**Used Car Price Prediction**

Data Science Project

------------------------------------------------------------------------------------------------------------

1. **Introduction**

The used automobile market is a growing business with a market value that has nearly doubled itself in previous years. The rise of online websites and other tools like it have made it easier for both buyers and sellers to get a better understanding of the factors that determine the market value of a used car. Based on a set of factors, Machine Learning algorithms may be used to forecast the price of any automobile.

The main purpose of this study is to compare the accuracy of various predictions for estimating a used car’s retail price in a manner to maximise the accuracy of prediction. In this research we have used models like Linear Regression, Decision Trees and Random Forest Regression and compared their accuracies.

1. **Review of Research Papers**

**FAIR PRICE PREDICTION SYSTEM FOR USED CARS IN SRI LANKA**

* This analysis was done on a used car dataset of used cars in Sri Lanka.
* 4 algorithms were compared here which were Multiple Linear Regression, Decision Tree, K Nearest Neighbours and Random Forest Regression, Gradient Boosting Algorithm
* All the algorithms except the K Nearest Neighbours were able to perform with an accuracy greater than 0.8. KNN showed an accuracy of 0.35.
* KNN gave an exceptionally high Mean Absolute Error (MAE) for which it was rejected. Random Forest Regression showed the least Mean Absolute Error (MAE).
* Followed are the obtained results

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Multiple Linear Regression | 0.84 |
| Random Forest Regression | 0.87 |
| K-Nearest Neighbours | 0.35 |
| Decision Tree | 0.81 |
| Gradient Boosting Algorithm | 0.88 |

**COMPARATIVE ANALYSIS OF MULTIPLE LINEAR REGRESSION AND RANDOM FOREST**

* The dataset used here has 19 columns/attributes.
* The number of samples and explanatory variables have been varied over 9 times thus covering all possible probabilities and NMSE (Normalized Mean Squared Error) for Linear Regression and Random Forest have been compared each time.
* Linear Regression proved to be a strong model 2/9 times especially when no. of samples were less but with a very small margin.
* Whereas Random Forest proved to be a stronger model 7/9 times and most of the times with big margins.
* Thus, when the no. of samples are less, both the models are comparable but as the no. of samples goes on increasing, random forest algorithm starts to show it’s clear advantage.

**HOW MUCH IS MY CAR WORTH? A METHODOLOGY FOR PREDICTING USED CARS PRICES USING RANDOM FOREST**

* The Author of this paper is Priya Arora and was published in the year 2018.
* The database used here is 'Used Car Database' from Kaggle.
* The dataset contains the prices and attributes of over 370,000 used cars and contains 20 unique attributes.
* In this paper, two models namely Random Forest and Linear Regression were used to predict the prices of used cars.
* It was found that random forest regression performed much better in comparison to linear Regression.The accuracy of the regression was less than 75%.
* A Random Forest with 500 Decision Trees were created to train the data.From experimental results, the training accuracy was found out to be 95.82%, and the testing accuracy was 83.63%.  
    
    
  **VEHICLE RESALE PRICE PREDICTION USING MACHINE LEARNING**
* This paper was published by B.Lavanya in the year 2021.
* In this paper, three distinctive algorithms have been implemented to figure the cost of pre-owned vehicles in Mauritius.
* These three algorithms are Linear Regression , Random Forest Regression and Decision Tree Regression and their accuracy was calculated by R-squared , Mean Absolute Error(MAE) and Root Mean Squared Error (RMSE).
* For Linear Regression we got an R-squared value of 76.65% , RMSE of 6907.93 and MAE of 3412.80.
* R-squared , RMSE and MAE for Decision tree Regression are 66.93% , 8221.38 and 4398.59 respectively And for Random forest are 86.14% , 5321.68 and 1725.18 respectively.
* Out of the three algorithms the best performance was observed by Random Forest Regression.

**USED CAR PRICE PREDICTION USING K-NEAREST NEIGHBOUR BASED MODEL:**

* This paper was published in the International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE) in September 2020.
* In this paper, a model is proposed to estimate the cost of used cars using the K-nearest neighbour algorithm which is simple and suitable for small datasets.
* The dataset used in this paper is the Used Cars dataset from Kaggle . The dataset contains 14 attributes which includes an unnamed serial number, Name, Location, Mileage, Fuel\_type, Engine, Transmission, Kilometers\_Driven, Power, New\_Price, Year, Seats, Owner\_type and last one is Price which is the target variable.
* Three steps have been performed in the data Preprocessing stage that are removal of non-numerical part from numerical features, converting categorical values into numerical, Separating the target variable
* For estimating the accuracy of the model different values of K from 2 to 10 are used to find the best performing value of K.
* The model is trained with 3 different ratios of train:test data for K values from 2 to 10.
* Without Cross-validation the accuracy of the model is 85%, Root-mean Squared error rate(RMSE) of 4.01 and mean absolute error(MAE) of 2.01 with K-value of 4.
* Using Cross-Validation with 5 folds (k=5) the accuracy is 81.1% and that with 10 folds is 82% with RMSE rate of 4.73 , MAE rate of 2.13 and K-Value of 4

**PREDICTING USED CAR PRICES:**

* The dataset used in this paper is 1.2 million Used car listings dataset from Kaggle.
* Performance of 7 algorithms along with their training time is compared. The algorithms used for comparison are Linear regression, Gradient boost, Random forest, Light GBM, XGBoost, KMeans+ Linear regression, Deep Neural network.
* Out of all the algorithms Random Forest model and K-Means clustering with linear regression yielded the best results but their training time was comparatively higher as compared to conventional linear regression model whose result was satisfactory.
* In the data pre-processing stage the dataset was pruned to three standard deviations around the mean in order to remove outliers and the categorical data was converted to numerical data using one-hot encoding.
* Out of 1.2 million examples only 500 thousand examples were used for reducing the training time.
* 90% of the data was used for training and 10% was used for testing
* Following are the obtained results:

|  |  |  |
| --- | --- | --- |
| **Learning Algorithm** | **Accuracy** | **Training time(in mins)** |
| Linear regression | 0.87 | 15 |
| Gradient Boost | 0.64 | 130 |
| Random Forest | 0.88 | 75 |
| Light GBM | 0.81 | 104 |
| XGBoost | 0.78 | 180 |
| K-mean+LinReg | 0.88 | 70 |
| Deep Neural Network | 0.85 | 10 hrs |

**OLD CAR PRICE PREDICTION WITH MACHINE LEARNING**

* This paper is published by the students of Parul Institute of Technology in 2021.
* A dataset containing 92386 records is used to train the model. Attributes in the dataset are kilometers traveled, year of registration, fuel type, car model, fiscal power, car brand and gear type
* Algorithms implemented are: K Nearest Neighbors (KNN) Regressor, Random Forest Regressor, Linear Regression, XGBoost Regressor and Decision Tree Regressor.
* Comparison of different models

|  |  |  |  |
| --- | --- | --- | --- |
| SN. | Model Name | Root Mean Squared Error (RMSE) | Test Accuracy |
| 1 | Random Forest Regressor | 3702.34 | 93.11% |
| 2 | KNN\_Regressor | 7771.91 | 69.66% |
| 3 | Decision Tree Regressor | 5590.43 | 84.30% |
| 4 | XGBoost Regressor | 3980.77 | 92.04% |
| 5 | Linear Regression | 6846.23 | 76.46% |

* Out of all Random Forest has lowest RMSE, and performed well with highest R Squared value: 0.93

**CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES**

* The paper was published by the students of International Burch University, Sarajevo, Bosnia and Herzegovina.
* Data is collected from a local web portal for selling and buying cars autopijaca.ba
* Attributes are brand, model, car condition, fuel, year of manufacturing, power in kilowatts, transmission type, millage, color, city, state, number of doors, four wheel drive (yes/no), damaged (yes/no), navigation (yes/no), leather seats (yes/no), alarm (yes/no), aluminum rims (yes/no), digital air condition (yes/no), parking sensors (yes/no), xenon lights (yes/no), remote unlock (yes/no), electric rear mirrors (yes/no), seat heat (yes/no), panorama roof (yes/no), cruise control (yes/no), abs (yes/no), esp (yes/no), asr (yes/no) and price expressed in BAM (Bosnian Mark).
* The following features were used to build model: brand, model, car condition, fuel, age, kilowatts, transmission, miles, color, doors, drive, leather seats, navigation, alarm, aluminum rims, digital AC, manual AC, parking sensors, xenon, remote unlock, seat heat, panorama roof, cruise control, abs, asr, esp and price.
* Applying a single machine algorithm on the data set accuracy was less than 50%. Therefore, the ensemble of SVM and ANN machine learning algorithms has been proposed and this combination of ML methods gains accuracy of 92.38%.

