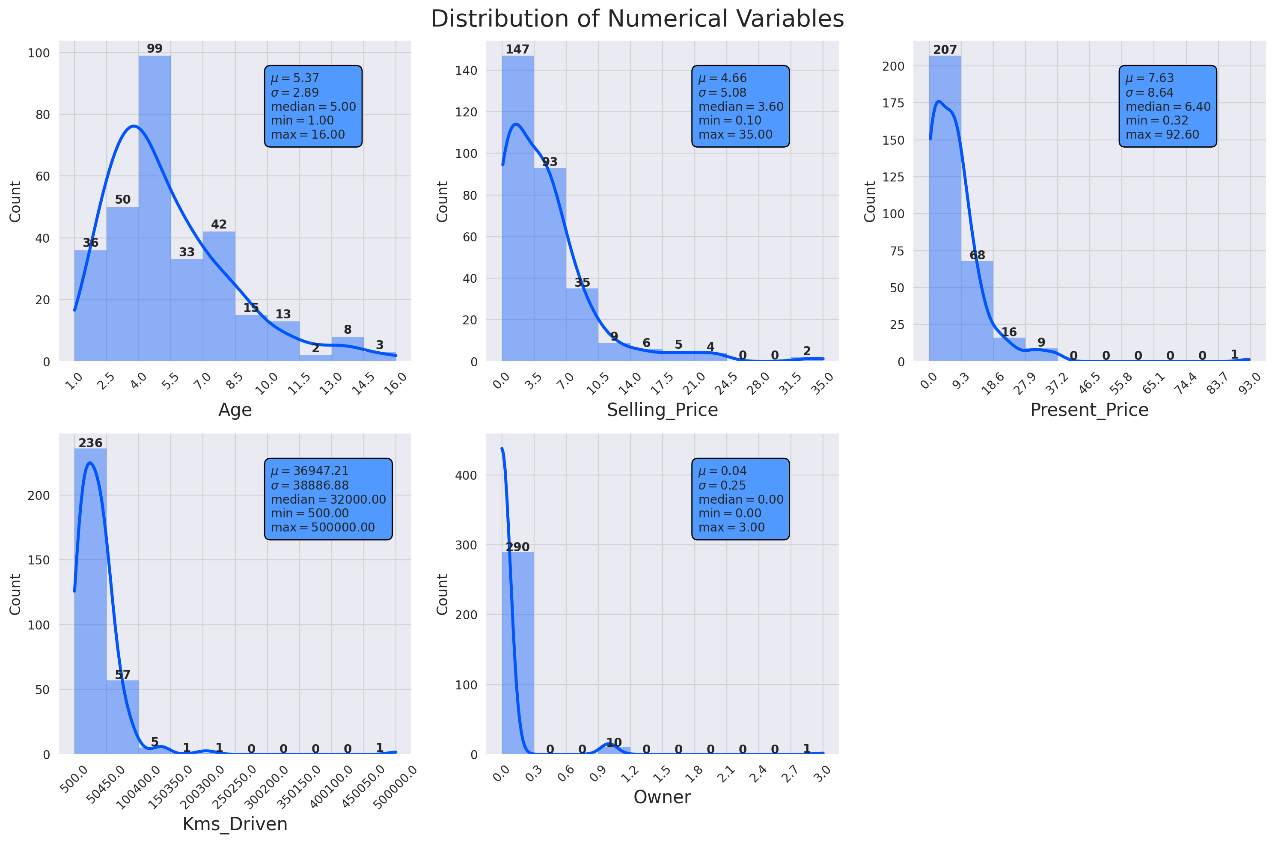




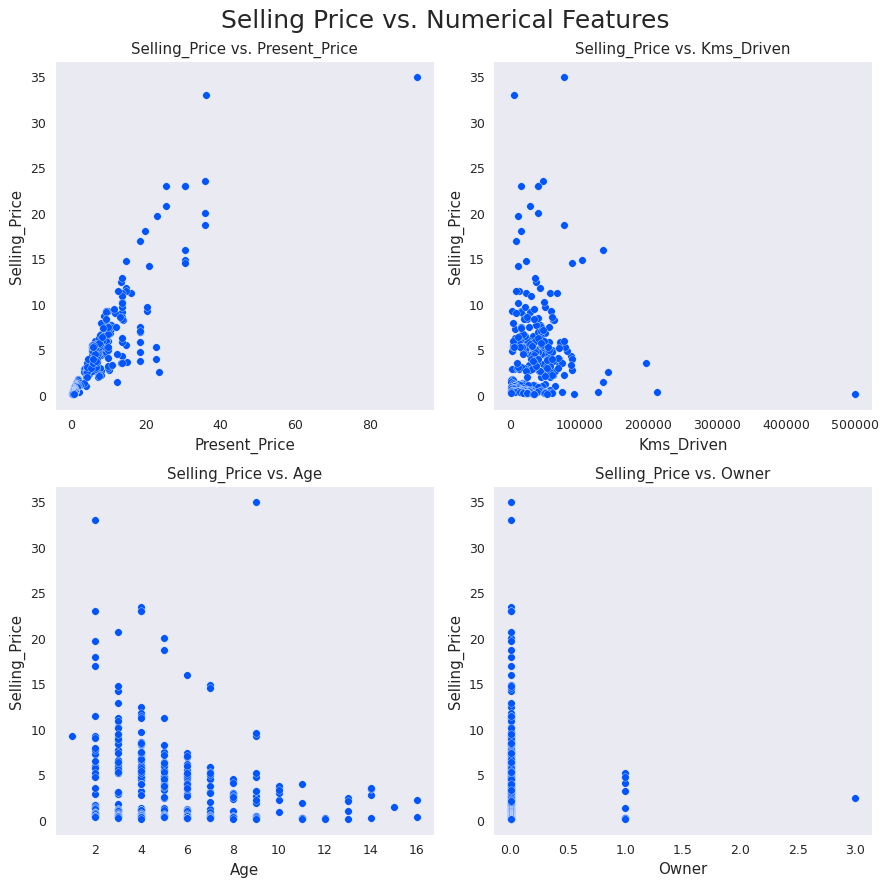
* After examining all the plots above regarding outliers, it was decided to replace the values of Selling\_Price above 25, Kms\_Driven above 300000 and Present\_Price above 50 with their corresponding  “q3 + (1.5 \* iqr)” values. where q3 is the 75th quantile and iqr is the difference between 75th and 25th quantiles.
