Abstract: Linear regression is a statistical process in which a relationship is drawn between value of a dependent variable from the value of independent variable. Linear regression finds out the relationship between dependent and independent variable. The value of dependent variable is calculated by taking one or more independent variables into account. This paper gives a deep look into Linear regression and how to use it. The paper also talks about types of Linear regression and explains it with examples of linear regression in real world usage. In this paper we will perform linear regression using python and its libraries. Python is a language which is commonly used for ML, Data Science and Statistics and Visualization.

Introduction: Automobile Industry is a big Product based industry where selling of products like cars, motorcycles, trucks, etc, happens. These sales are highly dependent on various variables. These market variables change time to time based on new trends, these trends are representation of change in technology used in making of a car or any vehicle or the representation of change in customer needs with respect to fuel, design and quality, etc. The trends in this industry can be highly volatile for example buying a car based on some kind of crude oil (i.e., petrol, diesel) few years back was the only option customer had but now with more climate friendly source of energy being used in newer models, the competition has increased. Looking at this when a new model or a new generation car/vehicle is launched a lot of market research is done in order to get pricing and specification right. Companies now have started using predictive algorithms for the same, As the predictive target also helps in generating investments requires in Research and development of that model.

Predictive analysis there are various models available like Random Forest, Linear Regression, Logistic Regression, K-means and many other algorithms are prepared according to the data available. In our case we will look to see if there is a linear relationship between our independent variable and dependent variable. If we have a linear relationship, we will use Linear regression.

The concept of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is test applied on a dataset to define and quantify the relation between considered variable. In today’s world linear regression is used by Insurance company or to calculate effect of supplement in weight loss or the impact of advertisement on sales or water bill and amount of water used or connection between work experience and salary, etc. All these applications of Linear regression models are to see how value of a dependent variable i.e., salary, sales, water used, weight loss, Insurance premiums changes with respect to independent variables like work experience, water bill, advertisements, Risk assessment.

Suppose we have a dataset of company’s monthly spend on advertisement and marketing and monthly sales, using this dataset in linear regression we can predict how much sales will be made with amount of money spend on marketing. In linear regression we will first divide dataset in two parts 80 percent of this dataset will be taken as testing dataset this dataset is used to train the model which will predict the data. The other 20 percent of the dataset will be used to cross-check the results of dependent variable in this case monthly sales with respect to independent variable in this case money spent on advertisement and marketing. This will also help us in finding the error percentage.

Firstly, we will see simple linear regression in which we have one dependent and one independent variable this regression is based on same formula as a line which defines relationship between x and y variable.

$$\boxed{\mathbf{Y = mX + C}}$$

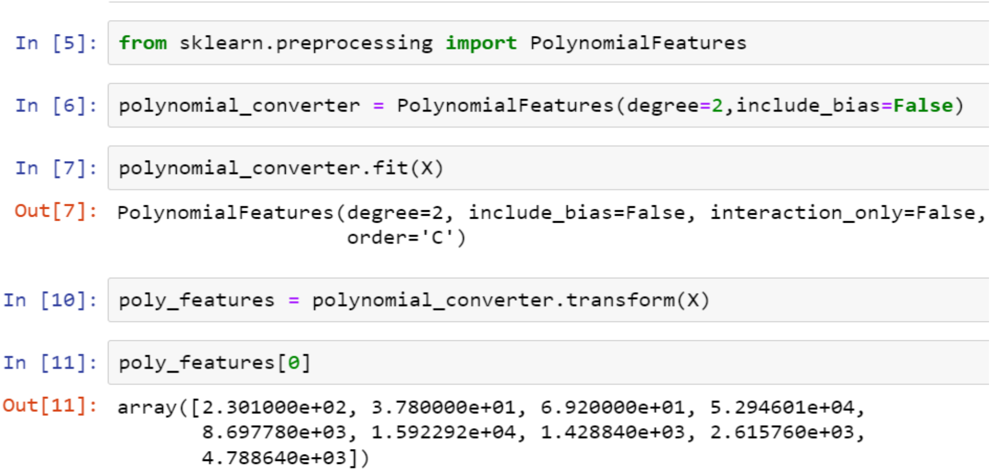
In above fig.1 we can see the formula used by linear regression when we have 1 dependent and 1 independent variable. This formula is used to describe a line with X and Y as points, m as slope of line and C as constant. For linear regression in this formula Y is a dependent variable and X is independent variable with b0 as y-intercept and b1 becoming beta coefficients or parameters and e is the error term also known as the residual errors. This formula is written as following.