1. **Experimental Setup**

**DATASET**

**Name** : usedCarPrices

**Source**: Kaggle

**Shape** : (6019,12)

**Target variable:** Price

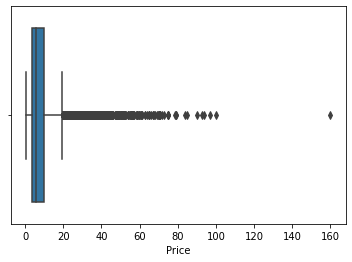
**Description of Attributes:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Data Type** | **Type of attribute** | **Description** |
| Name | object | Discrete | The brand and model of the car. |
| Location | object | Discrete | The location in which the car is being sold or is available for purchase. |
| Year | int64 | Discrete | The year or edition of the model. |
| Kilometers\_Driven | int64 | Continuous | The total kilometres driven in the car by the previous owner(s) in KM. |
| Fuel\_Type | object | Discrete | The type of fuel used by the car. |
| Transmission | object | Discrete | The type of transmission used by the car. |
| Owner\_Type | object | Discrete | Whether the ownership is Firsthand, Second hand or other. |
| Mileage | object | Continuous | The standard mileage offered by the car company in kmpl or km/kg |
| Engine | object | Continuous | The displacement volume of the engine in cc. |
| Power | object | Continuous | The maximum power of the engine in bhp. |
| Seats | float64 | Discrete | The number of seats in the car. |
| Price | float64 | Continuous | The price of the used car in INR Lakhs. |

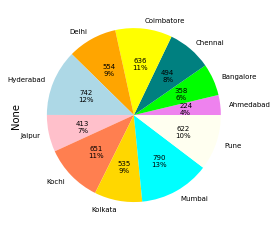
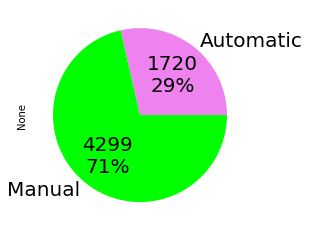
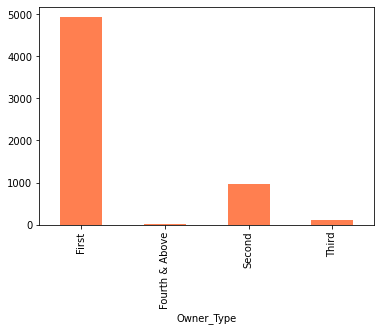
**Null values:**

|  |  |
| --- | --- |
| **Attribute** | **Total NULL values** |
| Name | 0 |
| Location | 0 |
| Year | 0 |
| Kilometers\_Driven | 0 |
| Fuel\_Type | 0 |
| Transmission | 0 |
| Owner\_Type | 0 |
| Mileage | 0 |
| Engine | 36 |
| Power | 36 |
| Seats | 42 |
| Price | 0 |

1. **Data Visualization**

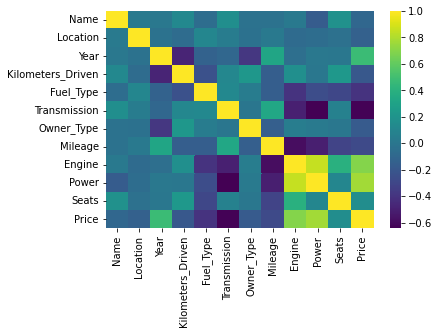
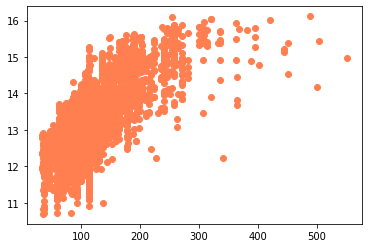
** **

We can see some outliers in Kilometers\_Driven and Price

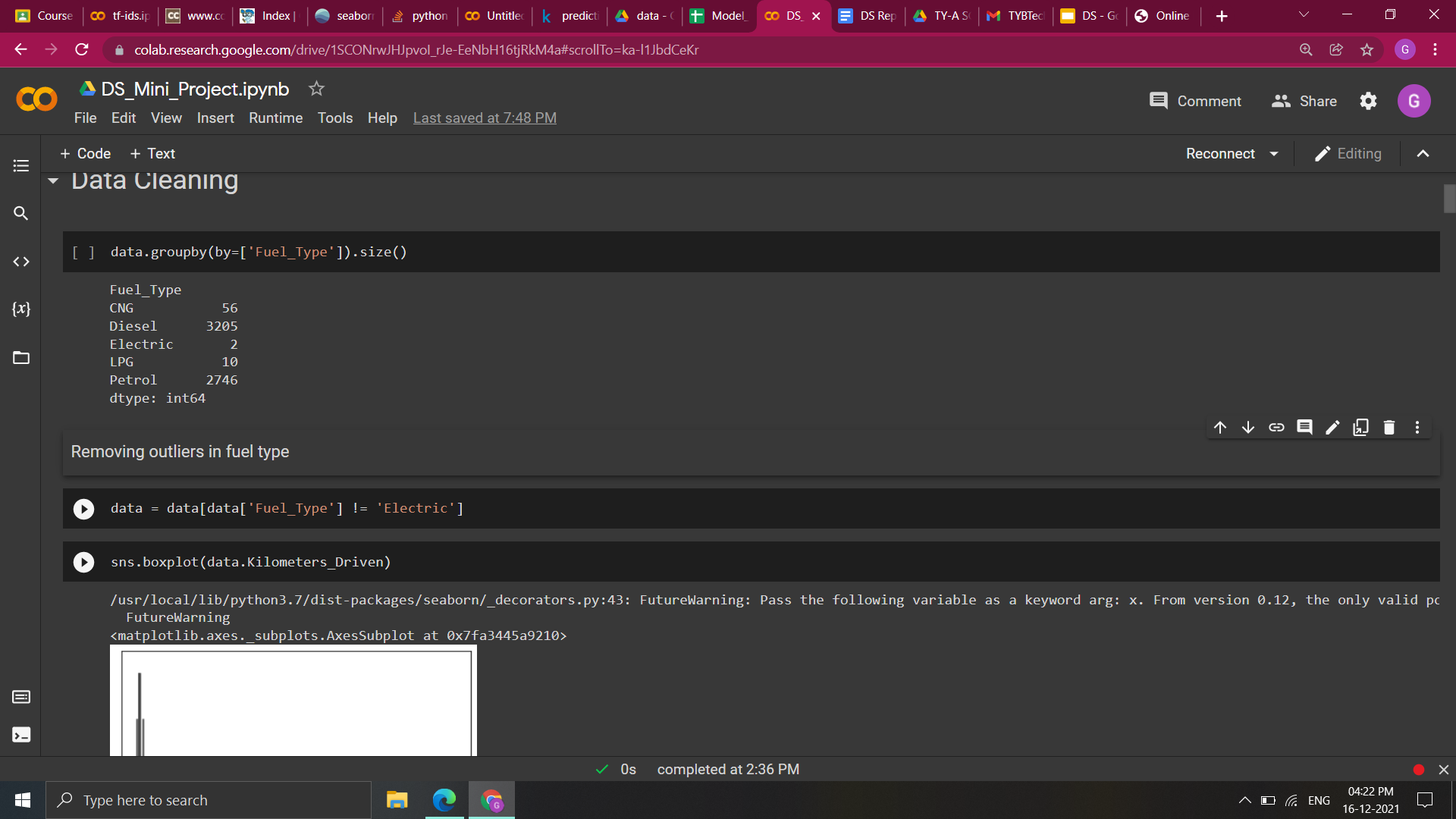
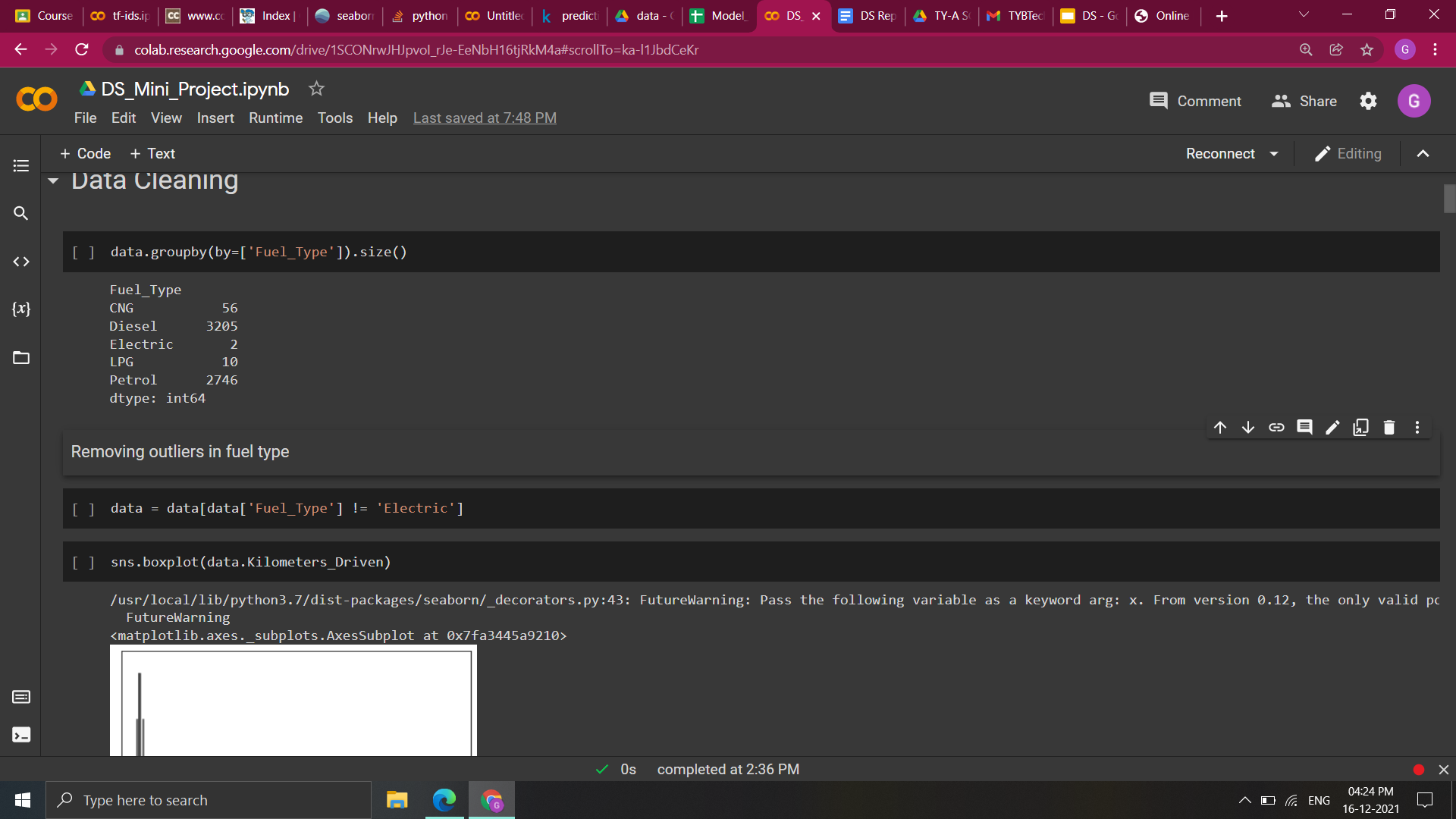
* 71% of cars have ‘Manual’ Transmission
* 13% of cars listed are from Mumbai
* ‘First’ tops the ‘Owner\_Type’ category

