# 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: DHEERAJ MISHRA Batch ID:**  DS\_01072021

**Topic: Text Mining and NLP**

**Grading Guidelines:**

**An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the image below:**



**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. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing.**
   1. **Data Cleaning, Feature Engineering, etc.**

**3.2. Outlier Treatment.**

1. **Exploratory Data Analysis (EDA):**

**4.1 Summary.**

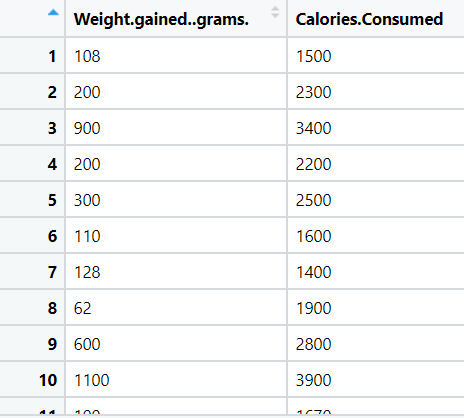
**4.2 Univariate analysis.**

**4.3 Bivariate analysis.**

1. **Model Building:**
   1. **Perform Simple Linear Regression on the given datasets.**
   2. **Apply different transformations such as exponential, log, polynomial, etc. transformations and calculate RMSE values, R-Squared values, and the correlation coefficient for each model.**
   3. **Build the models and choose the best fit model.**
   4. **Briefly explain the model output in the documentation.**
2. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided**

**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.



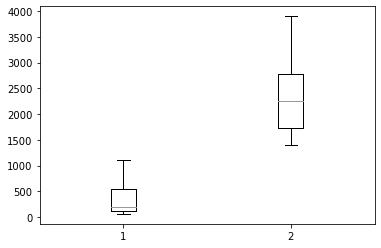
1. BUSINESS OBJECTIVE:-

Maximize relationship between variables

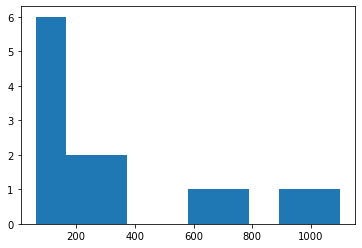
1. DATA UNDERSTANDING:-

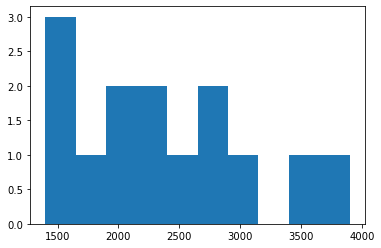
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Weight gained | Weight gained | Continuous | Relevant |
| Calories consumed | Calories consumed | Continuous | Relevant |

1. DATA CLEANSING :-
2. Dataset consists of 2 colums and 14 rows
3. Duplicate row does not exists
4. All data types are of form int64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. There is positive linear relationship between columns as correlation coefficient is high
9. Both columns are positively skewed
10. From histogram data is not normal
11. EDA:-
12. From box plot

