# Topic: Simple Linear Regression

**Instructions:**

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

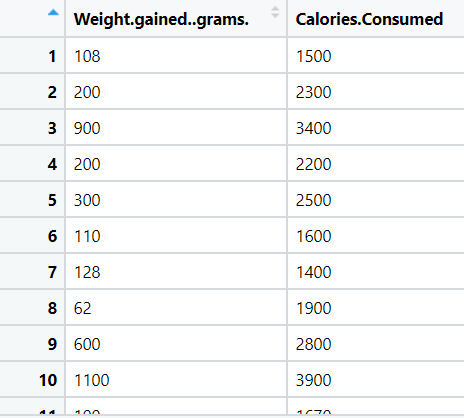
Please ensure you update all the details:

**Name: \_Prajay B. Urkude Batch ID: 16092021**

**Topic: Simple Linear Regression**

**Problem Statement: -**

A certain food-based company conducted a survey with the help of a fitness company to find the relationship between a person’s weight gain and the number of calories they consumed in order to come up with diet plans for these individuals. Build a Simple Linear Regression model with calories consumed as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



**Ans:- Business Objective :**

To find out the relationship between the weight gain vs the calories consumed

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Features** | **Description** | **Type** | **Relevance** |
| Weight.gain.grams | Weight in grams | Quantitative, Ratio | Relevant |
| Calories.Consumed | Calories consumed to gain weight | Quantitative, Ratio | Relevant |

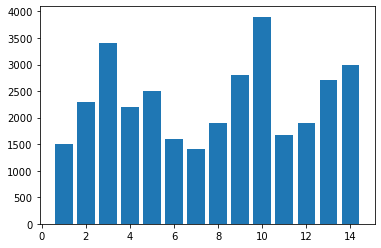
**Steps to find The Linear regression between two variables as below:**

* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

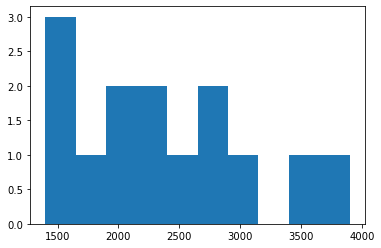
From Statsmodel library import formulae.api package.

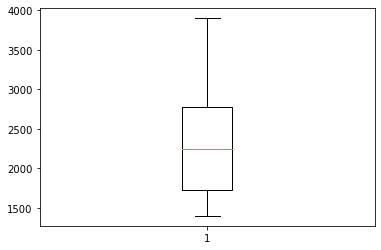
**Statsmodel** :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.

For Weight column:

From the barplot, 10th row has the highest weight

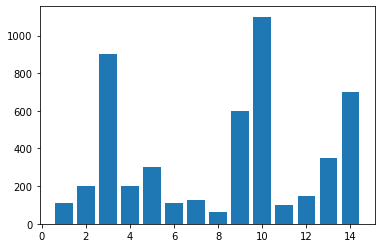


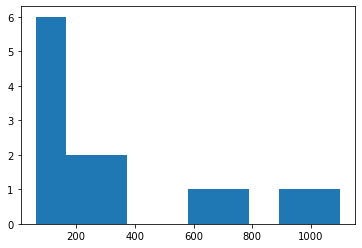
**The data is not normally distributed as it is right ske**

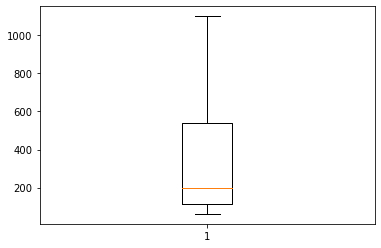
The data has no outlier.

For Calories.Consumed column:

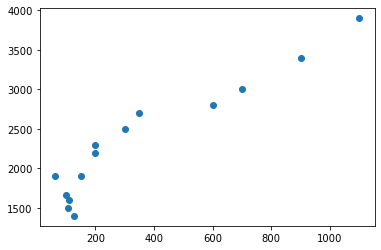
10th observation has the highest calories consumed





The data is normally distributed and the data is right skewed**.**

No outliers in the data.

