# Prediction of car prices for competitive analysis

Mohammadreza Abedi

## Introduction

A foreign automobile company wants to enter into a new market by setting up a manufacturing plant and producing cars locally. But before beginning with the production, they want to study their competition in the market. In this case, they want to study Toyota Corolla models in detail.

The company specifically wants to know about which variables (predictors) are significant in predicting the price of a car and how well those variables describe the price of the car. Based on secondary research, the consulting firm has gathered the dataset for Toyota Corolla car models.

The company wants to understand exactly how the price of a car varies with the independent variables in the new market. They can accordingly manipulate the car design, the business strategy to meet certain price levels. The model will help the company to understand the pricing dynamics of a new market.

Pricing of a car is dependent on many factors like age, KM driven, horsepower, automatic or manual, no. of car doors, weight in pounds, brand image etc. We are given the task of examining the predictors responsible for the pricing of a Toyota Corolla car.

In this study going to use multiple linear regression model due to its simplicity and comparatively. Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (*x*) and the single output variable (*y*). More specifically, *y* can be calculated from a linear combination of the input variables (*x*).

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}+e$$

When there is a single input variable (*x*), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression. Multiple linear regression model uses Ordinary Least Squares (OLS) to estimate the value of coefficients of independent variables. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares attempt to minimize.

## Model specification

This study using multiple linear regression model. With select, the dependent column data (Price) in the Input *y* Range, select the independent column data (Age, KM, HP, Automatic, CC, Doors, Weight) in the Input *x* Range.

## Results

The study was designed to determine the effects of Age, Distance travelled, Engine power, Transmission type, Engine capacity, no. of car doors and Weight of the used car in determining its present price. Our dataset included 1436 observations with both the dependent and independent factors.

*Performing OLS on the dataset, we were able to establish a strong relationship between the dependent variable (Price) and the independent variable (Age, Distance travelled, engine power, engine capacity and Weight). While the factors like Car transmission type and No. of doors in a car were rejected in the final model due to P-value being greater than 0.05.*

The predicted price of the car can be represented in the equation as:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_p x_{ip}+e$$

**Price of car = -4403.90 - 123.19*Age - 0.01*Distance Travelled (KM) + 30.43*Engine Power (HP) - 1.31*Engine capacity (CC) + 20.81*Weight of car.**

In other words, *for each unit increase in Age, Price decreases with 123.19 units; for each unit increase in KM, Price decreases with 0.01 units; for each unit increase in HP, Price increases with 30.43 units; for each unit increase in CC, Price decreases with 1.31 units and each unit increase in Weight, Price increases with 20.81 units. Price is equal to -4403 when all predictor variables are zero in the model.*

The following equation was derived from the fact that p-value of independent **variables was less than 0.05 (95% statistically significant)** and also from the **F stat 0** ***reflecting that the results are significantly valuable.***

The R-squared signifies the "percentage variation independent that is explained by independent variables". Here, ***85.6% variation*** **in Price** is explained by Age, distance Travelled (KM), Engine Power (HP), Transmission type (Automatic/Manual), Engine capacity (CC), No. of Doors, Weight of car. From the model we can easily confirm that ***Prices of Car tends to be higher if they are less driven, are new, have high engine power, have low engine capacity and are heavy in weight.***

*On average, our model had a forecast error of only* ***9.84%*** which typically denotes an average prediction model and best performing models will have ***lower MAPE values***.

## Discussion & conclusion

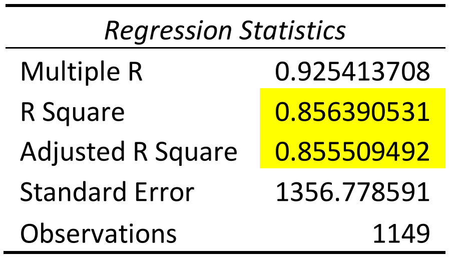
With this study, it purpose was to predict prices of car by using a dataset that has 1436 observations. With the help of the data visualizations and exploratory data analysis, the dataset was uncovered and features were explored deeply. The relation between features were examined.