$$\boxed{\mathbf{Y = \beta 0 + \beta 1x + \varepsilon}}$$

If we have more than one independent variable let’s say X1 and X2 then we have use one more beta coefficient b2 in the equation if we have X variables from 1 to p we will use beta parameters from 0 to p. The following formula shows how the equation will be with p independent variables in an equation.

$$Yi =\beta0+\beta1xi1+\beta2xi2+...+\beta pxip+\epsilon$$

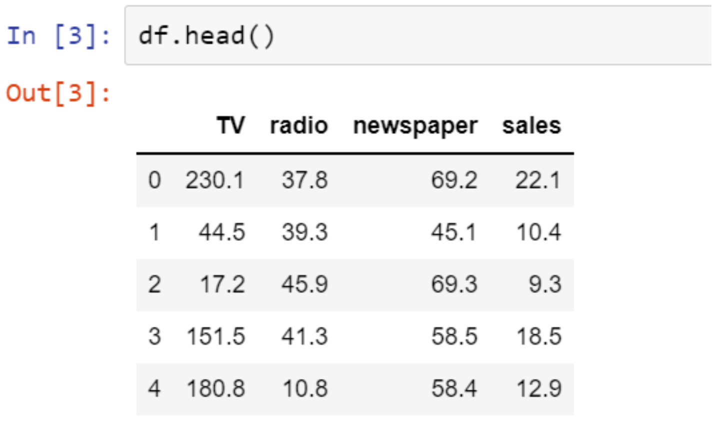
To find beta variables we will use built-in python functions from libraries that already exists in python.



```
In [5]:  from sklearn.preprocessing import PolynomialFeatures

In [6]:  polynomial_converter = PolynomialFeatures(degree=2,include_bias=False)

In [7]:  polynomial_converter.fit(X)

Out[7]:  PolynomialFeatures(degree=2, include_bias=False, interaction_only=False,
                            order='C')

In [10]: poly_features = polynomial_converter.transform(X)

In [11]: poly_features[0]

Out[11]: array([2.301000e+02, 3.780000e+01, 6.920000e+01, 5.294601e+04,
                8.697780e+03, 1.592292e+04, 1.428840e+03, 2.615760e+03,
                4.788640e+03])
```

Above figure is example of how we use transform function to generate values for the equation to find Yi or the value of dependent variable. This value will be used to train and test the model.

Linear Regression models in real life are used to predict different outcomes. For example, a company is nearing launch of a new generation car and they want to predict how budget should the marketing have to achieve a particular sale target, they have previous data of their cars that are in the market. The data shows how much money they have spent on likes of digital marketing, holdings and newspaper adverts in order to get a sale for example they had spent 20 million rupees on digital adverts, 15 million on holdings and 5 million on newspaper adverts in a month, the data also shows sales made from these advertisements in total. Similarly, there will be data for other months and other models of cars. This data can used by a company to decide how much money they should spend in order to achieve the sale target. In this example sale target is a dependent variable Y, money on digital advertisement, holdings and newspaper adverts are our independent variables X1, X2 and X3.



```
In [3]: df.head()
```

Out[3]:

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

As shown in above figure TV, Radio and Newspaper are our independent variables and sales is our dependent variable.

Literature Survey:

When we are making a model for a particular industry, we have to make sure that we are using the model that shows the data in best way possible with respect to accuracy, complexity and different variables.

To decide which model to use we can try if the relation between dependent and independent variable. If the relationship is linear, we can use Linear regression model as using other models like random forest will give a similar result. Random forest is used when we have a complex relationship and do not have any time constraint.

Sometimes the data we have is highly volatile, the trends shown by this data can confuse a lot of people on which model will fit the most. Using models like SVR, decision tree or random forest could help, even these models have limitations.

Support Vector Regression works in similar way to SVM where in a hyperplane is created and the points closest to the hyperplane are support vectors. In regression a error or margin of tolerance is set in approximation, but in support vector regression it would have requested from the problem.

