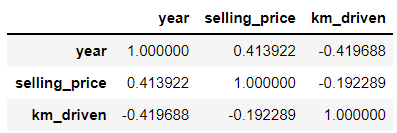
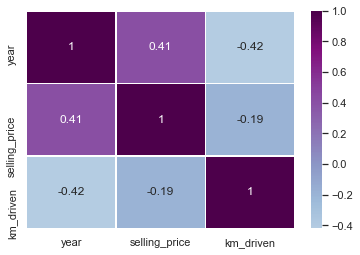
**DEON S**

**21BDA57**

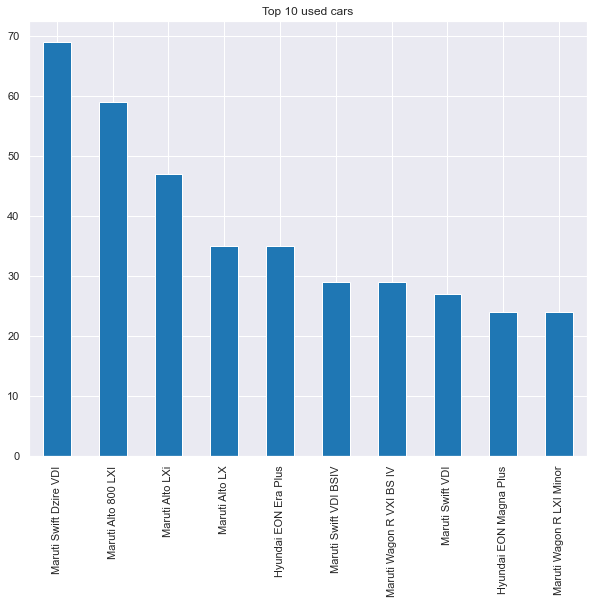
1. In a word file,
2. **1. Document 5-6 key insights from EDA and support each point with a visualization.**
3. Correlation between the numerical variables is calculated first.