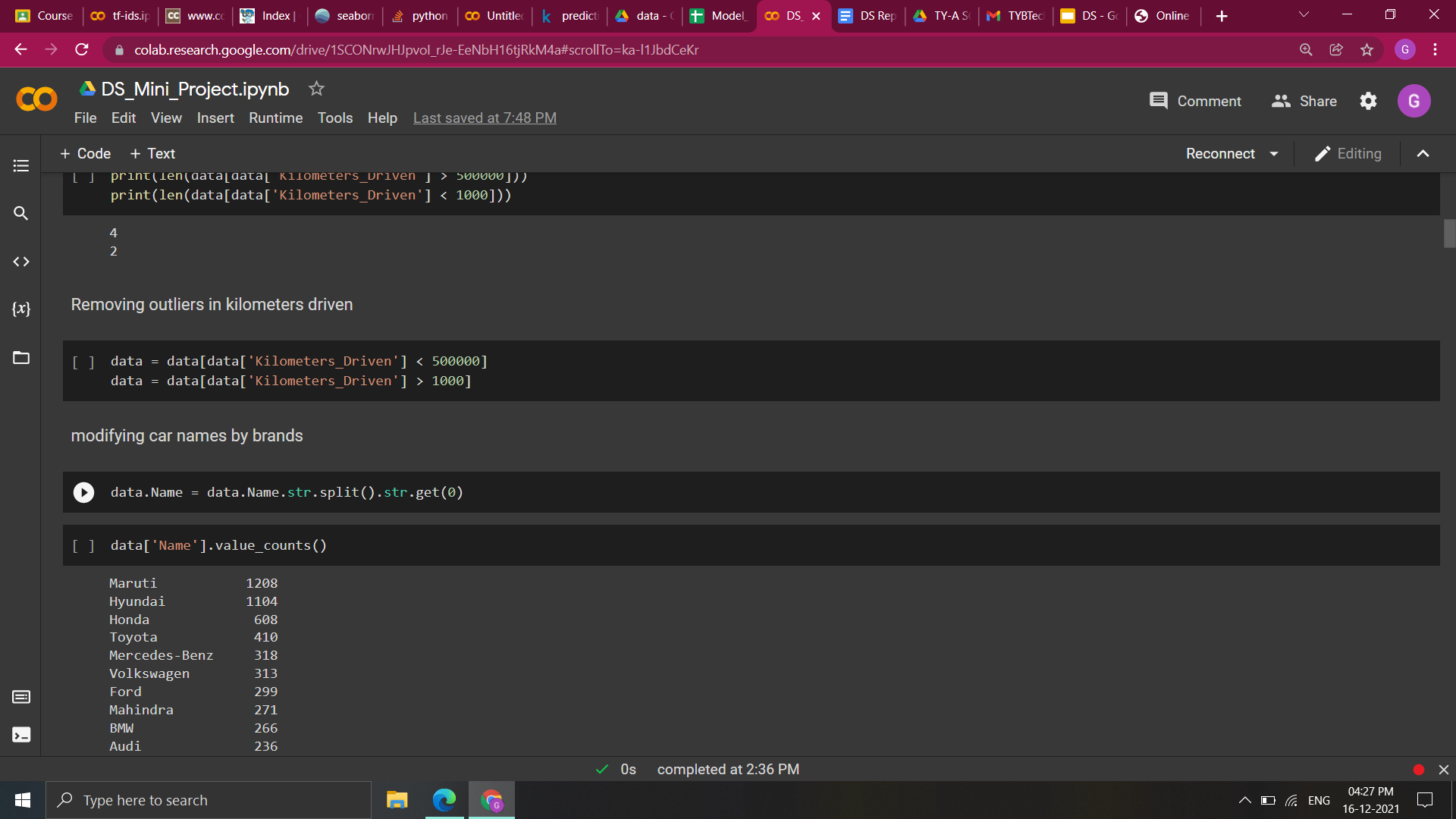
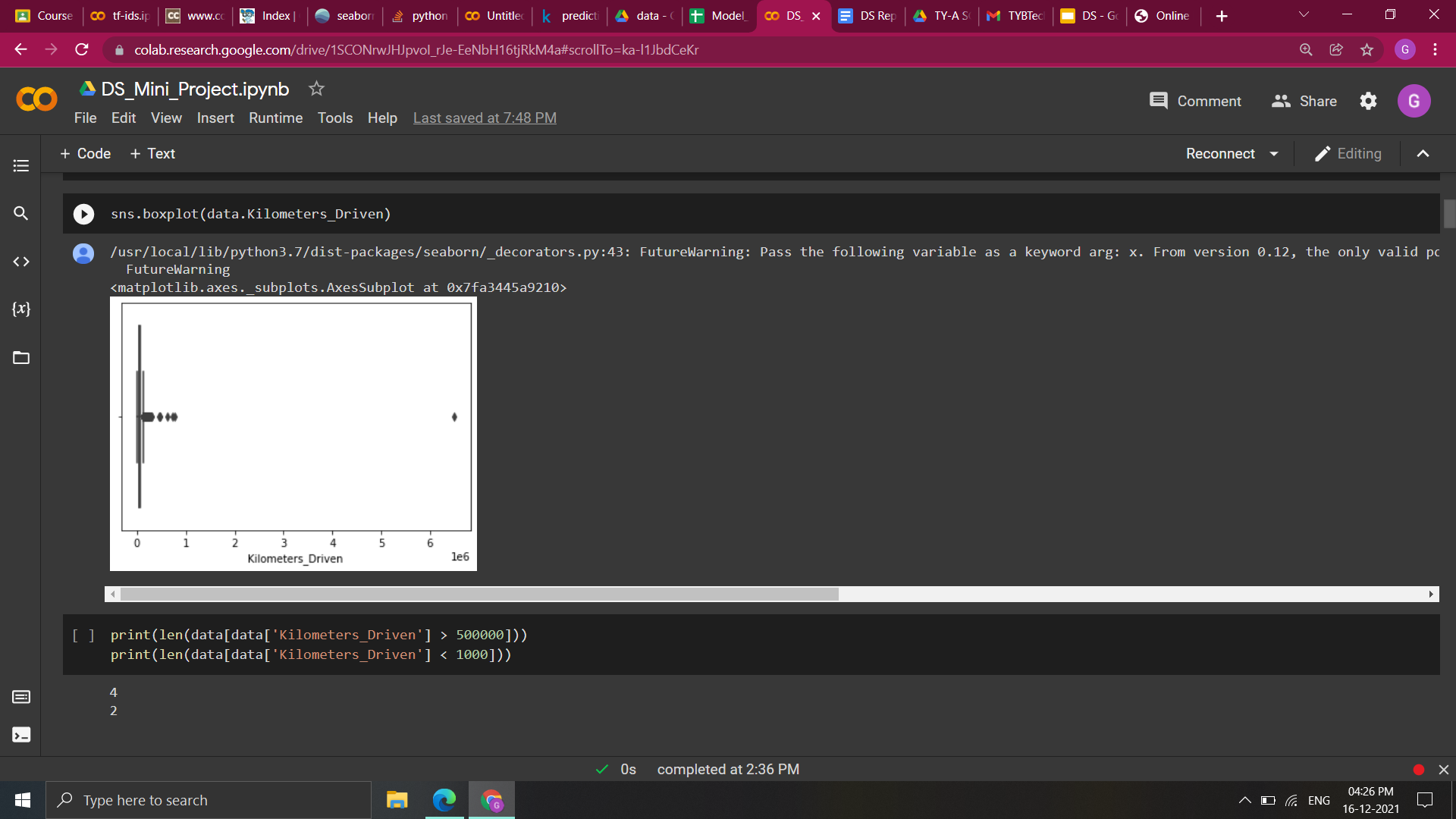
