**Multi Linear Regression**

**Problem Statement:** – To predict the price of a computer provided its features are entered by the user.

**Dataset**: – Computer\_Data

**Algorithm Used** – Multi Linear Regression Model

**Assumptions made while preparing a Multi linear regression is that:**

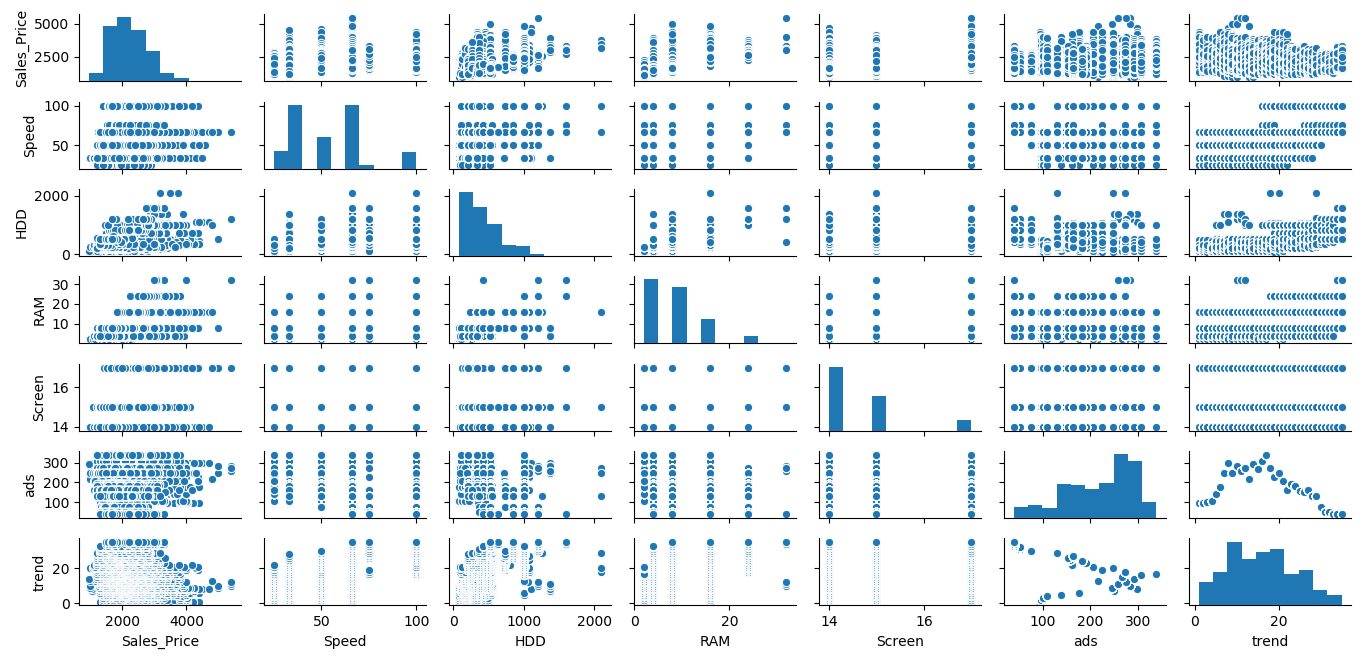
1. All features have a linear relationship with the outcome or output variable (L).
2. Features should not be dependent on each other, if so the phenomenon of multi-collinearity will exist and there by impacting the o/p prediction or estimates.
3. Multiple regression assumes that the residuals are normally distributed.
4. This assumption states that the variance of error terms are similar across the values of the independent variables.