****

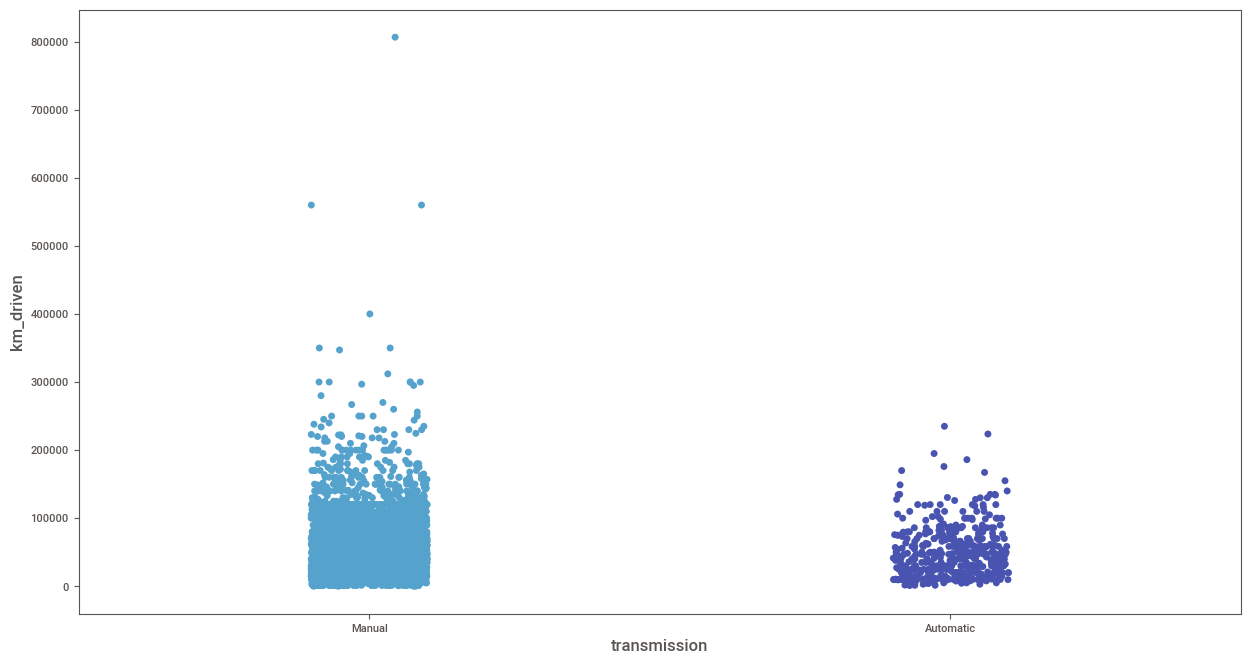
**Heatmap of correlation.**

****

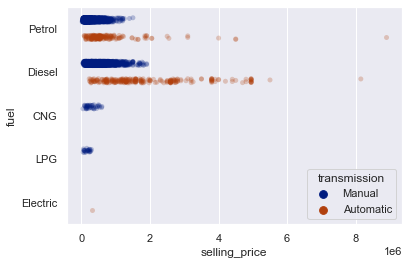
Top 10 used cars are found out from the data.



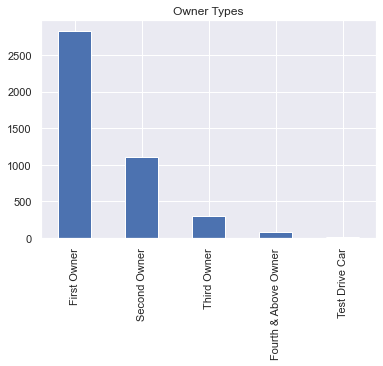
Stripplot between transmission and km\_driven.



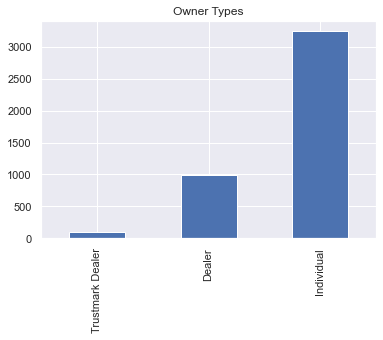
Stripplot of selling\_price vs various fuel types.



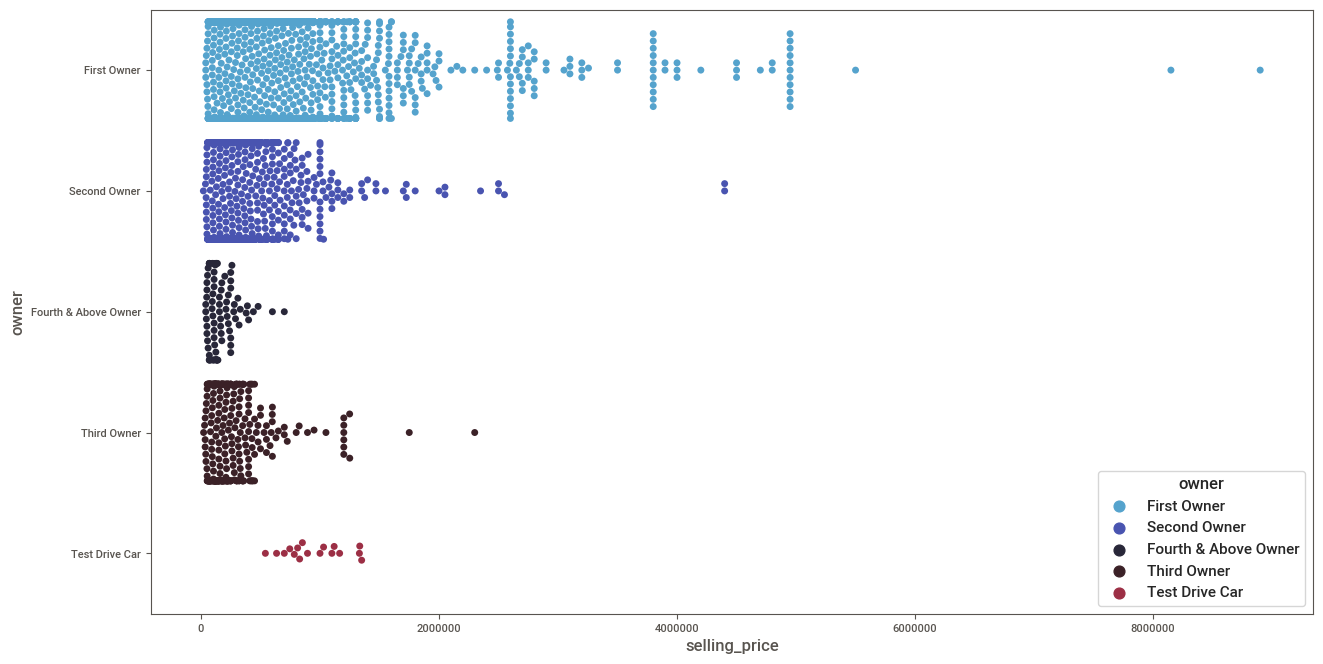
Bar chart showing the owner types. First owner type is the most.