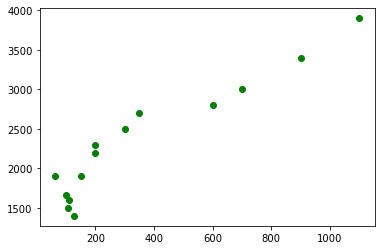


1. From histogram





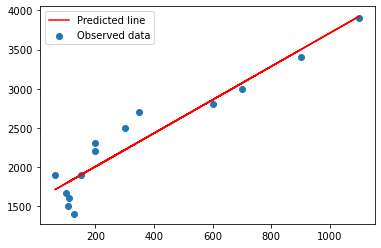
1. Scatter plot



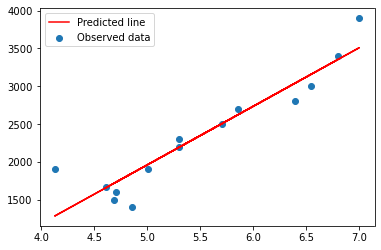
1. MODEL BUILDING:-
2. Model builded on simple , logarithmic , exponential and polynomial transformation
3. Output table for best transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Name of transformation | RMSE | Rsquare value | Correlation coeff |
| Simple linear regression | 232.8335 | 0.897 | 2.1344 |
| Logarathmic | 253.5580 | 0.878 | 774.1735 |
| Exponential | 272.4207 | 0.808 | 0.0009 |
| Polynomial degree two | 240.8277 | 0.852 | 0.0017 |

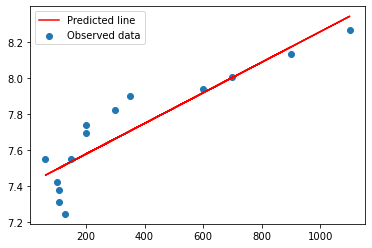
1. Best fit line for simple linear regression



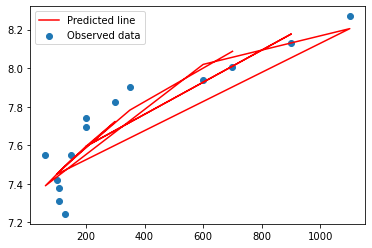
1. Best fit line for logarithmic transformation



1. Best fit line for exponential transformation



1. Best fit line for polynomial transformation



OUTPUT:-

1. Best fit model is simple linear regression with low RMSE
2. Splitting data to train 80% and test 20%
3. Train accuracy :- 214.5336903937176
4. Test accuracy:- 299.3533507505858
5. It is right fit model
6. BENEFITS :-

From above information we can predict for calories consumed against weight gained .

**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.



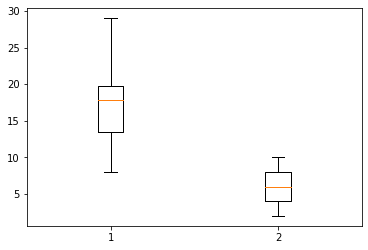
1. BUSINESS OBJECTIVE:-

Maximize relationship between variables

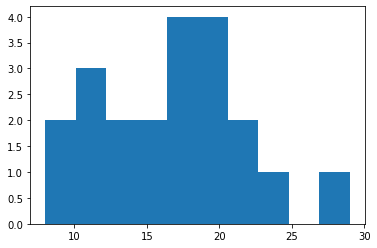
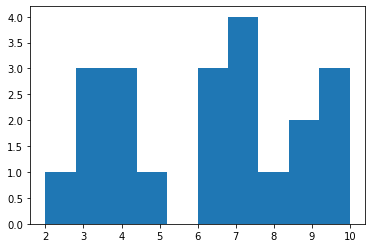
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Delivery time | Delivery time | Continuous | Relevant |
| Sorting time | Sorting time | Continuous | Relevant |

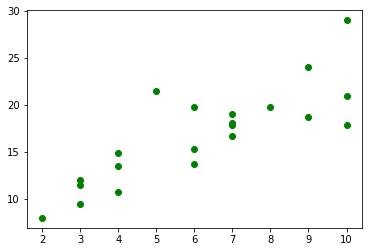
1. DATA CLEANSING :-
2. Dataset consists of 2 colums and 21 rows
3. Duplicate row does not exists
4. All data types are of form int64 and float64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. There is positive linear relationship between columns as correlation coefficient is high
9. Both columns are positively skewed
10. From histogram data is normal
11. EDA:-
12. From box plot



