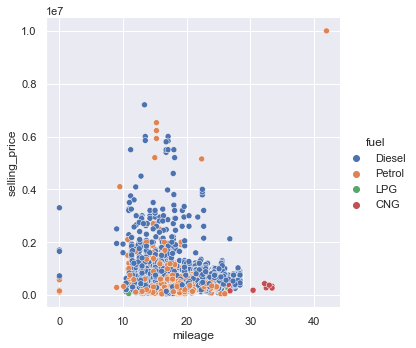
MACHINE LEARNING LAB-02 ASSIGNMENT

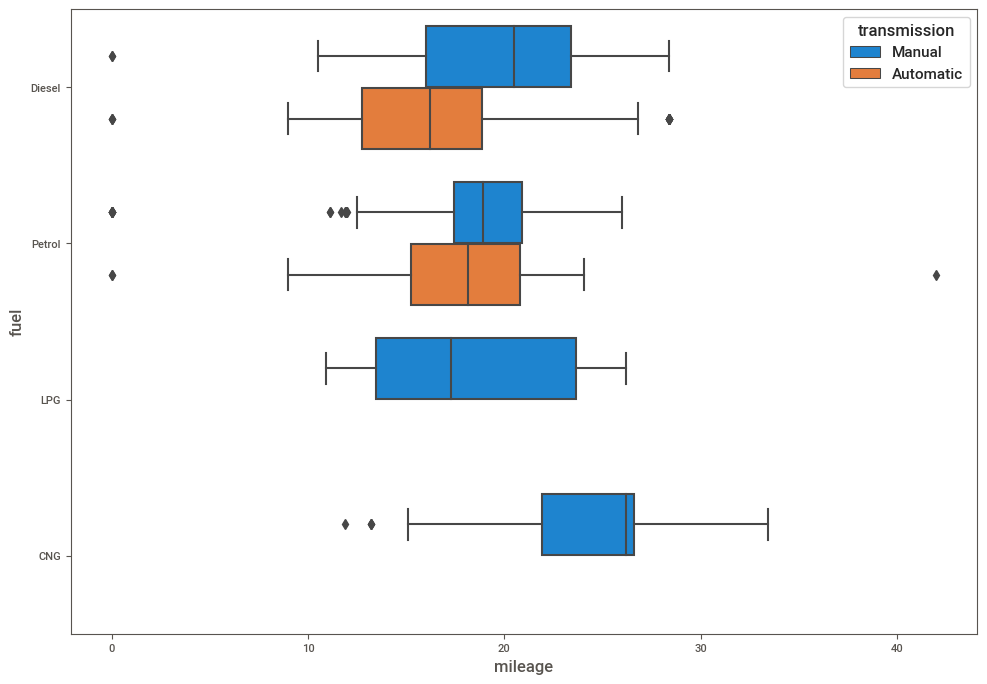
**NAME : ARYA CHANDRAN**

**ROLL NO : 21BDA34**

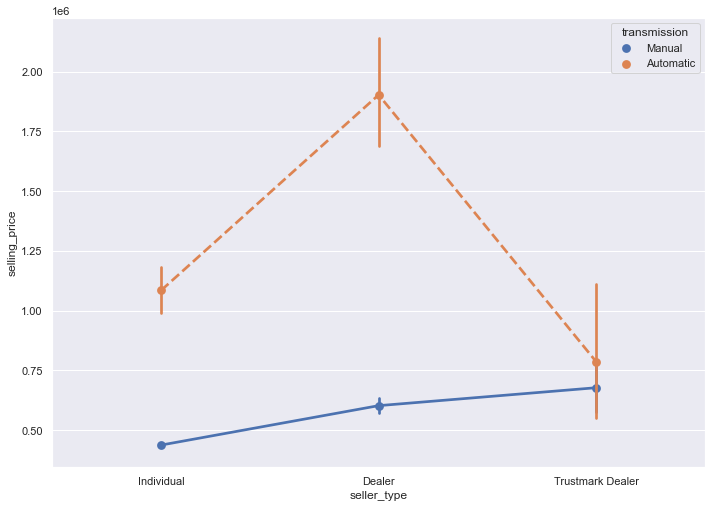
1. **EXPLORATORY DATA ANALYSIS**

****

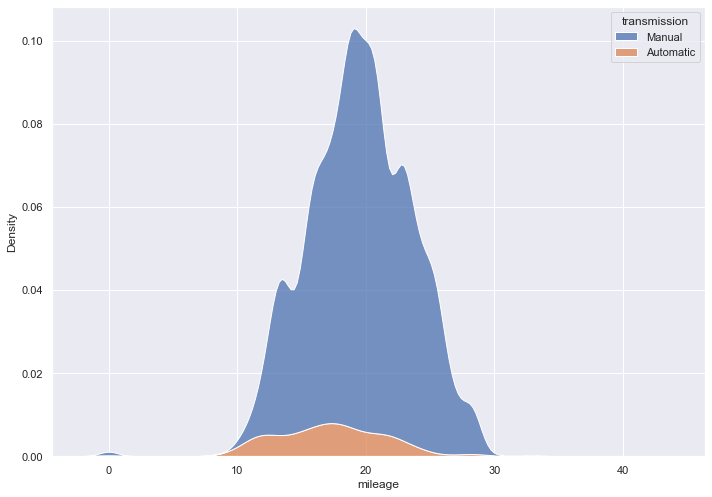
* Majority of the Petrol and Diesel cars have mileage between 10 and 30 kmpl.
* CNG vehicles have low selling price and mileage greater than 25.
* LPG vehicles are less in number and have low mileage and low selling price.