**Feature Description: -**

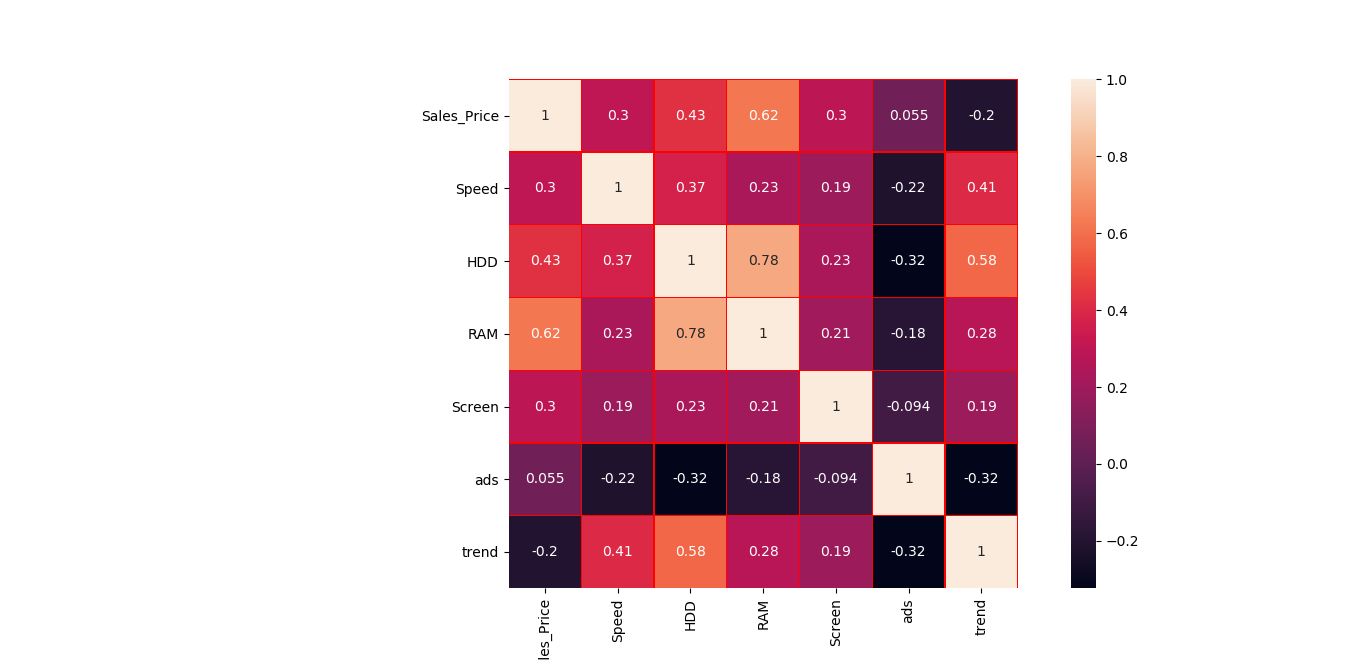
1. Price : price in US dollars of 486 PCs
2. Speed : clock speed in MHz
3. HD : size of hard drive in MB
4. ram : size of Ram in in MB
5. screen : size of screen in inches
6. cd: is a CD-ROM present? (Yes =1, No=0)
7. multi: is a multimedia kit (speakers, sound card) included? (Yes =1, No=0)
8. Premium: is the manufacturer was a "premium" firm (IBM, COMPAQ)? (Yes =1, No=0)
9. ads : number of 486 price listings for each month
10. Trend: time trend indicating month starting from January of 1993 to November of 1995.

