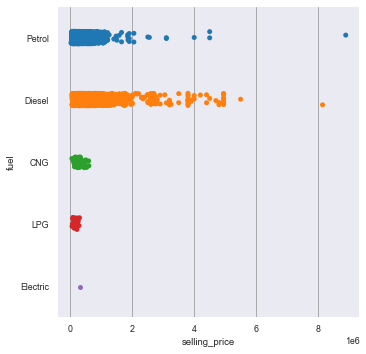
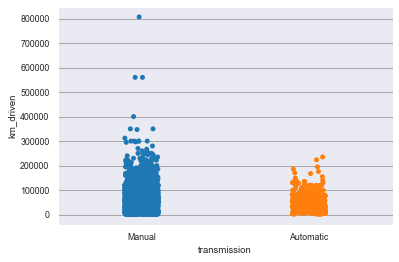
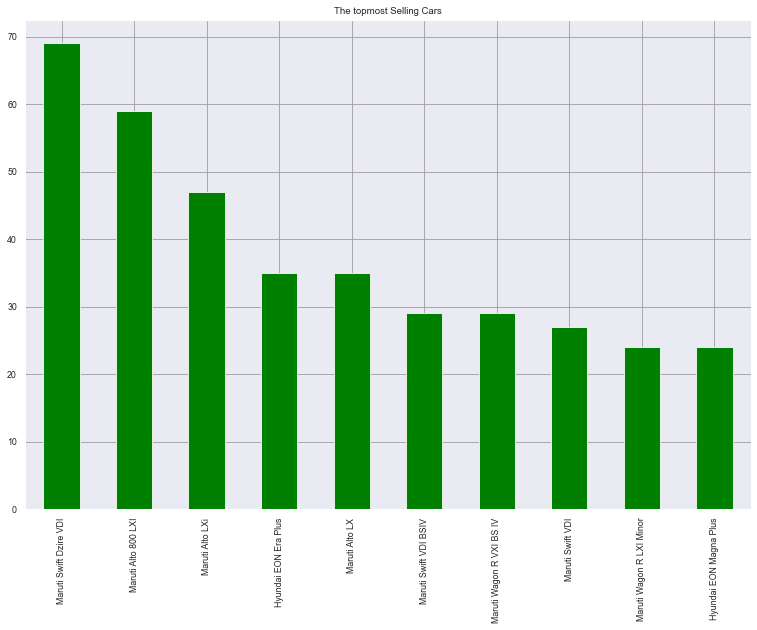
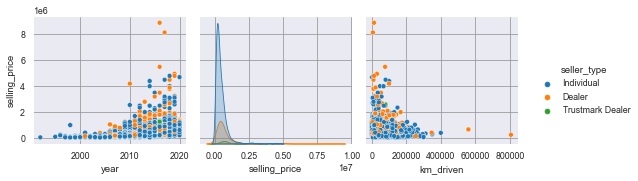
VEHICLE\_SALES EDA INSIGHTS-21BDA41

. 

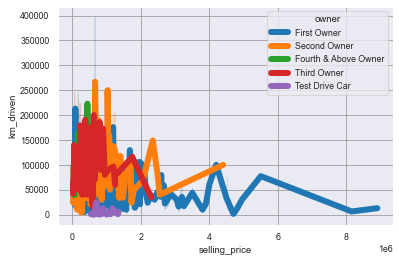
From this catplot it is clear that petrol cars and diesel cars are the cars mostly in sale and diesel cars are the most expensive cars,also electric cars are least in use and sale though it has affordable price.

From this stripplot it is conclusive that manual cars are mostly in use and prefer buying manual cars.Also as automatic cars are new to the market it had no records of a run over 30,000km 

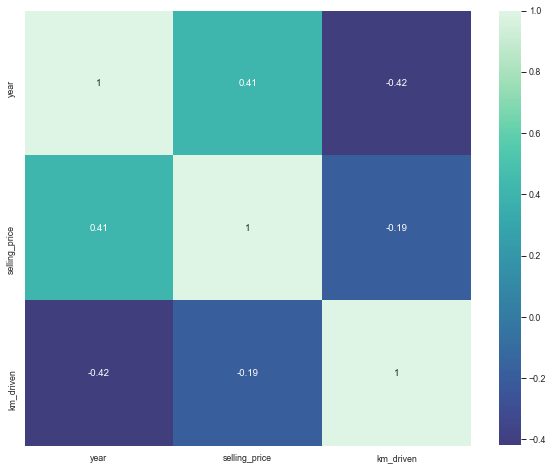
It shows the topmost sold cars in the market, we can see the cars of Maruti tops the list with its cars like Maruti Swift Dzire VDI at the top, followed by Alto models and so on. It shows Maruti is the most preferable car manufacturers in the market.



From this pairplot it is evident that the selling price of cars have increased significantly in recent years and recently a good portion of expensive cars are sold by dealers. Also, the most km\_driven cars have low price and the least km\_driven cars have high price.



from this linegraph we can see that ,selling price is high for first owner and second owner cars and km\_driven is high for third owner and fourth& above owner cars that being a reason for its low price compared to that of firs or second owner.



From the heatmap ,we can understand the correlation between each variables ,we can see year and km\_driven has -.42 correlation which means as year decreases the value of Km\_driven increases there exist a negative correlation, year and selling price have a positive correlation of .41.

Assumptions of linear regression

1.Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2.Independence:The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3.Homoscedasticity:The residuals have constant variance at every level of x.

4.Normality:The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

How can we evaluate a Regression model? Define each metric and its interpretation.

The main metrics in regression model are:

R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

Mean Square Error(MSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

iii. Can R squared be negative?

Yes, R-square is defined as the proportion of variance explained by the fit ,if the fit is actually worse than just fitting a horizontal line then R-square is negative.

iv. What is dummy variable trap?

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables.

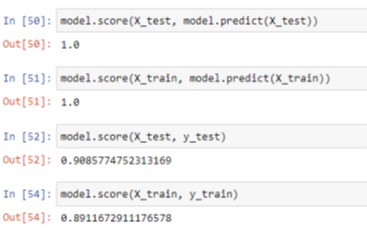
v. Is One Hot Encoding different from Dummy Variables?

The dummy (binary) variable just takes the value 0 or 1 to indicate the exclusion or inclusion of a category. In one-hot encoding, “Red” color is encoded as [1 0 0] vector of size 3. “Green” color is encoded as [0 1 0] vector of size 3.

vi. How is polynomial regression different from linear regression?

Polynomial Regression is a form of linear regression known as special case of multiple linear regression which estimates the relationship as nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance.

vii. Interpret the screenshot below from the notebook we discussed in class today:



We are using the regression model to predict, we use training and testing data to predict, the total score is 1,in this model we can see that the dependency of the variable is highly explained by the independent variable, the R square values are good indicating it is a good regression model and the prediction will be good.