**Regression and Prediction Using R**

Jiufeng Zhang

B00654927

**Abstract**

As part of daily life, machine learning is used to make decisions, especially by data scientists. This paper aimed to incorporate machine learning algorithms in the prediction of vehicle prices. First, the car.csv dataset was inspected, cleaned, and organized. A final dataset was arrived at and used for further analysis. The dataset was fitted with Present\_Price as the response variable and the rest as explanatory variables using the basic linear model. Three algorithms, linear regression, random forest, and support vector machines, were selected for modeling. Data were partitioned into two; training and testing set. The training data was used to predict the prices on the testing set. The models' performances were evaluated and improved using tuning, cross-validation, and checked for overfitting. Lastly, the models were compared against one another using a calculated RMSE (Root Mean Squared Error). The best performing model was chosen, hence leading to the arrival of meaningful conclusions.

***Keywords:*** Algorithms, linear regression, random forest, basic linear model, dataset, regression, and output.

**Introduction**

Every day, applications for machine learning are typical to come across. The algorithms help in making critical decisions in every field of work. For instance, media sites rely on machine learning to sift through millions of options to give you song or movie recommendations, and retailers use it to gain insight into their customers' purchasing behavior. Closer to home, data scientists using machine learning to advise on future data patterns and behavior that could be encouraged or discouraged. Besides, it entails building models that offer predictive power and can be used to understand data not yet collected.

In earlier statistics classes, machine learning has been used when running simple regression models. On the other hand, this is a complex topic with a wide range of possibilities and applications. Therefore, this study sought to present a basic understanding of regression modeling using linear, random forest, and support vector modeling, as well as to answer the following questions of interest;

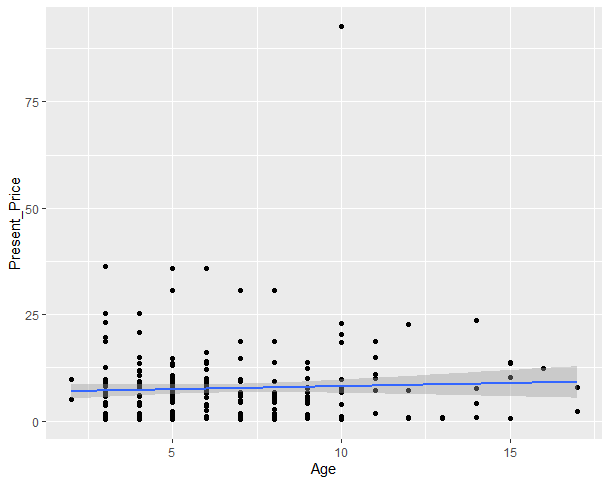
* What is the relationship among variables, especially between the car price variable with other variables?
* Is it possible to predict the price of a new car based on historical data?
* Which is the best model for use in prediction among the three?

**Data Description**

The car dataset contained information about cars and motorcycles listed on CarDekho.com. The car data was in a CSV file and included the following columns: model/ Car\_Name, year, selling price, showroom price/Present\_Price, kilometers driven, fuel type, seller type, transmission, and the number of previous owners. Using R, the dataset had nine columns and 301 rows/observations. After cleaning the dataset, there were four numerical and five categorical variables. Those categories were model, fuel type, seller type, and transmission. Categorical ones were namely, year, selling prices, showroom price, kilometers driven, and the number of previous owners. As instructed, an additional data point was added for a 2018 manual city selling at 4.34M but presently priced at 5.12M with a mileage of 37,000 kilometers. The model was being sold by a Dealer, thus no previous ownership, and ran on Diesel. During cleaning, the model variable was dropped as it was regarded unimportant for upcoming analysis. A new variable, age, was created after subtracting the year values from the current year (2020). This meant deletion of the year and current year variables. Notably, there were no missing values in the dataset. Finally, the car dataset contained 302 rows and eight columns; Selling\_Price, Present\_Price, Kms\_Driven Fuel\_Type, Seller\_Type, Transmission, Owner, and Age.

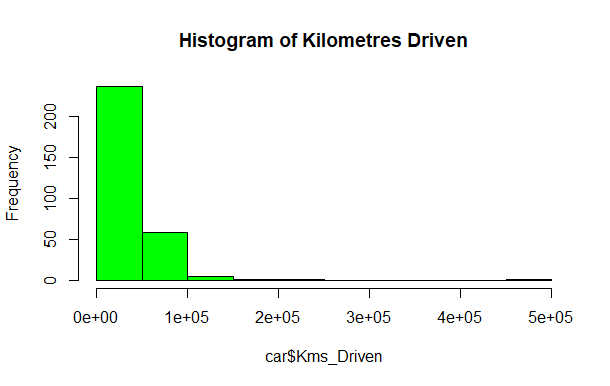
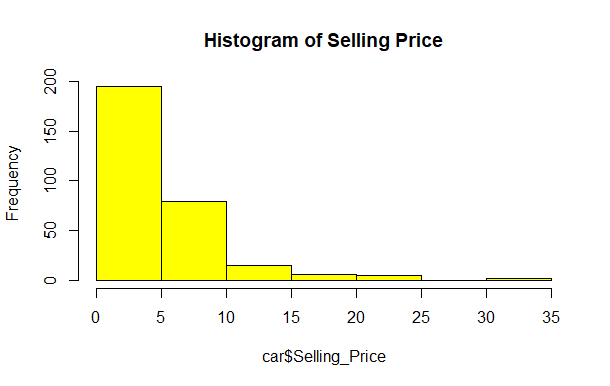
Further, visual presentations in ggplots, histograms, and plots were obtained for each variable.

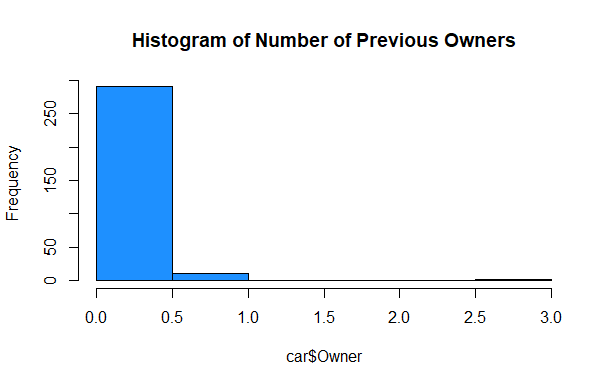
*A ggplot of Age and Present\_Price*



The above ggplot showed that the relationship between the ages of the cars was positive and linear. In other words, the newer the car, the more expensive it was.

