# Multiple Linear Regression

**Instructions**

Please share your answers filled in this word document. Submit code files wherever applicable.

Please ensure you update all the details:

**Name:**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name:RAKESH SETHU NP Batch ID: 11052022\_7.30PM**

**Topic: Multilinear Regression**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
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. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

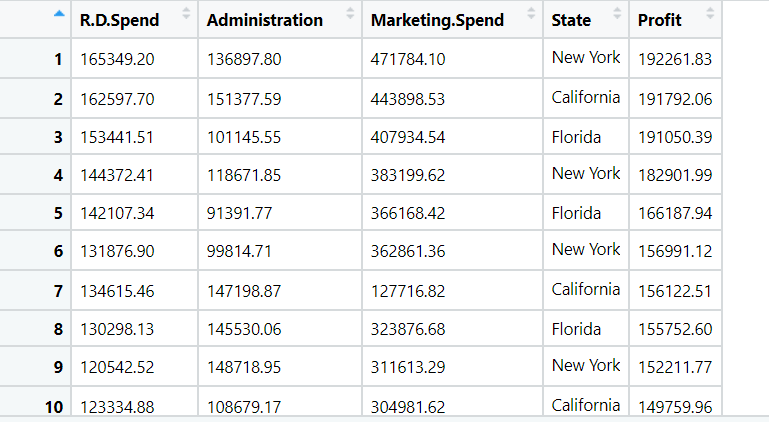
**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Treatment.**

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 and RMSE values for different models in the documentation and provide an explanation.**
   4. **Briefly explain the model output in the documentation.**
   5. **Tune the model and improve its accuracy.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements: -**

1. An analytics company has been tasked with the crucial job of finding out what factors affect a startup company and if it will be profitable or not. For this, they have collected some historical data and would like to apply multilinear regression to derive brief insights into 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. |

1. Perform multilinear regression with price as the output variable and document the different RMSE 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. |

1. 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 once the user inputs the required factors. Help the business achieve their 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, and 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 |

1. With the growing consumption of avocados in the USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and would like to come up with a prediction model for the price of avocados. 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 |