# Multiple Linear Regression (Module -7)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shreyas A B**

**Batch Id: 05012021-10AM**

**Topic: Multilinear Regression**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Multi linear regression model and check for VIF, AvPlots, Influence Index Plots.**
   3. **Train and Test the data and compare RMSE values tabulate R-Squared values , RMSE for different models in documentation and provide your explanation on it.**
   4. **Briefly explain the model output in the documentation.**



1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

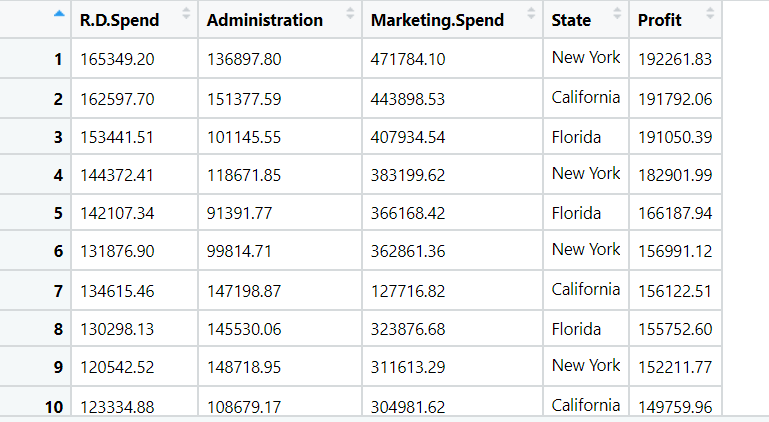
# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

An Analytics Company has been tasked by a crucial job of finding out what factors does affect a startup company and will it be profitable to do so or not. For this they have collected some historical data and would like to apply supervised predictive learning algorithm such as Multilinear regression on it and provide brief insights about their data. Predict Profit, given different attributes for various startup companies.



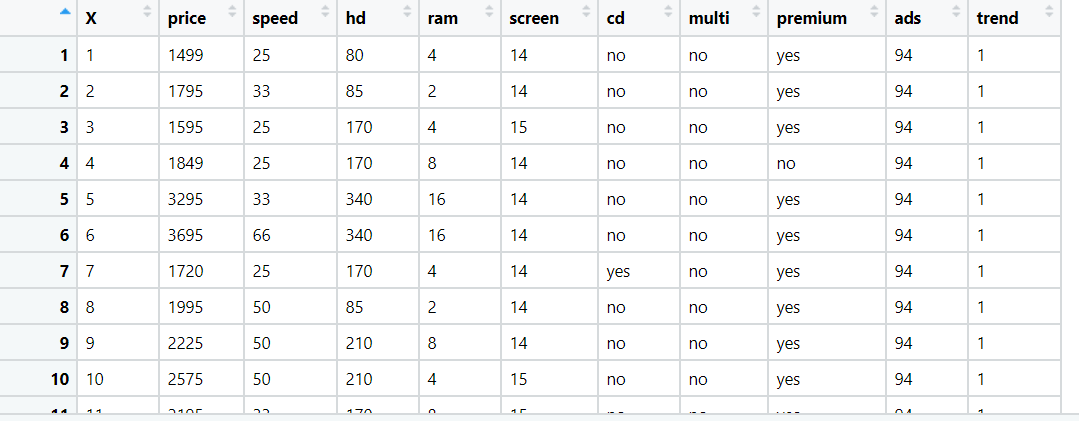
**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable.

**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| R.D.Spend | Research and Development Expense | Continuous, Ratio | Relevant, Provides useful information. |
| Administration | Administration expense | Continuous, Ratio | Relevant, Provides useful information. |
| Marketing.Spend | Marketing expense | Continuous, Ratio | Relevant, Provides useful information. |
| State | Place of company | Nominal | Irrelevant, Doesn’t Provides useful information. |
| Profit | Profit of company | Continuous, Ratio | Relevant, Provides useful information. |

**Problem Statement: -**

Officeworks, is a leading retail store in Australia, with numerous outlets around the country. The manager would like to improve their customer experience by providing them online predictive prices about their gadgets/ Laptops if they wants to sell them. To improve this experience the manager would like us to build a model which is sustainable and accurate enough, to get the objective achieved. Apply multilinear model on the dataset and predict Price, given other attributes and tabulate R squared ,RMSE and correlation values.