In this paper we have taken to different datasets into consideration to show how regression works. The second dataset is more complex than the first, the number of independent variables is more in second dataset. We have used Linear regression model as to draw a comparison on how it answers to different datasets. Linear regression is used on various datasets like temperature, humidity or rainfall prediction

Daniela Şchiopu [1] and his team in his publication used SPSS 13.0 tool and forecasted temperature from data collected from the Hong Kong Observatory website. They used factor analysis technique in the SPSS tool to reduce the complexity in calculations the temperature using correlation and regression.

Paras and Sanjay [4] developed a forecasting model using mathematical regression. The weather data is collected for a period of 3 years and this model can predict max and min temperatures for a period of 15 to 45 weeks into the future.
Goutami [5] used Multiple Linear Regression to estimate average summer – monsoon rainfall on the data from 1871 to 1999. She analysed the monthly rainfall of Indian summer monsoon months.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Let's say you are concerned about climate change and wants to study the weather condition to know what parameters have an impact on the temperature. we can use humidity, air pressure, wind speed to predict the temperature that day. we will use linear regression here. Linear regression is the simplest yet very powerful way to model linear relationship between scalar dependent and one or more independent variable.

Methodology:

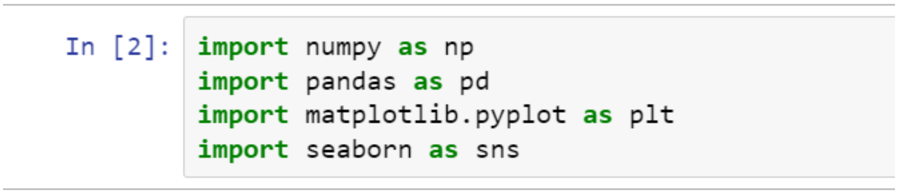
There are multiple types of Linear Regression that can be used in order to answer the problem statement with the help of previous data that was recorded or stored by an organisation.

Simple Linear Regression is used to answer problems where there is only one independent variable like a company has launched a product exclusively on a Online platform but it has not received the sales they wanted only single mode of digital marketing was done. Data set will be created in order to make a comparison between money spent on digital marketing and sales. Sales will become our dependent variable and marketing will become aur independent variable.

As shown in above diagram a table representation of amount spent a by a organisation on marketing to the amount generated from sales.

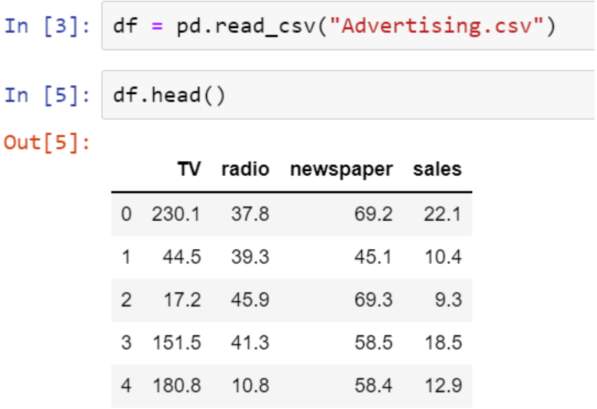
To perform Simple Linear Regression, we will have to follow following steps using python programming language:

1. Firstly, in Python we will have to import libraries in order to perform regression. Libraries like NumPy, Pandas, Matplotlib, and Seaborn are imported into our Jupyter Notebook. These libraries are used in making dataset, importing data, analysis of data, Visualisation of data and running different ML algorithms on data.
    A. NumPy: NumPy is a Python programming language library that is used for making multi-dimensional arrays and matrices, along with this it is also used for mathematical functions.
    B. Pandas: Pandas is another Python library that is used for data manipulation and analysis. It also helps in Data Structures and in manipulation of numerical tables and time series.
    C. Matplotlib: Matplotlib is a python library used for plotting python programming language and its numerical mathematics extension NumPy. It also has “pylab” interface based on a state machine.
    D. Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.



```
In [2]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

2. Second step is to read the data file into our notebook and save it to a variable, to read a file we have to Pandas library’s ‘read_csv’ function.

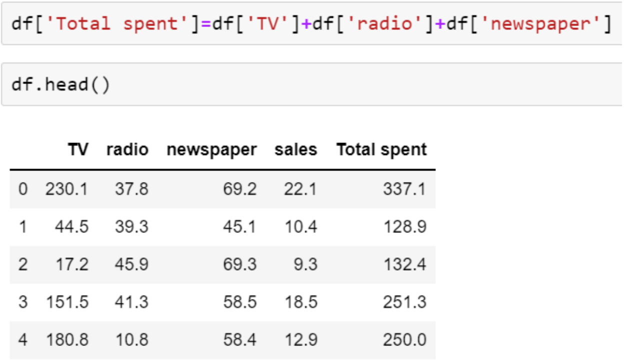