****

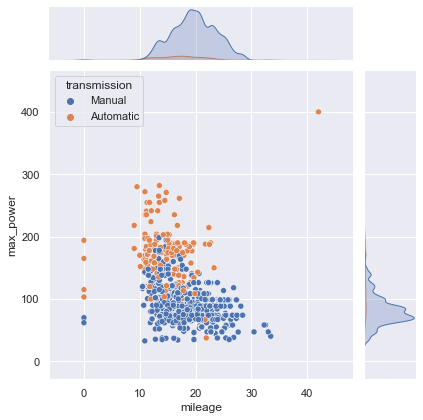
* LPG and CNG cars have no automatic transmission cars.
* CNG cars have high mileage compared to other fuel types.
* For Diesel, manual transmission cars have higher mileage compared to automatic. Whereas in Petrol, the median value of manual is close to that of automatic.

****

* The total selling price of Trustmark dealers are the highest for manual cars and the lowest for automatic cars.
* The selling price of automatic cars are highest for dealers.
* Indivitual seller type has a very low total selling price for manual cars but has fairly high selling price for automatic cars.
* Overall data shows that automatic cars has high selling prices compared to manual.

****

* Majority of the cars have mileage between 10 and 30 kmpl.
* Most of the manual cars have range close to 20 kmpl.
* Mileage of automatic cars are almost evenly spread in the range 10 to 25kmpl.

****

* Automatic cars have higher power compared to manual cars.

**2.**

1. What are the assumptions of Linear Regression?

There are four assumptions associated with a linear regression model:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

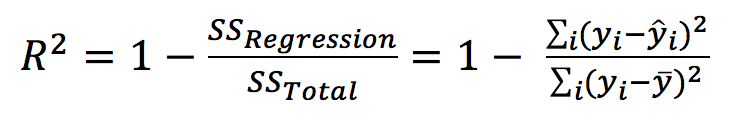
Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

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

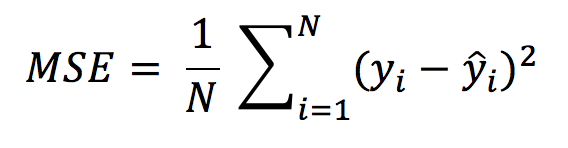
R Square/Adjusted R Square

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.



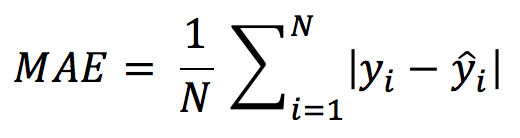
Mean Square Error(MSE)/Root Mean Square Error(RMSE)

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.



1. Can R squared be negative?

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.

1. 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.

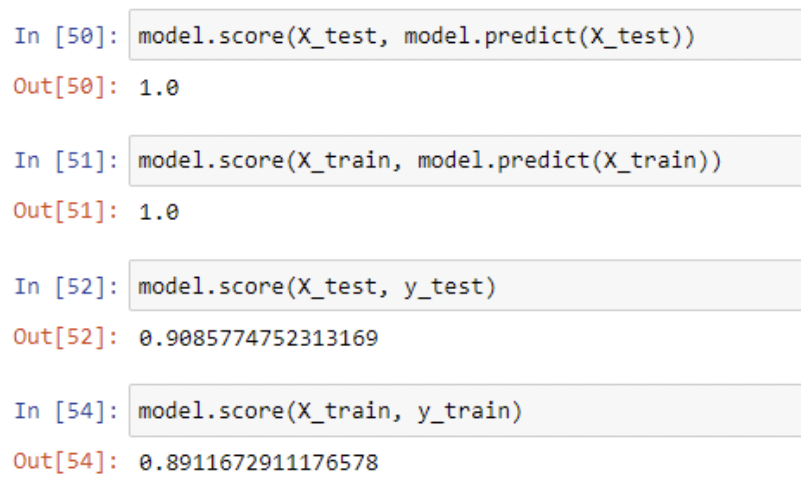
1. Is One Hot Encoding different from Dummy Variables?

A 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

1. 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.

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



Every estimator or model in Scikit-learn has a score method after being trained on the data, usually X\_train, y\_train.

When you call score on classifiers like LogisticRegression, RandomForestClassifier, etc. the method computes the accuracy score of the input and its target value.

1. 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.

Other options for Sweetviz are:

1. dtale

2. pandas profiling

3. sweetviz

4. autoviz

DataPrep provides much more functionality than simple EDA. It can help you ingest more data sources and can help you get through large data sets faster.

In addition, the clean API in DataPrep can help you clean your data set without many hurdles.