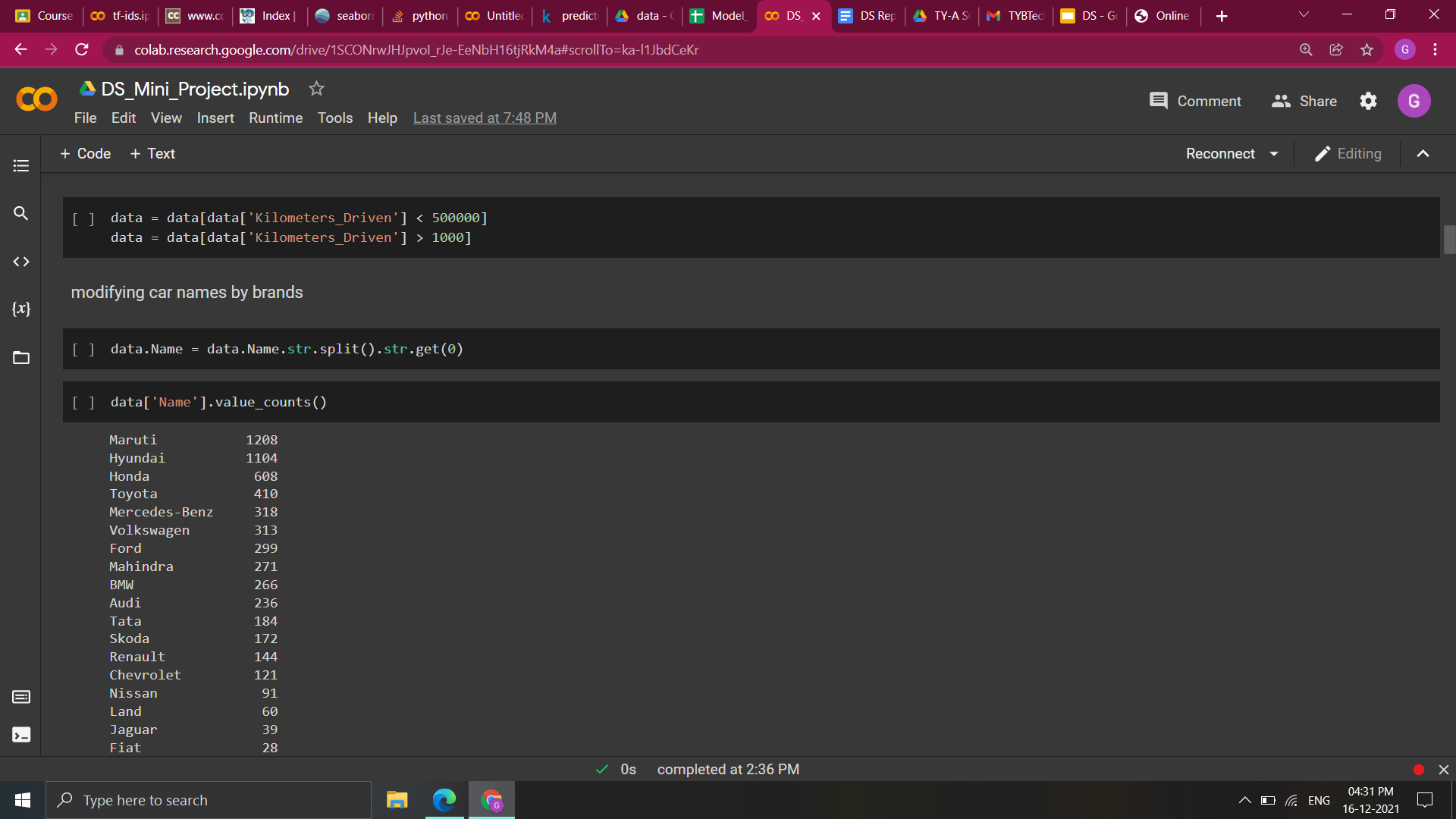
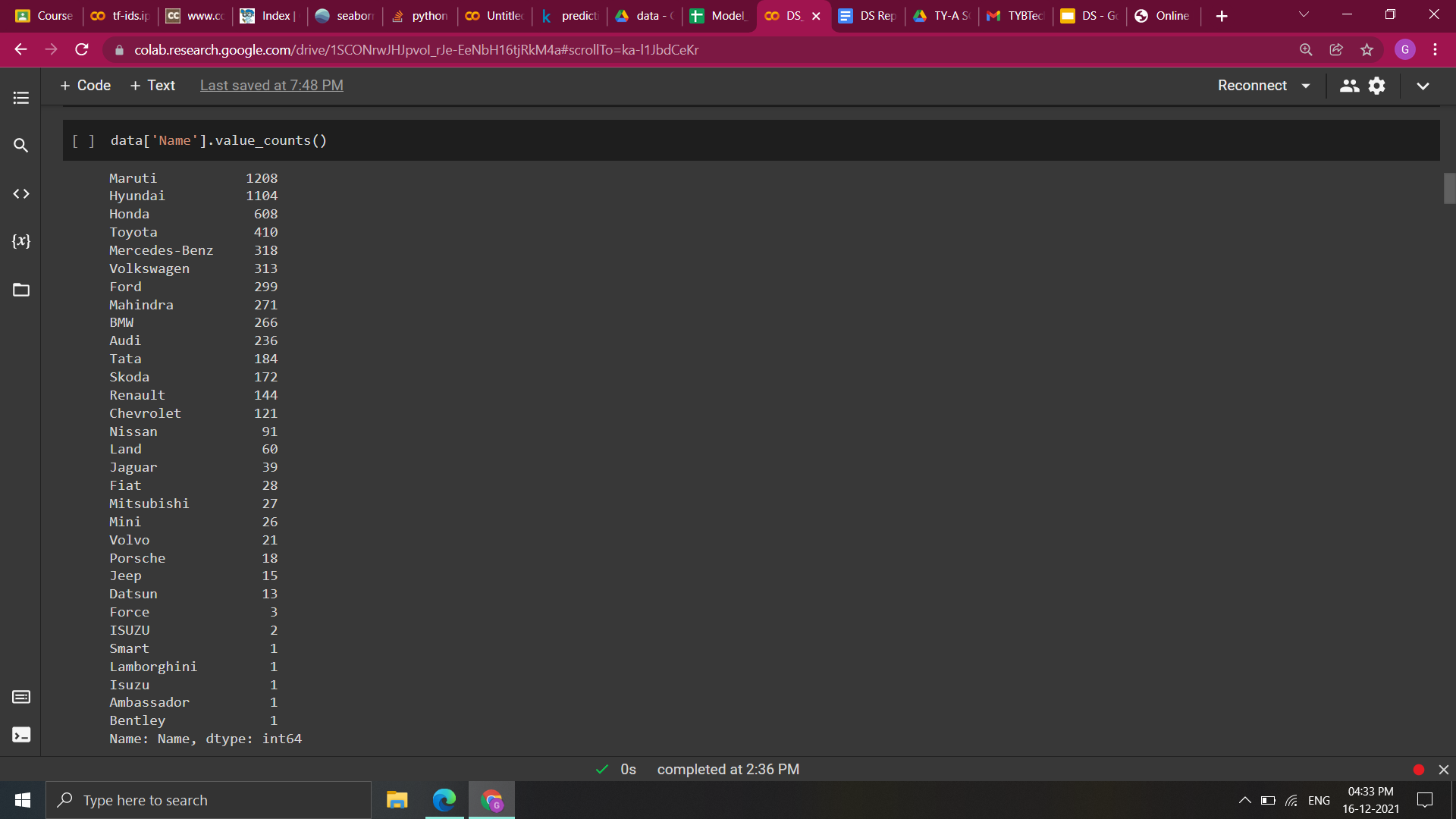
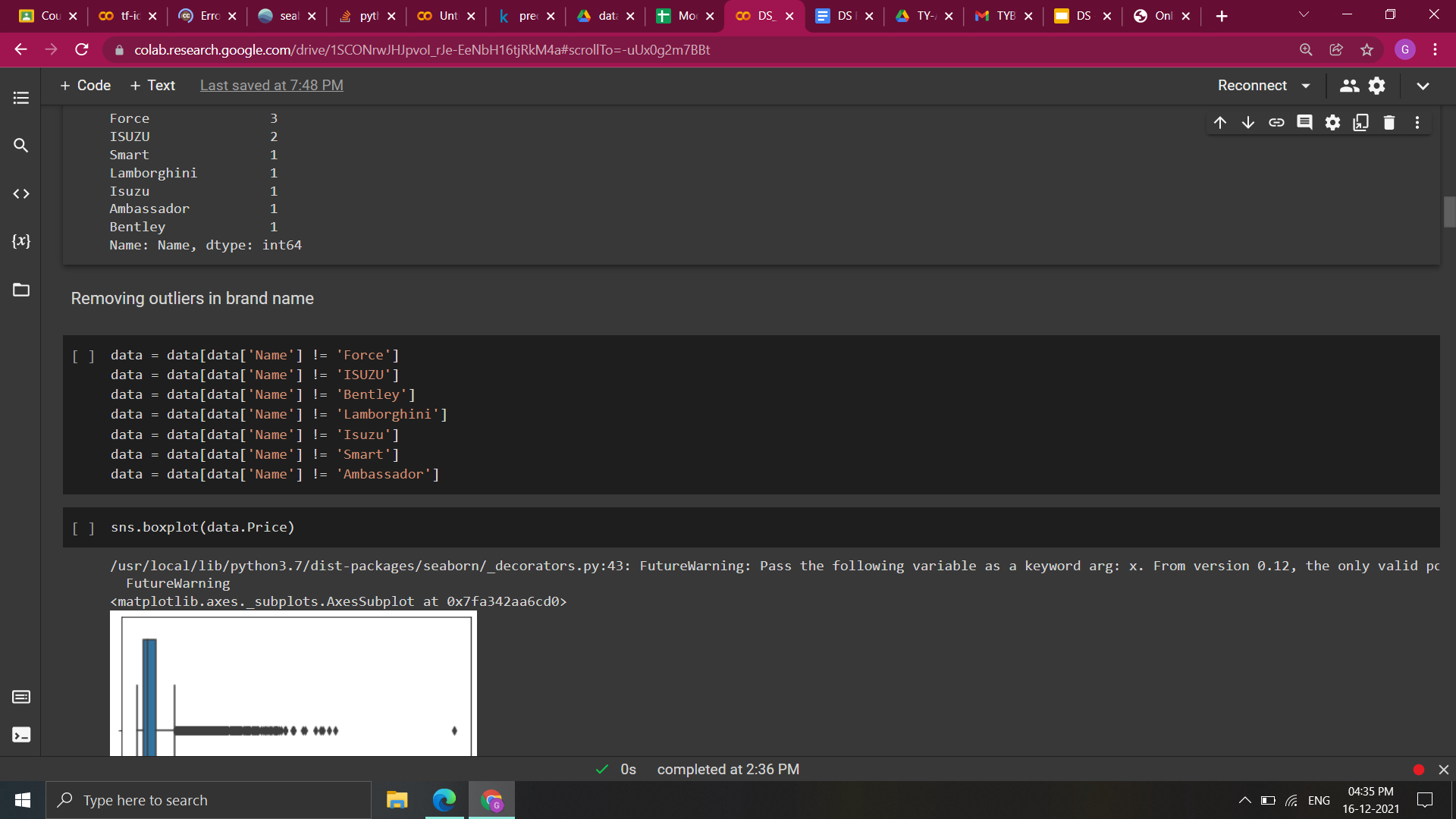