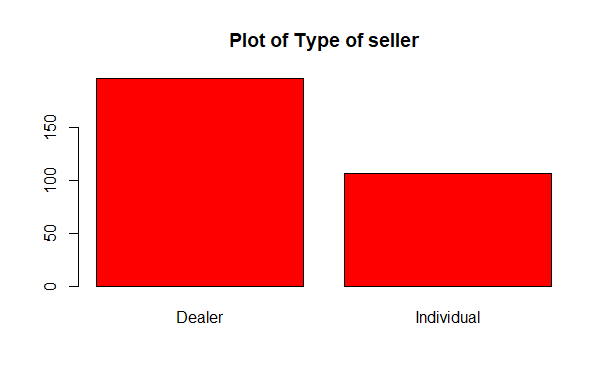
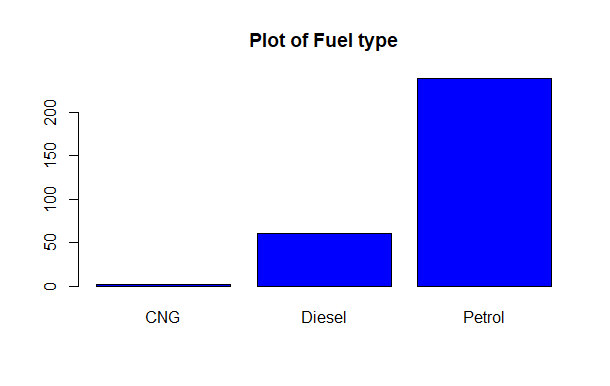
*Histograms*

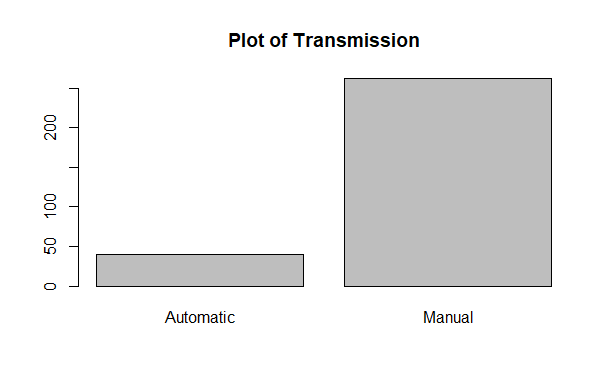




Most cars' selling prices were between 0 and 5M while a few cars were priced between 30 and 35M. Also, most cars had the least number of kilometers driven/mileage. Most cars had close to no previous ownership, with a few having about three previous owners. Expectedly, most cares were either new or almost new.

*Plots*





The majority of the cars ran on Petrol while less than 50 cars used CNG. Dealers were selling more cars than individuals. Many operated on manual while less than 50 or a few were automatic. With all these explanations, most of the cars featured in the dataset were new.

**Methods**

As mentioned earlier, a linear regression was fit using Present\_Price as the response variable. Overall, from the R output, the model's predictors explained 84.71% R2 of the price variation. The model was a good fit, *F* (8, 293) = 202.9, *p-value* < 2.2e-16. The following model equation was obtained.

Present\_Price = Selling\_Price + Kms\_Driven + Fuel\_Type + Seller\_Type + Transmission + Owner + Age

However, some coefficients were more statistically significant than others as ranked by importance below.

|  |  |
| --- | --- |
| Overall | |
| Selling\_Price | 27.4908128 |
| Kms\_Driven | 2.1504483 |
| Fuel\_Type Diesel | 0.6907296 |
| Fuel\_Type Petrol | 0.2347870 |
| Seller\_Type Individual | 0.3083687 |
| Transmission Manual | 0.5142432 |
| Owner | 1.9049391 |
| Age | 8.1736365 |

For example, selling price, kilometers are driven, and age was statistically significant in predicting showroom price (*p* < 0.05). Fuel type and seller type were negative and statistically insignificant, while transmission and owner were positive but insignificant (*p* > 0.05). As expected, the more miles a car had, the cheaper it was. However, this was not the case because the more kilometers drove, the more the showroom price. The transmission manual did not have much of an influence on Prices.

Based on the above results, it was generally not a noble idea to use the entire data sample to fit the model. Hence, the decision was made to train the model on a sample of the data. Then, its performance would be observed outside of the training sample. Thus, the dataset was partitioned into training and test set, where the latter was used to evaluate the performance of the three models with unseen data. See the next section.

**Results**

On a ratio of 0.8: 0.2, the dataset was divided to form training and testing sets, 241 rows and eight columns and 61 rows and eight columns. Next, the caret package was loaded to obtain the train function used for model training. A glimpse of the first six rows of the predicted showroom prices of cars on the testing set was obtained using the predict function. Evaluation metrics, RMSE, R2, MAE, and many others were calculated for each algorithm. All models were compared against one another based on the metrics.

1. *Linear Regression algorithm*

The first algorithm, linear regression, was trained using the train function where Present\_Price/showroom price was the outcome variable. The following results were obtained.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Intercept | RMSE | R Squared | MAE | RMSESD | R Squared SD | MAESD |
| TRUE | 3.425228 | 0.831936 | 2.098598 | 1.871193 | 0.1224851 | 0.5642068 |

On the testing set, a trained model was used to predict car prices at the showroom.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 6 | 10 | 17 | 21 | 22 | 23 |
| 11.320564 | 10.548235 | 10.184033 | 2.766763 | 5.617042 | 9.694914 |

The predictive ability of the model was tested on the testing set, and the evaluation metrics were derived as shown below;

|  |  |  |
| --- | --- | --- |
| RMSE | R squared | MAE |
| 4.4259303 | 0.8192709 | 2.5507228 |

The above output showed that RMSE, one of the two evaluation metrics, was 3.43M for the training set and 4.43M for the testing set. On the other hand, the R-squared value was around 83% and 82% for both training and testing sets, respectively, which indicated good performance.

1. *Support Vector Machine Algorithm*

Similarly, the cross-validation methods were specified with k-folds and leave one out CV. Since SVM has hyperparameters, they were tuned by inspecting and modifying specific algorithm parameters. Specific values were fed into the model training function using the expand.grid command.

Then, the model was tuned, trained on the training set.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | C | RMSE | R squared | MAE | RMSES | R squared SD | MAESD |
| 1e-03 | | 6.112354 | 0.7257655 | 3.376103 | 3.913042 | 0.1337751 | 1.0075231 |
| 1e-02 | | 3.961975 | 0.8439960 | 1.988603 | 3.079260 | 0.1179277 | 0.7879450 |
| 1e-01 | | 3.435029 | 0.8595457 | 1.818873 | 2.524667 | 0.1268926 | 0.6702223 |
| 1e+00 | | 3.365431 | 0.8604723 | 1.822928 | 2.328030 | 0.1310797 | 0.6151965 |
| 1e+01 | | 3.365003 | 0.8607997 | 1.822493 | 2.323299 | 0.1307758 | 0.6101864 |
| 1e+02 | | 3.365878 | 0.8607631 | 1.823303 | 2.324530 | 0.1307558 | 0.6107806 |

Afterward, it was used to obtain the predicted car prices at the showroom.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 6 | 10 | 17 | 21 | 22 | 23 |
| 10.798856 | 10.37632 | 10.0967 | 3.925298 | 5.803843 | 8.91126 |

The model's predictability was evaluated on the testing set. See the table below.

|  |  |  |
| --- | --- | --- |
| RMSE | R squared | MAE |
| 3.1847648 | 0.8468279 | 1.7856504 |

The tables above showed that RMSE, one of the two evaluation metrics, was between 3.3 and 6.1M for training data and 3.2M for the testing set. On the other hand, the R-squared value was around 72.3 to 86.1% and 85% for both training and testing sets, respectively, which indicated good performance.