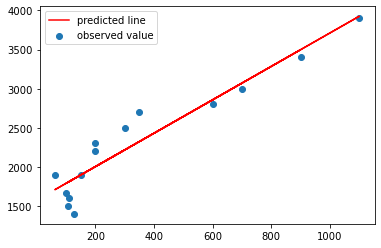


Coefficient of correlation r = 0.94 and from the above scatter plot we can say that the both the variable are strongly correlated with each other in the positive direction.

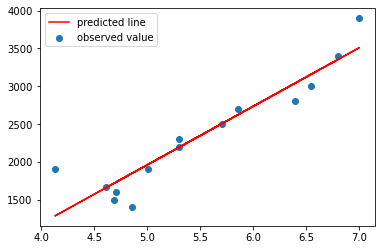
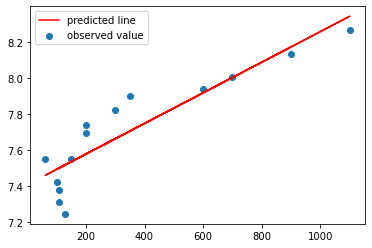
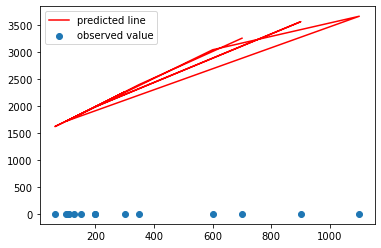
* Create a model by using the ols function by taking calories consumed as the target variable and weight gain as the independent variable and check the summary.

Coefficient of determination R2 should be more that 0.85 and the p value also should be less than 0.05. otherwise, we use transformation technique to best fit the model and for better prediction by using log transformation, exponential transformation, polynomial transformation until we get the R2 value > 0.85 and P value <0.05

* Prediction to be done by using the above model and calculate the ROOT MEAN SQUARE ERROR.

from the above graph, actual values are away from the predicted and actual values are not normally distributed and the RMSE is found out to be 232.83.

Model building by using the transformation

* We use log transformation, exponential transformation and polynomial transformation and from these three methods we found out the RMSE values are 253.55, 272.42, 240.82 and for different transformation technique we get different graph as below.

Log Transformation exponential Transformation Polynomial Transformation

* From the above RMSE value we get the less RMSE value by using simply applying the model without transformation and as the less the RMSE better will be the model.
* Splitting the data into training data and testing data.
* Fit the finalized model on the training datasets and do prediction on the test datasets and calculate the RMSE value for test datasets.
* Again evaluate the same model on train datasets and calculate the RMSE value for train datasets and compare the RMSE value for both train and test datasets. If both are nearly equal then the model is perfectly fit.

**Problem Statement: -**

A logistics company recorded the time taken for delivery and the time taken for the sorting of the items for delivery. Build a Simple Linear Regression model to find the relationship between delivery time and sorting time with delivery time as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



Ans:- **:- Business Objective :**

To find out the relationship between the sorting time vs the delivery time

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Features** | **Description** | **Type** | **Relevance** |
| Delivery.time | Time required for delivery | Quantitative, Ratio | Relevant |
| Sorting.time | Sorting time for the parcels | Quantitative, Ratio | Relevant |

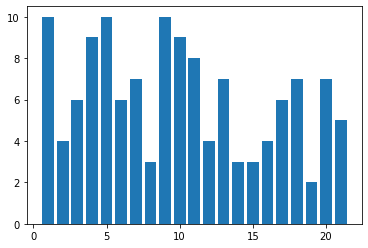
**Steps to find The Linear regression between two variables as below:**

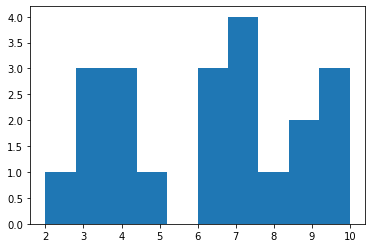
* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

From Statsmodel library import formulae.api package.