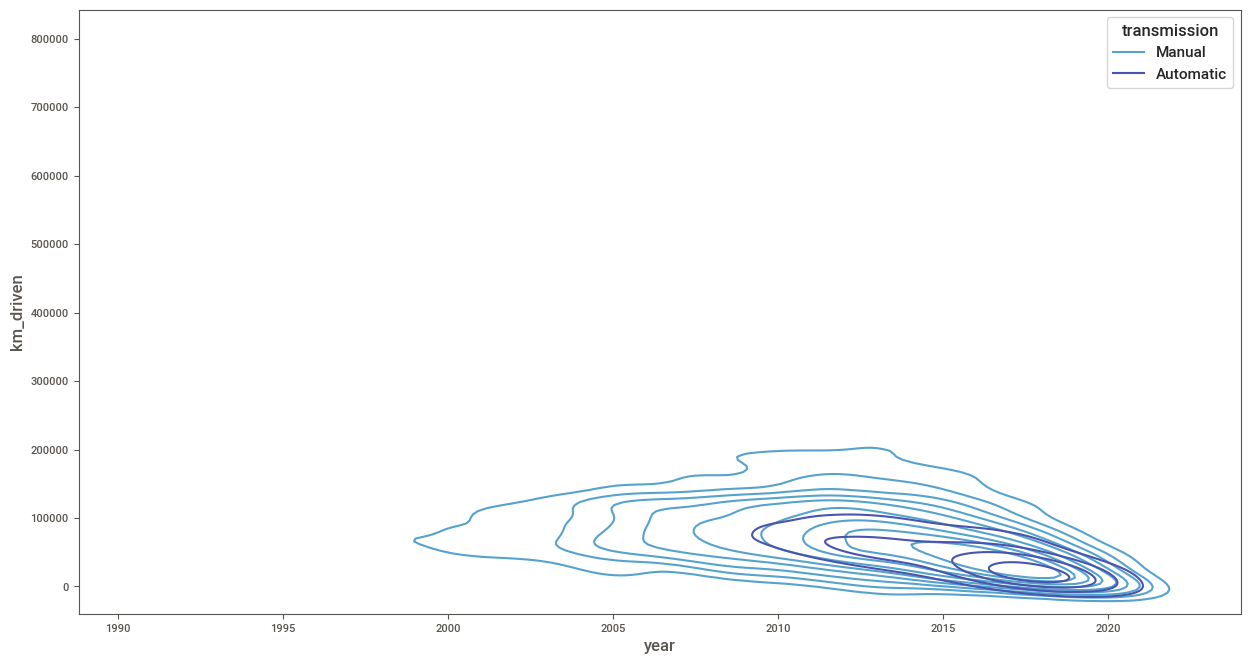
Bar chart of various seller\_types.