1. *RandomForest Algorithm*

Like -wise, the cross-validation method was specified before tuning and later training the model. The table below shows the model's results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mtry | RMSE | R squared | MAE | RMSESD | R squared SD | MAESD |
| 1 | 2 | 3.691223 | 0.8484474 | 1.956376 | 3.842158 | 0.12096760 | 0.9986551 |
| 2 | 5 | 2.997213 | 0.9142653 | 1.343094 | 3.646090 | 0.08030471 | 0.8501472 |
| 3 | 8 | 2.983392 | 0.9152494 | 1.305836 | 3.402356 | 0.06566592 | 0.7632234 |

As usual, the model was used to obtain the first six rows of the predicted prices as below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 6 | 10 | 17 | 21 | 22 | 23 |
| 10.634268 | 10.822836 | 10.779255 | 4.185635 | 5.818534 | 11.641299 |

The predictive ability of the model was assessed on the testing set using the three metrics.

|  |  |  |
| --- | --- | --- |
| RMSE | R squared | MAE |
| 3.2610962 | 0.8614406 | 1.4489065 |

Lastly, all tables above showed that RMSE, one of the two evaluation metrics, ranged from 3 to 3.7M for the training set and 3.3M for the testing set. On the other hand, the R-squared value was around 84.8 to 91.5% and 86.1% for both training and testing sets, respectively, which indicated good performance.

All three models were compared to each other using the RMSE and R2 to determine the best model for predicting car prices at the showroom. The data frame containing all values for the three models on the two metrics was obtained, as shown in the R output. Also, note that the evaluation metrics could not be discussed per model because there would not be any values to compare with.

Linear regression had the lowest RMSE of 3.4M on the training set, while the support vector machine model had the highest at 6.11M. In terms of R2 values on the training set, the random forest model had the highest (84.8%), while the support vector machine had the least at 72.6%. The latter also had the lowest RMSE on the testing set at 3.18M, while linear regression had the highest at 4.43M. For R2 on the testing set, linear regression had the least at 81.9%, while random forest had the highest (86.14%). Typically, a lower RMSE and higher R2 values are indicative of a good model. Therefore, the best models for predicting car prices were the support vector machine (low RMSE of 3.18M) and random forest due to a higher R2 value of 86.14%.

**Conclusion**

This study aims to help understand the predictability of regression models in R using the car dataset. The critical points of interest were addressed via regression, which showed a relationship between showroom price and predictors. Similarly, predicting a car's price based on the already known data in each model was possible. See the predicted prices above in tabular form. Besides, the RMSE and R2 helped to determine the best performing model. The regularized regression models, support vector machine, and random forest performed better than the linear regression model. Both models had a lower RMSE and a higher R2 than the linear regression model. Overall, all the models performed well with decent R2 above 80% and stable RMSE values below 7. In reality, the ideal result would be an RMSE value of zero and an R-squared value of 1, but that's almost impossible in real economic datasets.