At the last stage, predictive models were applied to predict price of cars in an order: Multiple linear regression, Train data, test data, error check, EDA and according to EDA, here are the most important features: *Age, Distance travelled, engine power, engine capacity and Weight.*

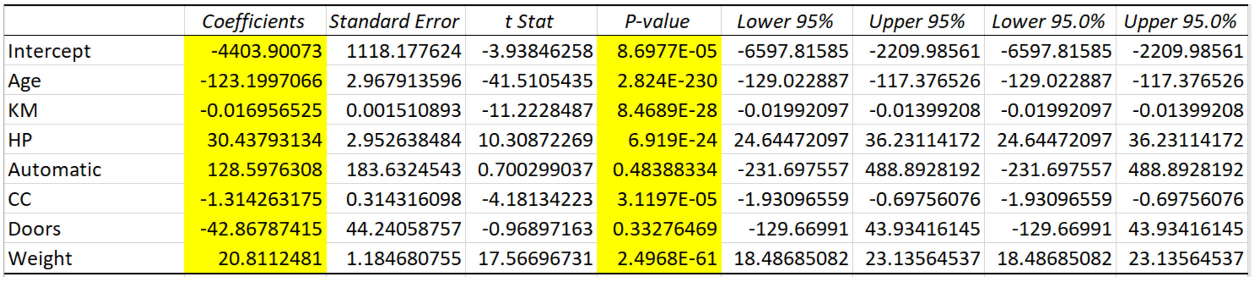
Also, Multiple linear regression is a commonly used predictive model for numerical response variable. This study demonstrates how multiple linear regression analysis can be done using from data processing all the way to testing the generalizablity of the selected model.

The model can be extended to other business use cases like predicting the Car resale value which would be helpful for a car leasing company. At present, our model focused on a specific car make which was Toyota Corolla. This model could be enhanced by including data sets of other car models. During car purchase, not only the car features but brand value also plays into the mind of the consumers and at present, we have not included in our model which leaves us with future scope of improvement.

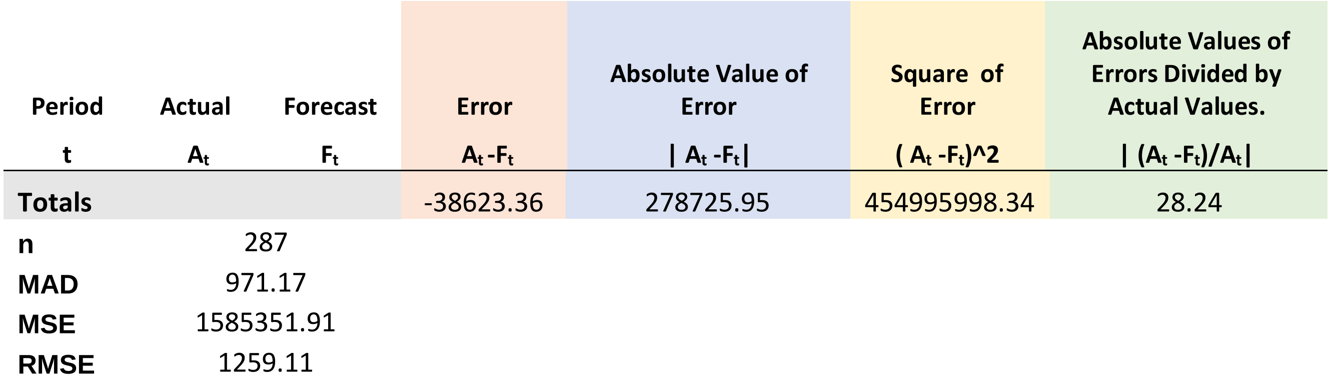
**Appendix:**



| *Regression Statistics* | |
|---|---|
| Multiple R | 0.925413708 |
| R Square | 0.856390531 |
| Adjusted R Square | 0.855509492 |
| Standard Error | 1356.778591 |
| Observations | 1149 |