**Modelling Process followed –**

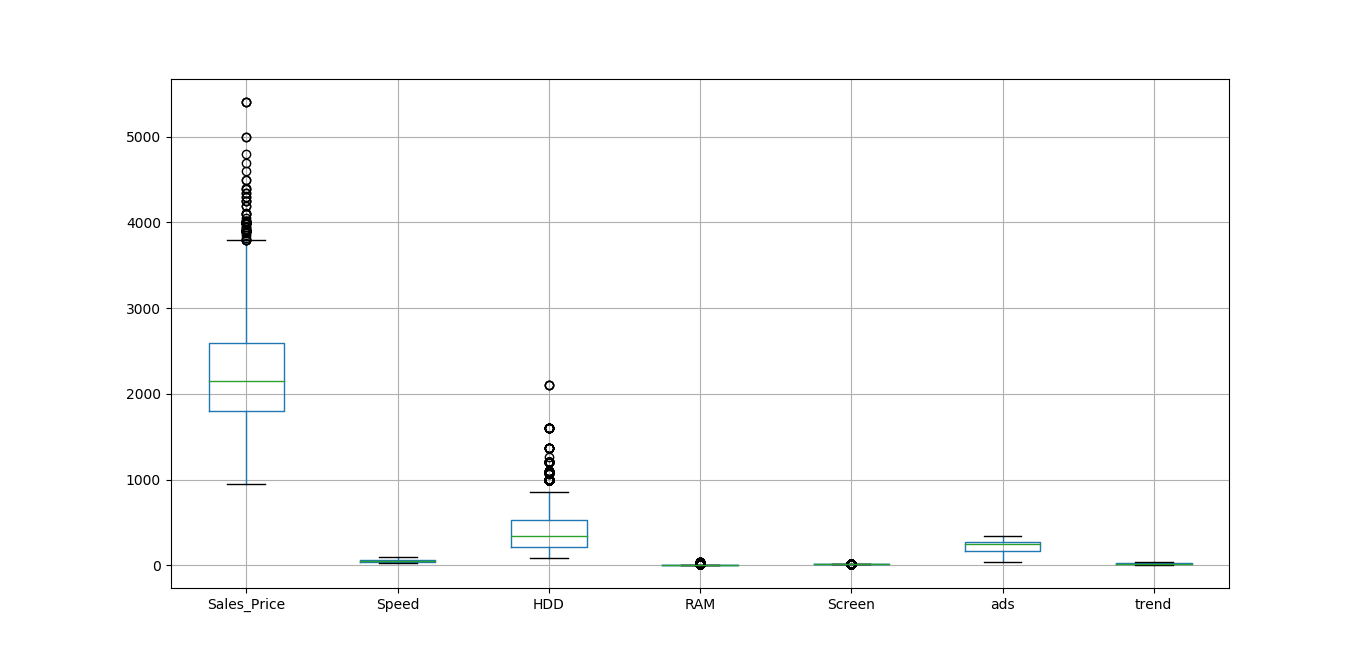
1. Reading the data (using pandas – pd)
2. EDA or Exploratory Data Analysis (Involves finding influencers, treating the outliers, removal of the same and data cleansing)
3. Model input – to test the data after cleansing the data
4. Splitting the data – To split the data into test and train
5. Checking the statistics and accuracy of the model.
6. Final model presentation

**EDA:**

**Correlation Plot:**



Boxplot – for checking the outlier presence in each feature.



After building the model taking raw input into account (without any data wrangling), we get the below features.

**Regression Results:**

R-squared: 0.776

OLS Adj. R-squared: 0.775

F-statistic: 2399.

Prob (F-statistic): 0.00

Log-Likelihood: -44039.