```
In [3]: df = pd.read_csv("Advertising.csv")

In [5]: df.head()

Out[5]:
```

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

3. After reading this data we will assign independent and dependent variables to ‘x’ and ‘y’. In this case total money spent on ‘TV’, ’radio’ and ‘newspaper’

of all will be ‘x’ independent variable and sales will be dependent variable ‘y’.

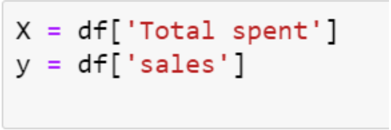


```
df['Total spent']=df['TV']+df['radio']+df['newspaper']
```

```
df.head()
```

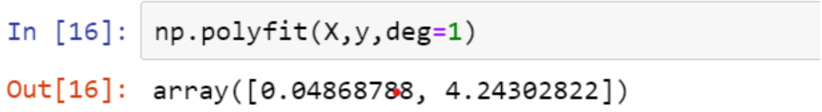
	TV	radio	newspaper	sales	Total spent
0	230.1	37.8	69.2	22.1	337.1
1	44.5	39.3	45.1	10.4	128.9
2	17.2	45.9	69.3	9.3	132.4
3	151.5	41.3	58.5	18.5	251.3
4	180.8	10.8	58.4	12.9	250.0

As show in the above figure total spent which is addition of TV, radio and newspaper is our independent variable ‘x’.



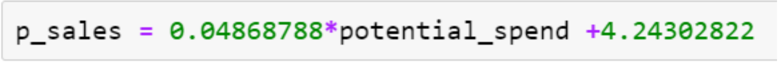
```
X = df['Total spent']
y = df['sales']
```

4. Now we will find the coefficient of the variables or parameters using polyfit function. We will give ‘x’ and ‘y’ as inputs to the function

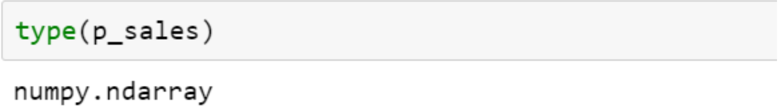


```
In [16]: np.polyfit(X,y,deg=1)

Out[16]: array([0.04868788, 4.24302822])
```

5. Now to calculate the sales with help of independent variable we will use Simple linear regression formula.



```
p_sales = 0.04868788*potential_spend +4.24302822
```

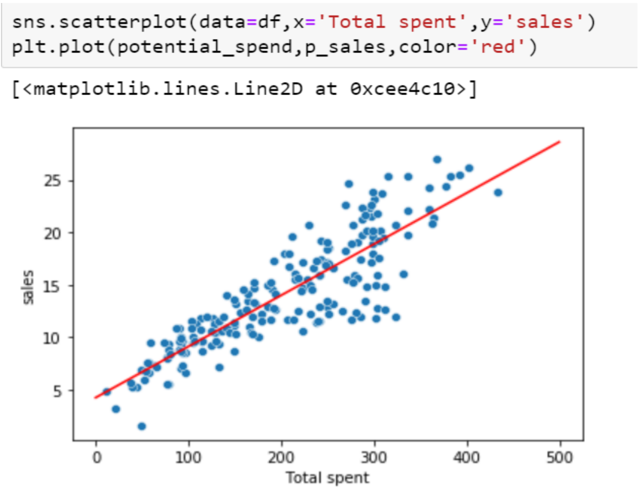


```
type(p_sales)
```

```
numpy.ndarray
```

Using this formula, we will generate an array of ‘p_sales’ which will be stored in an array dataset.

6. Lastly, we will make a plot showing Linear regression.



