Topic(s): Simple Linear Regression

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: Simple Linear 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 the following: -

3. Exploratory Data Analysis (EDA):

3.1. Summary

3.2. Univariate analysis

3.3. Bivariate analysis

4. Data Pre-processing

4.1 Data Cleaning, Feature Engineering, etc.

4.2 Outlier Imputation

5. Model Building:

5.1 Perform Simple Linear Regression on the given datasets

5.2 Apply different transformations such as exponential, log, polynomial transformations and calculate RMSE values, R-Squared values,

Correlation Coefficient for each model

5.3 Build the models and choose the best fit model

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: -

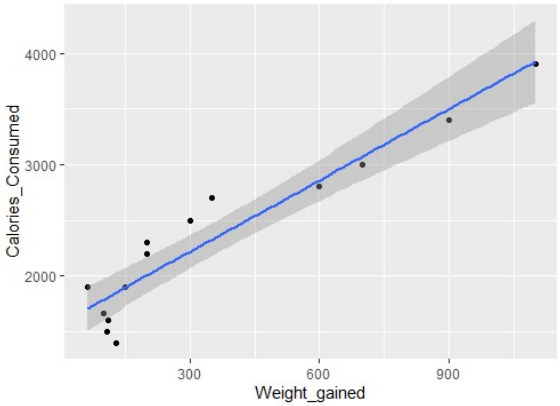
A certain food based company conducted a survey with the help of a fitness company spread across the country to find relationship between a person's weight gain and the no of calories consumed by them in order to come up a diet plan for individuals that fall under different weight groups. Approach - A Simple Linear regression model needs to be built with target variable 'Calories.Consumed'. Apply necessary transformations and record the RMSE values, Correlation coefficient values for different transformation models.

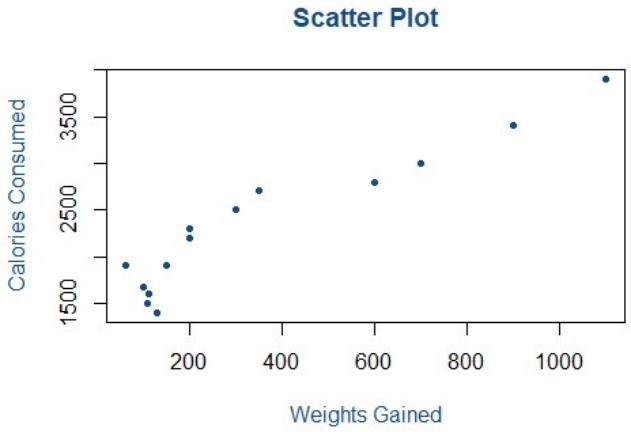
Sol:

Business Objective: To find the relationship between the persons weight gained and calories consumed.

Data Type: the given data is a numeric data which is continuous and the complete data is used for doing the analysis.

Exploratory data analysis: The normal distribution for the data is observed by using the box plots, histograms and QQ-Plots, after knowing the normal distribution in the data then the data is further used for simple regression models. Scatter plots and the GG plot for the given data is as follows.





Simple Regression: the given data is analyzed by taking output variable as calories consumed and the values of R^2, Co-relation coefficient, RMSE are recorded and then the transformations are done on the variables of the data. Finally polynomial linear regression gives the better output and the values of R^2, Co-relation coefficient, RMSE for the different models are as follows:

Output	Input	Correlation Co-efficient	R^2	RMSE	Method
Calories Consumed	Weight Gained	0.996	0.8882	232.8325	Simple regression
Calories Consumed	log(Weight Gained)	0.9368	0.8674	253.558	Logarthemic Regression
log(Calories Consumed)	Weight Gained	0.9306	0.7919	272.42	Exponential Regression
Calories Consumed	WG, WG^2	0.9207	0.8911	220.04	Quadratic regression

	Weight.gained..grams.	Calories.Consumed
1	108	1500
2	200	2300
3	900	3400
4	200	2200
5	300	2500
6	110	1600
7	128	1400
8	62	1900
9	600	2800
10	1100	3900

