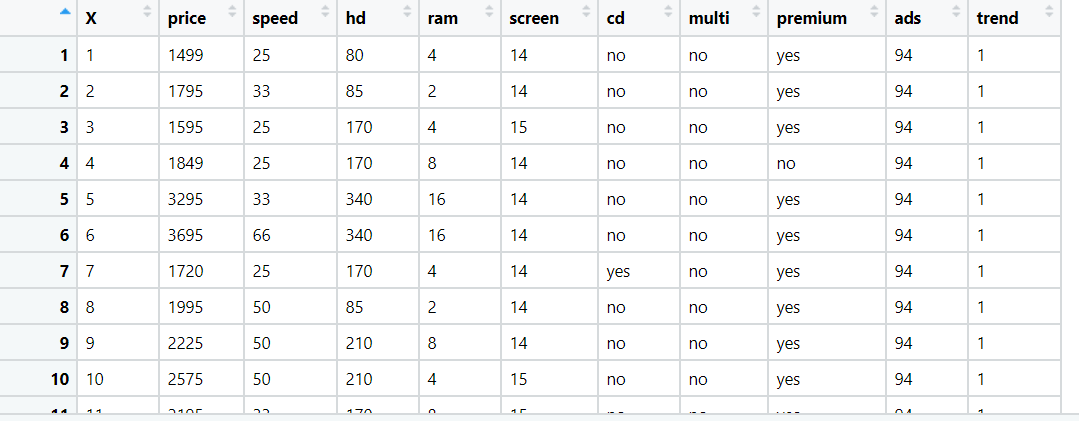
# Multiple Linear Regression (Module -7)

**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.

**Business Objective- build a ML model that predict price** .

**Python code details :**

Data Frame name is computer. It has 6259 entries and 11 features.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

1st column is index so we removed it .Now we check info and describe for df .Check for data types ,unique value and variance.we have done label encoder for “cd”, ”multi”, ”premium” columns.

Then we check for unique value in each columns

:-

price 808

speed 6

hd 59

ram 6

screen 3

cd 2

multi 2

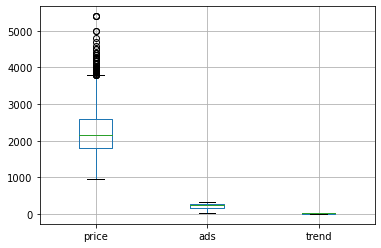
premium 2

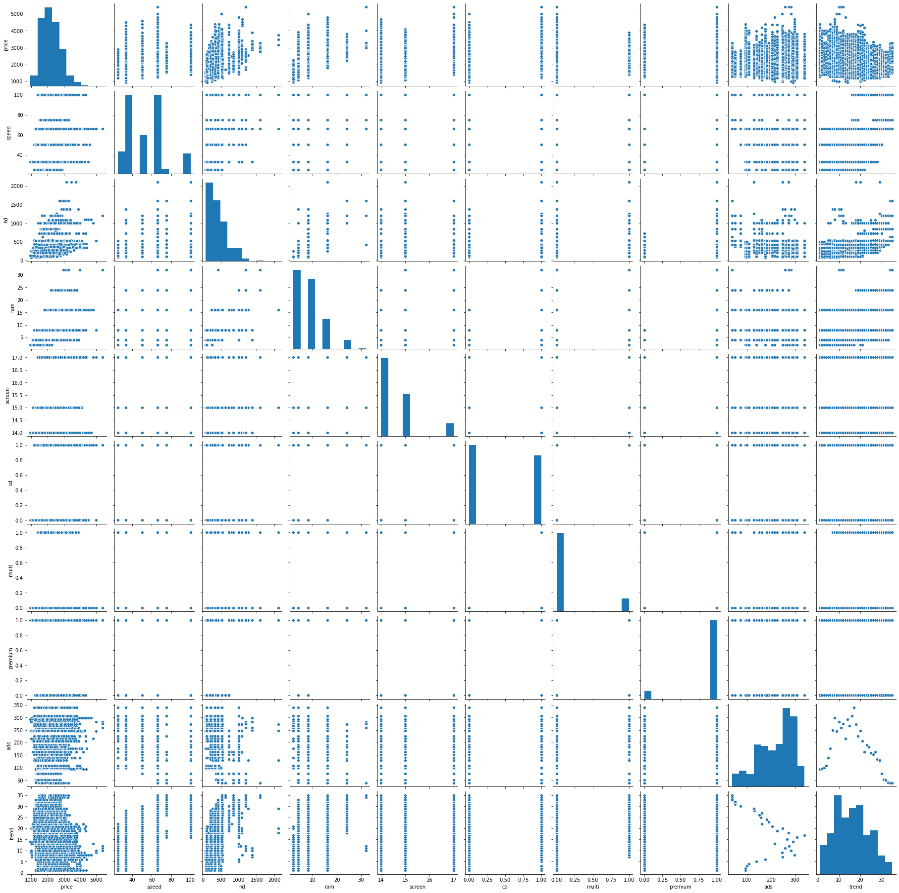
ads 34

trend 35

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .there is no outliers present except Profit.

Boxplot:- 

Histogram and Scatter plot:- 

If we take a look at our Dataset we can clearly see that State is a String type variable and like we have discussed,We cannot feed String type variables into our Machine Learning model as it can only work with numbers.To overcome this problem we use the one hot encoding object and create Dummy Variables using the grt dummy .

Using seaborn (pairplot) in python we can check for distribution and correlation between each other. According scatter plot strong positive correlation between price and ram and also some relation between ram and hd. Correlation save as co dataframe .

**Model Building**

**Build the model on the scaled data (try multiple options)**

**Perform Multi linear regression model and check for VIF, AvPlots, Influence Index Plots.**

**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.**

**Briefly explain the model output in the documentation.**

Using library statsmodels.formula.api as smf # for regression model

**Model name R^2 value Adj. R-squared: Intercept**

**Ml1 0.77 0.77 8.810e+04**

**Ml\_new 0.74 0.74 8.885e+04**

**Ml2 0.70 0.70 8.987e+04**

Model ml1 right model .

and then Checking whether data has any influential values

and So we Check for Colinearity to decide to remove a variable using VIF

**Variables VIF**

**0 hd 4.20**

**1 ram 2.97**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a Simple linear regression on training data set

R-squared: 0.77

Adj. R-squared: 0.77

,then test on test data ,

Evaluation on Test Data as result root mean square error=288

Evaluation on Train Data also as result root mean square error=271

Used library –

pandas for data manipulations

Numpy for Numerical Calculatations

Sklearn for Data mining / Machine learning

Matplotlib for Data visualization

Seaborn for Advance data visualization

Scipy for Advance data visualization

Statsmodels for Regression model