* After cleaning the data and modifying and removing the features, a detailed EDA is performed. The figure below shows that most of the cars in the dataset has manual transmission, sold by a dealer and operate on regular gas (petrol).



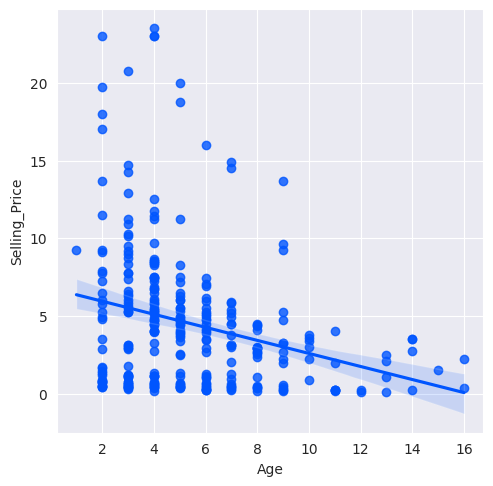
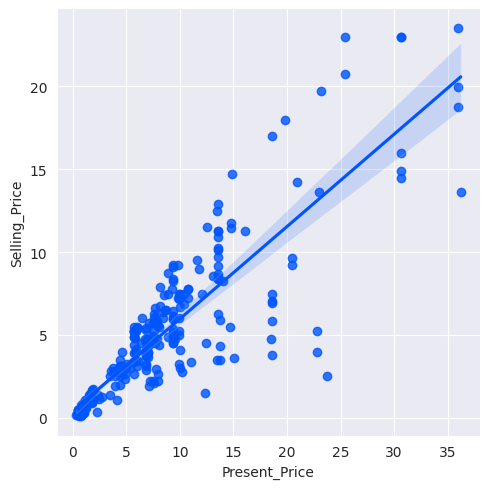
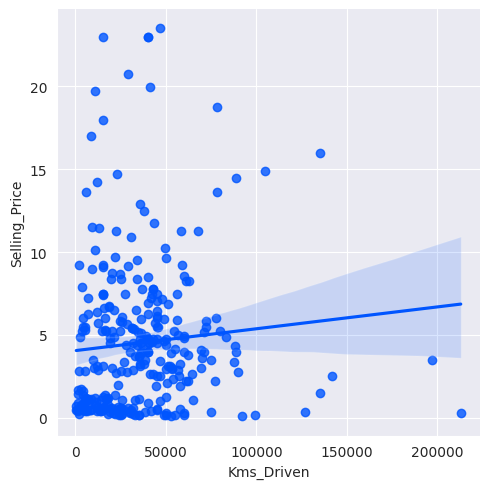
* The figure below demonstrates the distribution of numerical features after outliers were removed.