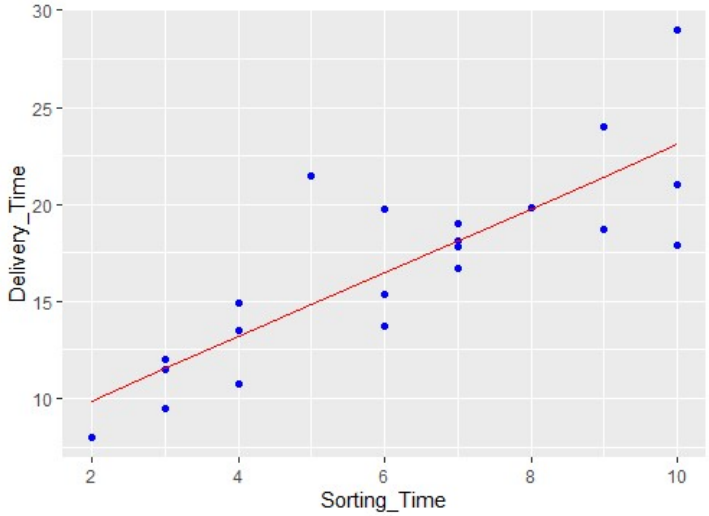
Problem Statement: -

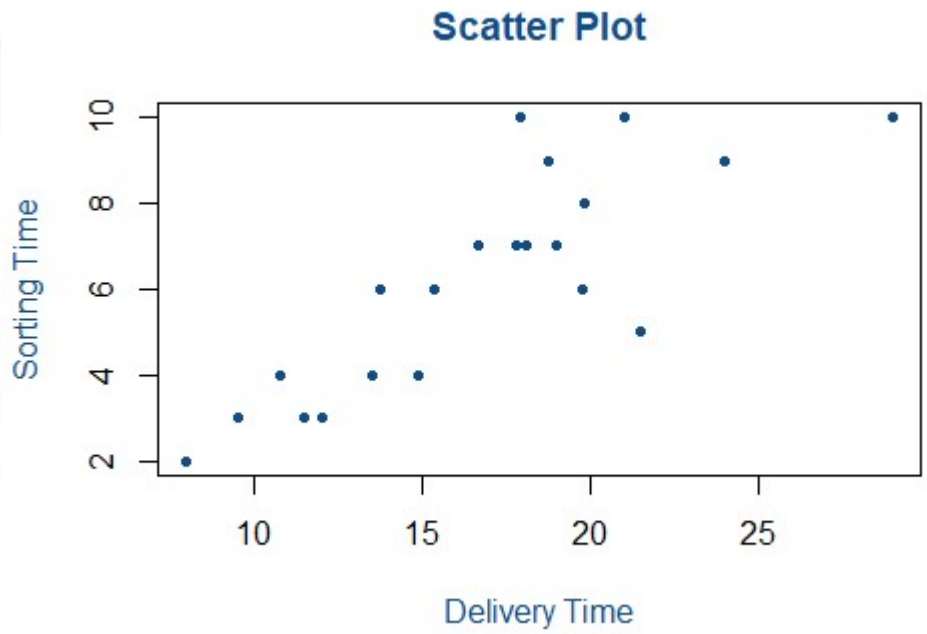
A food delivery service recorded the data of delivery time taken and the time taken for the deliveries to be sorted by the restaurants in order to improve their delivery services. Approach – A Simple Linear regression model needs to be built with target variable 'Delivery.Time'. Apply necessary transformations and record the RMSE values, Correlation coefficient values for different transformation models.

Business Objective: To find the relationship between the delivery time taken and service time for the package.

Data Type: the given data is a numeric data which is continuous and the complete data is used for doing the analysis.

Exploratory data analysis: The normal distribution for the data is observed by using the box plots, histograms and QQ-Plots, after knowing the normal distribution in the data then the data is further used for simple regression models. Scatter plots and the GG plot for the given data is as follows.