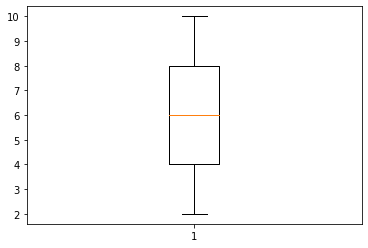
**Statsmodel** :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.

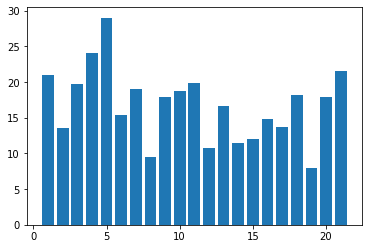
For sort time :

From the barplot, 1st row has the highest sorting time

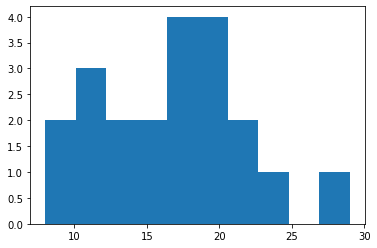
The data is not normally distributed



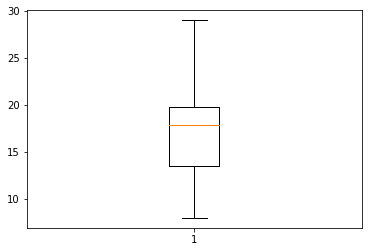
The sort\_time variable has no outlier

For delivery column:

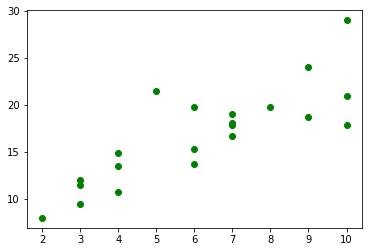
The observation 5 takes the highest delivery time.



The data is not normally distributed and it is right skwed.



No outliers in the data.



Coefficient of correlation r = 0.82 and from the above scatter plot we can say that the both the variable moderately correlated in positive direction.

Calculate the covariance which is come out 10.

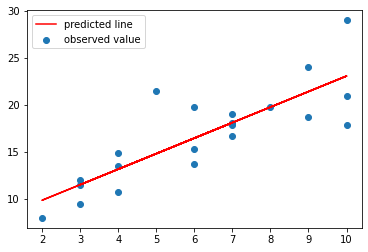
* Create a model by using the ols function by taking calories consumed as the target variable and weight gain as the independent variable and check the summary.

Coefficient of determination R2 should be more that 0.85 and the p value also should be less than 0.05. otherwise, we use transformation technique to best fit the model and for better prediction by using log transformation, exponential transformation, polynomial transformation until we get the R2 value > 0.85 and P value <0.05

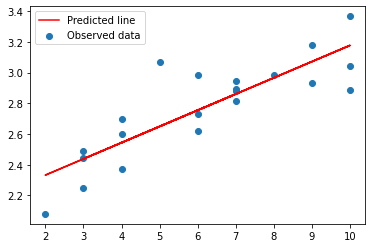
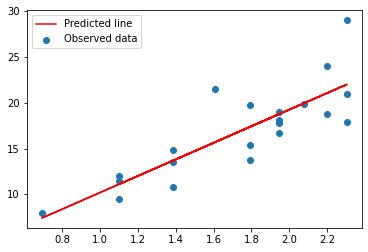
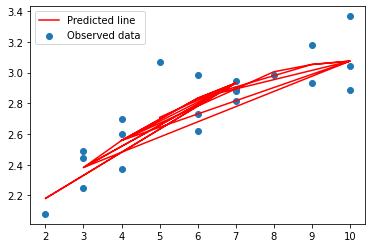
* Prediction to be done by using the above model and calculate the ROOT MEAN SQUARE ERROR.

from the above graph, actual values are away from the predicted and actual values are not normally distributed and the RMSE is found out to be 2.79.