**Appendix**

1. *The Data*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Selling\_Price** | **Present\_Price** | **Kms\_Driven** | **Fuel\_Type** | **Seller\_Type** | **Transmission** | **Owner** | **Age** |
| 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 | 6 |
| 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 | 7 |
| 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 | 3 |
| 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 | 9 |
| 4.6 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 | 6 |
| 9.25 | 9.83 | 2071 | Diesel | Dealer | Manual | 0 | 2 |
| 6.75 | 8.12 | 18796 | Petrol | Dealer | Manual | 0 | 5 |
| 6.5 | 8.61 | 33429 | Diesel | Dealer | Manual | 0 | 5 |
| 8.75 | 8.89 | 20273 | Diesel | Dealer | Manual | 0 | 4 |
| 7.45 | 8.92 | 42367 | Diesel | Dealer | Manual | 0 | 5 |
| 2.85 | 3.6 | 2135 | Petrol | Dealer | Manual | 0 | 3 |
| 6.85 | 10.38 | 51000 | Diesel | Dealer | Manual | 0 | 5 |
| 7.5 | 9.94 | 15000 | Petrol | Dealer | Automatic | 0 | 5 |
| 6.1 | 7.71 | 26000 | Petrol | Dealer | Manual | 0 | 5 |
| 2.25 | 7.21 | 77427 | Petrol | Dealer | Manual | 0 | 11 |
| 7.75 | 10.79 | 43000 | Diesel | Dealer | Manual | 0 | 4 |
| 7.25 | 10.79 | 41678 | Diesel | Dealer | Manual | 0 | 5 |
| 7.75 | 10.79 | 43000 | Diesel | Dealer | Manual | 0 | 4 |
| 3.25 | 5.09 | 35500 | CNG | Dealer | Manual | 0 | 5 |
| 2.65 | 7.98 | 41442 | Petrol | Dealer | Manual | 0 | 10 |
| 2.85 | 3.95 | 25000 | Petrol | Dealer | Manual | 0 | 4 |
| 4.9 | 5.71 | 2400 | Petrol | Dealer | Manual | 0 | 3 |
| 4.4 | 8.01 | 50000 | Petrol | Dealer | Automatic | 0 | 9 |
| 2.5 | 3.46 | 45280 | Petrol | Dealer | Manual | 0 | 6 |
| 2.9 | 4.41 | 56879 | Petrol | Dealer | Manual | 0 | 7 |
| 3 | 4.99 | 20000 | Petrol | Dealer | Manual | 0 | 9 |
| 4.15 | 5.87 | 55138 | Petrol | Dealer | Manual | 0 | 7 |
| 6 | 6.49 | 16200 | Petrol | Individual | Manual | 0 | 3 |
| 1.95 | 3.95 | 44542 | Petrol | Dealer | Manual | 0 | 10 |
| 7.45 | 10.38 | 45000 | Diesel | Dealer | Manual | 0 | 5 |
| 3.1 | 5.98 | 51439 | Diesel | Dealer | Manual | 0 | 8 |
| 2.35 | 4.89 | 54200 | Petrol | Dealer | Manual | 0 | 9 |
| 4.95 | 7.49 | 39000 | Diesel | Dealer | Manual | 0 | 6 |
| 6 | 9.95 | 45000 | Diesel | Dealer | Manual | 0 | 6 |
| 5.5 | 8.06 | 45000 | Diesel | Dealer | Manual | 0 | 6 |
| 2.95 | 7.74 | 49998 | CNG | Dealer | Manual | 0 | 9 |
| 4.65 | 7.2 | 48767 | Petrol | Dealer | Manual | 0 | 5 |
| 0.35 | 2.28 | 127000 | Petrol | Individual | Manual | 0 | 17 |
| 3 | 3.76 | 10079 | Petrol | Dealer | Manual | 0 | 4 |
| 2.25 | 7.98 | 62000 | Petrol | Dealer | Manual | 0 | 17 |
| 5.85 | 7.87 | 24524 | Petrol | Dealer | Automatic | 0 | 4 |
| 2.55 | 3.98 | 46706 | Petrol | Dealer | Manual | 0 | 6 |
| 1.95 | 7.15 | 58000 | Petrol | Dealer | Manual | 0 | 12 |
| 5.5 | 8.06 | 45780 | Diesel | Dealer | Manual | 0 | 6 |
| 1.25 | 2.69 | 50000 | Petrol | Dealer | Manual | 0 | 8 |
| 7.5 | 12.04 | 15000 | Petrol | Dealer | Automatic | 0 | 6 |
| 2.65 | 4.89 | 64532 | Petrol | Dealer | Manual | 0 | 7 |
| 1.05 | 4.15 | 65000 | Petrol | Dealer | Manual | 0 | 14 |
| 5.8 | 7.71 | 25870 | Petrol | Dealer | Manual | 0 | 5 |
| 7.75 | 9.29 | 37000 | Petrol | Dealer | Automatic | 0 | 3 |
| 14.9 | 30.61 | 104707 | Diesel | Dealer | Automatic | 0 | 8 |
| 23 | 30.61 | 40000 | Diesel | Dealer | Automatic | 0 | 5 |
| 18 | 19.77 | 15000 | Diesel | Dealer | Automatic | 0 | 3 |
| 16 | 30.61 | 135000 | Diesel | Individual | Automatic | 0 | 7 |
| 2.75 | 10.21 | 90000 | Petrol | Individual | Manual | 0 | 15 |
| 3.6 | 15.04 | 70000 | Petrol | Dealer | Automatic | 0 | 11 |
| 4.5 | 7.27 | 40534 | Petrol | Dealer | Manual | 0 | 5 |
| 4.75 | 18.54 | 50000 | Petrol | Dealer | Manual | 0 | 10 |
| 4.1 | 6.8 | 39485 | Petrol | Dealer | Manual | 1 | 6 |
| 19.99 | 35.96 | 41000 | Diesel | Dealer | Automatic | 0 | 6 |
| 6.95 | 18.61 | 40001 | Petrol | Dealer | Manual | 0 | 7 |
| 4.5 | 7.7 | 40588 | Petrol | Dealer | Manual | 0 | 5 |
| 18.75 | 35.96 | 78000 | Diesel | Dealer | Automatic | 0 | 6 |
| 23.5 | 35.96 | 47000 | Diesel | Dealer | Automatic | 0 | 5 |
| 33 | 36.23 | 6000 | Diesel | Dealer | Automatic | 0 | 3 |
| 4.75 | 6.95 | 45000 | Diesel | Dealer | Manual | 0 | 6 |
| 19.75 | 23.15 | 11000 | Petrol | Dealer | Automatic | 0 | 3 |
| 9.25 | 20.45 | 59000 | Diesel | Dealer | Manual | 0 | 10 |
| 4.35 | 13.74 | 88000 | Petrol | Dealer | Manual | 0 | 9 |
| 14.25 | 20.91 | 12000 | Petrol | Dealer | Manual | 0 | 4 |
| 3.95 | 6.76 | 71000 | Diesel | Dealer | Manual | 0 | 6 |
| 4.5 | 12.48 | 45000 | Diesel | Dealer | Manual | 0 | 9 |
| 7.45 | 18.61 | 56001 | Petrol | Dealer | Manual | 0 | 7 |