**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable.

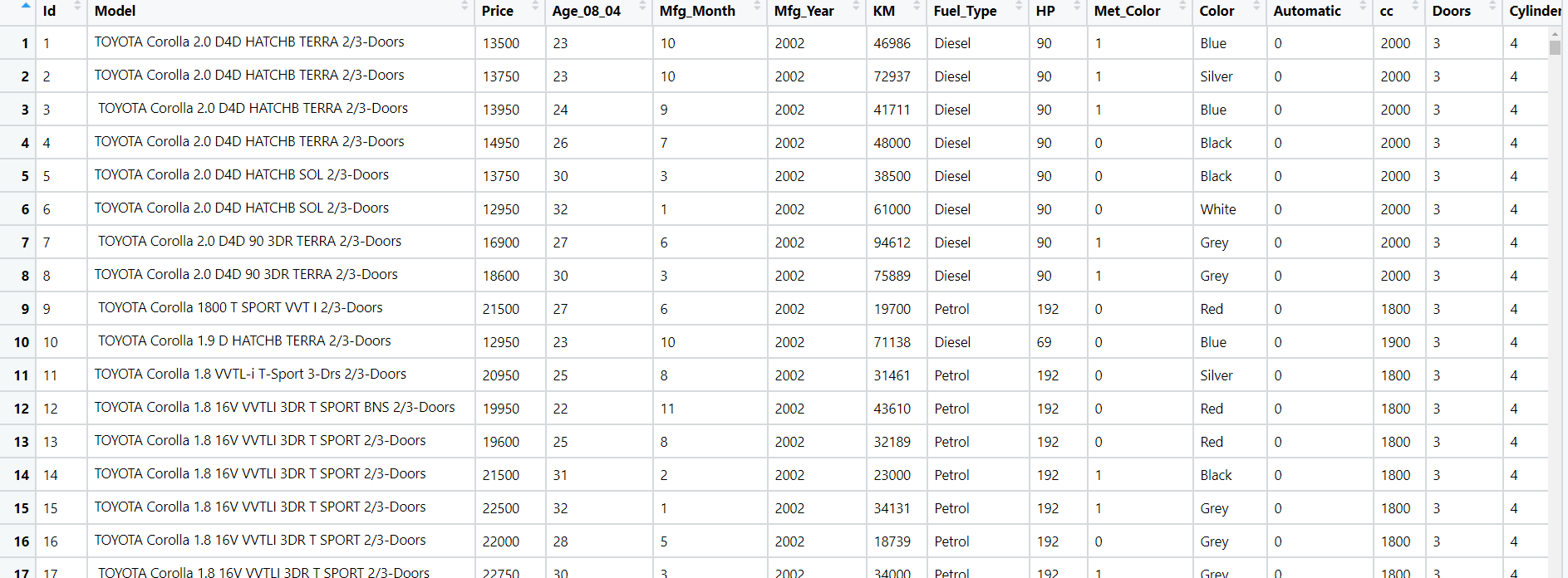
**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| X | Index Values | Nominal | Irrelevant, Doesn’t Provides useful information. |
| price | Price of gadget | Continuous, Ratio | Relevant, Provides useful information. |
| speed | Gadget speed | Continuous, Ratio | Relevant, Provides useful information. |
| hd | Scale of HD | Discrete, count | Relevant, Provides useful information. |
| ram | RAM of device | Discrete, count | Relevant, Provides useful information. |
| screen | Screen size | Discrete, count | Relevant, Provides useful information. |
| cd | CD supported | Categorical, Ordinal | Relevant, Provides useful information. |
| multi | Multimedia supported | Categorical, Ordinal | Relevant, Provides useful information. |
| premium | Premium device | Categorical, Ordinal | Relevant, Provides useful information. |
| ads | ads | Discrete, count | Relevant, Provides useful information. |
| trend | Trend | Discrete, count | Relevant, Provides useful information. |

**Problem Statement: -**

An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like an automated model which can predict the price of the car if user inputs the required factors. Help the business achieve the objective by applying multilinear regression on the given dataset.

Please use the below columns for the analysis purpose.

Price,Age\_08\_04, KM,HP,cc,Doors,Gears,Quarterly\_Tax,Weight



**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable

**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Price | Price of the car | Continuous, Ratio | It provides useful information |
| Age\_08\_04 | Age of a person | Continuous, Ratio | It is not useful |
| KM | Total Travelled Distance | Continuous, Ratio | It provides useful information |
| HP | Horse Power | Continuous, Ratio | It is a useful information |
| Cc | Cubic Capacity | Continuous, Ratio | Useful information |
| Doors | Number of doors in car | Discrete, Count | It is useful information |
| Gears | Number of gears | Discrete, Count | It provides useful information |
| Quarterly Tax | Insurance Tax for car | Continuous, Ratio | It provides useful information |
| Weight | Total Weight of car | Continuous, Ratio | It provides useful information |

**Problem Statement: -**

With the growing consumption of Avacado, in USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and also would like to come up with a prediction model of price for Avocado. For this to be implemented build a prediction model using multilinear regression and provide your insights on it.

Snapshot of the dataset is given below: -

A close up of a piece of paper

Description automatically generated

**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable

**Constraints:** Limited to linear Relationships

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Average Price | Mean Price of Avocado | Continuous, Ratio | It provides useful information |
| Total Volume | Total Volume of avocados | Continuous, Ratio | It provides useful information |
| tot\_ava1 | Terms of trade of avacado1 | Continuous, Ratio | It provides useful information |
| tot\_ava2 | Terms of trade of avacado2 | Continuous, Ratio | It provides useful information |
| tot\_ava3 | Terms of trade of avacado3 | Continuous, Ratio | It provides useful information |
| Total Bags | Total Number of bags required | Discrete, Ordinal | It provides useful information |
| Small Bags | Total Number of small bags required | Discrete, Ordinal | It provides useful information |
| Large Bags | Total Number of large bags required | Discrete, Ordinal | It provides useful information |
| Large Bags | Total Number of large bags required | Discrete, Ordinal | It provides useful information |
| Type | Method of growing plants | Discrete, Nominal | It is not useful |
| Year | Year at which the fruit grown | Continuous, Ratio | It provides useful information |
| Region | Region at which the fruits grown | Discrete, Ordinal | It provides useful information |