Model building by using the transformation



* we use log transformation, exponential transformation and polynomial transformation and from this three methods we found out the RMSE values are 2.73, 2.94, 2.79 and for different transformation technique we get different graph as below.

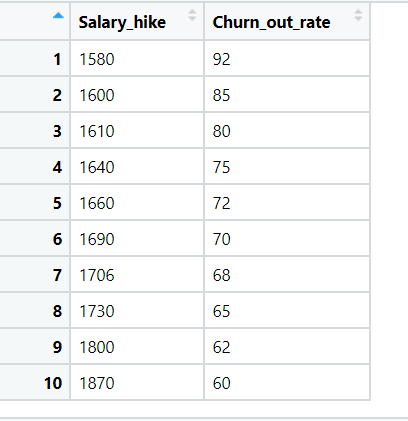


Log Transformation exponential Transformation Polynomial Transformation

* From the above RMSE value we get the less RMSE value by using simply applying the model without transformation and as the less the RMSE better will be the model.
* Splitting the data into training data and testing data.
* Fit the finalized model on the training datasets and do prediction on the test datasets and calculate the RMSE value for test datasets.
* Again, evaluate the same model on train datasets and calculate the RMSE value for train datasets and compare the RMSE value for both train and test datasets. If both are nearly equal then the model is perfectly fit.

**Problem Statement: -**

A certain organization wants an early estimate of their employee churn out rate. So the HR department gathered the data regarding the employee’s salary hike and the churn out rate in a financial year. The analytics team will have to perform an analysis and predict an estimate of employee churn based on the salary hike. Build a Simple Linear Regression model with churn out rate as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



**Ans: Business Objective :-**

To find out the relationship between the salary hike and the churning rate of the employee.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Features** | **Description** | **Type** | **Relevance** |
| Salary\_hike | Hike in the salary | Quantitative, Ratio | Relevant |
| Churn\_out\_ rate | Churning rate of the employee | Quantitative, Ratio | Relevant |

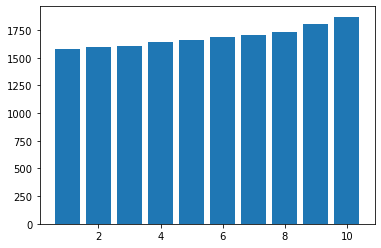
**Steps to find The Linear regression between two variables as below:**

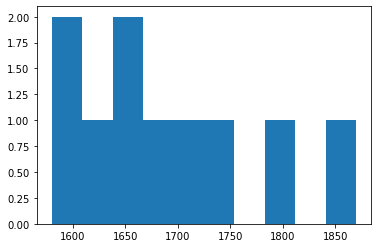
* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

From Statsmodel library import formulae.api package.

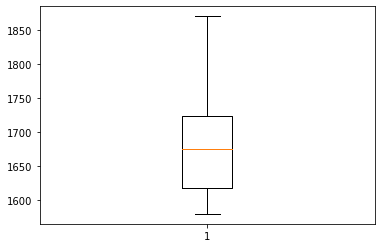
**Statsmodel** :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.

For Salary Hike:

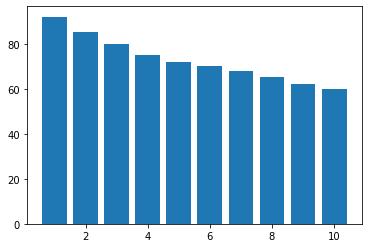
From the bar plot, 10th row has the highest Salary.

The data is not normally distributes and it is right skwed.

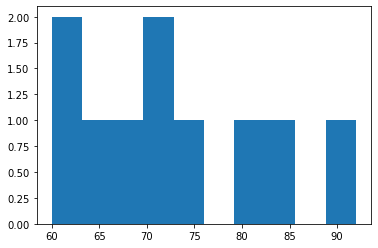


The data has no outlier.

