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: Nukala Ayyappa Bharthwaj

Batch Id: DSWDMCOS 21012022

Topic: Multilinear Regression

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

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

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

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:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

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

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

5.3 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.

5.4 Briefly explain the model output in the documentation.

6. 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.

Sol:

Business Objective: To predict the profits of the company with other factors by using Multilinear regression model.

Constraints: Lack of analysis of the company data.

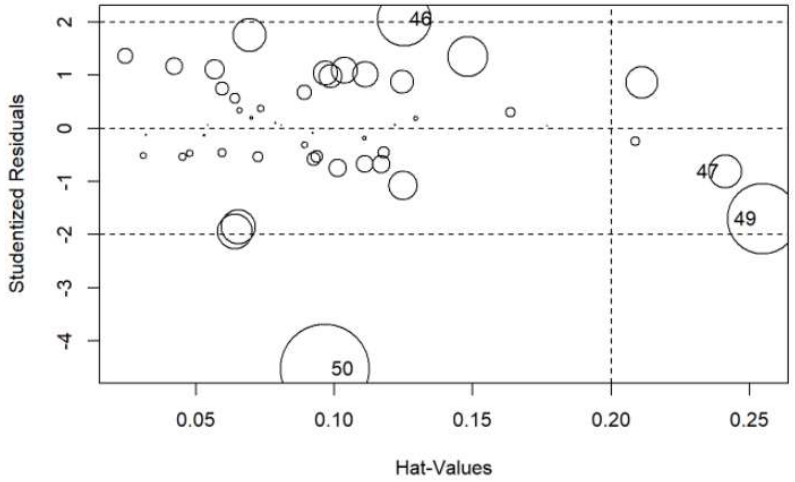
Data Types: All the given data is in numeric format except the states column and the complete data can be used for the analysis.

Data Cleaning: Since the states column is in Non-Numeric format the same is converted to numeric data for doing the analysis.

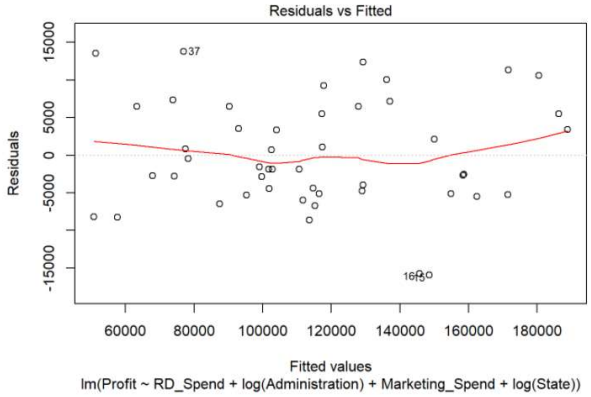
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as profit of the company.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.9496.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 49,50 and that are found by using the following influential plot.



After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.
After applying Logarithmic transformations the R^2 value is 0.9451 and for the exponential transformations R^2 value is 0.9557 and for the quadratic model is 0.9567.
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.
The following residual plot is done to know the accuracy of the model.



	R.D.Spend	Administration	Marketing.Spend	State	Profit
1	165349.20	136897.80	471784.10	New York	192261.83
2	162597.70	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.90	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.60
9	120542.52	148718.95	311613.29	New York	152211.77
10	123334.88	108679.17	304981.62	California	149759.96

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.

Sol:

Business Objective: To predict the Price of the car with other factors by using Multilinear regression model.

Constraints: Lack of analysis of the car sales data of the company

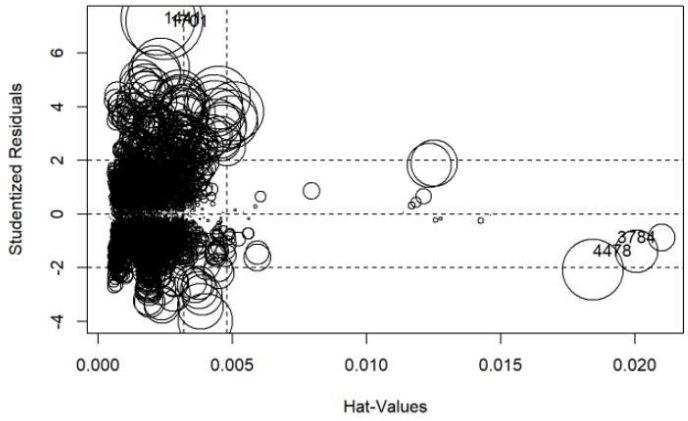
Data Types: The first column in the given data is I'd which is not useful for doing the analysis and the remaining data is used for doing the analysis.