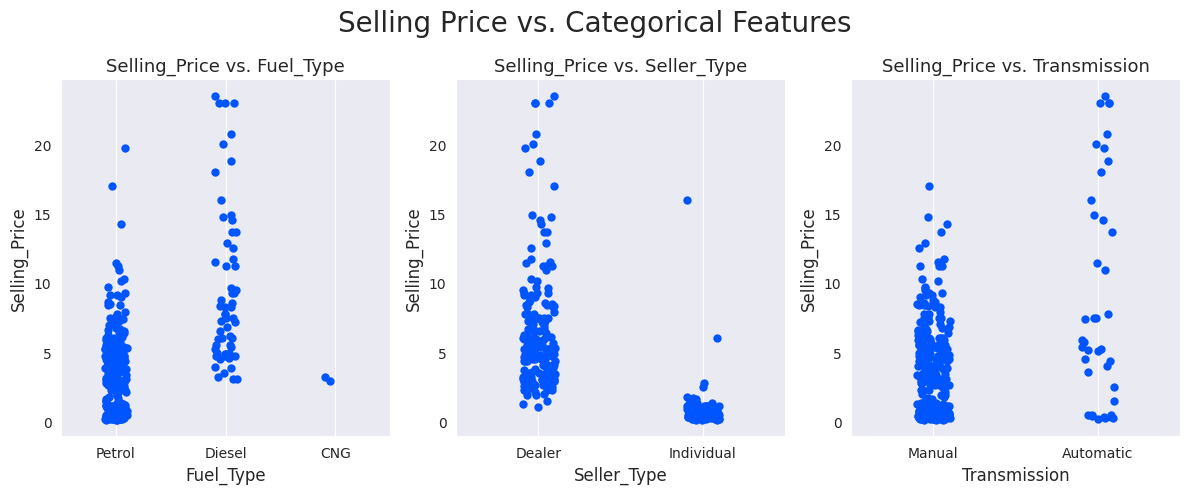
* The figure below plots the Selling\_Price with other numerical values using the cleaned dataset. The figures show that (i) Present\_Price is directly correlated to Selling\_Price, (ii) Selling\_Price decreases as the Kms\_Driven increases, but it can be still high at low Kms\_Driven values, (iii) Selling\_Price decreases as the Age increases, but it can be still high up to age “10”, and (iv) there is a clear correlation between number of owners and selling price: the price decreases as the number of the owners increases.



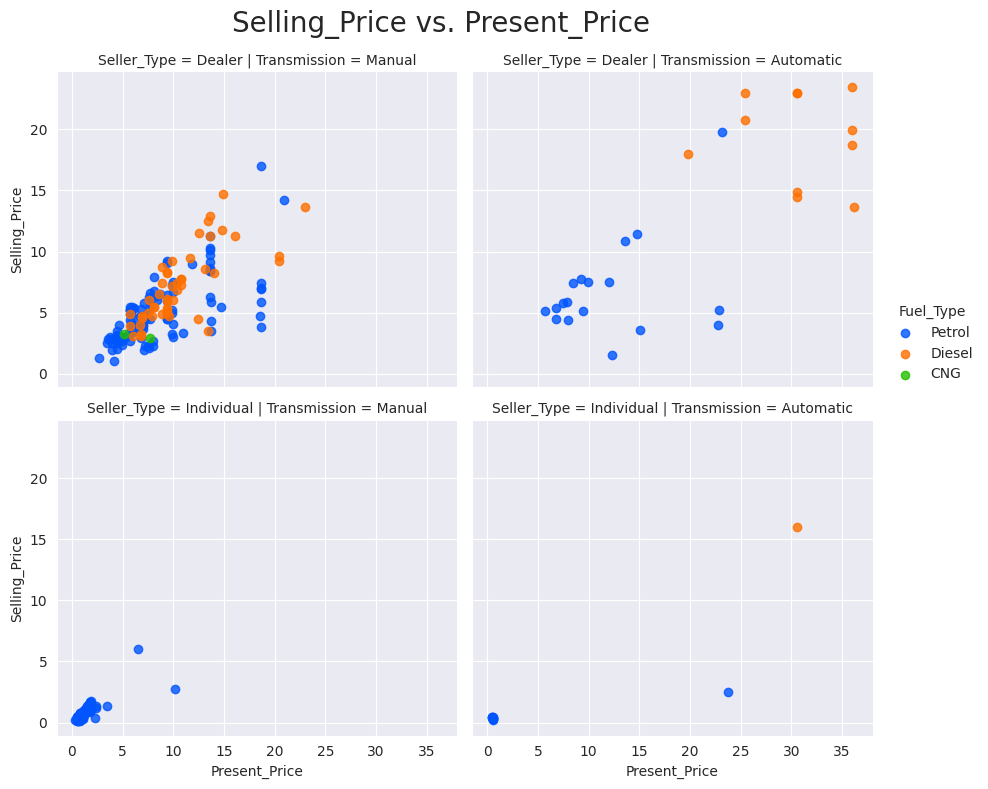
* The figure below is just a repetition of the previous fundings, but a linear regression is placed on the figures.