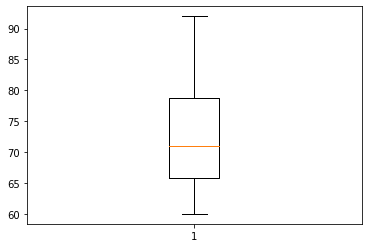
For Churning rate column:



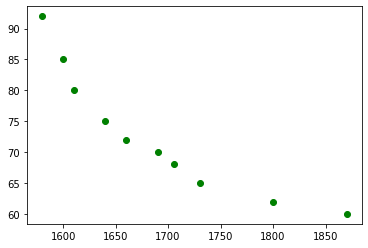
The bar 1 has the highest churning rate.



The data is right skewed and is not normally distributed.

The data is normally distributed and the data is right skewed**.**

There is no outliers in the datasets



Coefficient of correlation r = 0.91 and from the above scatter plot we can say that the both the variable are strongly correlated with each other in the negative direction. As the salary hike increses churning rate automatically decreases.

There is very high covariance between two variables. i.e -861

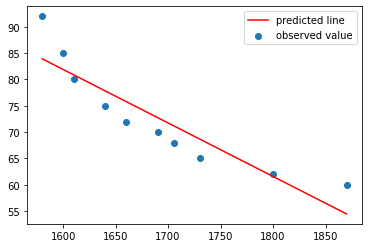
* Create a model by using the ols function by taking calories consumed as the target variable and weight gain as the independent variable and check the summary.

Coefficient of determination R2 should be more that 0.82 and the p value also should be less than 0.05. otherwise, we use transformation technique to best fit the model and for better prediction by using log transformation, exponential transformation, polynomial transformation until we get the R2 value > 0.85 and P value <0.05

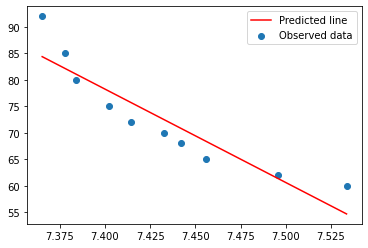
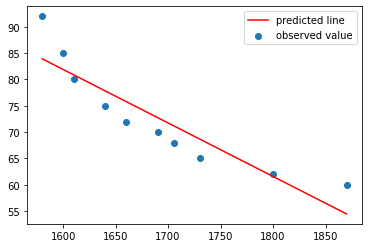
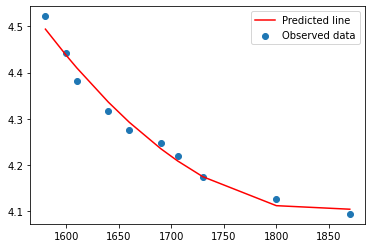
* Prediction to be done by using the above model and calculate the ROOT MEAN SQUARE ERROR.

from the above graph, actual values are away from the predicted and actual values are not normally distributed and the RMSE is found out to be 3.99

Model building by using the transformation



* We use log transformation, exponential transformation and polynomial transformation and from these three methods we found out the RMSE values are 3.78, 3.54, 1.32and for different transformation technique we get different graph as below.



Log Transformation exponential Transformation Polynomial Transformation

* From the above RMSE value we get the less RMSE value by using polynomial transformation and as the less the RMSE better will be the model.
* Splitting the data into training data and testing data.
* Fit the finalized model on the training datasets and do prediction on the test datasets and calculate the RMSE value for test datasets.
* Again, evaluate the same model on train datasets and calculate the RMSE value for train datasets and compare the RMSE value for both train and test datasets. If both are nearly equal then the model is perfectly fit.

**Problem Statement: -**

## The head of HR of a certain organization wants to automate their salary hike estimation. The organization consulted an analytics service provider and asked them to build a basic prediction model by providing them with a dataset that contains the data about the number of years of experience and the salary hike given accordingly. Build a Simple Linear Regression model with salary as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



**Ans:- Business Objective :**

To build a model which gives the relationship between the no. of years of experience and Salary hike

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Features** | **Description** | **Type** | **Relevance** |
| Years of experience | No. years of experience employee have | Quantitative, Ratio | Relevant |
| Salary | Calories consumed to gain weight | Quantitative, Ratio | Relevant |