Simple Regression: the given data is analyzed by taking output variable as delivery time and the values of R^2, Co-relation coefficient, RMSE are recorded and then the transformations are done on the variables of the data. Finally polynomial linear regression gives the better output and the values of R^2, Co-relation coefficient, RMSE for the different models are as follows:

Output	Input	Correlation Co-efficient	R^2	RMSE	Method
Delivery Time	Sorting Time	0.8259973	0.6655	2.79165	Simple regression
Delivery Time	log(Sorting Time)	0.8339325	0.6794	2.733171	Logarthemic Regression
log(Delivery Time)	Sorting Time	0.808578	0.6957	2.94025	Exponential Regression
Delivery Time	ST, ST^2	0.8327302	0.6594	2.742148	Quadratic regression

	Delivery.Time	Sorting.Time
1	21.00	10
2	13.50	4
3	19.75	6
4	24.00	9
5	29.00	10
6	15.35	6
7	19.00	7
8	9.50	3
9	17.90	10
10	18.75	9
11	19.83	9

Problem Statement: -

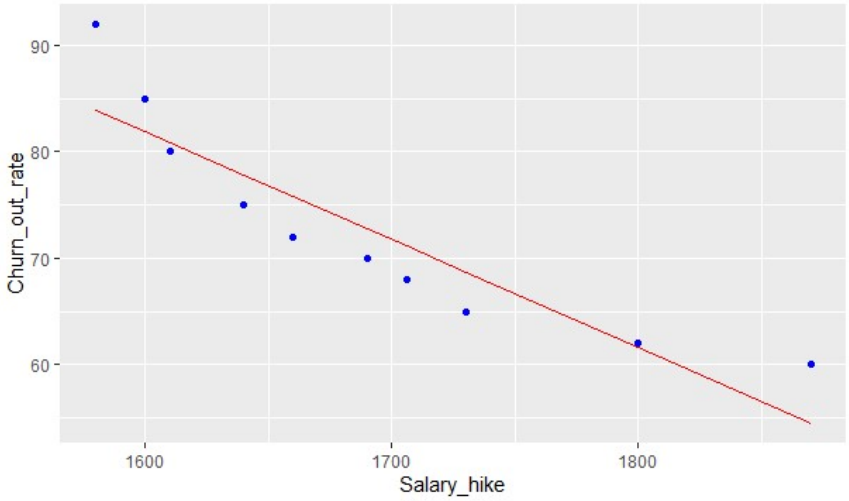
A certain organization wanted an early estimate of their employee churn out rate. So, the HR department came up with data regarding the employee's salary hike and churn out rate for a financial year. The analytics team will have to perform a deep analysis and predict an estimate of employee churn and present the statistics. Approach –A Simple Linear regression model needs to be built with target variable 'Churn_out_rate'. Apply necessary transformations and record the RMSE values, Correlation coefficient values for different transformation models.

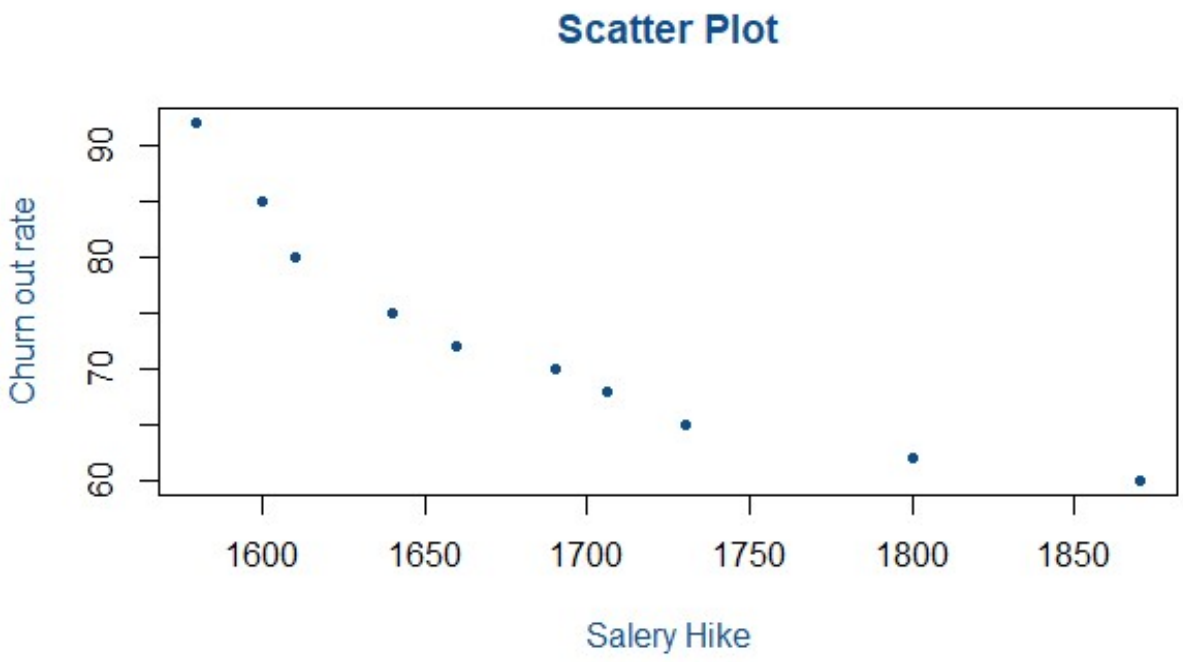
Sol:

Business Objective: To find the relationship between the salary hike and churn out rate of the employee.

Data Type: the given data is a numeric data which is continuous and the complete data is used for doing the analysis.

Exploratory data analysis: The normal distribution for the data is observed by using the box plots, histograms and QQ-Plots, after knowing the normal distribution in the data then the data is further used for simple regression models. Scatter plots and the GG plot for the given data is as follows.





Simple Regression: the given data is analyzed by taking output variable as churn out rate and the values of R^2, Co-relation coefficient, RMSE are recorded and then the transformations are done on the variables of the data. Finally polynomial linear regression gives the better output and the values of R^2, Co-relation coefficient, RMSE for the different models are as follows:

Output	Input	Correlation Co-efficient	R^2	RMSE	Method
Churn out rate	Salary hike	0.9117216	0.8101	3.997528	Simple regression
Churn out rate	log(Salary hike)	0.9212077	0.8297	3.786004	Logarthemic Regression
log(Churn out rate)	Salary hike	0.9334219	0.8577	3.541549	Exponential Regression
log(Churn out rate)	SH, SH^2	0.9907286	0.9789	1.32679	Quadratic regression

	Salary_hike	Churn_out_rate
1	1580	92
2	1600	85
3	1610	80
4	1640	75
5	1660	72
6	1690	70
7	1706	68
8	1730	65
9	1800	62
10	1870	60

Problem Statement: -

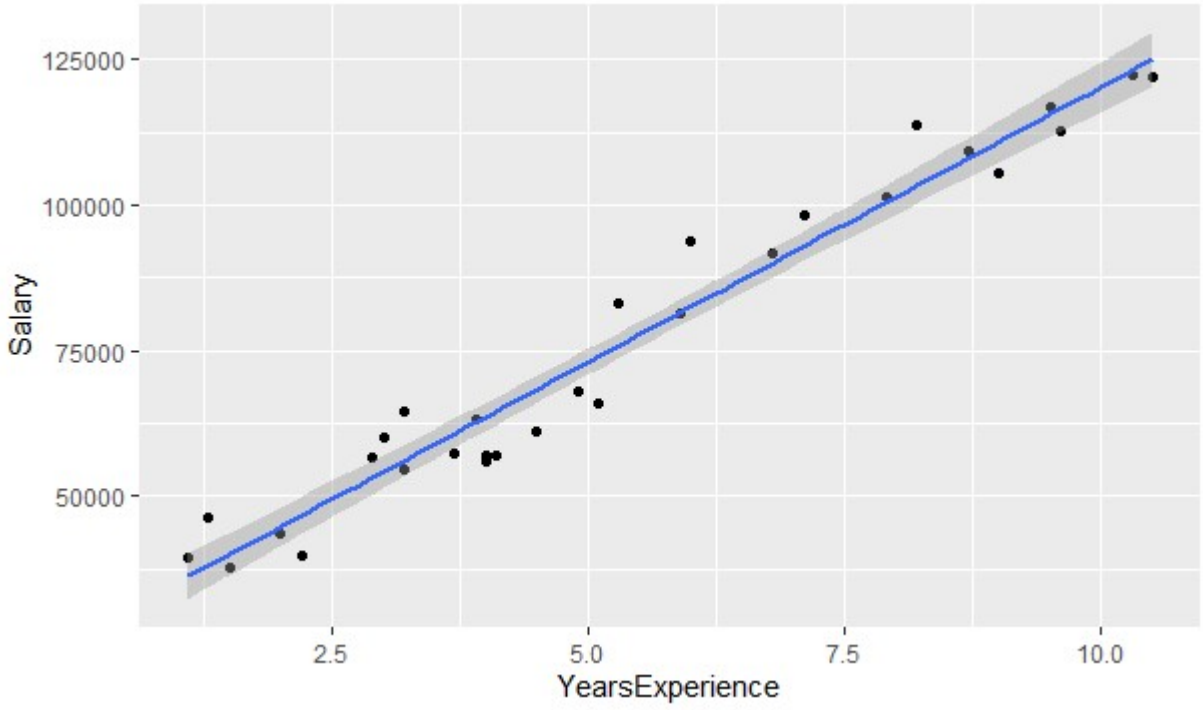
The Head HR of a certain organization wants to automate their salary hike estimation. The organization consulted an analytics service provider and asked them to build a basic prediction model by providing them with a sample data that contains historic data of the years of experience and the salary hike given accordingly over the past years. Approach - A Simple Linear regression model needs to be built with target variable 'Salary' to predict the salary hikeapply necessary transformations and record the RMSE values, Correlation coefficient values for different transformation models.