```
sns.scatterplot(data=df,x='Total spent',y='sales')
plt.plot(potential_spend,p_sales,color='red')
```

```
[<matplotlib.lines.Line2D at 0xcee4c10>]
```

In Multiple Linear Regression we can increase the number of independent variables for a dependent variable. We can use Polynomial function in Multiple Linear Regression to find the coefficients for variables. But in Multiple Linear Regression we have to make sure that the variables are not showing multicollinearity, because when independent variables are showing multicollinearity, it will be hard to predict which specific variable is contributing to variance in dependent variable.

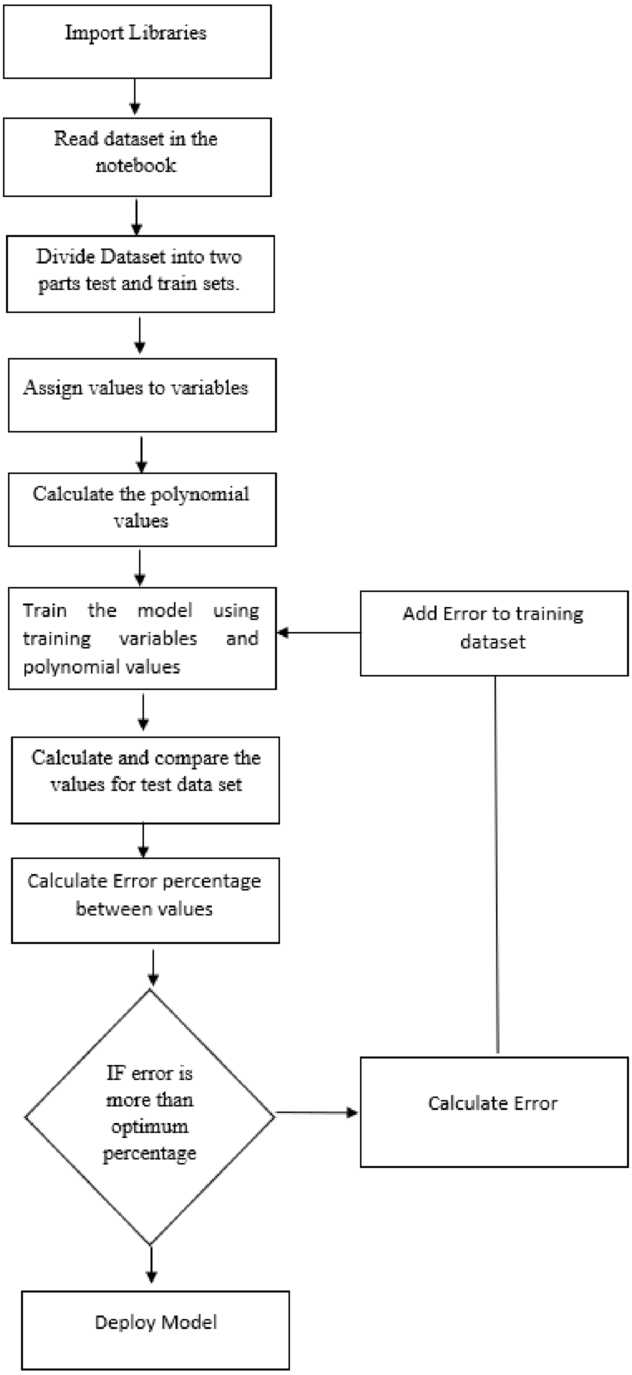
In MLR we have to keep bias variance in account while increasing the number of independent variables as they also effect the error rate. Bias variable is kept in check in order to keep low error rate. Depending on the contribution of the variables in predicting the dependent variable, these variables have the effect on error rate. There are scenarios where increasing number of variables till a certain number may reduce the error rate and then shoot up because of 1 variable. While selecting the Sets of independent variables we have to be careful with the changes in bias variable. This is important as when we will use this model to test or predict dependent variable in particular scenario this model will take error rate in consideration so bigger the error rate less accurate the result will be.

In MLR one assumption is that relationship between dependent and independent variable should be linear. Best practice is to do a scatterplot and then use different models to find best model that fits the data.

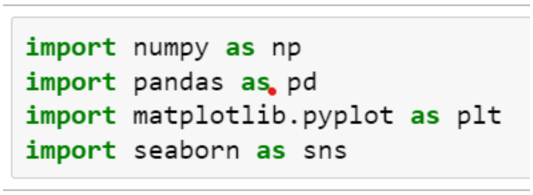