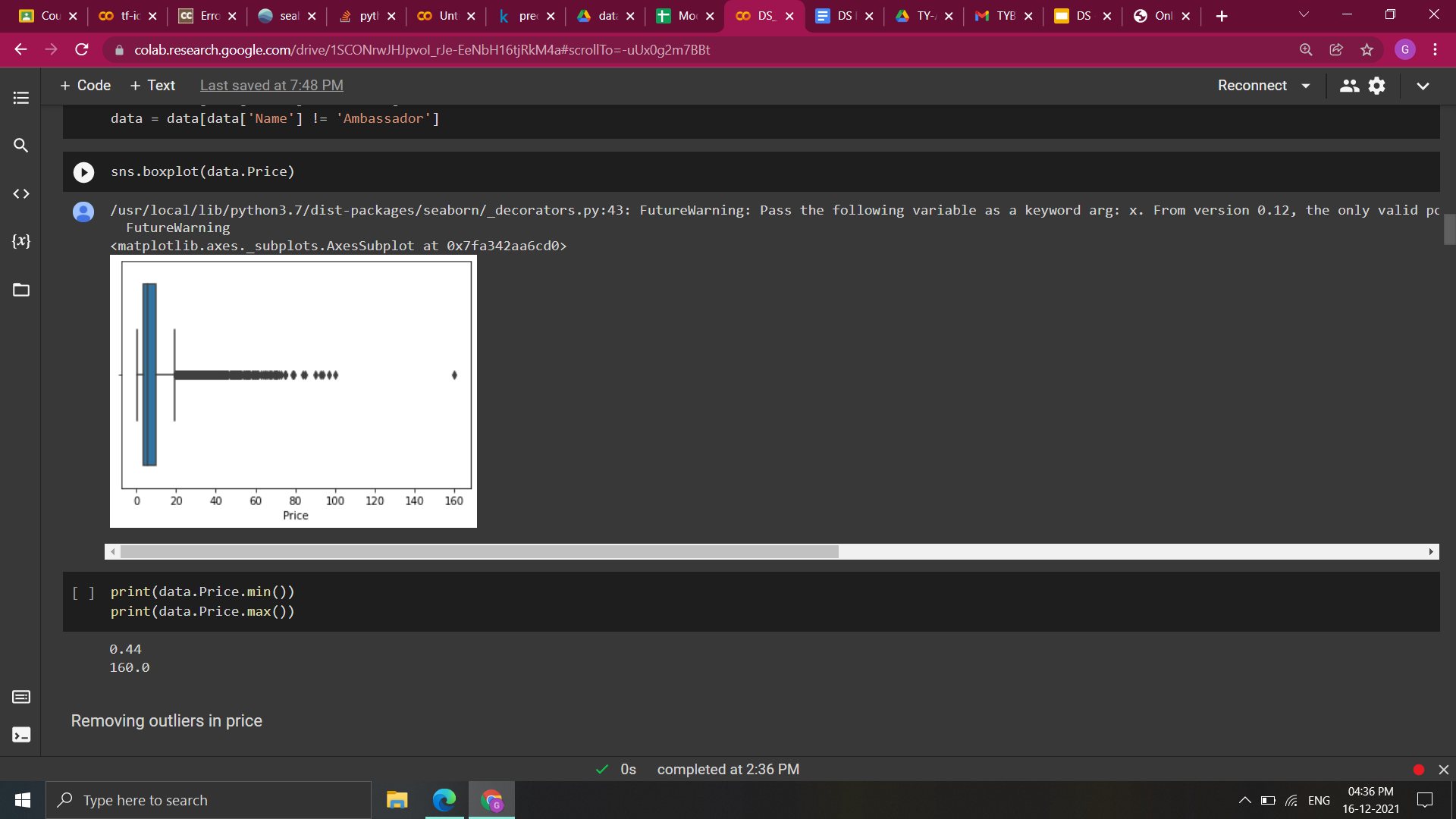
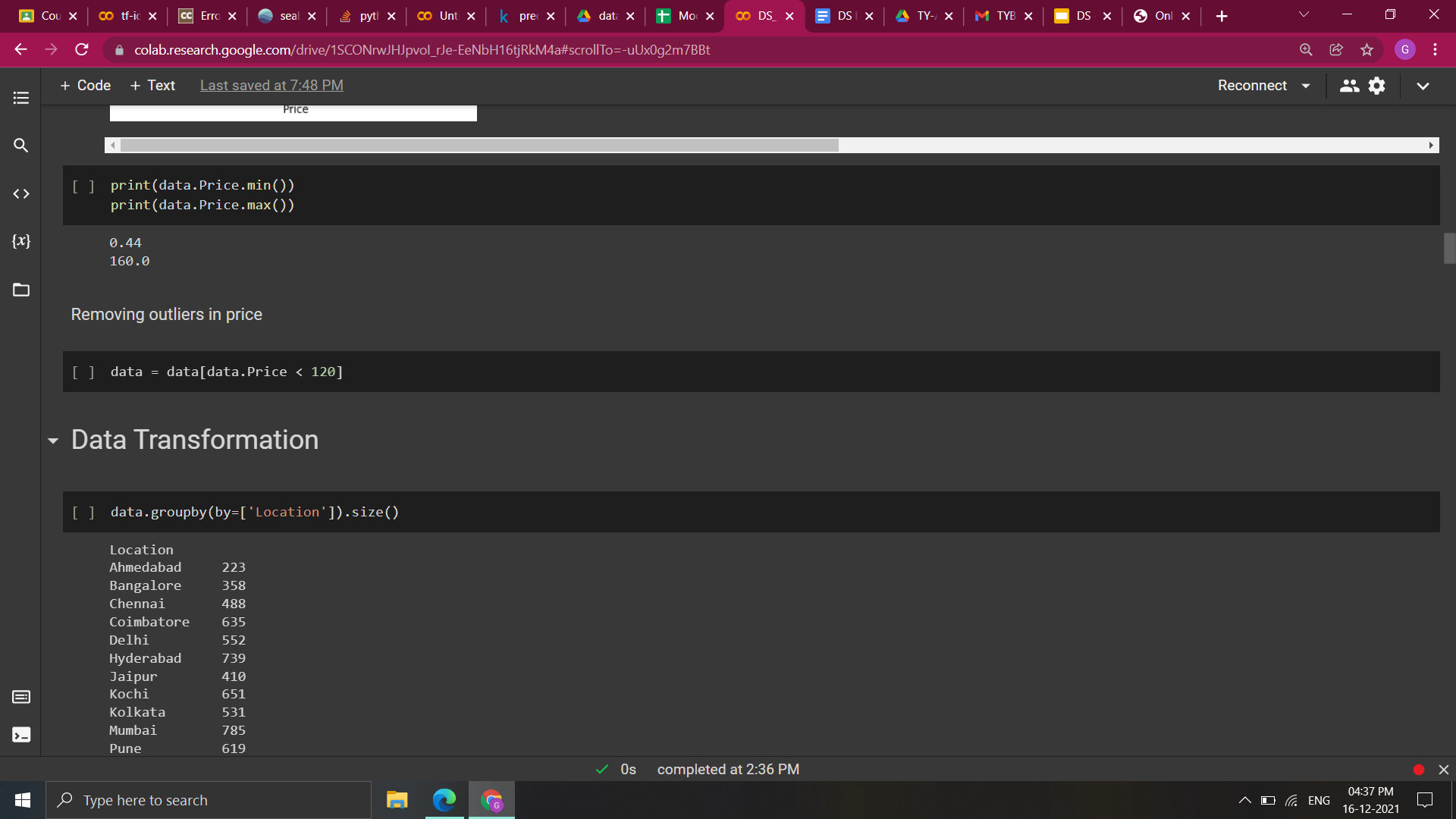