1. From histogram

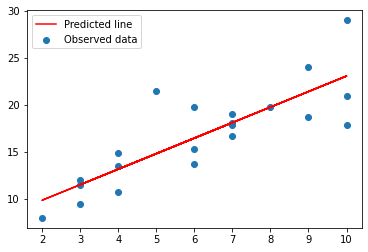
1. Scatter plot



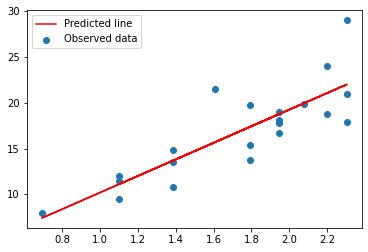
1. MODEL BUILDING:-
2. Model builded on simple , logarithmic , exponential and polynomial transformation
3. Output table for best transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Name of transformation | RMSE | Rsquare value | Correlation coeff |
| Simple linear regression | 2.7916 | 0.682 | 1.6490 |
| Logarathmic | 2.7331 | 0.695 | 9.0434 |
| Exponential | 2.9402 | 0.711 | 0.1056 |
| Polynomial degree two | 2.7990 | 0.765 | 0.2659 |

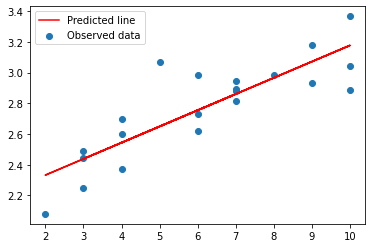
1. Best fit line for simple linear regression



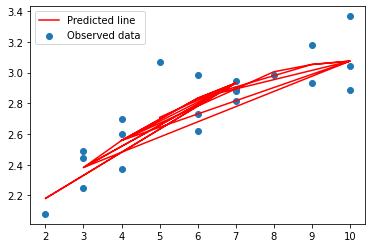
1. Best fit line for logarithmic transformation



1. Best fit line for exponential transformation



1. Best fit line for polynomial transformation



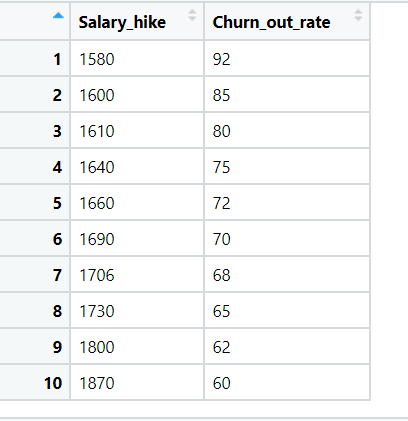
OUTPUT:-

1. Best fit model is logarithmic trasformation with low RMSE
2. Splitting data to train 80% and test 20%
3. Train accuracy :- 2.8311098908383734
4. Test accuracy:- 2.4329533238572654
5. It is right fit model
6. BENEFITS :-

From above information we can predict for sorting time against delivery time .

**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.



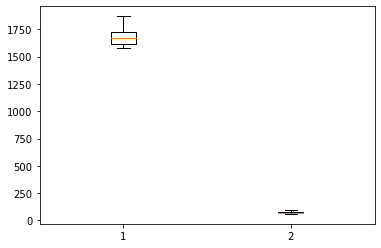
1. BUSINESS OBJECTIVE:-

Maximize relationship between variables

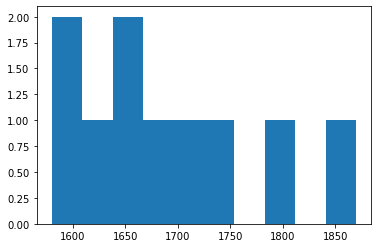
1. DATA UNDERSTANDING:-

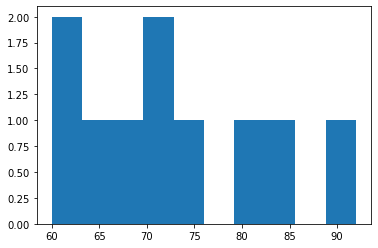
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Salary hike | Hike in salary | Continuous | Relevant |
| Churn rate | Churn rate | Continuous | Relevant |