If we have categorical data for multiple regression model we can use dummy variables for this data, these variables consist of values such as 0 or 1 representing the presence and absence of categorical data.

For MLR we will be using a dataset of 50 startups. Features include R&D spend, Administration, Marketing spend, State and finally profit.

In MLR we will follow similar steps to prepare the model.

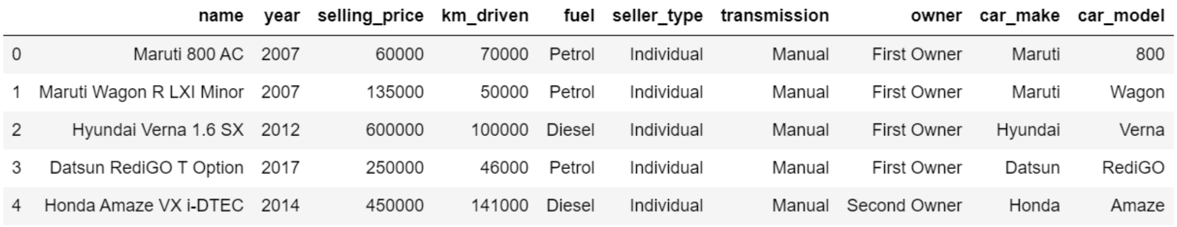

Import Libraries
Read dataset in the notebook
Divide Dataset into two parts test and train sets.
Assign values to variables
Calculate the polynomial values
Train the model using training variables and polynomial values
Add Error to training dataset
Calculate and compare the values for test data set
Calculate Error percentage between values
IF error is more than optimum percentage
Calculate Error
Deploy Model

Step 1: Step one is same as Simple linear regression importing the libraries in the notebook. NumPy, Pandas, Matplotlib.pyplot and Seaborn.



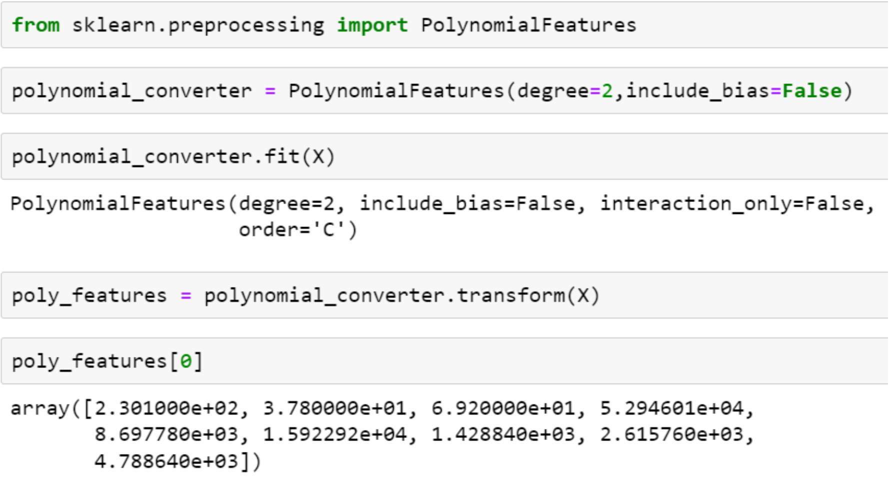
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Step 2: Step two is to look at all the variables and decides which variables effect the dependent variable.



	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	car_make	car_model
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner	Maruti	800
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner	Maruti	Wagon
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner	Hyundai	Verna
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner	Datsun	RediGO
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner	Honda	Amaze

Step 3: In this step find the values of polynomial degree and load it to the our model.