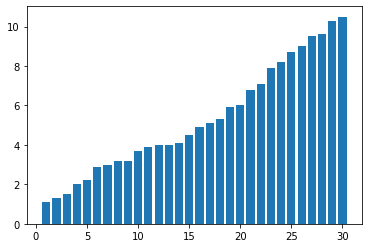
**Steps to find The Linear regression between two variables as below:**

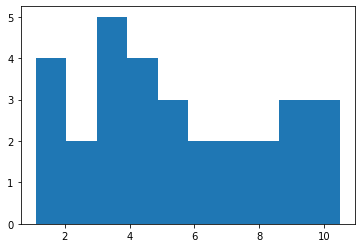
* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

From Statsmodel library import formulae.api package.

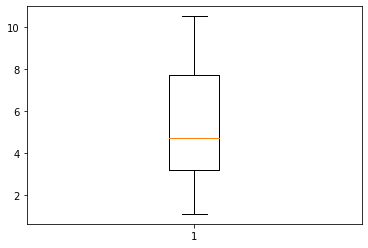
**Statsmodel** :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.

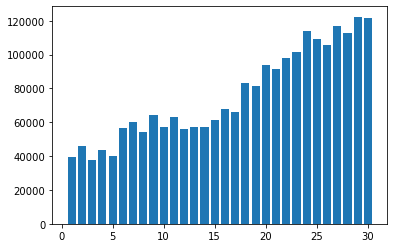
For years of experience column:

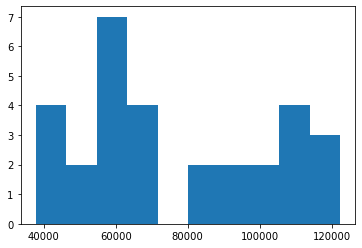
From the barplot we can see experience is arranged in the ascending order.

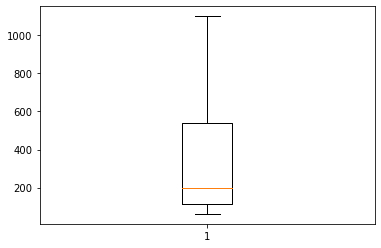
The data is not normally distributes and right skewed.

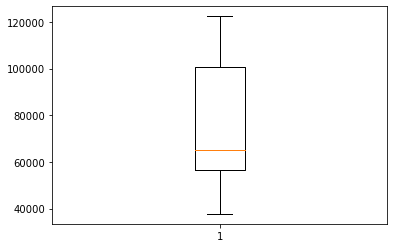


The data has no outlier.

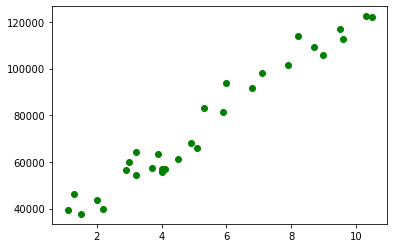
For Salary Hike column:

last observation has the salary hike.

The data is normally distributed and the data is right skewed**.**



The data has no outlier.



Coefficient of correlation r = 0.98 and from the above scatter plot we can say that the both the variable are strongly correlated with each other in the positive direction.

The covariance between two variable is very high i.e. 76106.3

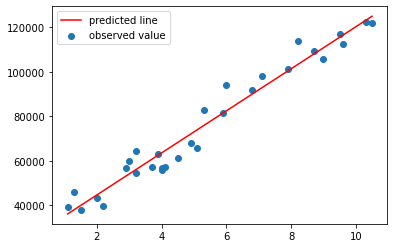
* Create a model by using the ols function by taking calories consumed as the target variable and weight gain as the independent variable and check the summary.