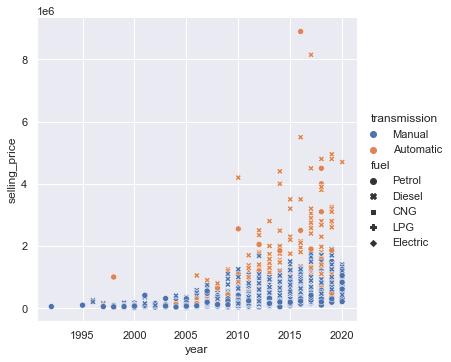
Swarmplot of selling\_price vs owner type.



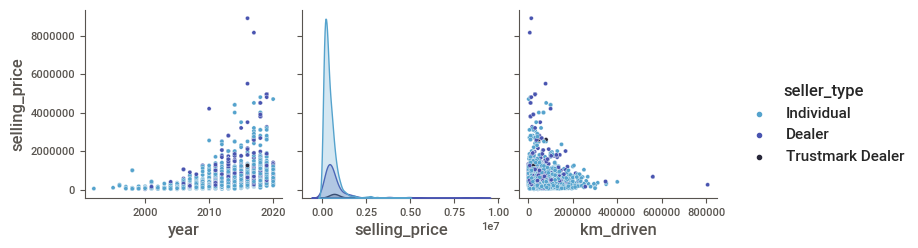
Kdeplot of year vs km\_driven.



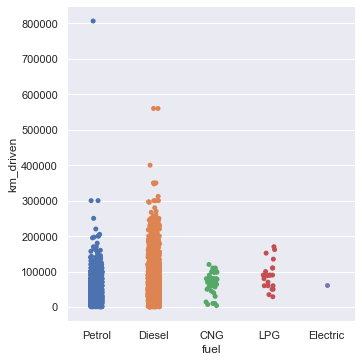
Relplot of years vs selling\_price. Selling price has increased over the years.



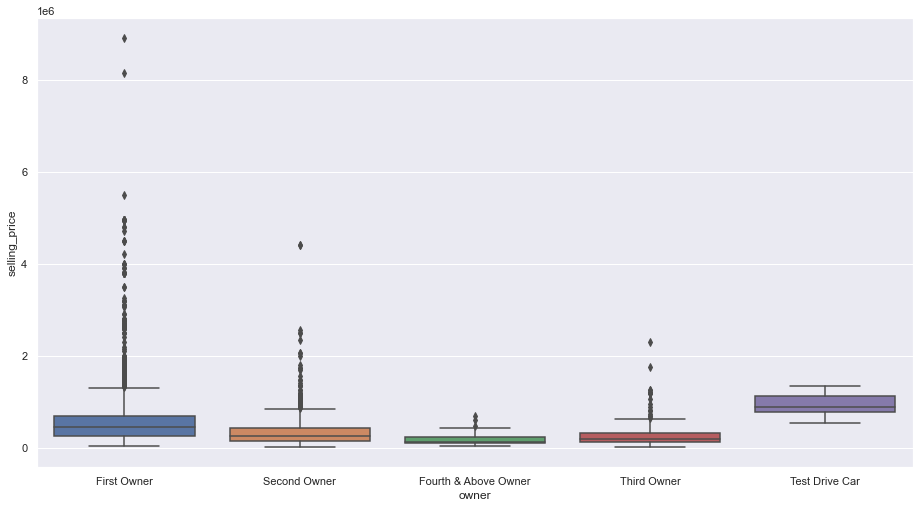
Pairplot of selling\_price vs year, selling\_price, km\_driven. People mostly prefer cars which were less driven.