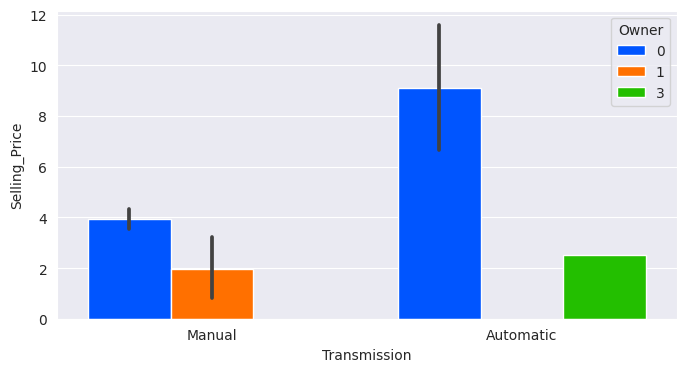
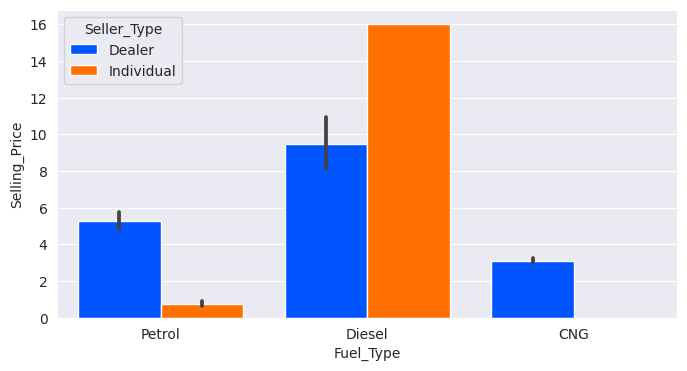
* The figure below shows that (i) the price of Diesel cars is higher compared to cars functioning with other type of fuels, (ii) it is more expensive to purchase a car from a dealer, and (iii) cars with automatic transmission can be more expensive than the ones with manual transmission.



* The figures below plot Selling\_Price by Present\_Price, by grouping based on fuel type, seller\_Type and Transmission. The main outcomes of these figures are (i) Individual sellers with transmission cars sell only cars that operate with regular gas (petrol) and the price of such cars are low, (ii) Only a few individual sellers have cars with automatic transmission, (iii) cars sold by dealers usually have manual transmission and the ones operating on diesel can have higher prices, and (iv) cars with manual transmissions sold by dealers have higher prices if the cars operate on Diesel.



* The figures below show that (i) Selling\_Price is higher for cars both with manual transmission and automatic transmission as the number of owners decreases to zero, (ii) cars with automatic transmission seem to have higher Selling\_Prices, and (iii) individual sellers hold more cars working with diesel.

* The density graph below also supports the findings that Selling\_Price gets lower as the number of owners increases.