AIC: 8.810e+04

BIC: 8.817e+04

**Significance stats:**

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------------

Intercept 307.9880 60.353 5.103 0.000 189.675 426.301

Speed 9.3203 0.185 50.364 0.000 8.958 9.683

HDD 0.7818 0.028 28.311 0.000 0.728 0.836

RAM 48.2560 1.066 45.265 0.000 46.166 50.346

Screen 123.0890 3.999 30.776 0.000 115.249 130.929

ads 0.6573 0.051 12.809 0.000 0.557 0.758

trend -51.8496 0.629 -82.470 0.000 -53.082 -50.617

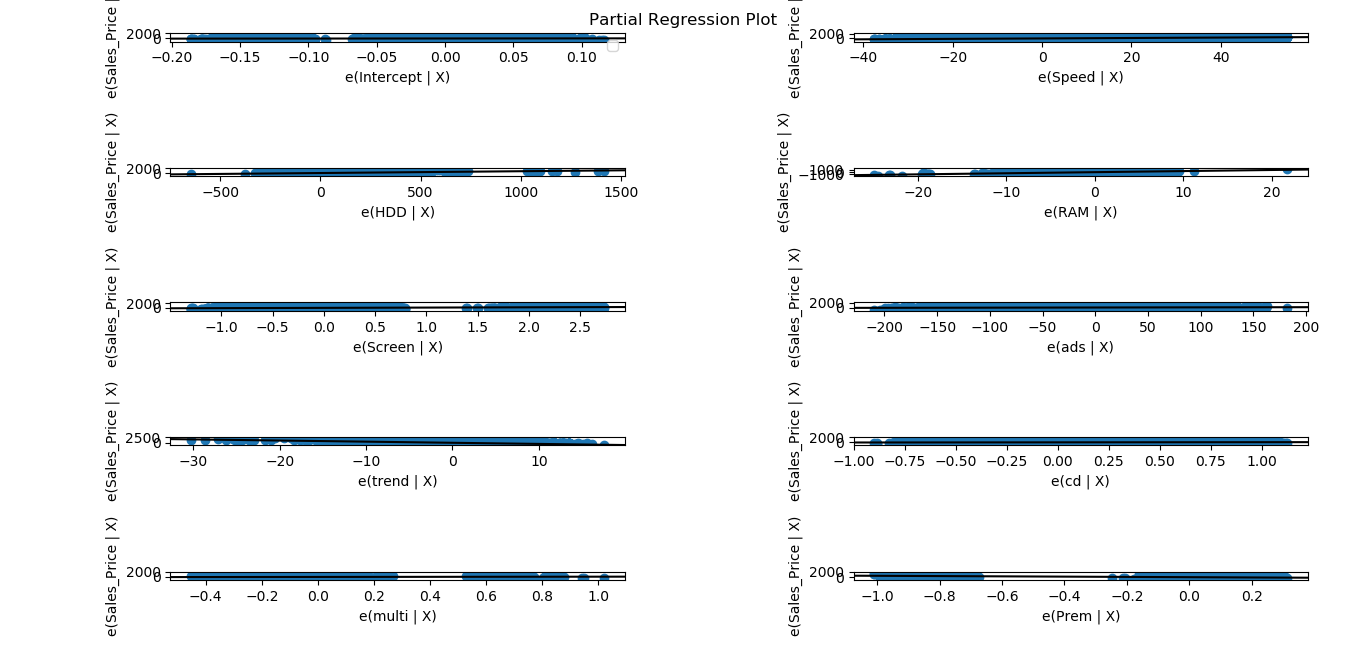
cd 60.9167 9.516 6.402 0.000 42.263 79.571

multi 104.3238 11.413 9.141 0.000 81.951 126.697

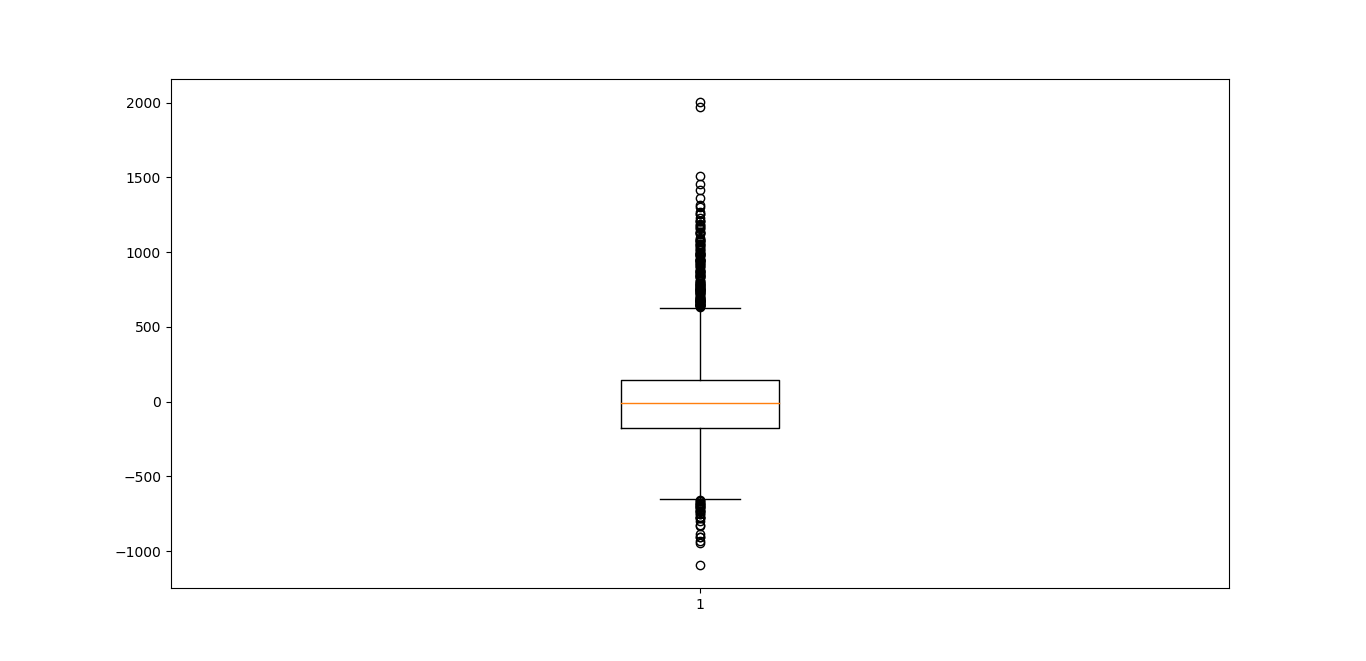
Prem -509.2247 12.342 -41.259 0.000 -533.420 -485.030