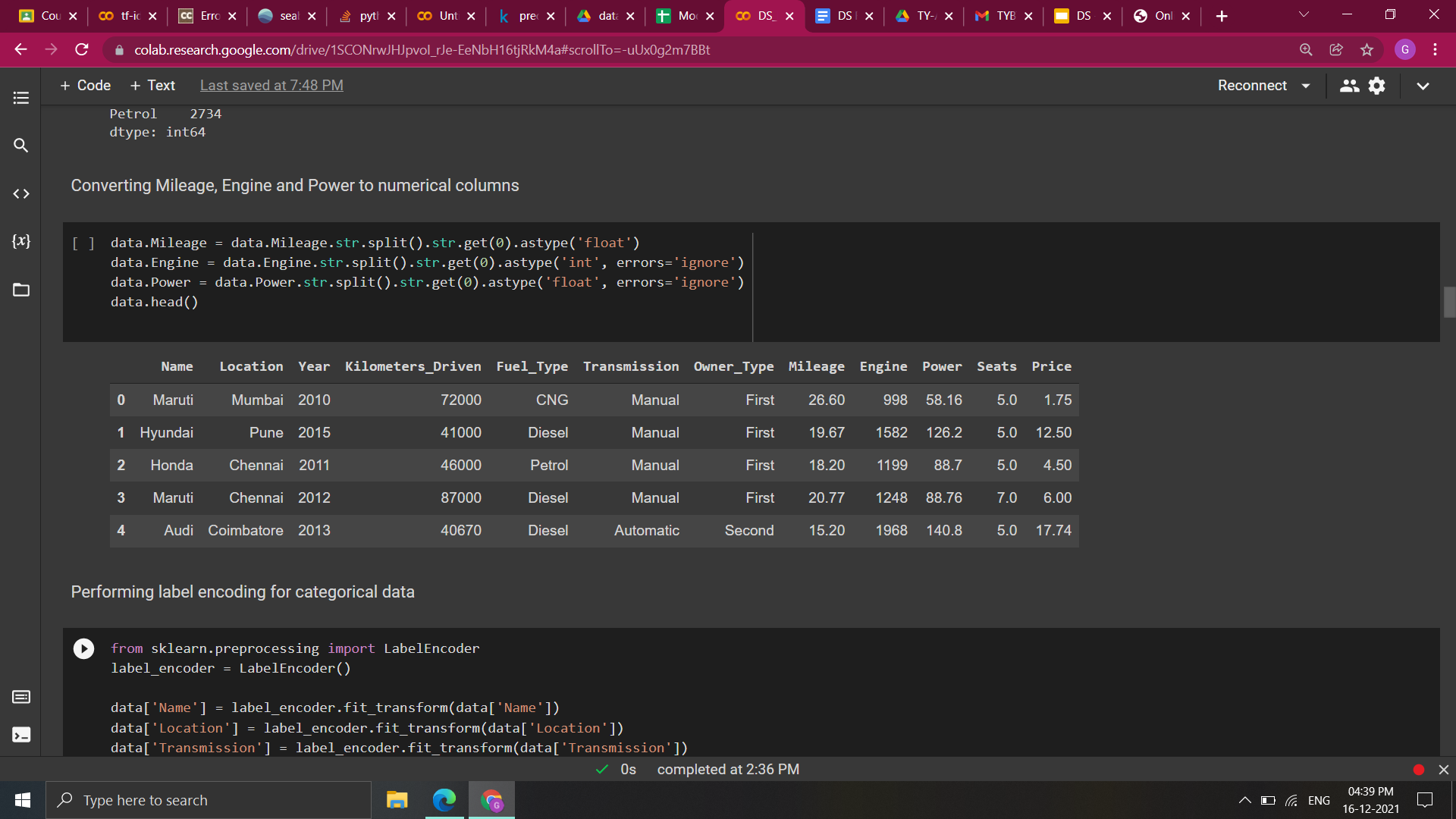
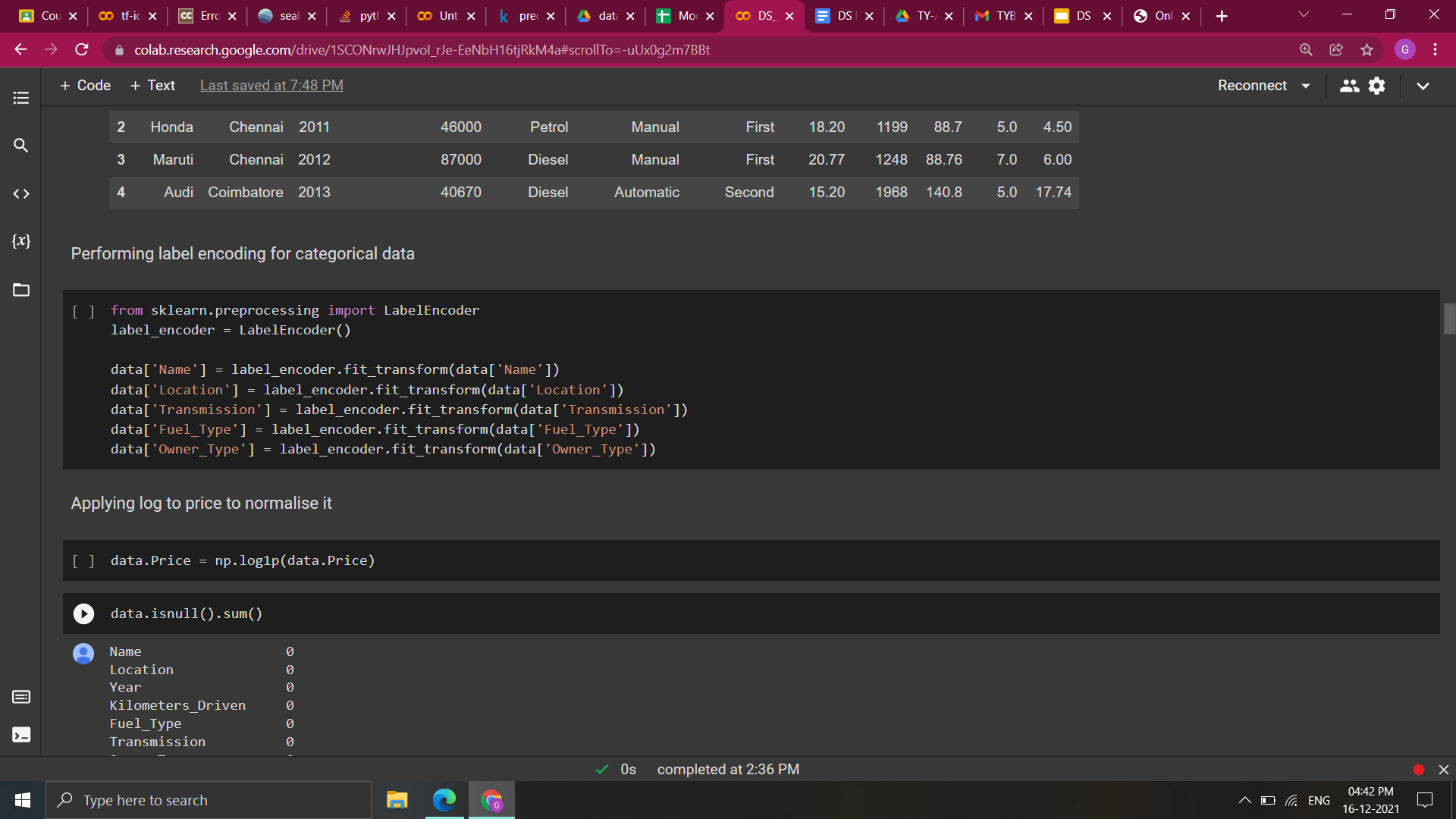
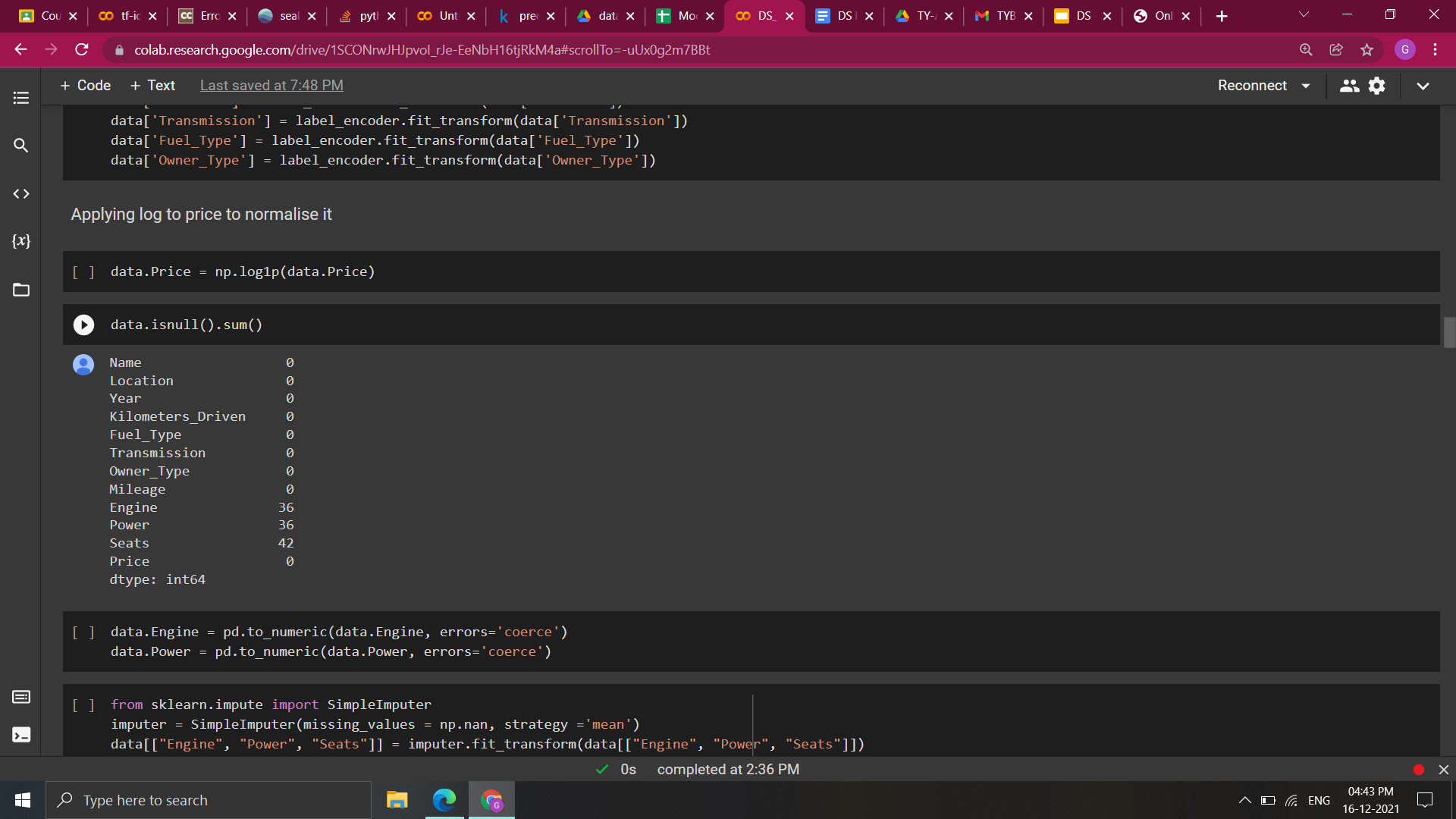
