# 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:RAKESH SETHU NP Batch ID: 11052022\_7.30PM**

**Topic: Simple Linear Regression**

**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.**

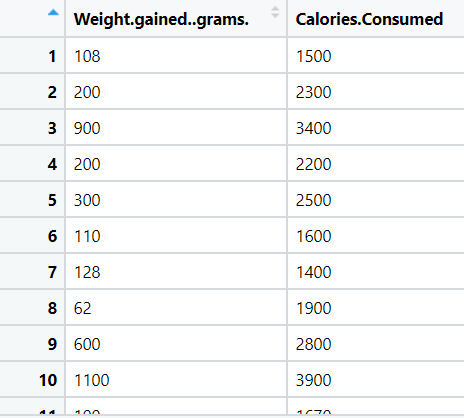
**4.2 Univariate analysis.**

**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.

 **Object:** Maximize the accuracy in modelling relationship between 2 variables.

**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Weight Gained grams | Persons weight gain | Ratio | Relevant, Provides useful info. |
| Calories consumed | Number of calories consumed by the people | Ratio | Relevant, Provides useful info. |

**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.



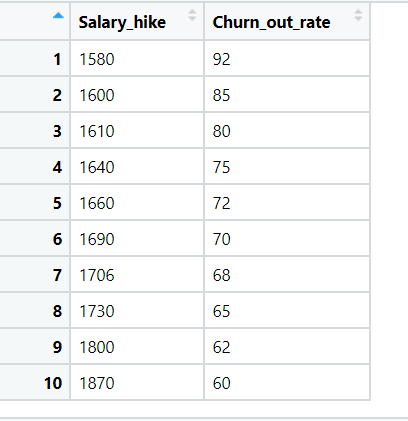
**Object:** Maximize the accuracy in modelling relationship between 2 variables.

**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Delivery time | Time needed to deliver an item. | Ratio | Relevant, Provides useful info. |
| Sorting time | Time taken to sort deliveries by restaurant. | Ratio | Relevant, Provides useful info. |

**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.



**Object:** Maximize the accuracy in modelling relationship between 2 variables.

**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Salary hike | Hike in salary of employees | Ratio | Relevant, Provides useful info. |
| Churn rate out | Number of employees leaving the company over a period of time. | Discrete | Relevant, Provides useful info. |

**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.



**Object:** Maximize the accuracy in modelling relationship between 2 variables.

**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Years of Experience | Historic data of years of experience of a candidate. | Ratio | Relevant, Provides useful info. |
| Salary | Historic data of Salary of a candidate. | Ratio | Relevant, Provides useful info. |

## **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.

**Object:** Maximize the accuracy in modelling relationship between 2 variables.

**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| SAT Scores | SAT scores of a Student. | Ratio | Relevant, Provides useful info. |
| GPA | GPA scores of a student | Ratio | Relevant, Provides useful info. |