1. DATA CLEANSING :-
2. Dataset consists of 2 colums and 10 rows
3. Duplicate row does not exists
4. All data types are of form int64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. There is negative linear relationship between columns as correlation coefficient is high
9. Both columns are positively skewed
10. From histogram data is normal
11. EDA:-
12. From box plot

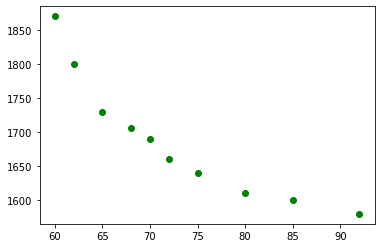


1. From histogram





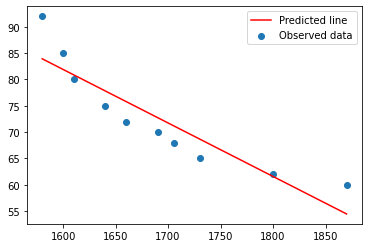
1. Scatter plot



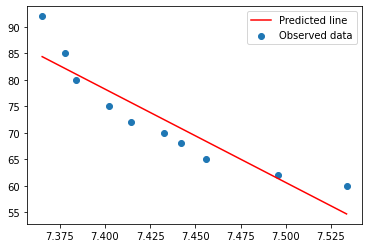
1. MODEL BUILDING:-
2. Model builded on simple , logarithmic , exponential and polynomial transformation
3. Output table for best transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Name of transformation | RMSE | Rsquare value | Correlation coeff |
| Simple linear regression | 3.9975 | 0.831 | -0.1015 |
| Logarathmic | 3.78600 | 0.849 | -176.1097 |
| Exponential | 3.5415 | 0.874 | -0.0014 |
| Polynomial degree two | 1.3267 | 0.984 | -0.0207 |

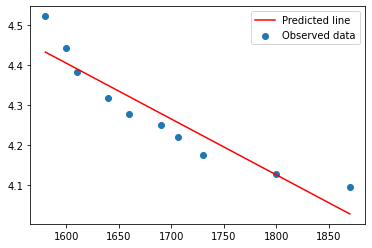
1. Best fit line for simple linear regression



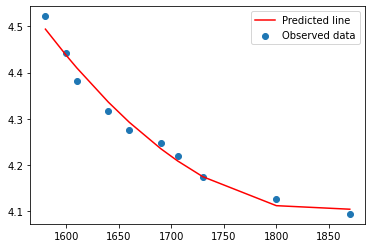
1. Best fit line for logarithmic transformation



1. Best fit line for exponential transformation



1. Best fit line for polynomial transformation



OUTPUT:-

1. Best fit model is polynomial transformation with low RMSE
2. Splitting data to train 80% and test 20%
3. Train accuracy :- 68.6511858242197
4. Test accuracy:- 71.74854848553356
5. It is right fit model
6. BENEFITS :-

From above information we can predict for churn rate against salary hike . .

**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.



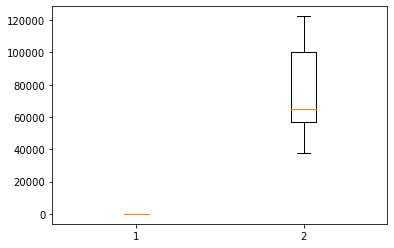
1. BUSINESS OBJECTIVE:-

Maximize relationship between variables

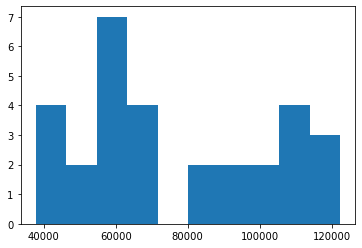
1. DATA UNDERSTANDING:-

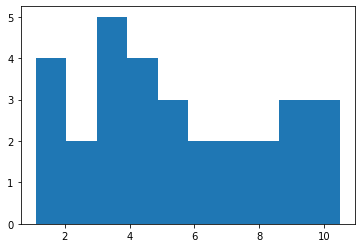
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Years experience | Years of experiance | Continuous | Relevant |
| Salary | Salary | Continuous | Relevant |

