# 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: Nithin DSouza**

**Batch Id: 05012021-10AM**

**Topic: Simple Linear Regression**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**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: -**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Data Pre-processing**

**4.1 Data Cleaning, Feature Engineering, etc.**

**4.2 Outlier Imputation**

1. **Model Building:**
   1. **Perform Simple Linear Regression on the given datasets**
   2. **Apply different transformations such as exponential, log, polynomial transformations and calculate RMSE values, R-Squared values, 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. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

# Note:

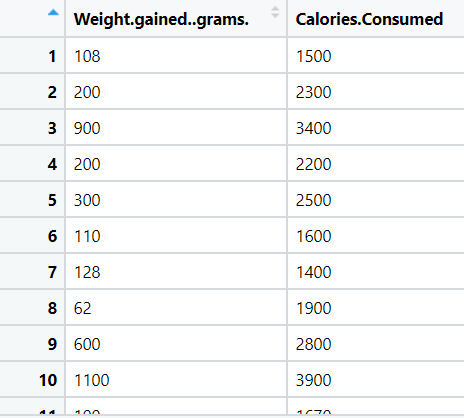
**The assignment should be submitted in the following format:**

* **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.



**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 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.



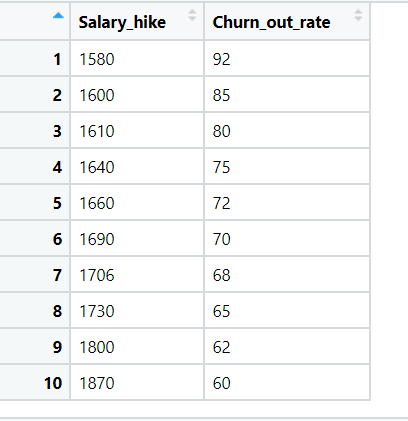
**Object:** Maximize the accuracy in modelling relationship between 2 variables.

**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 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.



**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 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.



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

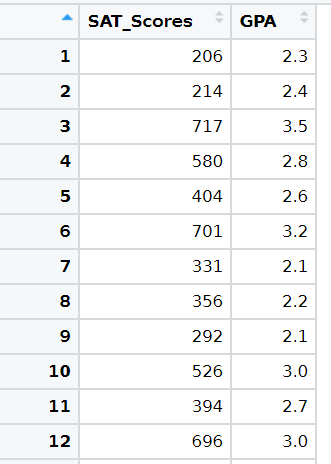
**Constraint:** Minimal availability of linear problems in real world.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 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.

## 



**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. |