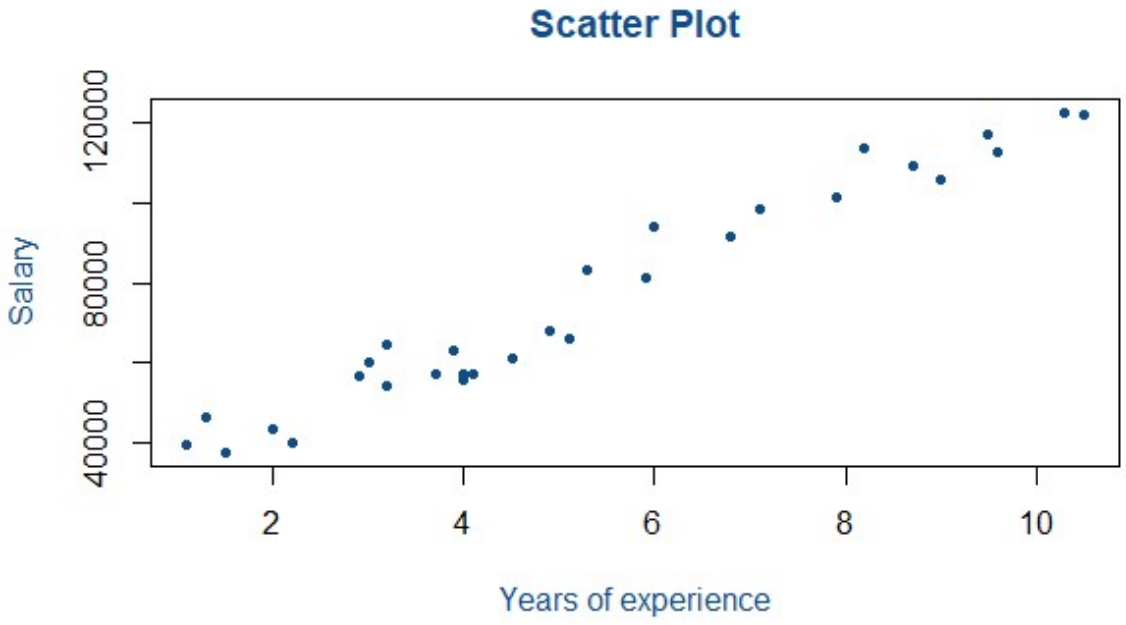
Sol:

Business Objective: To find the relationship between the number of years of experience and salary of the employee.

Data Type: the given data is a numeric data which is continuous and the complete data is used for doing the analysis.