1. DATA CLEANSING :-
2. Dataset consists of 2 colums and 30 rows
3. Duplicate row does not exists
4. All data types are of form int64 and float 64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. There is positive linear relationship between columns as correlation coefficient is high
9. Both columns are positively skewed
10. From histogram data is normal
11. EDA:-
12. From box plot

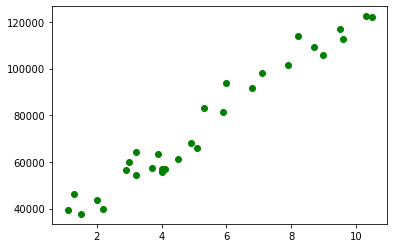


1. From histogram





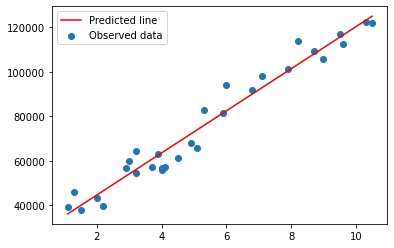
1. Scatter plot



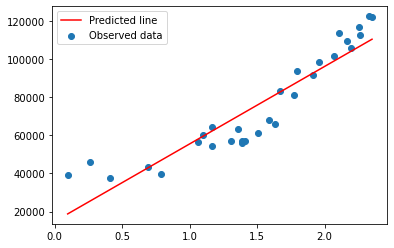
1. MODEL BUILDING:-
2. Model builded on simple , logarithmic , exponential and polynomial transformation
3. Output table for best transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Name of transformation | RMSE | Rsquare value | Correlation coeff |
| Simple linear regression | 5592.0436 | 0.957 | 9449.9623 |
| Logarathmic | 10302.893 | 0.854 | 4.058e+04 |
| Exponential | 7213.2350 | 0.932 | 0.1255 |
| Polynomial degree two | 5391.0815 | 0.949 | 0.2024 |

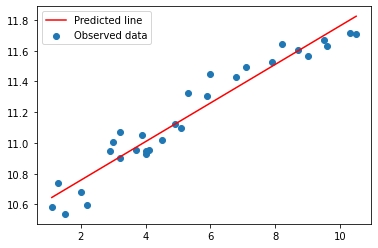
1. Best fit line for simple linear regression



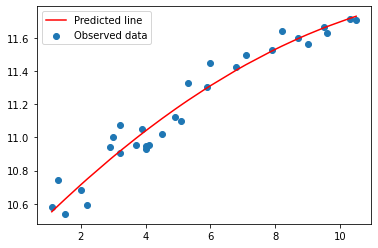
1. Best fit line for logarithmic transformation



1. Best fit line for exponential transformation



1. Best fit line for polynomial transformation



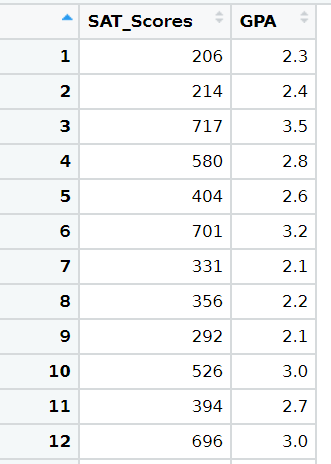
OUTPUT:-

1. Best fit model is polynomial transformation with low RMSE
2. Splitting data to train 80% and test 20%
3. Train accuracy :- 80271.3558
4. Test accuracy:- 82050.1506
5. It is right fit model
6. BENEFITS :-

From above information we can predict for salary against years of experience .