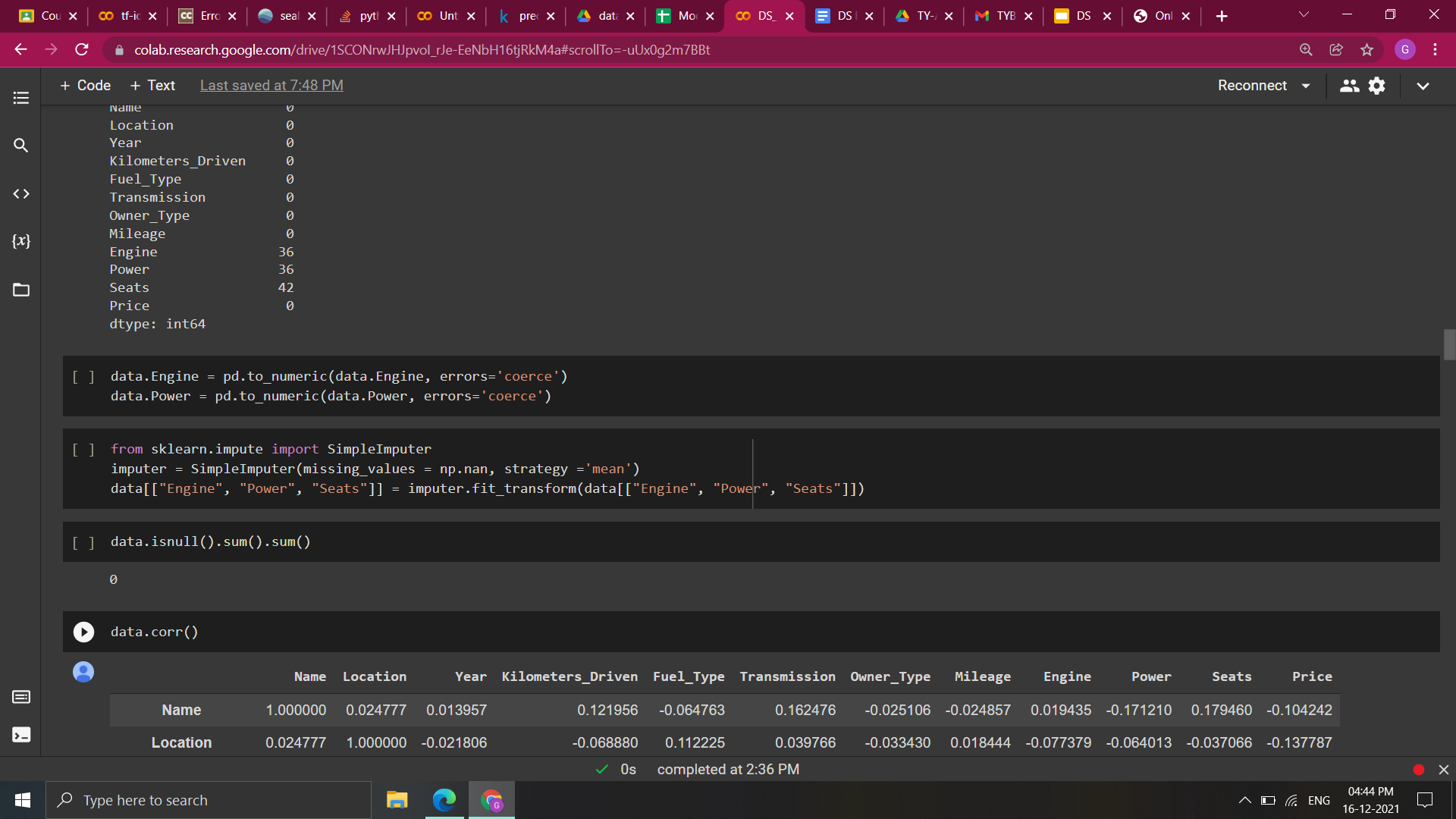
**Correlation Matrix and Scatter Plot**