Coefficient of determination R2 should be more that 0.85 and the p value also should be less than 0.05. otherwise, we use transformation technique to best fit the model and for better prediction by using log transformation, exponential transformation, polynomial transformation until we get the R2 value > 0.85 and P value <0.05

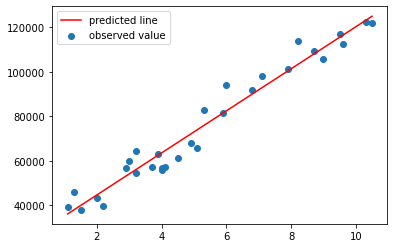
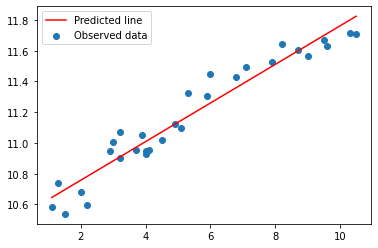
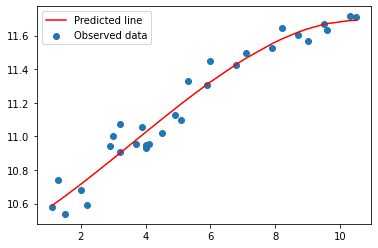
* Prediction to be done by using the above model and calculate the ROOT MEAN SQUARE ERROR.

from the above graph, actual values are away from the predicted and actual values are not normally distributed and the RMSE is found out to be 5592.

Model building by using the transformation



* We use log transformation, exponential transformation and polynomial transformation and from these three methods we found out the RMSE values are 10302, 7213, 5186 and for different transformation technique we get different graph as below.

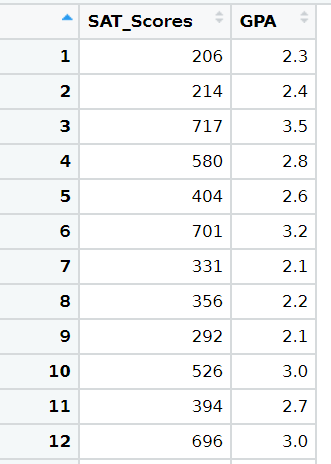


Log Transformation exponential Transformation Polynomial Transformation

* From the above RMSE value we get the less RMSE value by using polynomial transformation and as the less the RMSE better will be the model.
* Splitting the data into training data and testing data.
* Fit the finalized model on the training datasets and do prediction on the test datasets and calculate the RMSE value for test datasets.
* Again evaluate the same model on train datasets and calculate the RMSE value for train datasets and compare the RMSE value for both train and test datasets. If both are nearly equal then the model is perfectly fit.

## **Problem Statement: -**

## A certain university wants to understand the relationship between students’ SAT scores and their GPA. Build a Simple Linear Regression model with GPA as the target variable and record the RMSE and correlation coefficient values for different models.



**Business Objective :-**

To find out the relationship between the salary hike and the churning rate of the employee.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Of Features** | **Description** | **Type** | **Relevance** |
| Sat score | Sat score of the student | Quantitative, Ratio | Relevant |
| GPA | GPA got by the student | Quantitative, Ratio | Relevant |

**Steps to find The Linear regression between two variables as below:**

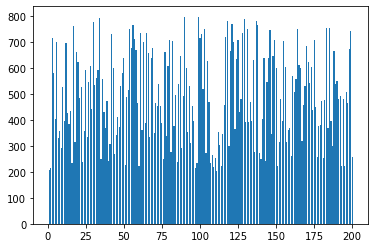
* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

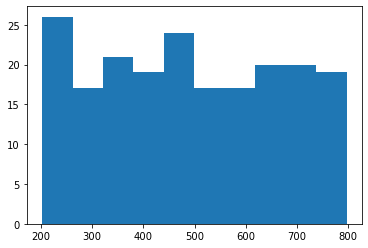
From Statsmodel library import formulae.api package.

**Statsmodel** :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

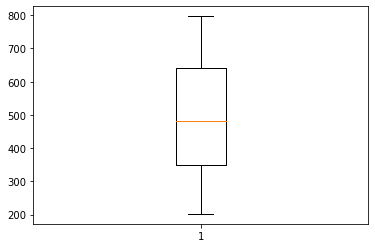
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.