## **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.



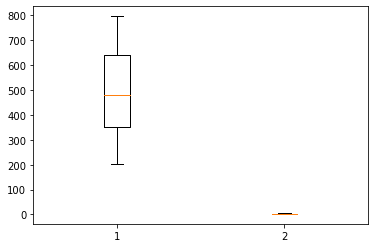
1. BUSINESS OBJECTIVE:-

Maximize relationship between variables

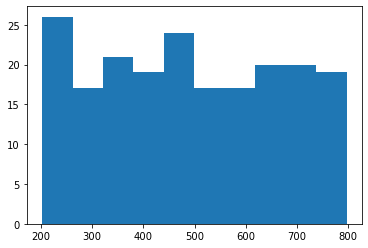
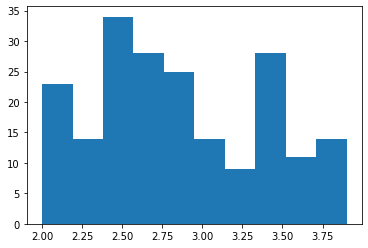
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Sat scores | Sat scores | Discrete | Relevant |
| Gpa | Gpa scores | Continuous | Relevant |

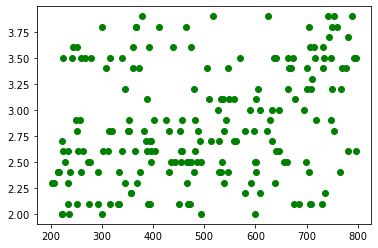
1. DATA CLEANSING :-
2. Dataset consists of 2 colums and 200 rows
3. Duplicate row exists and removed
4. All data types are of form int64 and float 64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. There is weak positive linear relationship between columns as correlation coefficient is low
9. Both columns are positively skewed
10. From histogram data is normal
11. EDA:-
12. From box plot



1. From histogram

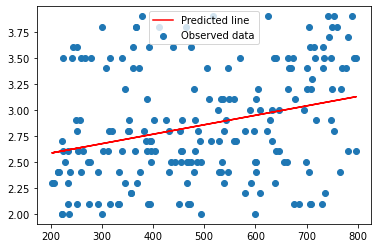
1. Scatter plot



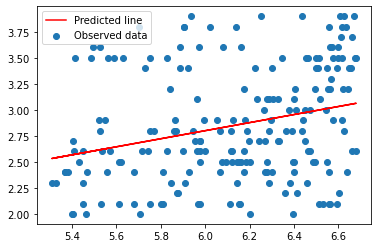
1. MODEL BUILDING:-
2. Model builded on simple , logarithmic , exponential and polynomial transformation
3. Output table for best transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Name of transformation | RMSE | Rsquare value | Correlation coeff |
| Simple linear regression | 0.5159 | 0.086 | 0.0009 |
| Logarathmic | 0.5184 | 0.077 | 0.3868 |
| Exponential | 0.5175 | 0.086 | 0.0003 |
| Polynomial degree two | 0.5144 | 0.094 | 6.142e-07 |

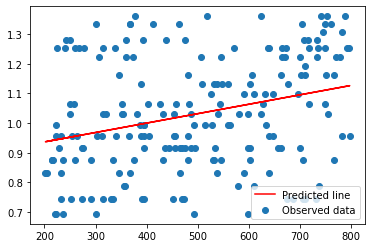
1. Best fit line for simple linear regression



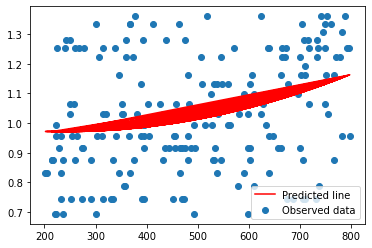
1. Best fit line for logarithmic transformation



1. Best fit line for exponential transformation



1. Best fit line for polynomial transformation



OUTPUT:-

1. Best fit model is polynomial transformation with low RMSE
2. Splitting data to train 80% and test 20%
3. Train accuracy :- 1.8906
4. Test accuracy:- 1.9130
5. It is not right fit model
6. BENEFITS :-

From above information we can predict for GPA against sat scores .