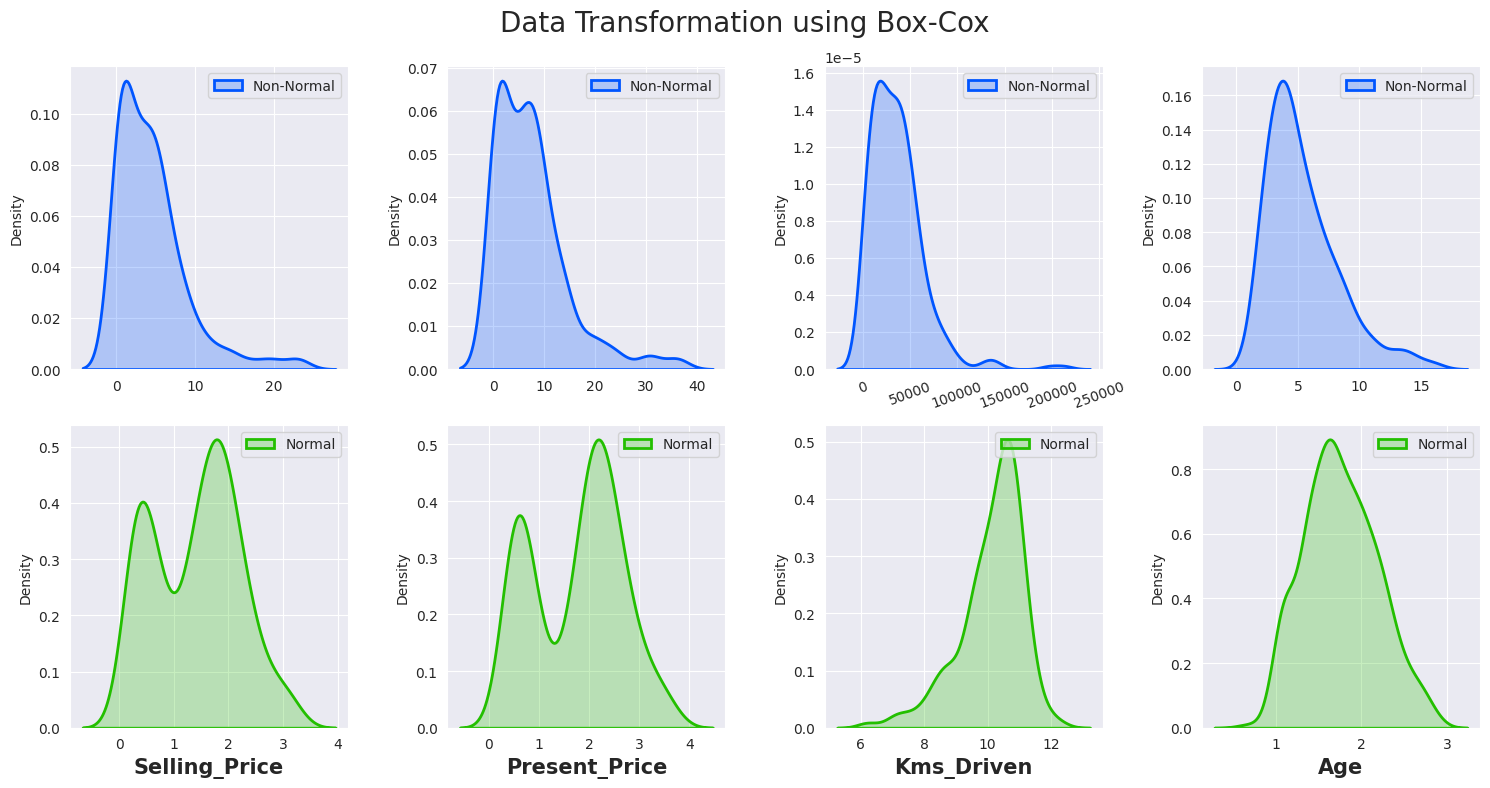


* The correlation map shows that almost all the features are fairly correlated with Selling\_Price except Owner number and Kms\_Driven.