```
from sklearn.preprocessing import PolynomialFeatures
```

```
polynomial_converter = PolynomialFeatures(degree=2,include_bias=False)
```

```
polynomial_converter.fit(X)
```

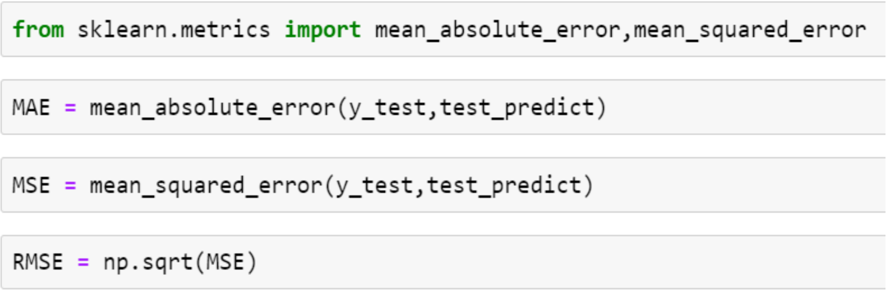
```
PolynomialFeatures(degree=2, include_bias=False, interaction_only=False,
                   order='C')
```

```
poly_features = polynomial_converter.transform(X)
```

```
poly_features[0]
```

```
array([2.301000e+02, 3.780000e+01, 6.920000e+01, 5.294601e+04,
       8.697780e+03, 1.592292e+04, 1.428840e+03, 2.615760e+03,
       4.788640e+03])
```

Step 4: We import different functions to find errors in our model.



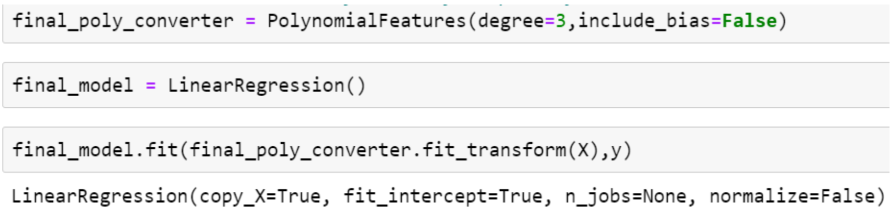
```
from sklearn.metrics import mean_absolute_error,mean_squared_error
```

```
MAE = mean_absolute_error(y_test,test_predict)
```

```
MSE = mean_squared_error(y_test,test_predict)
```

```
RMSE = np.sqrt(MSE)
```

Step 5: Use the values we make changes in our independent variables and make a new model with less error.



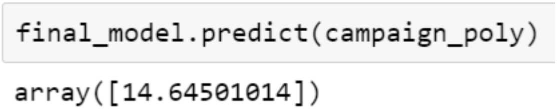
```
final_poly_converter = PolynomialFeatures(degree=3,include_bias=False)
```

```
final_model = LinearRegression()
```

```
final_model.fit(final_poly_converter.fit_transform(X),y)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Step 6: The final model is then use to predict the values.



```
final_model.predict(campaign_poly)
```

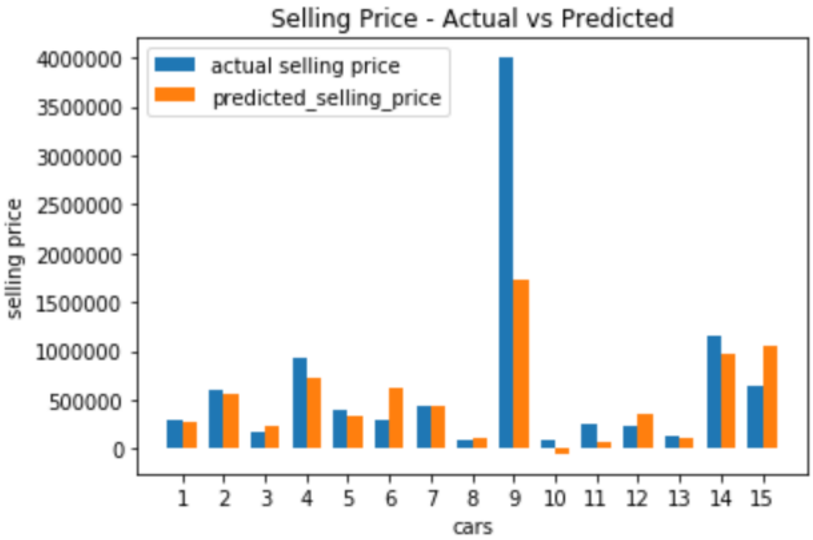
```
array([14.64501014])
```

Results

In simple Linear regression we have learnt how dependent variables are affected by the independent variable. The result we got is a model that can be deployed on data that has a single independent variable. This model can be used to predict values of dependent variable with respect to the values of independent variable. The Example of the simple linear regression model given is based on sales and advertisement data, so we can use this model by giving money spent on advertisement as input of independent variable and get output value of expected sales. To Create model for any other problem statement we will have to make new model which is to be trained on data for that specific problem statement, we will use the same steps just the data used to train the model will be different.