* Positive correlation between Power and Price: 0.77
* Negative correlation between Transmission and Price: -0.69

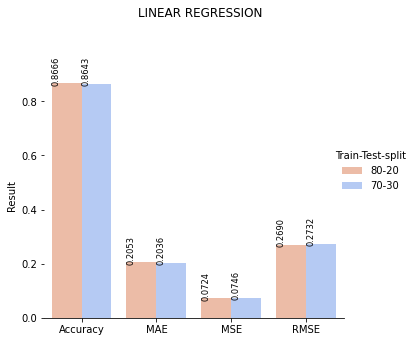
1. **Data Pre-Processing**

* Removing outliers in ‘**Fuel\_Type’** :  
   **  
  **

* Removing outliers in **‘Kilometers\_Driven’** :****
* Modifying car names by brand :   
  
* Removing outliers in car **‘Name’** :  
    
  
* Removing outliers in car **‘Price’** : **  
  **
* Converting **‘Mileage’** , **‘Engine’** and  **‘Power’**  to numerical columns :  
  ****
* Applying **Label Encoding**  for categorical data : ****
* Normalizing **‘Price’** using log :   
  
* Filling Null values by  **mean** :  
  

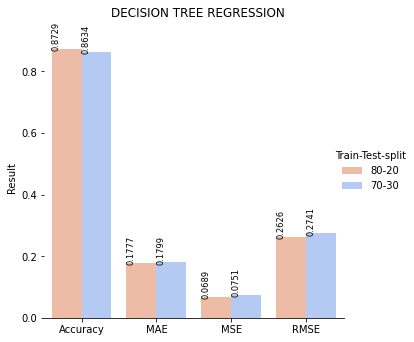
1. **Result and Discussion**

* **Performance Analysis**
* **Linear Regression**

****

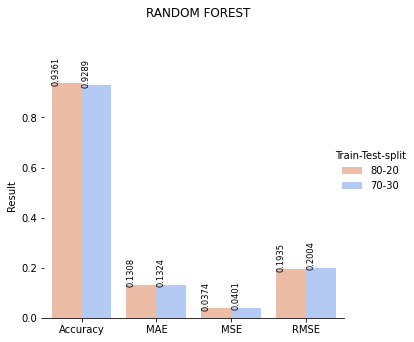
Linear regression Model performs Better for a Train: Test Split of 80:20

* **Decision Tree regression**

****

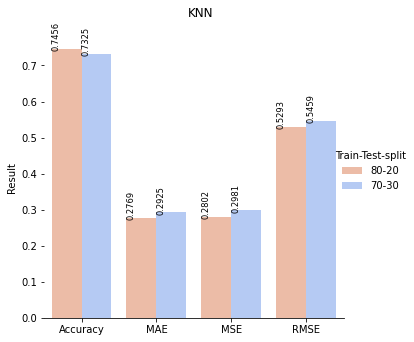
Decision tree performs better when the train:test split is 80:20 as compared to 70:30.

* **Random Forest Regression**

****

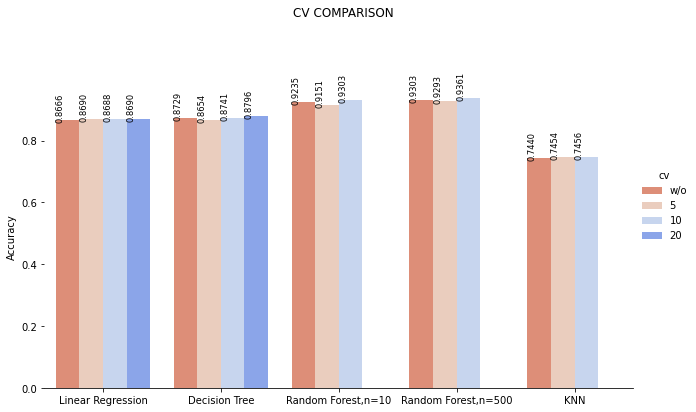
Random forest has a better accuracy for 80:20 split of train and test data.

* **K-Nearest Neighbor Regression**

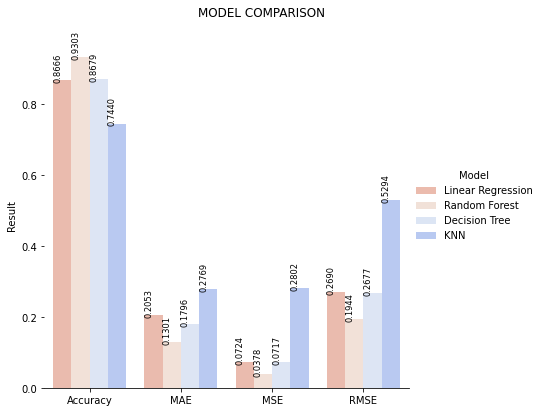
****

The model has higher accuracy and lower error score for train:test split of 80:20 as compared to 70:30

* **Performance comparison of Models after applying Cross-Validation**

****

* Linear Regression and Decision tree has the highest accuracy for cv=20
* Random Forest and KNN has the highest accuracy for cv=10
* **Overall Comparison of Models**

****

Random Forest has an accuracy of **93.03%** which is highest among all 4 models.

1. **Inference**

* We studied different research papers and analyzed different algorithms and techniques that have been used .
* The dataset which we have chosen named ‘Used Car Prices’ has 6019 rows and 12 unique attributes, of which ‘Price’ is our target attribute.
* There are 7 discrete and 5 continuous attributes in our dataset of which our target attribute ‘Price’ is continuous.
* We preprocessed our data by removing outliers in ‘Fuel\_Type’ , ‘Kilometers\_Driven’ , ‘Name’ and ‘Price’ .
* Label Encoding is performed for categorical attributes which are ‘Name’ , ’Location’ , ‘Transmission’ , ‘Fuel\_Type’ and ‘Owner\_type’ .
* ‘Price’ is normalized by the use of log.
* Null values in ‘Engine’ , ‘Power’ and ‘Seats’ fields are replaced by the ‘mean’ of that attribute .
* By some visualization performed we can see that 71% of cars have ‘Manual’ Transmission ,13% of cars listed are from Mumbai and ‘First’ tops the ‘Owner\_Type’ category .
* From the correlation heatmap we can see that ‘Price’ is positively correlated with ‘Power’ with correlation index of 0.77 and negatively correlated with ‘Transmission’ with correlation index of -0.69.
* After all the preprocessing we have built four models for prediction which are ‘Linear regression’, ‘Decision Tree’, ‘Random Forest’ and ‘K Nearest Neighbor’.
* After implementing ‘Linear Regression’ for 80-20 and 70-30 split, better results were observed for 80-20, which were accuracy of 0.8666, MAE of 0.2053, MSE of 0.0724 and RMSE of 0.2690.
* In the ‘Decision Tree’ implementation for 80-20 and 70-30 split, better results were observed for 80-20 , which were accuracy of 0.8729, MAE of 0.1777, MSE of 0.0689 and RMSE of 0.2626.
* For the ‘Random-Forest’ model , among 80-20 and 70-30 split, 80-20 gave better results with accuracy of 0.9361, MAE of 0.1308, MSE of 0.0374 and RMSE of 0.1935.
* In the ‘K-nearest neighbor’ implementation for 80-20 and 70-30 split, better results were observed for 80-20 , which were accuracy of 0.7456, MAE of 0.2769, MSE of 0.2802 and RMSE of 0.5293.
* Linear Regression and Decision tree has the highest accuracy for cv=20 while Random Forest and KNN has the highest accuracy for cv=10.
* So the final results for our four models are as accuracy=0.86896 , MAE = 0.20529 , MSE=0.072369 and RMSE=0.26901 for Linear Regression.  
  For the Decision Tree we have accuracy=0.87963, MAE=0.17769, MSE=0.06894 and RMSE=0.26257.  
  Random Forest has accuracy=0.9361247, MAE=0.13018, MSE=0.03745 and RMSE=0.19353  
  For K nearest neighbor we have accuracy=0.745621, MAE=0.27689, MSE=0.28023 and RMSE=0.5293.
* So it is clearly seen, best Performance is given by ‘Random-Forest’ and least by ‘K-Nearest Neighbor’.

1. **Conclusion**

We analyzed and summarized multiple research papers, based on which we were able to selectively choose multiple algorithms to maximise the accuracy and among various algorithms like Linear Regression, Random Forest, K Nearest Neighbor and Decision Trees we were able to obtain the maximum accuracy of 93.61 using the Random Forest Algorithm.

1. **References**

* [**https://www.nsbm.ac.lk/wp-content/uploads/2021/08/ICOBI\_2019\_Fair-Price-Prediction-System-for-Used-Cars-in-Sri-Lanka-Using-Machine-Learning-and-Robotic-Process-Automation.pdf**](https://www.nsbm.ac.lk/wp-content/uploads/2021/08/ICOBI_2019_Fair-Price-Prediction-System-for-Used-Cars-in-Sri-Lanka-Using-Machine-Learning-and-Robotic-Process-Automation.pdf)
* [**https://aip.scitation.org/doi/pdf/10.1063/1.4982530**](https://aip.scitation.org/doi/pdf/10.1063/1.4982530)
* [**https://www.ijirase.com/assets/paper/issue\_1/volume\_4/V4-Issue-3-686-689.pdf**](https://www.ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-3-686-689.pdf)
* [**http://cs229.stanford.edu/proj2019aut/data/assignment\_308832\_raw/26612934.pdf**](http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf)
* [**https://www.irjmets.com/uploadedfiles/paper/volume3/issue\_3\_march\_2021/6681/1628083284.pdf**](https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf)
* [**https://www.temjournal.com/content/81/TEMJournalFebruary2019\_113\_118.pdf**](https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf)
* [**http://junikhyatjournal.in/no\_1\_Online\_21/68.pdf**](http://junikhyatjournal.in/no_1_Online_21/68.pdf)
* [**https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf**](https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf)
* [**https://www.ijirt.org/master/publishedpaper/IJIRT151705\_PAPER.pdf**](https://www.ijirt.org/master/publishedpaper/IJIRT151705_PAPER.pdf)
* [**https://www.ijcseonline.org/pub\_paper/74-IJCSE-07051.pdf**](https://www.ijcseonline.org/pub_paper/74-IJCSE-07051.pdf)
* [**https://www.academia.edu/13579173/Predicting\_the\_Price\_of\_Second\_hand\_Cars\_using\_Artificial\_Neural\_Networks**](https://www.academia.edu/13579173/Predicting_the_Price_of_Second_hand_Cars_using_Artificial_Neural_Networks)