Data Cleaning: Since some of the columns in the given data is no-numeric the same is converted into numeric data so that they can be used for doing the analysis.

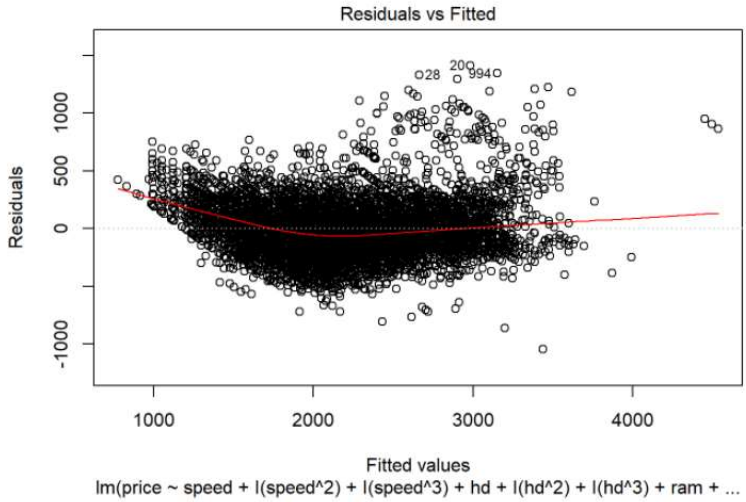
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer part.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.7752.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.



After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.
After applying Logarithmic transformations the R^2 value is 0.7441and for the exponential transformations R^2 value is 0.7833 and for the quadratic model is 0.8049.
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.
The following residual plot is done to know the accuracy of the model.



	X	price	speed	hd	ram	screen	cd	multi	premium	ads	trend
1	1	1499	25	80	4	14	no	no	yes	94	1
2	2	1795	33	85	2	14	no	no	yes	94	1
3	3	1595	25	170	4	15	no	no	yes	94	1
4	4	1849	25	170	8	14	no	no	no	94	1
5	5	3295	33	340	16	14	no	no	yes	94	1
6	6	3695	66	340	16	14	no	no	yes	94	1
7	7	1720	25	170	4	14	yes	no	yes	94	1
8	8	1995	50	85	2	14	no	no	yes	94	1
9	9	2225	50	210	8	14	no	no	yes	94	1
10	10	2575	50	210	4	15	no	no	yes	94	1

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

	Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type	HP	Met_Color	Color	Automatic	cc	Doors	Cylinder
1	1	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13500	23	10	2002	46986	Diesel	90	1	Blue	0	2000	3	4
2	2	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13750	23	10	2002	72937	Diesel	90	1	Silver	0	2000	3	4
3	3	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13950	24	9	2002	41711	Diesel	90	1	Blue	0	2000	3	4
4	4	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	14950	26	7	2002	48000	Diesel	90	0	Black	0	2000	3	4
5	5	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	13750	30	3	2002	38500	Diesel	90	0	Black	0	2000	3	4
6	6	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	12950	32	1	2002	61000	Diesel	90	0	White	0	2000	3	4
7	7	TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors	16900	27	6	2002	94612	Diesel	90	1	Grey	0	2000	3	4
8	8	TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors	18600	30	3	2002	75889	Diesel	90	1	Grey	0	2000	3	4
9	9	TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors	21500	27	6	2002	19700	Petrol	192	0	Red	0	1800	3	4
10	10	TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors	12950	23	10	2002	71138	Diesel	69	0	Blue	0	1900	3	4
11	11	TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors	20950	25	8	2002	31461	Petrol	192	0	Silver	0	1800	3	4
12	12	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors	19950	22	11	2002	43610	Petrol	192	0	Red	0	1800	3	4
13	13	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	19600	25	8	2002	32189	Petrol	192	0	Red	0	1800	3	4
14	14	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	21500	31	2	2002	23000	Petrol	192	1	Black	0	1800	3	4
15	15	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22500	32	1	2002	34131	Petrol	192	1	Grey	0	1800	3	4
16	16	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22000	28	5	2002	18739	Petrol	192	0	Grey	0	1800	3	4
17	17	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22750	30	3	2002	34000	Petrol	192	1	Grey	0	1800	3	4

Sol:

Business Objective: To predict the Price of the computer parts with other factors by using Multilinear regression model.