For Sat score :

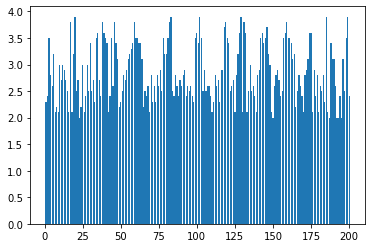




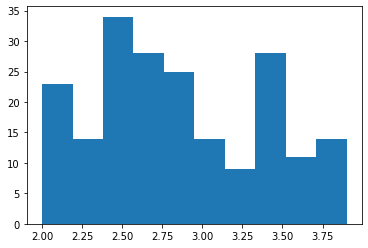
The data is not normally distributed



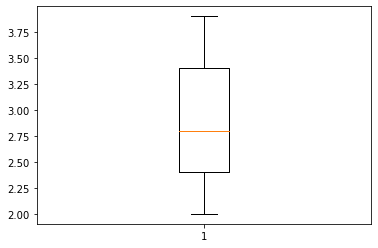
The satscore variable has no outlier

For gpa column:

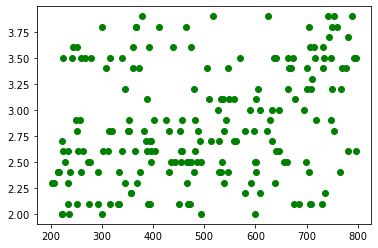
The observation 5 takes the highest delivery time.



The data is not normally distributed and it is right skwed.



No outliers in the data.



Coefficient of correlation r = 0.29, both the variables are not correlates with each other.

Calculate the covariance which is come out 27.22.

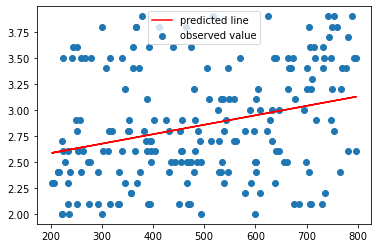
* Create a model by using the ols function by taking calories consumed as the target variable and weight gain as the independent variable and check the summary.

Coefficient of determination R2 should be more that 0.85 and the p value also should be less than 0.05. otherwise, we use transformation technique to best fit the model and for better prediction by using log transformation, exponential transformation, polynomial transformation until we get the R2 value > 0.85 and P value <0.05

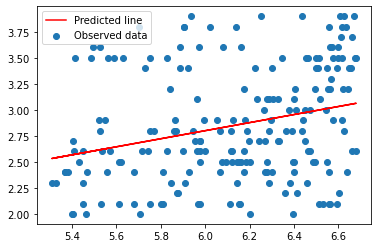
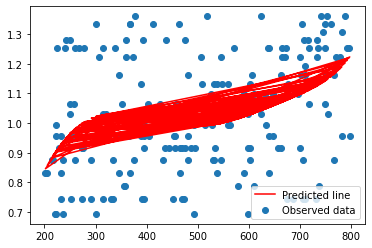
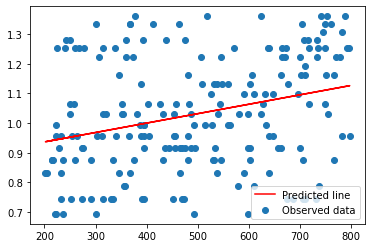
* Prediction to be done by using the above model and calculate the ROOT MEAN SQUARE ERROR.

from the above graph, actual values are away from the predicted and actual values are not normally distributed and the RMSE is found out to be 0.51

Model building by using the transformation



* we use log transformation, exponential transformation and polynomial transformation and from these three methods we found out the RMSE values are0.51, 0.51, 0.50 and for different transformation technique we get different graph as below.



Log Transformation exponential Transformation Polynomial Transformation

* From the above RMSE value we get the less RMSE value by using polynomial transformationand as the less the RMSE better will be the model.
* Splitting the data into training data and testing data.
* Fit the finalized model on the training datasets and do prediction on the test datasets and calculate the RMSE value for test datasets.
* Again, evaluate the same model on train datasets and calculate the RMSE value for train datasets and compare the RMSE value for both train and test datasets. If both are nearly equal then the model is perfectly fit.