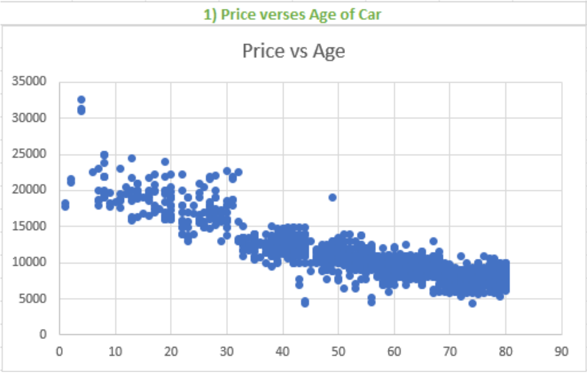


| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -4403.90073 | 1118.177624 | -3.93846258 | 8.6977E-05 | -6597.81585 | -2209.98561 | -6597.81585 | -2209.98561 |
| Age | -123.1997066 | 2.967913596 | -41.5105435 | 2.824E-230 | -129.022887 | -117.376526 | -129.022887 | -117.376526 |
| KM | -0.016956525 | 0.001510893 | -11.2228487 | 8.4689E-28 | -0.01992097 | -0.01399208 | -0.01992097 | -0.01399208 |
| HP | 30.43793134 | 2.952638484 | 10.30872269 | 6.919E-24 | 24.64472097 | 36.23114172 | 24.64472097 | 36.23114172 |
| Automatic | 128.5976308 | 183.6324543 | 0.700299037 | 0.48388334 | -231.697557 | 488.8928192 | -231.697557 | 488.8928192 |
| CC | -1.314263175 | 0.314316098 | -4.18134223 | 3.1197E-05 | -1.93096559 | -0.69756076 | -1.93096559 | -0.69756076 |
| Doors | -42.86787415 | 44.24058757 | -0.96897163 | 0.33276469 | -129.66991 | 43.93416145 | -129.66991 | 43.93416145 |
| Weight | 20.8112481 | 1.184680755 | 17.56696731 | 2.4968E-61 | 18.48685082 | 23.13564537 | 18.48685082 | 23.13564537 |

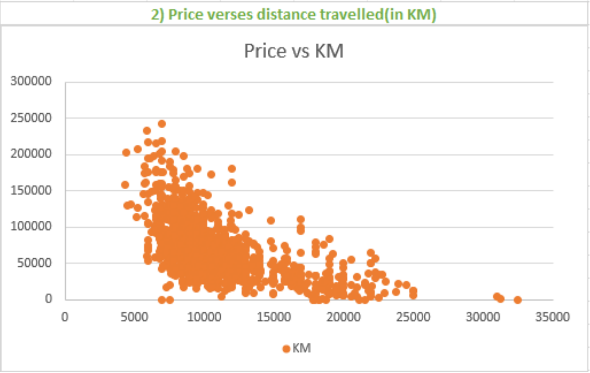


| **Period** | **Actual** | **Forecast** | **Error** | **Absolute Value of Error** | **Square of Error** | **Absolute Values of Errors Divided by Actual Values.** |
|---|---|---|---|---|---|---|
| **t** | **$A_t$** | **$F_t$** | **$A_t - F_t$** | **$\| A_t - F_t \|$** | **$( A_t - F_t)$^2** | **$\| (A_t - F_t)/A_t \|$** |
| **Totals** | | | -38623.36 | 278725.95 | 454995998.34 | 28.24 |
| **n** | 287 | | | | | |
| **MAD** | 971.17 | | | | | |
| **MSE** | 1585351.91 | | | | | |
| **RMSE** | 1259.11 | | | | | |