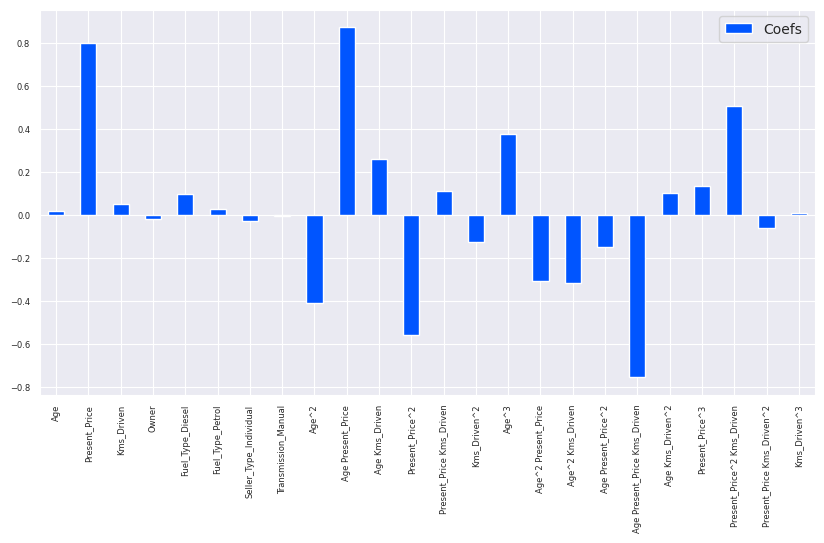
1. LOG TRANSFORMATION – FEATURE ENCODING – MINMAX SCALING

* As shown below, since the numeric features did not exhibit normal distribution required for the performance of linear regression models, Box-Cox transformation was applied to them to create more normally distributed numerical features.
* Moreover, the categorical variables such as Seller\_Type, Fuel\_Type and Transmission were converted to 0s and 1s using get\_dummies Panda function.
* For the scaling of numerical variables, standard scaling was used. However, since polynomial features will be used during regression analysis, standard scaling was performed after the poly features were obtained.

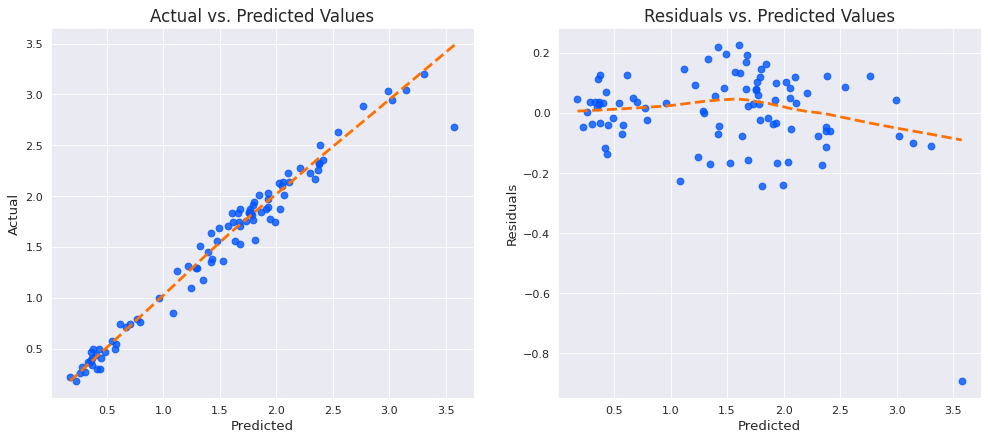


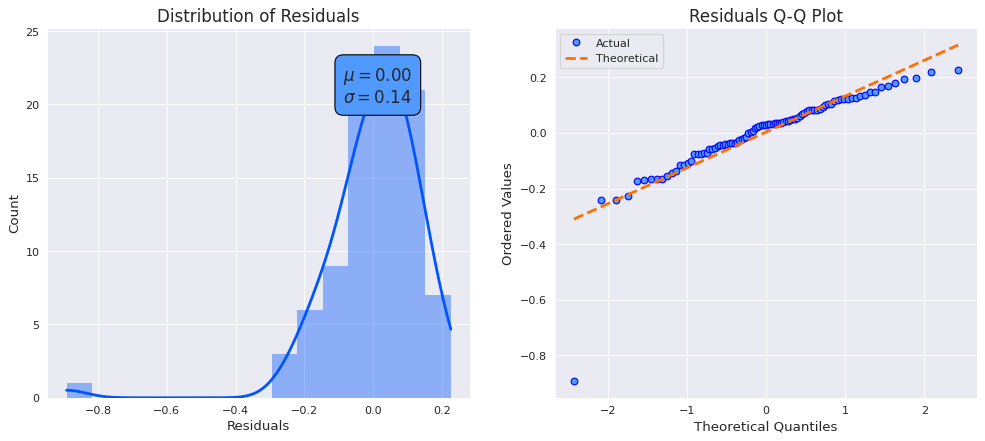
1. **LINEAR REGRESSION MODELS**

* In all regression models, polynomial orders of Age, Present\_Price and Kms\_Driven were considered. After the poly orders were created, numerical values were scaled with standard scaler as mentioned above. Selling\_Price was used as the target variable to predict and a 0.7-0.3 ratio was used to separate the dataset into training and validation datasets.
* At first, Ridge regression (poly order-3) with cross-validation was used to create the Ridge regression model with the optimum hyperparameters. The goal was to see the performance of an model before comparing different regression methods.
* The figure below shows the feature contributions. The coefficients including Age and Present\_Price features have large feature importances as expected since they showed a large correlation with Selling\_Price. However, features including Kms\_Driven have relatively moderate contributions as well while Kms\_Driven itself had a low correlation with Selling\_Price. So, the positive effect of including poly orders is seen herein.