Exploratory data analysis: The normal distribution for the data is observed by using the box plots, histograms and QQ-Plots, after knowing the normal distribution in the data then the data is further used for simple regression models. Scatter plots and the GG plot for the given data is as follows.





Simple Regression: the given data is analyzed by taking output variable as salary and the values of R^2, Co-relation coefficient, RMSE are recorded and then the transformations are done on the variables of the data. Finally polynomial linear regression gives the better output and the values of R^2, Co-relation coefficient, RMSE for the different models are as follows:

Output	Input	Correlation Co-efficient	R^2	RMSE	Method
Salary	Yesrs of Experience	0.9782416	0.9554	5592.044	Simple regression
Salary	log(Yesrs of Experience)	0.9240611	0.8487	10302.89	Logarthemic Regression
log(Salary)	Yesrs of Experience	0.966047	0.9295	7213.235	Exponential Regression
Salary	YE,YE^2	0.9782511	0.9538	5590.841	Quadratic regression

	YearsExperience	Salary
1	1.1	39343
2	1.3	46205
3	1.5	37731
4	2.0	43525
5	2.2	39891
6	2.9	56642
7	3.0	60150
8	3.2	54445
9	3.2	64445
10	3.7	57189
11	3.9	63219

Problem Statement: -

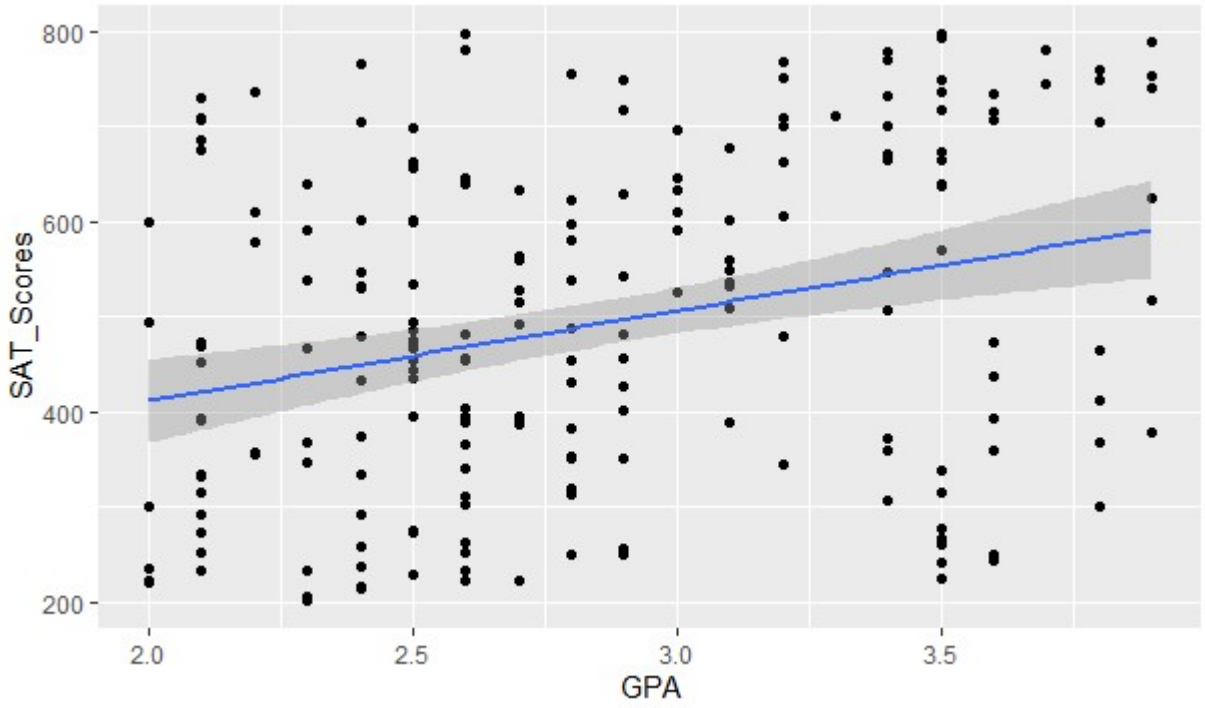
A student from a certain University was asked to prepare a dataset and build a prediction model for predicting SAT scores based on the exam giver's GPA. Approach - A regression model needs to be built with target variable 'SAT_Scores'and record the RMSE values, Correlation coefficient values for different transformation models.