From above stats it can be observed that all the variables have significance on the output and hence cannot be ignored. Moreover the adj R^2 is found to be 0.775 which can still be considered as a valid R score.

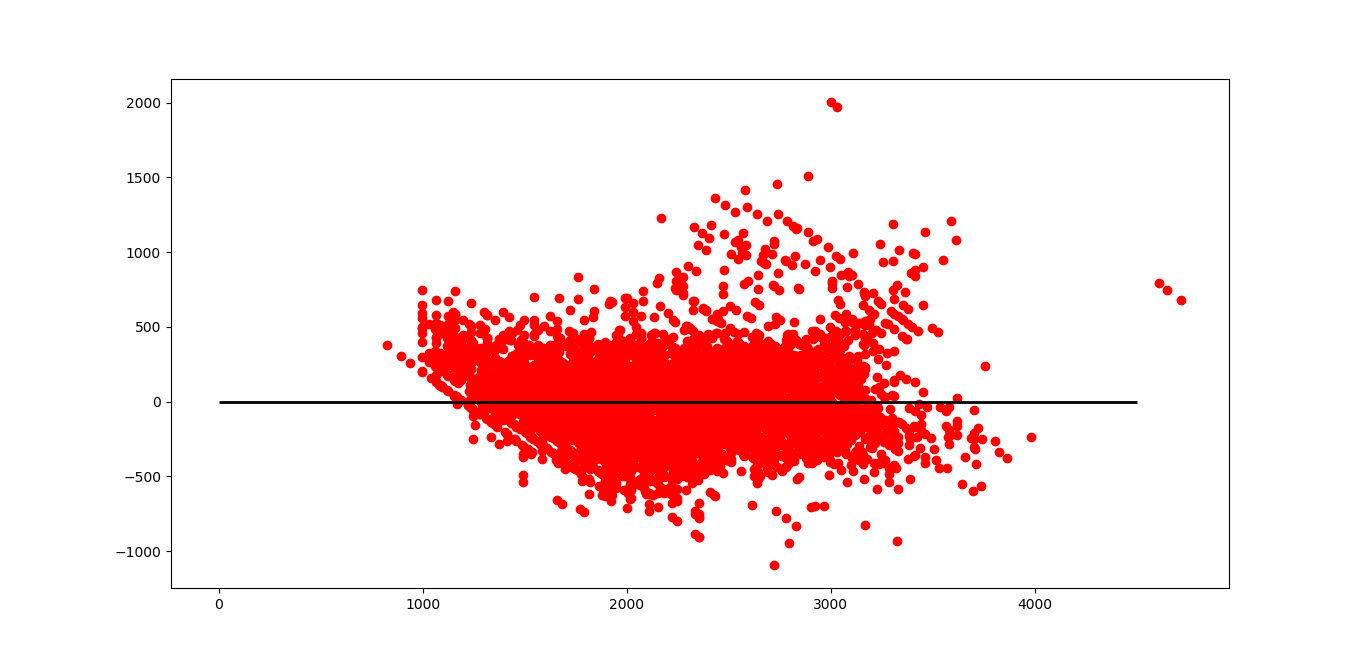
Partial Regression Plot



Error Distribution box plot – The plot shows that the error distribution is normal and the same can be identified by the mean/median/mode which lies at the centre.



We got to know that the mean of the residue (y-y\_pred) is almost equal to 0 (1.609420e-11)



Plotting the actual vs fitted values in the hyperplane