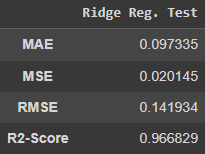


* Actual vs Predicted values seems to be close to each other except for one data point which could not be predicted well. Accordingly, Residuals also seem to be normally distributed if the high residual value due to “not-well-predicted” point is not considered.

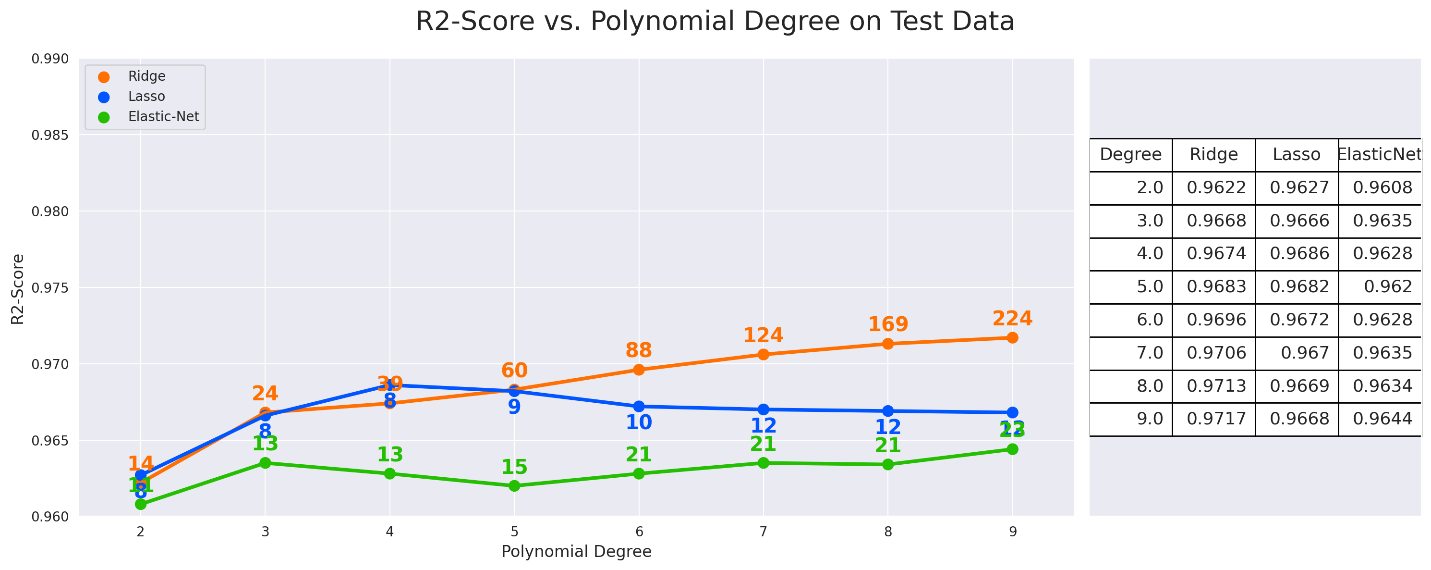




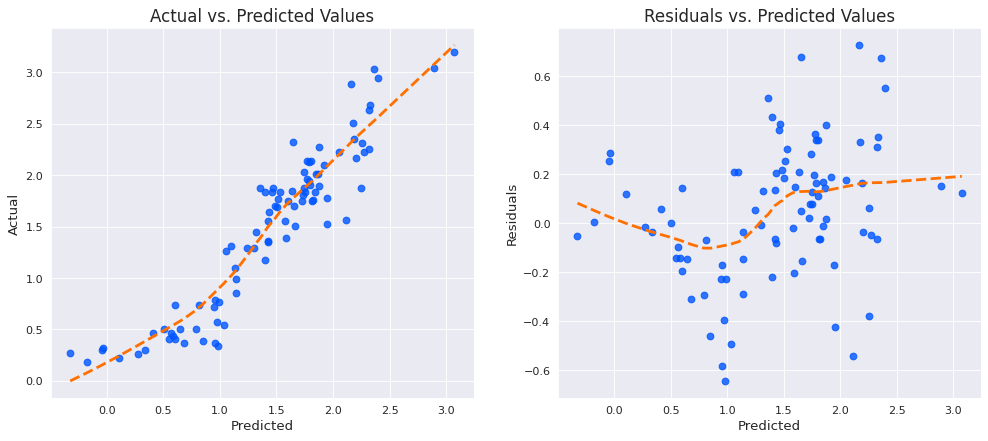
* The below table shows the prediction scores for the Ridge Regression with poly order 3 (with optimized hyper parameters):