| 2.65 | 5.71 | 43000 | Petrol | Dealer | Manual | 0 | 9 |
| 4.9 | 8.93 | 83000 | Diesel | Dealer | Manual | 0 | 6 |
| 3.95 | 6.8 | 36000 | Petrol | Dealer | Manual | 0 | 5 |
| 5.5 | 14.68 | 72000 | Petrol | Dealer | Manual | 0 | 7 |
| 1.5 | 12.35 | 135154 | Petrol | Dealer | Automatic | 0 | 16 |
| 5.25 | 22.83 | 80000 | Petrol | Dealer | Automatic | 0 | 10 |
| 14.5 | 30.61 | 89000 | Diesel | Dealer | Automatic | 0 | 8 |
| 14.73 | 14.89 | 23000 | Diesel | Dealer | Manual | 0 | 4 |
| 4.75 | 7.85 | 40000 | Diesel | Dealer | Manual | 0 | 5 |
| 23 | 25.39 | 15000 | Diesel | Dealer | Automatic | 0 | 3 |
| 12.5 | 13.46 | 38000 | Diesel | Dealer | Manual | 0 | 5 |
| 3.49 | 13.46 | 197176 | Diesel | Dealer | Manual | 0 | 15 |
| 2.5 | 23.73 | 142000 | Petrol | Individual | Automatic | 3 | 14 |
| 35 | 92.6 | 78000 | Diesel | Dealer | Manual | 0 | 10 |
| 5.9 | 13.74 | 56000 | Petrol | Dealer | Manual | 0 | 8 |
| 3.45 | 6.05 | 47000 | Petrol | Dealer | Manual | 0 | 7 |
| 4.75 | 6.76 | 40000 | Petrol | Dealer | Manual | 0 | 6 |
| 3.8 | 18.61 | 62000 | Petrol | Dealer | Manual | 0 | 11 |
| 11.25 | 16.09 | 58242 | Diesel | Dealer | Manual | 0 | 6 |
| 3.51 | 13.7 | 75000 | Petrol | Dealer | Manual | 0 | 15 |
| 23 | 30.61 | 40000 | Diesel | Dealer | Automatic | 0 | 5 |
| 4 | 22.78 | 89000 | Petrol | Dealer | Automatic | 0 | 12 |
| 5.85 | 18.61 | 72000 | Petrol | Dealer | Manual | 0 | 8 |
| 20.75 | 25.39 | 29000 | Diesel | Dealer | Automatic | 0 | 4 |
| 17 | 18.64 | 8700 | Petrol | Dealer | Manual | 0 | 3 |
| 7.05 | 18.61 | 45000 | Petrol | Dealer | Manual | 0 | 7 |
| 9.65 | 20.45 | 50024 | Diesel | Dealer | Manual | 0 | 10 |
| 1.75 | 1.9 | 3000 | Petrol | Individual | Manual | 0 | 4 |
| 1.7 | 1.82 | 1400 | Petrol | Individual | Manual | 0 | 3 |
| 1.65 | 1.78 | 4000 | Petrol | Individual | Manual | 0 | 3 |
| 1.45 | 1.6 | 1200 | Petrol | Individual | Manual | 0 | 3 |
| 1.35 | 1.47 | 4100 | Petrol | Individual | Manual | 0 | 3 |
| 1.35 | 2.37 | 21700 | Petrol | Individual | Manual | 0 | 5 |
| 1.35 | 3.45 | 16500 | Petrol | Individual | Manual | 1 | 6 |
| 1.25 | 1.5 | 15000 | Petrol | Individual | Manual | 0 | 7 |
| 1.2 | 1.5 | 18000 | Petrol | Individual | Manual | 0 | 4 |
| 1.2 | 1.47 | 11000 | Petrol | Individual | Manual | 0 | 3 |
| 1.2 | 1.78 | 6000 | Petrol | Individual | Manual | 0 | 4 |
| 1.15 | 1.5 | 8700 | Petrol | Individual | Manual | 0 | 4 |
| 1.15 | 2.4 | 7000 | Petrol | Individual | Manual | 0 | 6 |
| 1.15 | 1.4 | 35000 | Petrol | Individual | Manual | 0 | 4 |
| 1.15 | 1.47 | 17000 | Petrol | Individual | Manual | 0 | 5 |
| 1.11 | 1.47 | 17500 | Petrol | Individual | Manual | 0 | 5 |
| 1.1 | 1.47 | 33000 | Petrol | Individual | Manual | 0 | 7 |
| 1.1 | 1.9 | 14000 | Petrol | Individual | Manual | 0 | 5 |
| 1.1 | 1.47 | 26000 | Petrol | Individual | Manual | 0 | 5 |
| 1.05 | 1.9 | 5400 | Petrol | Individual | Manual | 0 | 7 |
| 1.05 | 1.26 | 5700 | Petrol | Individual | Manual | 0 | 4 |
| 1.05 | 1.5 | 6900 | Petrol | Individual | Manual | 0 | 9 |
| 1.05 | 1.17 | 6000 | Petrol | Individual | Manual | 0 | 4 |
| 1 | 1.47 | 46500 | Petrol | Individual | Manual | 0 | 7 |
| 0.95 | 1.75 | 11500 | Petrol | Individual | Manual | 0 | 8 |
| 0.9 | 1.75 | 40000 | Petrol | Individual | Manual | 0 | 11 |
| 0.9 | 0.95 | 1300 | Petrol | Individual | Manual | 0 | 3 |
| 0.75 | 0.8 | 7000 | Petrol | Individual | Manual | 0 | 4 |
| 0.8 | 0.87 | 3000 | Petrol | Individual | Manual | 0 | 3 |
| 0.78 | 0.84 | 5000 | Petrol | Individual | Manual | 0 | 3 |
| 0.75 | 0.87 | 11000 | Petrol | Individual | Manual | 0 | 3 |
| 0.75 | 0.82 | 18000 | Petrol | Individual | Manual | 0 | 5 |
| 0.75 | 0.95 | 3500 | Petrol | Individual | Manual | 0 | 3 |
| 0.72 | 0.95 | 500 | Petrol | Individual | Manual | 0 | 4 |
| 0.65 | 0.81 | 11800 | Petrol | Individual | Manual | 0 | 3 |
| 0.65 | 0.74 | 5000 | Petrol | Individual | Manual | 0 | 5 |
| 0.65 | 1.2 | 23500 | Petrol | Individual | Manual | 0 | 6 |
| 0.65 | 0.787 | 16000 | Petrol | Individual | Manual | 0 | 7 |
| 0.6 | 0.87 | 15000 | Petrol | Individual | Manual | 0 | 4 |
| 0.6 | 0.95 | 16600 | Petrol | Individual | Manual | 0 | 5 |
| 0.6 | 1.2 | 32000 | Petrol | Individual | Manual | 0 | 7 |
| 0.6 | 0.8 | 20000 | Petrol | Individual | Manual | 0 | 4 |
| 0.6 | 0.84 | 29000 | Petrol | Individual | Manual | 0 | 5 |
| 0.6 | 0.84 | 25000 | Petrol | Individual | Manual | 0 | 4 |
| 0.6 | 0.99 | 25000 | Petrol | Individual | Manual | 0 | 6 |
| 0.6 | 0.81 | 19000 | Petrol | Individual | Manual | 0 | 8 |
| 0.55 | 0.787 | 15000 | Petrol | Individual | Manual | 0 | 6 |
| 0.55 | 0.84 | 58000 | Petrol | Individual | Manual | 0 | 5 |