Constraints: Lack of analysis of the sales data of the company

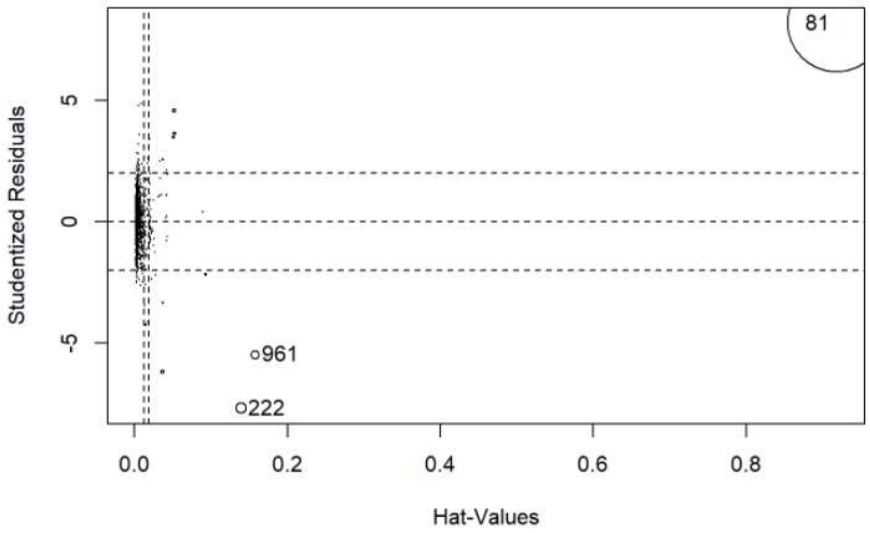
Data Types: The first column in the given data is I'd which is not useful for doing the analysis and the remaining data is used for doing the analysis.

Data Cleaning: Since some of the columns in the given data is no-numeric the same is converted into numeric data so that they can be used for doing the analysis.

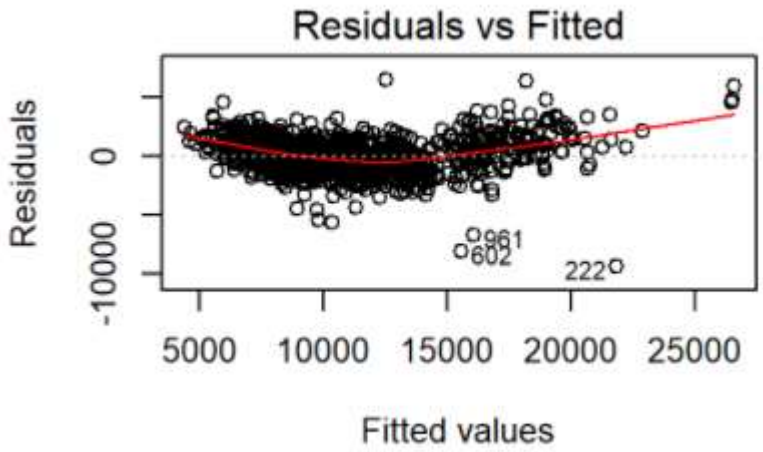
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.863.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.



After observing the influential plot it is noted that 81st observation is effecting the complete analysis so the same observation is removed and then the basic regression model is done then the R^2 values is observed as 0.9291 so the same model is used for the future prediction in the data . The residual plot for the model is as follows.