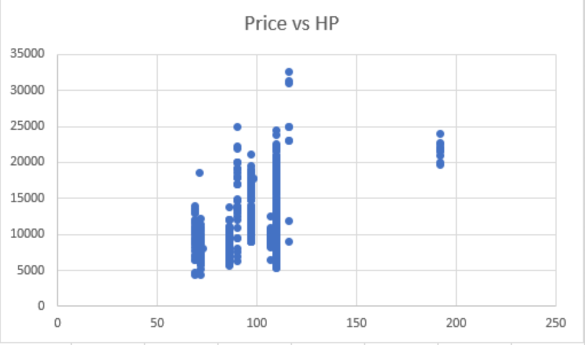
1) Price verses Age of Car



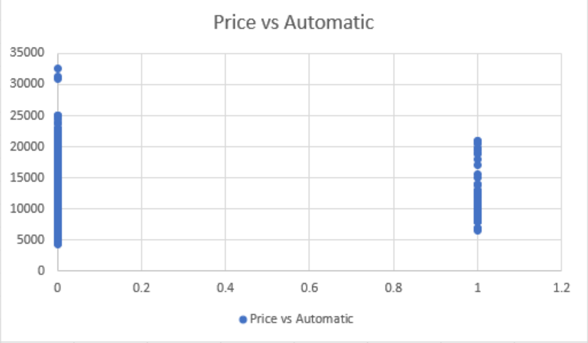
2) Price verses distance travelled(in KM)



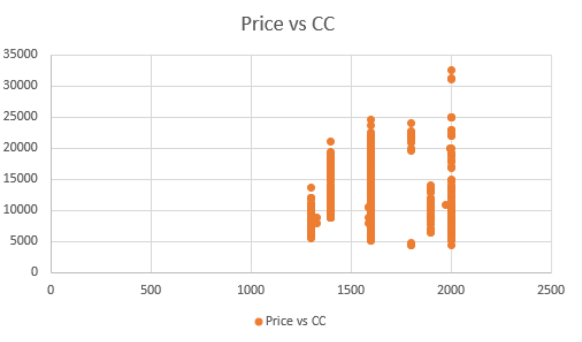
3) Price verses Engine Power (in HorsePower)



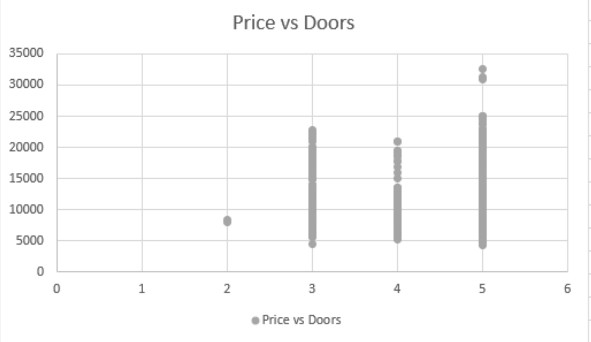
4) Price verses Automatic/Manual cars



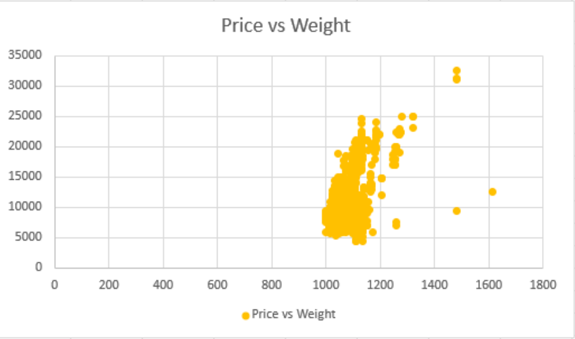
5) Price verses Engine Capacity(in CC)



6) Price verses no. of car doors



7) Price verses car weight



## Reference:

- ToyotaCorolla.csv https://www.kaggle.com/
- Used Car Price Prediction using Machine Learning: Panwar Abhash Anil, towardsdatascience.com
- Model Interpretation: Prediction and Validation of Pre-owned Car Price using Linear Regession Analysi, http://datascienceandme.com/topics/RMultipleLinearRegression.html