In Multiple Linear Regression we have made a model that takes more than one dependent variable into account. The model we have made is for automobile industry, it takes different factors in to account of different cars to find what factors effects the most. In multiple linear regression we have to be careful with the number of dependent variables we are using as it can also increase the error percentage.

The figure given below is representation of difference between the value predicted versus the real value. This is a comparison of test data on model trained on 80 percent of Data that was divided at the start. This comparison is used in order to decide what variables effect the model the most. This is useful in order to find more accurate model.



Conclusion: Linear regression is a very powerful tool for data science or machine learning, we use this regression model in order to make the assumption of the relationship between the dependent and independent variables. It is assumed that in this model the relation is always a linear relation, but in some cases the relation cannot be justified in a linear format so we sometimes see a non-linear relation like a curve. If the result is not a linear relation, we should try a different model like decision tree or random forest algorithm as they are better suited for more complex models.

Simple Linear can be used at start in order to see what trends are, if the trend is simple and not complex, we can try different combination of independent variables to in-order to find the one that predicts the values accurately and has lower error rate. Linear regression should be used when there are less complex variables as the rate of accuracy increases and it also takes less amount of time as compared to random forest which more complex. Linear regression is widely used in various industries where predictive analysis is required like weather prediction or sale analysis.

References:

[1]Daniela Şchiopu, Elia Georgiana Petre, Catalina Negoina "Weather Forecast using SPSS Statistical Methods" This paper presents a case study of using SPSS 13.0 in weather prediction. Vol. LXI No. 1/2009 Gas University of Ploiesti, Bd. Bucuresti, No. 39, 100680, Romania.

[4] Paras and Sanjay Mathur "A Simple Weather Forecasting Model Using Mathematical Regression" Department of Electronics & Communication Engineering, College of Technology, G.B. Pant University of Agriculture & Technology, Pantnagar, (India) 263 145.

[5] Goutami Bandyopadhyay "The Prediction of Indian Monsoon Rainfall: A Regression Approach" 1/19 Dover Place Kolkata-700 019 West Bengal India.

[2] Wuthrich, Mario V. and Buser, Christoph, Data Analytics for Non-Life Insurance Pricing (October 27, 2021). Swiss Finance Institute Research Paper No. 16-68, Available at SSRN: https://ssrn.com/abstract=2870308 or http://dx.doi.org/10.2139/ssrn.2870308

[3] Abdulraheem Sal, Mohammed Raja, The Impact of Training and Development on Employees Performance and Productivity (july 10, 2016). International Journal of Management Sciences and Business Research, Vol. 5, Issue 7, July 2016, Available at SSRN: https://ssrn.com/abstract=2849769

[6] Uddin, Gazi and Alam, Md. Mahmudul, The Impacts of Interest Rate on Stock Market: Empirical Evidence from Dhaka Stock Exchange (2010). South Asian Journal of Management Sciences, Vol. 4(1), pp. 21-30, (2010), Available at SSRN: https://ssrn.com/abstract=2941287

[7] Kaushal, Anirudh and Shankar, Achyut, House Price Prediction Using Multiple Linear Regression (April 25, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3833734 or http://dx.doi.org/10.2139/ssrn.3833734

[8] Fuller, James R., Investment Forecasting With Multivariate Linear Regression (January 12, 2007). Available at SSRN: https://ssrn.com/abstract=532002 or http://dx.doi.org/10.2139/ssrn.532002

[9] Timothy DelSole and J. Shukla "Linear Prediction of Indian Monsoon Rainfall".

[10] "Multiple Linear Regression Analysis Research Paper", n.d. https://studentshare.org/human-resources/1683996-multiple-linear-regression-analysis.

[11] Multiple Linear Regression - Research Prospect

[12] Linear Regression. This article is about Linear regression… | by Renu Khandelwal | DataDrivenInvestor

[13] ML | Linear Regression - GeeksforGeeks

[14] Linear Regression (Definition, Examples) | How to Interpret? (wallstreetmojo.com)

[15] Linear regression review (article) | Khan Academy

[16] jeas_0617_6115.pdf (arpnjournals.org)