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: -

AveragePrice	Total_Volume	tot_ava1	tot_ava2	tot_ava3	Total_Bags	Small_Bags	Large_Bags	XLarge Bags	type	year	region
1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0	conventional	2015	Albany
1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0	conventional	2015	Albany
0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0	conventional	2015	Albany
1.08	78992.15	1132	71976.41	72.58	5811.16	5677.4	133.76	0	conventional	2015	Albany
1.28	51039.6	941.48	43838.39	75.78	6183.95	5986.26	197.69	0	conventional	2015	Albany
1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0	conventional	2015	Albany
0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0	conventional	2015	Albany
0.98	109428.33	703.75	101815.36	80	6829.22	6266.85	562.37	0	conventional	2015	Albany
1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0	conventional	2015	Albany
1.07	74338.76	842.4	64757.44	113	8625.92	8061.47	564.45	0	conventional	2015	Albany
1.12	84843.44	924.86	75595.85	117.07	8205.66	7877.86	327.8	0	conventional	2015	Albany
1.28	64489.17	1582.03	52677.92	105.32	10123.9	9866.27	257.63	0	conventional	2015	Albany
1.31	61007.1	2268.32	49880.67	101.36	8756.75	8379.98	376.77	0	conventional	2015	Albany
0.99	106803.39	1204.88	99409.21	154.84	6034.46	5888.87	145.59	0	conventional	2015	Albany
1.33	69759.01	1028.03	59313.12	150.5	9267.36	8489.1	778.26	0	conventional	2015	Albany
1.28	76111.27	985.73	65696.86	142	9286.68	8665.19	621.49	0	conventional	2015	Albany
1.11	99172.96	879.45	90062.62	240.79	7990.1	7762.87	227.23	0	conventional	2015	Albany
1.07	105693.84	689.01	94362.67	335.43	10306.73	10218.93	87.8	0	conventional	2015	Albany
1.34	79992.09	733.16	67933.79	444.78	10880.36	10745.79	134.57	0	conventional	2015	Albany
1.33	80043.78	539.65	68666.01	394.9	10443.22	10297.68	145.54	0	conventional	2015	Albany
1.12	111140.93	584.63	100961.46	368.95	9225.89	9116.34	109.55	0	conventional	2015	Albany
1.45	75133.1	509.94	62035.06	741.08	11847.02	11768.52	78.5	0	conventional	2015	Albany
1.11	106757.1	648.75	91949.05	966.61	13192.69	13061.53	131.16	0	conventional	2015	Albany
1.26	96617	1042.1	82049.4	2238.02	11287.48	11103.49	183.99	0	conventional	2015	Albany

Sol:

Business Objective: To predict the Price of the Avacado with other factors by using Multilinear regression model.

Constraints: Lack of analysis of the sales data of the avocado.

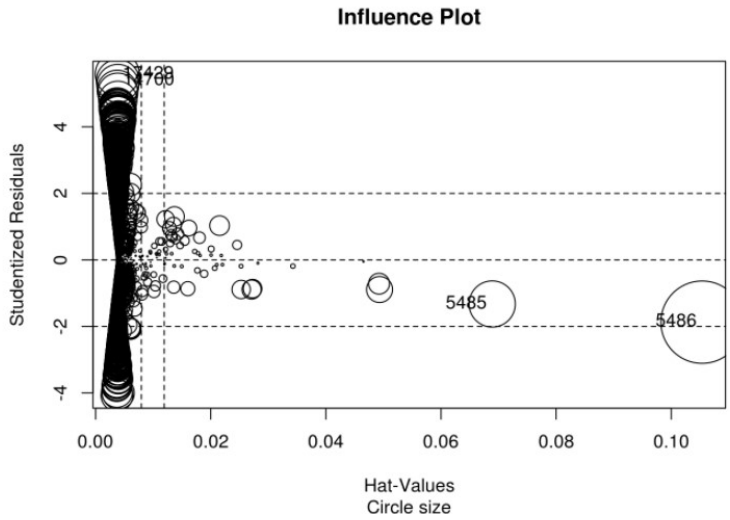
Data Types: All the data in the given data set is used for doing the analysis.

Data Cleaning: Since some of the columns in the given data is no-numeric the same is converted into numeric data so that they can be used for doing the analysis.

Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as Avg. price of the Avacado.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.7211.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 5485,5486 and that are found by using the following influential plot.



After removing the influential observations the multiple regression is done and the R^2 value is 0.8244 and the same model can be used for the prediction of the price of the Avacado. The residual plot to know the accuracy of the model is as follows.