| 0.52 | 0.94 | 45000 | Petrol | Individual | Manual | 0 | 10 |
| 0.51 | 0.94 | 24000 | Petrol | Individual | Manual | 0 | 4 |
| 0.5 | 0.826 | 6000 | Petrol | Individual | Manual | 0 | 9 |
| 0.5 | 0.55 | 31000 | Petrol | Individual | Manual | 0 | 4 |
| 0.5 | 0.99 | 13000 | Petrol | Individual | Manual | 0 | 8 |
| 0.5 | 0.99 | 45000 | Petrol | Individual | Manual | 0 | 7 |
| 0.5 | 0.88 | 8000 | Petrol | Individual | Manual | 0 | 6 |
| 0.48 | 0.51 | 4300 | Petrol | Individual | Automatic | 0 | 3 |
| 0.48 | 0.52 | 15000 | Petrol | Individual | Manual | 0 | 3 |
| 0.48 | 0.84 | 23000 | Petrol | Individual | Manual | 0 | 5 |
| 0.48 | 0.54 | 8600 | Petrol | Individual | Manual | 0 | 3 |
| 0.45 | 0.51 | 4000 | Petrol | Individual | Automatic | 0 | 3 |
| 0.45 | 0.95 | 24000 | Petrol | Individual | Manual | 0 | 9 |
| 0.45 | 0.826 | 23000 | Petrol | Individual | Manual | 0 | 6 |
| 0.45 | 0.99 | 14500 | Petrol | Individual | Manual | 0 | 8 |
| 0.45 | 0.95 | 27000 | Petrol | Individual | Manual | 0 | 10 |
| 0.45 | 0.54 | 14000 | Petrol | Individual | Manual | 0 | 4 |
| 0.45 | 0.54 | 500 | Petrol | Individual | Automatic | 0 | 4 |
| 0.45 | 0.55 | 1000 | Petrol | Individual | Manual | 0 | 4 |
| 0.42 | 0.81 | 42000 | Petrol | Individual | Manual | 0 | 6 |
| 0.42 | 0.73 | 12000 | Petrol | Individual | Manual | 0 | 7 |
| 0.4 | 0.54 | 14000 | Petrol | Individual | Manual | 0 | 5 |
| 0.4 | 0.83 | 5500 | Petrol | Individual | Manual | 0 | 8 |
| 0.4 | 0.55 | 6700 | Petrol | Individual | Manual | 0 | 5 |
| 0.4 | 0.64 | 13700 | Petrol | Individual | Manual | 0 | 6 |
| 0.4 | 0.51 | 1300 | Petrol | Individual | Automatic | 0 | 3 |
| 0.38 | 0.72 | 38600 | Petrol | Individual | Manual | 0 | 5 |
| 0.38 | 0.787 | 75000 | Petrol | Individual | Manual | 0 | 9 |
| 0.35 | 1.05 | 30000 | Petrol | Individual | Manual | 0 | 9 |
| 0.35 | 0.57 | 24000 | Petrol | Individual | Automatic | 0 | 4 |
| 0.35 | 0.52 | 19000 | Petrol | Individual | Automatic | 0 | 6 |
| 0.31 | 1.05 | 213000 | Petrol | Individual | Manual | 0 | 10 |
| 0.3 | 0.51 | 60000 | Petrol | Individual | Manual | 0 | 8 |
| 0.3 | 0.48 | 50000 | Petrol | Individual | Manual | 0 | 4 |
| 0.3 | 0.58 | 30000 | Petrol | Individual | Manual | 0 | 7 |
| 0.27 | 0.47 | 21000 | Petrol | Individual | Manual | 0 | 7 |
| 0.25 | 0.75 | 26000 | Petrol | Individual | Manual | 1 | 12 |
| 0.25 | 0.58 | 1900 | Petrol | Individual | Automatic | 0 | 12 |
| 0.25 | 0.52 | 22000 | Petrol | Individual | Automatic | 0 | 10 |
| 0.25 | 0.51 | 32000 | Petrol | Individual | Manual | 0 | 7 |
| 0.25 | 0.57 | 18000 | Petrol | Individual | Manual | 0 | 7 |
| 0.2 | 0.57 | 55000 | Petrol | Individual | Manual | 0 | 15 |
| 0.2 | 0.75 | 60000 | Petrol | Individual | Manual | 0 | 12 |
| 0.2 | 0.57 | 25000 | Petrol | Individual | Manual | 1 | 8 |
| 0.2 | 0.75 | 49000 | Petrol | Individual | Manual | 1 | 13 |
| 0.2 | 0.65 | 24000 | Petrol | Individual | Manual | 1 | 7 |
| 0.2 | 0.787 | 50000 | Petrol | Individual | Manual | 0 | 12 |
| 0.18 | 0.32 | 35000 | Petrol | Individual | Manual | 0 | 5 |
| 0.17 | 0.52 | 500000 | Petrol | Individual | Automatic | 0 | 12 |
| 0.16 | 0.51 | 33000 | Petrol | Individual | Manual | 0 | 10 |
| 0.15 | 0.57 | 35000 | Petrol | Individual | Manual | 1 | 9 |
| 0.12 | 0.58 | 53000 | Petrol | Individual | Manual | 0 | 13 |
| 0.1 | 0.75 | 92233 | Petrol | Individual | Manual | 0 | 14 |
| 3.25 | 6.79 | 58000 | Diesel | Dealer | Manual | 1 | 10 |
| 4.4 | 5.7 | 28200 | Petrol | Dealer | Manual | 0 | 5 |
| 2.95 | 4.6 | 53460 | Petrol | Dealer | Manual | 0 | 9 |
| 2.75 | 4.43 | 28282 | Petrol | Dealer | Manual | 0 | 5 |
| 5.25 | 5.7 | 3493 | Petrol | Dealer | Manual | 1 | 4 |
| 5.75 | 7.13 | 12479 | Petrol | Dealer | Manual | 0 | 3 |
| 5.15 | 5.7 | 34797 | Petrol | Dealer | Automatic | 0 | 5 |
| 7.9 | 8.1 | 3435 | Petrol | Dealer | Manual | 0 | 3 |
| 4.85 | 5.7 | 21125 | Diesel | Dealer | Manual | 0 | 5 |
| 3.1 | 4.6 | 35775 | Petrol | Dealer | Manual | 0 | 8 |
| 11.75 | 14.79 | 43535 | Diesel | Dealer | Manual | 0 | 5 |
| 11.25 | 13.6 | 22671 | Petrol | Dealer | Manual | 0 | 4 |
| 2.9 | 6.79 | 31604 | Petrol | Dealer | Manual | 0 | 9 |
| 5.25 | 5.7 | 20114 | Petrol | Dealer | Manual | 0 | 3 |
| 4.5 | 9.4 | 36100 | Petrol | Dealer | Manual | 0 | 8 |
| 2.9 | 4.43 | 12500 | Petrol | Dealer | Manual | 0 | 4 |
| 3.15 | 4.43 | 15000 | Petrol | Dealer | Manual | 0 | 4 |
| 6.45 | 9.4 | 45078 | Petrol | Dealer | Manual | 0 | 6 |
| 4.5 | 9.4 | 36000 | Petrol | Dealer | Manual | 0 | 8 |
| 3.5 | 4.43 | 38488 | Petrol | Dealer | Manual | 0 | 3 |
| 4.5 | 6.79 | 32000 | Petrol | Dealer | Automatic | 0 | 7 |
