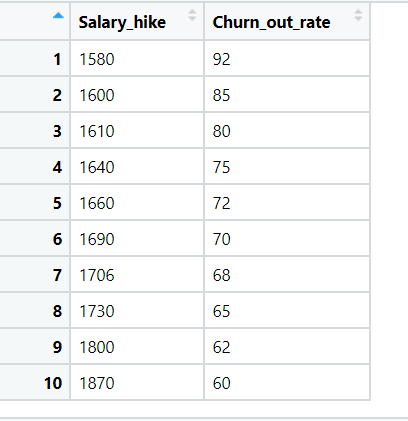
# Topic(s): Simple Linear Regression

**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.



**Business Problem**

**Objective :-** predict Churn\_out\_rate’

**Python code details :**

Data Frame name is churn. It has 10 rows and 2 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

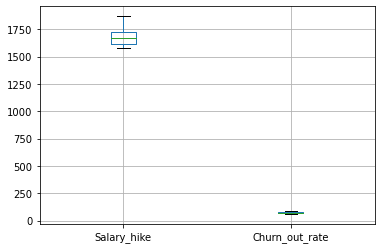
Salary\_hike 10

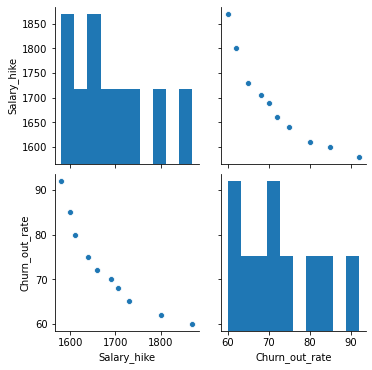
Churn\_out\_rate 10

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .there is no outliers present.

Boxplot:-



Histogram and Scatter plot:- 

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**

We build ML models that predict calories consumed gained using weight. here y is continuous .so we tried simple linear regression with different transformation.

**Result = MODEL RMSE R-squared Adj. R-squared Correlationcoefficient**

**SLR 3.99 0.831 0.810 0.911**

**Log model 3.78 0.848 0.829 0.934**

**Exp model 3 .54 0.873 0.857 0.921**

**Poly model 1.39 0.983 0.978 0.89**

**Model details saved Model\_details.**

Applying transformation is decreasing Multiple R Squared Value. So model does not need further transformation, Multiple R-squared: 0.98.

Model reg is best fit model= SLR

,root mean square error 1.39

R\_squared 0.983

Adj. R-squared 0.978 .

We split the data 50-50 in train and test.

So , we build our final model on SLR build on train data and test on test data .

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 50% data on train and 50% test . Preparing a Simple linear regression on training data set

R-squared: 0.998

Adj. R-squared: 0.995

,then test on test data , Evaluation on Test Data as result root mean square error=2.59.

Evaluation on Train Data also as result root mean square error=0.48.

* used library :- pandas, numpy, sk learn, matplotlib, statsmodels