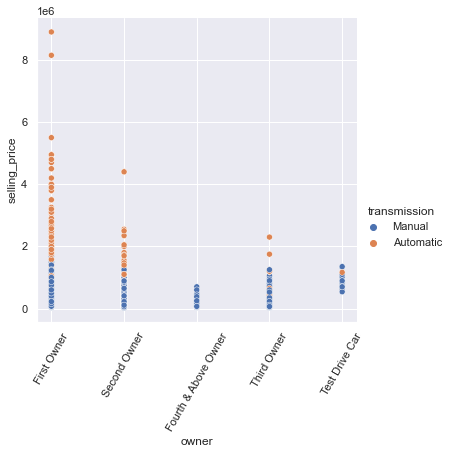
Catplot of fuel types vs km\_driven.



Boxplot of owner types and selling\_price.



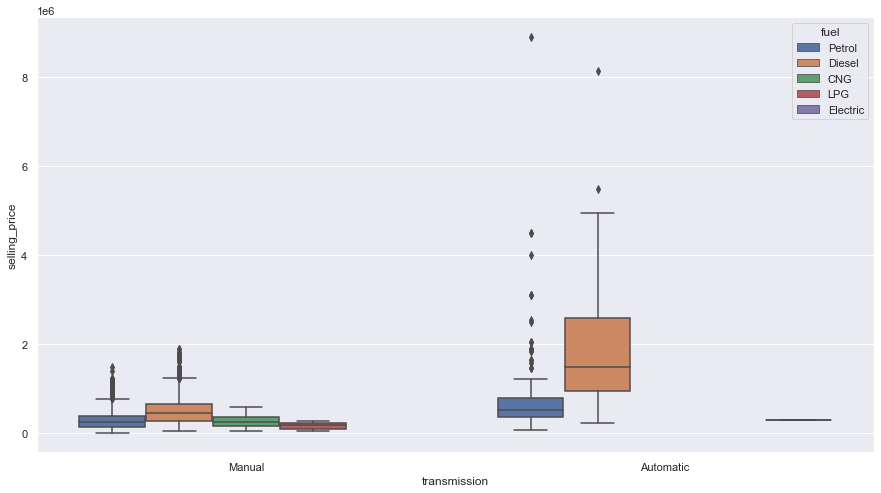
relplot of owner types and selling\_price.



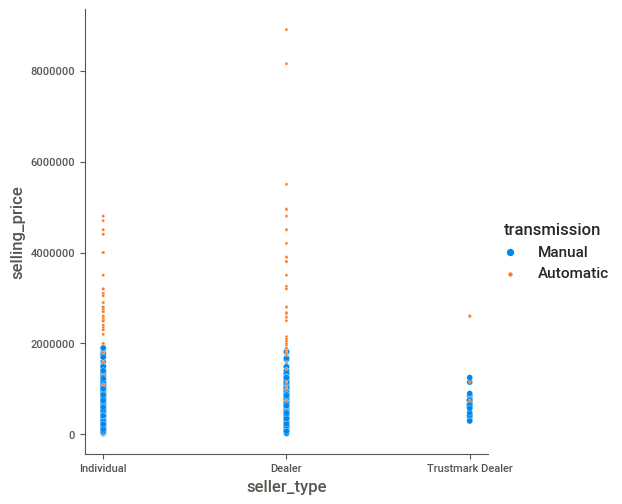
Boxplot of owner types and selling\_price and categorising them into manual and automatic types.

