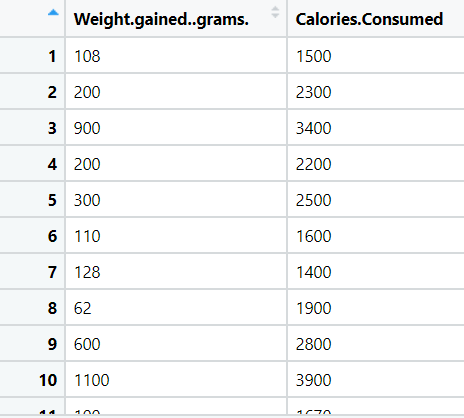
# Simple Linear Regression (Module - 6)

## Do the necessary transformations for input variables for getting better R^2 value for the model prepared. Build the model and predict for the output variables.

1.) Calories\_consumed -> predict weight gained using calories consumed



**Ans:**

**Inferences from Graphical Representation:**

* From the **boxplot** representation we can say that there are no outliers exist.
* From the histogram the data is right skewed distribution for weight gain

**Inferences from different models:**

* Coefficient Correlation = 0.95

1. **Simple Linear Regression / Base Model:**

* R^2 = 0.896 and Adjusted R^2 = 0.88
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.95
* RMSE = 103.30

1. **input = log(x); output = y**

* R^2 = 0.80 and Adjusted R^2 = 0.79
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.89
* RMSE = 144.00

1. **input = x; output = log(y)**

* R^2 = 0.87 and Adjusted R^2 = 0.86
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.94
* RMSE = 118.04

1. **input = x & x^2 (2-degree) and output = log(y)**

* R^2 = 0.87 and Adjusted R^2 = 0.85
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.94
* RMSE = 117.41

From the above observation the base model is the best fit model as RMSE value is less

**Data Partition:**

Here the data is divided into train and test in the ratio of 80:20

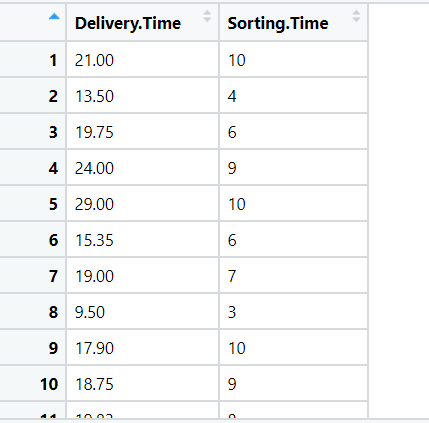
With the 95% confidence interval level the below are the train and test RMSE

Test RMSE = 97.77

Train RMSE = 104.27

With the above observation we can say that the model is **right fit**.

1. Delivery\_time -> Predict delivery time using sorting time



**Ans:**

**Inferences from Graphical Representation:**

* From the **boxplot** representation we can say that there are no outliers exist.
* From the histogram the data is normally distribution

**Inferences from different models:**

* Coefficient Correlation = 0.83

1. **Simple Linear Regression / Base Model:**

* R^2 = 0.68 and Adjusted R^2 = 0.66
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.825
* RMSE = 2.79

1. **input = log(x); output = y**

* R^2 = 0.70 and Adjusted R^2 = 0.68
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.83
* RMSE = 2.73

1. **input = x; output = log(y)**

* R^2 = 0.71 and Adjusted R^2 = 0.69
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.80
* RMSE = 2.94

1. **input = x & x^2 (2-degree) and output = log(y)**

* R^2 = 0.76 and Adjusted R^2 = 0.73
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.82
* RMSE = 2.79

From the above observation the base log(sorting time) and delivery time(transformation model) is the best fit model as RMSE value is less

**Data Partition:**

Here the data is divided into train and test in the ratio of 80:20

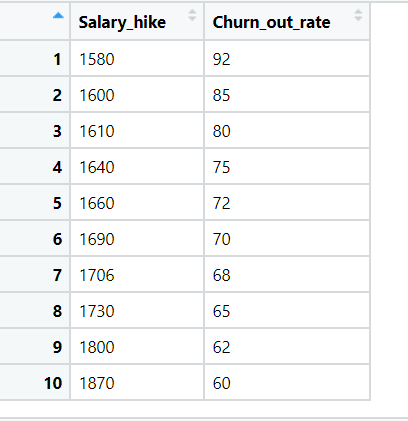
With the 95% confidence interval level the below are the train and test RMSE

Test RMSE = 3.3

Train RMSE = 2.8

With the above observation as there is slight variation in the RMSE, we can say that the model is **right fit**.

1. Emp\_data -> Build a prediction model for Churn\_out\_rate



**Ans:**

**Inferences from Graphical Representation:**

* From the **boxplot** representation we can say that there are no outliers exist.
* From the histogram the data is right skewed for Churn out rate
* From scatter plot representation its shows negatively correlated

**Inferences from different models:**

* Coefficient Correlation = -0.911

1. **Simple Linear Regression / Base Model:**

* R^2 = 0.83 and Adjusted R^2 = 0.81
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.911
* RMSE = 3.99

1. **input = log(x); output = y**

* R^2 = 0.85 and Adjusted R^2 = 0.83
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.92
* RMSE = 3.78

1. **input = x; output = log(y)**

* R^2 = 0.87 and Adjusted R^2 = 0.85
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.93
* RMSE = 3.54

1. **input = x & x^2 (2-degree) and output = log(y)**

* R^2 = 0.98 and Adjusted R^2 = 0.97
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.99
* RMSE = 1.32

From the above observation the base Polynomial Model(transformation model) is the best fit model as RMSE value is less

**Data Partition:**

Here the data is divided into train and test in the ratio of 80:20

With the 95% confidence interval level the below are the train and test RMSE

Test RMSE = 0.977

Train RMSE = 1.39

With the above observation as there is slight variation in the RMSE, we can say that the model is **right fit**.

## Salary\_hike -> Build a prediction model for Salary\_hike



**Ans:**

**Inferences from Graphical Representation:**

* From the **boxplot** representation we can say that there are no outliers exist.
* From the **histogram** the data is normally distributed
* From scatter plot representation its shows positively correlated

**Inferences from different models:**

* Coefficient Correlation = 0.97

1. **Simple Linear Regression / Base Model:**

* R^2 = 0.95 and Adjusted R^2 = 0.95
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.97
* RMSE = 5592.044

1. **input = log(x); output = y**

* R^2 = 0.85 and Adjusted R^2 = 0.84
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.92
* RMSE = 10302.89

1. **input = x; output = log(y)**

* R^2 = 0.93 and Adjusted R^2 = 0.92
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.96
* RMSE = 7213.23

1. **input = x & x^2 (2-degree) and output = log(y)**

* R^2 = 0.94 and Adjusted R^2 = 0.93
* As p-value < 0.05 the coefficients are significant
* Coefficient Correlation = 0.97
* RMSE = 6817.017

From the above observation the base Simple Linear Regression model is the best fit model as RMSE value is less

**Data Partition:**

Here the data is divided into train and test in the ratio of 80:20

With the 95% confidence interval level the below are the train and test RMSE

Test RMSE = 6680.911

Train RMSE = 5293.044

With the above observation as there is variation in the RMSE, we can say that the model is **Over fit as train error is low and test error is high**

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without errors
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.