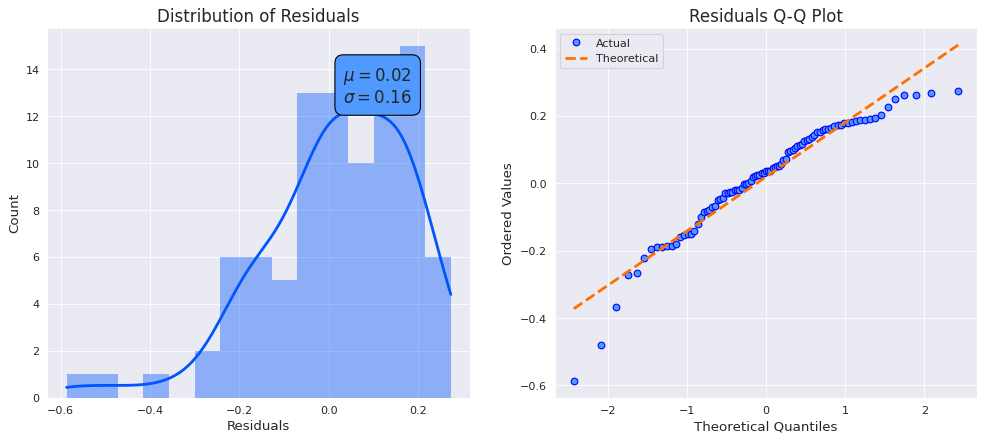


* Now, **Ridge, Lasso, and Elastic-Net regression models and well as XGBoost Regression will be evaluated.**
* An automated loop was created to evaluate the R2 scores of **Ridge, Lasso, and Elastic-Net regression models** (where the hyperparameters were optimized with cross-validation). The figure below shows that R2 score of Ridge regression seems to be improved as poly orders are increased, 4 poly order seems to be optimum for Lasso Regression, and 3 for Elastic-Net if the increase at the 9th poly order is not considered.

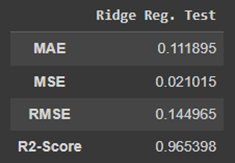


* It seemed a bit odd that R2 scores improved at very high poly orders. Therefore, to investigate this, 4th-order Lasso regression model outcomes are plotted below. As the figures show, the uniformity of the residuals is lost, and the model seems to be performing overfit in order to estimate the Selling\_Price values. It is not shown here for the sake of brevity, but other methods with poly orders of more than 3 also behaved similarly. Therefore, **Ridge Regression with poly order 3** (whose hyperparameters are optimized using cross validation) seems to be the best performing model so far. The outcomes of this model were already presented above.





* In addition, XGBoost with poly order 4 (whose hyper parameters were optimized using cross validation) was also evaluated to see if using higher poly orders with XGBoost can perform better since XGBoost can be more robust to overfitting due to using ensemble approach. However, the results demonstrated below are not better than the Ridge Regression with poly order 3.

****

* **As the conclusion of this section, Ridge Regression with poly order 3 is the best-performing model.**

1. **KEY FINDINGS AND RECOMMENDATIONS**

* Ridge Regression with poly order 3 was the best model to estimate the Selling\_Price. This can be explained:

1. Adding poly features to data which has only 7 features to estimate the target variable increased the performance of the model
2. Since the feature selection plot showed several features that affect the prediction of the target variable. Accordingly, Ridge regression performed better since it works better when the target variable is correlated to several features.
3. Adding more poly orders caused overfitting probably due to creating many features that are correlated with each other.

**Overall,** the prediction with Ridge regression with poly order 3 seems good considering that the dataset was very small. It is expected that the predictions would get better if the data set was bigger including more features.

**The main recommendation for the current analysis** is to include “owner” in the poly orders as well since it was a variable with a low correlation to Selling\_Price. Its higher poly order variables multiplied with other features can create more features that can help estimate Selling\_Price better.