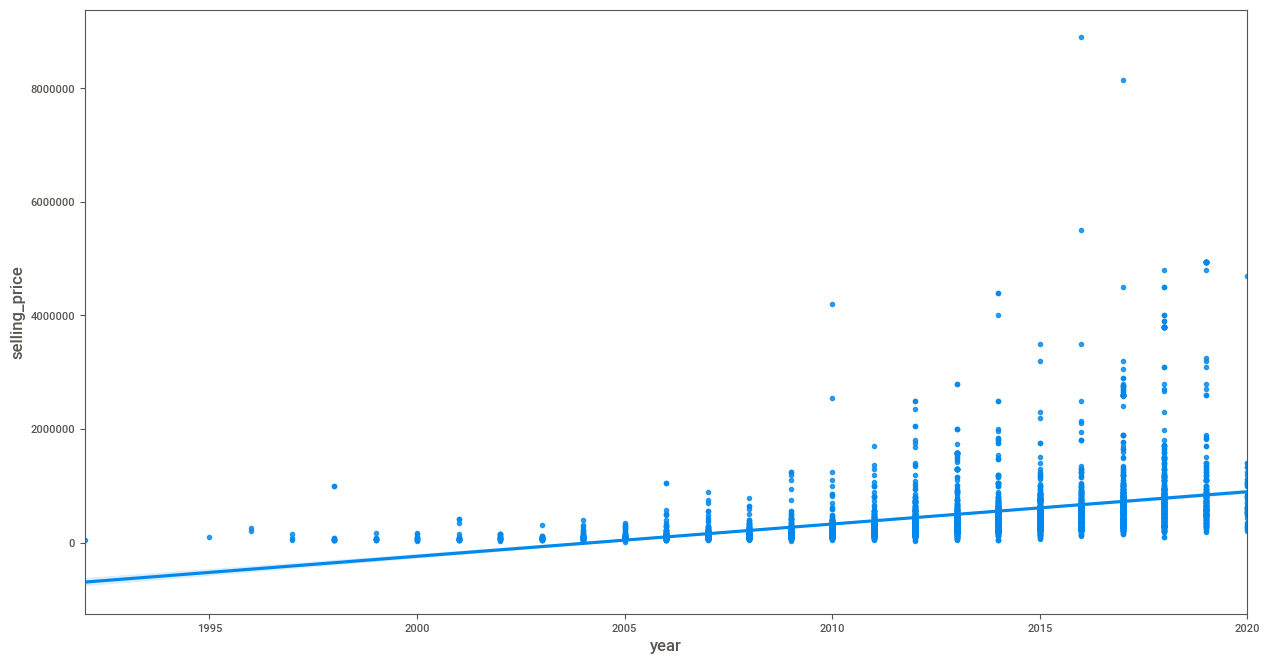
****

boxplot of transmission and selling\_price with hue as fuel types.

****

relplot of seller\_types and selling\_price.

****

****

Regression graph on year vs the selling price. Clearly a positive regression can be observed.

**2. Answer the following questions:**

**i. What are the assumptions of linear regression?**

[**Linear regression**](https://www.statology.org/introduction-to-simple-linear-regression/) is a useful statistical method we can use to understand the relationship between two variables, x and y. However, before we conduct linear regression, we must first make sure that four assumptions are met:

**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.

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

**Mean Squared Error (MSE)**

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

**R-squared or Coefficient of Determination**

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

**Root Mean Squared Error (RMSE)**

This is the square root of the average of the squared difference of the predicted and actual value.

R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure).

Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line.

**iii. Can R squared be negative?**

The ratio of the regression error against the total error tells you how much of the total error remains in your regression model.  Subtracting that ratio from 1.0 gives how much error you removed using the regression analysis. This is called R2

It is possible to get a negative R-square for equations that do not contain a constant term. Because 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. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap.

**v. Is One Hot Encoding different from Dummy Variables?**

There are two different ways to encoding categorical variables. Say, one categorical variable has n values. One-hot encoding converts it into n variables, while dummy encoding converts it into n-1 variables. If we have k categorical variables, each of which has n values. One hot encoding ends up with kn variables, while dummy encoding ends up with kn-k variables.

**vi. How is polynomial regression different from linear regression?**

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

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



Our model fits perfectly because the R2 value is 0.89 ie. 89%

**viii. Bonus: We saw *Sweetviz* as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.**

1. dtale

2. pandas profiling

3. autoviz