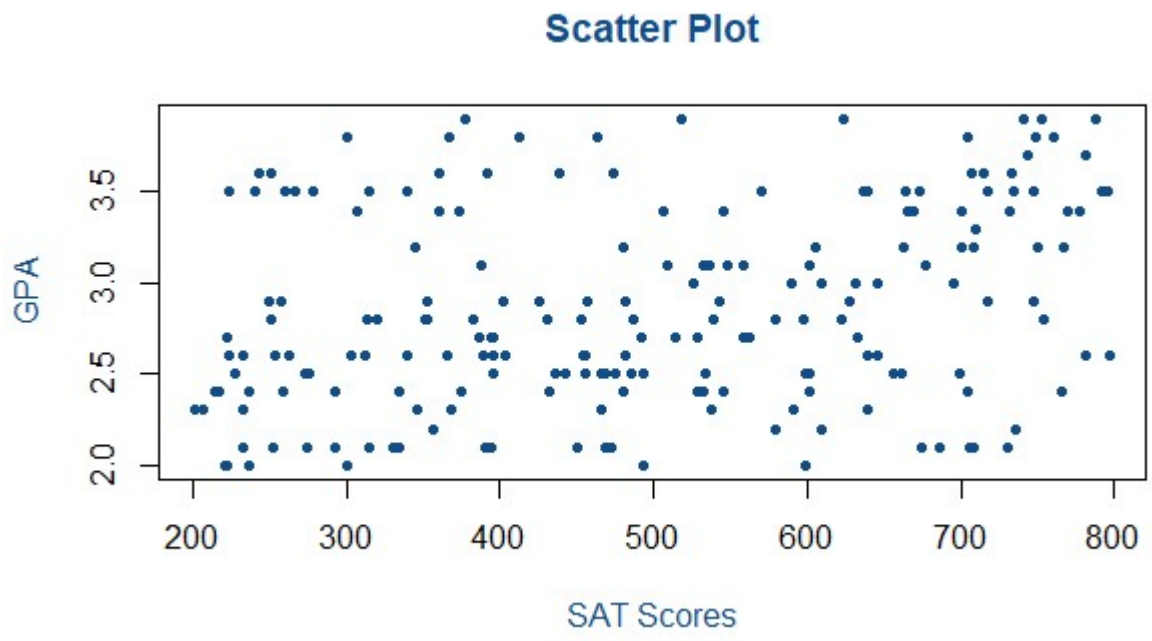
Sol:

Business Objective: To find the relationship between the sat scores and GPA of the students.

Data Type: the given data is a numeric data which is continuous and the complete data is used for doing the analysis.

Exploratory data analysis: The normal distribution for the data is observed by using the box plots, histograms and QQ-Plots, after knowing the normal distribution in the data then the data is further used for simple regression models. Scatter plots and the GG plot for the given data is as follows.





Simple Regression: the given data is analyzed by taking output variable as sat scores and the values of R^2, Co-relation coefficient, RMSE are recorded and then the transformations are done on the variables of the data. Finally polynomial linear regression gives the better output and the values of R^2, Co-relation coefficient, RMSE for the different models are as follows:

Output	Input	Correlation Co-efficient	R^2	RMSE	Method
Sat Scores	GPA	0.2935383	0.08155	166.7708	Simple regression
Sat Scores	log(GPA)	0.2940842	0.08187	166.7415	Logarthemic Regression
log(Sat Scores)	GPA	0.2918209	0.07247	169.6691	Exponential Regression
Sat Scores	GPA,GPA^2	0.9782511	0.07096	169.4982	Quadratic regression

	SAT_Scores	GPA
1	206	2.3
2	214	2.4
3	717	3.5
4	580	2.8
5	404	2.6
6	701	3.2
7	331	2.1
8	356	2.2
9	292	2.1
10	526	3.0
11	394	2.7
12	696	3.0