| 6 | 7.6 | 77632 | Diesel | Dealer | Manual | 0 | 6 |
| 8.25 | 9.4 | 61381 | Diesel | Dealer | Manual | 0 | 5 |
| 5.11 | 9.4 | 36198 | Petrol | Dealer | Automatic | 0 | 7 |
| 2.7 | 4.6 | 22517 | Petrol | Dealer | Manual | 0 | 9 |
| 5.25 | 5.7 | 24678 | Petrol | Dealer | Manual | 0 | 5 |
| 2.55 | 4.43 | 57000 | Petrol | Dealer | Manual | 0 | 9 |
| 4.95 | 9.4 | 60000 | Diesel | Dealer | Manual | 0 | 8 |
| 3.1 | 6.79 | 52132 | Diesel | Dealer | Manual | 0 | 8 |
| 6.15 | 9.4 | 45000 | Diesel | Dealer | Manual | 0 | 7 |
| 9.25 | 9.4 | 15001 | Petrol | Dealer | Manual | 0 | 3 |
| 11.45 | 14.79 | 12900 | Petrol | Dealer | Automatic | 0 | 5 |
| 3.9 | 5.7 | 53000 | Diesel | Dealer | Manual | 0 | 7 |
| 5.5 | 5.7 | 4492 | Petrol | Dealer | Manual | 0 | 5 |
| 9.1 | 9.4 | 15141 | Petrol | Dealer | Manual | 0 | 3 |
| 3.1 | 4.43 | 11849 | Petrol | Dealer | Manual | 0 | 4 |
| 11.25 | 13.6 | 68000 | Diesel | Dealer | Manual | 0 | 5 |
| 4.8 | 9.4 | 60241 | Petrol | Dealer | Manual | 0 | 7 |
| 2 | 4.43 | 23709 | Petrol | Dealer | Manual | 0 | 8 |
| 5.35 | 9.4 | 32322 | Diesel | Dealer | Manual | 0 | 8 |
| 4.75 | 7.13 | 35866 | Petrol | Dealer | Manual | 1 | 5 |
| 4.4 | 7.13 | 34000 | Petrol | Dealer | Manual | 0 | 6 |
| 6.25 | 7.6 | 7000 | Petrol | Dealer | Manual | 0 | 4 |
| 5.95 | 9.4 | 49000 | Diesel | Dealer | Manual | 0 | 7 |
| 5.2 | 9.4 | 71000 | Diesel | Dealer | Manual | 0 | 8 |
| 3.75 | 6.79 | 35000 | Petrol | Dealer | Manual | 0 | 8 |
| 5.95 | 9.4 | 36000 | Petrol | Dealer | Manual | 0 | 5 |
| 4 | 4.6 | 30000 | Petrol | Dealer | Manual | 0 | 7 |
| 5.25 | 7.6 | 17000 | Petrol | Dealer | Manual | 0 | 4 |
| 12.9 | 13.6 | 35934 | Diesel | Dealer | Manual | 0 | 4 |
| 5 | 9.9 | 56701 | Petrol | Dealer | Manual | 0 | 7 |
| 5.4 | 6.82 | 31427 | Petrol | Dealer | Automatic | 0 | 5 |
| 7.2 | 9.9 | 48000 | Diesel | Dealer | Manual | 0 | 6 |
| 5.25 | 9.9 | 54242 | Petrol | Dealer | Manual | 0 | 7 |
| 3 | 5.35 | 53675 | Petrol | Dealer | Manual | 0 | 8 |
| 10.25 | 13.6 | 49562 | Petrol | Dealer | Manual | 0 | 4 |
| 8.5 | 13.6 | 40324 | Petrol | Dealer | Manual | 0 | 5 |
| 8.4 | 13.6 | 25000 | Petrol | Dealer | Manual | 0 | 5 |
| 3.9 | 7 | 36054 | Petrol | Dealer | Manual | 0 | 6 |
| 9.15 | 13.6 | 29223 | Petrol | Dealer | Manual | 0 | 4 |
| 5.5 | 5.97 | 5600 | Petrol | Dealer | Manual | 0 | 4 |
| 4 | 5.8 | 40023 | Petrol | Dealer | Manual | 0 | 5 |
| 6.6 | 7.7 | 16002 | Petrol | Dealer | Manual | 0 | 4 |
| 4 | 7 | 40026 | Petrol | Dealer | Manual | 0 | 5 |
| 6.5 | 8.7 | 21200 | Petrol | Dealer | Manual | 0 | 3 |
| 3.65 | 7 | 35000 | Petrol | Dealer | Manual | 0 | 6 |
| 8.35 | 9.4 | 19434 | Diesel | Dealer | Manual | 0 | 4 |
| 4.8 | 5.8 | 19000 | Petrol | Dealer | Manual | 0 | 3 |
| 6.7 | 10 | 18828 | Petrol | Dealer | Manual | 0 | 5 |
| 4.1 | 10 | 69341 | Petrol | Dealer | Manual | 0 | 9 |
| 3 | 10 | 69562 | Petrol | Dealer | Manual | 0 | 11 |
| 7.5 | 10 | 27600 | Petrol | Dealer | Manual | 0 | 5 |
| 2.25 | 7.5 | 61203 | Petrol | Dealer | Manual | 0 | 10 |
| 5.3 | 6.8 | 16500 | Petrol | Dealer | Manual | 0 | 6 |
| 10.9 | 13.6 | 30753 | Petrol | Dealer | Automatic | 0 | 4 |
| 8.65 | 13.6 | 24800 | Petrol | Dealer | Manual | 0 | 5 |
| 9.7 | 13.6 | 21780 | Petrol | Dealer | Manual | 0 | 5 |
| 6 | 8.4 | 4000 | Petrol | Dealer | Manual | 0 | 4 |
| 6.25 | 13.6 | 40126 | Petrol | Dealer | Manual | 0 | 6 |
| 5.25 | 5.9 | 14465 | Petrol | Dealer | Manual | 0 | 5 |
| 2.1 | 7.6 | 50456 | Petrol | Dealer | Manual | 0 | 14 |
| 8.25 | 14 | 63000 | Diesel | Dealer | Manual | 0 | 6 |
| 8.99 | 11.8 | 9010 | Petrol | Dealer | Manual | 0 | 4 |
| 3.5 | 5.9 | 9800 | Petrol | Dealer | Manual | 0 | 7 |
| 7.4 | 8.5 | 15059 | Petrol | Dealer | Automatic | 0 | 4 |
| 5.65 | 7.9 | 28569 | Petrol | Dealer | Manual | 0 | 4 |
| 5.75 | 7.5 | 44000 | Petrol | Dealer | Automatic | 0 | 5 |
| 8.4 | 13.6 | 34000 | Petrol | Dealer | Manual | 0 | 5 |
| 10.11 | 13.6 | 10980 | Petrol | Dealer | Manual | 0 | 4 |
| 4.5 | 6.4 | 19000 | Petrol | Dealer | Manual | 0 | 6 |
| 5.4 | 6.1 | 31427 | Petrol | Dealer | Manual | 0 | 5 |
| 6.4 | 8.4 | 12000 | Petrol | Dealer | Manual | 0 | 4 |
| 3.25 | 9.9 | 38000 | Petrol | Dealer | Manual | 0 | 10 |
| 3.75 | 6.8 | 33019 | Petrol | Dealer | Manual | 0 | 6 |
| 8.55 | 13.09 | 60076 | Diesel | Dealer | Manual | 0 | 5 |
| 9.5 | 11.6 | 33988 | Diesel | Dealer | Manual | 0 | 4 |
| 4 | 5.9 | 60000 | Petrol | Dealer | Manual | 0 | 5 |
| 3.35 | 11 | 87934 | Petrol | Dealer | Manual | 0 | 11 |
| 11.5 | 12.5 | 9000 | Diesel | Dealer | Manual | 0 | 3 |
| 5.3 | 5.9 | 5464 | Petrol | Dealer | Manual | 0 | 4 |
| 4.34 | 5.12 | 37000 | Diesel | Dealer | Manual | 0 | 2 |

1. *Linear Regression Results*

Call:

lm(formula = Present\_Price ~ ., data = car)

Residuals:

Min 1Q Median 3Q Max

-13.950 -1.472 0.088 1.099 32.989

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.428e+00 2.607e+00 -1.699 0.0904 .

Selling\_Price 1.642e+00 5.971e-02 27.491 < 2e-16 \*\*\*

Kms\_Driven 1.346e-05 6.257e-06 2.150 0.0323 \*

Fuel\_TypeDiesel -1.719e+00 2.489e+00 -0.691 0.4903

Fuel\_TypePetrol -5.729e-01 2.440e+00 -0.235 0.8145

Seller\_TypeIndividual -1.579e-01 5.122e-01 -0.308 0.7580

TransmissionManual 3.373e-01 6.559e-01 0.514 0.6075

Owner 1.557e+00 8.175e-01 1.905 0.0578 .

Age 6.917e-01 8.462e-02 8.174 9.1e-15 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.421 on 293 degrees of freedom

Multiple R-squared: 0.8471, Adjusted R-squared: 0.8429

F-statistic: 202.9 on 8 and 293 DF, p-value: < 2.2e-16

1. *R-Markdown*

> ###1. DATA EXPLORATION

> #Load the dataset into R

> cardata<-read.csv("C:/Users/user/AppData/Local/Temp/Temp1\_archive.zip/car data.csv")

> #View the first 6 rows of the dataset

> head(cardata)

> #Get dimension

> dim(cardata)

> #Add a data point

> card<- rbind(cardata, "302" = c("city", "2018", "4.34", "5.12", "37000", "Diesel", "Dealer", "Manual", "0"))

> #Get the summary of the dataset

> summary(card)

> #View its dimension

> dim(card)

> #Check for missingness in the data

> colSums(is.na(card))

> #Remove the car name column

> car=card[, -1]

> #View the dataset

> View(car)

> #Check its structure

> str(card)

> #Prepare the dataset

> car$Year<-as.numeric(car$Year)

> car$Selling\_Price<-as.numeric(car$Selling\_Price)

> car$Present\_Price<-as.numeric(car$Present\_Price)

> car$Kms\_Driven<-as.numeric(car$Kms\_Driven)

> car$Owner<-as.numeric(car$Owner)

> #Check the structure again

> str(car)

> #Obtain the age of the vehicles

> car$Present\_Year <- 2020

> car$Age <- car$Present\_Year - car$Year

> #Drop the year and present year columns

> car<-subset(car,select = -c(Year,Present\_Year))

> #View the column names of the final dataset

> colnames(car)

> #Visualizations

> library(ggplot2)

> ggplot(car, aes(x=Age, y=Present\_Price))+geom\_point()+geom\_smooth(method = lm)

> hist(car$Selling\_Price,col="yellow", main="Histogram of Selling Price")

> hist(car$Kms\_Driven,col="green", main="Histogram of Kilometres Driven")

> plot(car$Fuel\_Type, col="blue", main="Plot of Fuel type")

> plot(car$Seller\_Type, col="red", main="Plot of Type of seller")

> plot(car$Transmission, col="gray", main="Plot of Transmission")

> hist(car$Owner, col="dodgerblue", main="Histogram of Number of Previous Owners")

> ###2. REGRESSION AND PREDICTION

> #Predicting price

> fit<-lm(Present\_Price~., data=car)

> #Get the model's summary

> summary(fit)

> #Check for variables importance

> library(caret)

> varImp(fit)

> #Partition the dataset

> #Setting seed to reproduce results of random sampling

> set.seed(100)

> trainingRowIndex <- sample(1:nrow(car), 0.8\*nrow(car))

> trainingData <- car[trainingRowIndex, ]

> testData <- car[-trainingRowIndex, ]

> #Linear Regression algorithm

> train.control <- trainControl(method = "cv", number = 10) # k-folds CV with k=10

> train.control2 <- trainControl(method = "LOOCV") # leave-one-out CV

> linear.model <- train(Present\_Price ~., data = trainingData,method = "lm", trControl = train.control, preProc = c("center"))

> linear.model

> linear.model$results

> #Predict values in the testing set

> linear.predict <- predict(linear.model, testData)

> #The accuracy of the model

> postResample(linear.predict, testData$Present\_Price)

> ##Support Vector Machine Linear modeling

> # k-folds CV with k=10

> train.control <- trainControl(method = "cv", number = 10)

> #Leave-one-out CV

> train.control2 <- trainControl(method = "LOOCV")

> #Inspect a specific ML algorithm

> svmL.info <- getModelInfo("svmLinear")

> #Look at the algorithm parameters that can be modified

> svmL.info$svmLinear$parameters

> #Specify values that you want to feed into the model training function

> tune.grid <- expand.grid(C = c(0.001, 0.01, 0.1, 1, 10, 100))

> svmL.model <- train(Present\_Price ~., data = trainingData, method = "svmLinear", trControl = train.control, tuneGrid = tune.grid, preProc = c("center"))

> svmL.model

> svmL.model$results

> #Testing predictive ability of model in test data

> svmL.predict <- predict(svmL.model, testData)

> postResample(svmL.predict, testData$Present\_Price)

> #Random Forest

> train.control <- trainControl(method = "cv", number = 10) # k-folds CV with k=10

> train.control2 <- trainControl(method = "LOOCV") # leave-one-out CV

> #Model tuning

> rf.info <- getModelInfo("rf")

> rf.info$rf$parameters

> tune.grid <- expand.grid(mtry = c(2, 3, 4))

> #Model training

> rf.model <- train(Present\_Price ~., data = trainingData, method = "rf", trControl = train.control, preProc = c("center"))

> rf.model

> rf.model$results

> #Testing predictive ability of model in test data

> rf.predict <- predict(rf.model, testData)

> postResample(rf.predict, testData$Present\_Price)

> ##Model Comparisons

> RMSE\_Training <- c(linear.model$results$RMSE, svmL.model$results$RMSE[1]

+ , rf.model$results$RMSE[1])

> RSq\_Training <- c(linear.model$results$Rsquared, svmL.model$results$Rsquared[1]

+ , rf.model$results$Rsquared[1])

> RMSE\_Testing <- c(postResample(linear.predict, testData$Present\_Price)[1], postResample(svmL.predict, testData$Present\_Price)[1], postResample(rf.predict, testData$Present\_Price)[1])

> RSq\_Testing <- c(postResample(linear.predict, testData$Present\_Price)[2], postResample(svmL.predict, testData$Present\_Price)[2], postResample(rf.predict, testData$Present\_Price)[2])

> Algorithm <- c("Linear Regression", "Support Vector Machine", "Random Forest")

> data.frame(cbind(Algorithm, RMSE\_Training, RSq\_Training, RMSE\_Testing, RSq\_Testing)

> #Saving the final dataset to Excel

> library(writexl)

> write\_xlsx(car,"C:/Users/user/Desktop/my work/car data.xlsx")

> ###END OF CODE