**Business Objective: Calories Consumed -> predict weight gained using calories consumed**

* Input variable x = Calories consumed; Output variable y = Weight gained
* Scatter Plot: Direction is Positive, Strength is Moderate, Linear
* Correlation Coefficient r =0.94699 >0.85
* There is a strong Correlation with input variable. Calories consumed are a good input to predict the output Weight gained.
* Linear model regression and P value is 4.542e-05<0.05
* R Squard= 0.8968 >0.8
* Y= -625.75236+ 0.42016 (1500) this gives only 1 value so probability of 1 variable is zero. So we
* Need to get the range of values to compare the predicted value with the actual value.
* 4.482599-1500 = error.
* Residual value = 103.3025
* Transform the variables to check whether the predicted values are better using sqrt function.
* R- Squared = 0.8567
* 2.5 % 97.5 %
* (Intercept) -2080.95760 -1073.80123
* sqrt(Calories. Consumed) 30.05875 50.8759
* Transform the data using log of calories consumed and calculate the correlation and plot the data and
* do the linear regression method.
* Residuals:
* Min 1Q Median 3Q Max
* -0.86562 -0.10529 0.02403 0.13541 0.42759
* Coefficients:
* Estimate
* (Intercept) 2.829e+00
* Calories. Consumed 1.142e-03

**Business Observation: Delivery\_time -> Predict delivery time using sorting time**

* X = Sorting time Y = Delivery time
* Input variable x = Calories consumed; Output variable y = Weight gained
* According to Boxplot there are no outliers in Delivery time and Sorting time.
* Histogram: dataset in Sorting time is random; data is Delivery time dataset is got positively Skewed. Concentrated towards one side.
* QQ plot : dataset of Sorting time is following normal distribution moderately ; dataset of Delivery time is following normal distribution
* Scatter Plot: Direction is Positive, Strength is Moderate, Linear
* Correlation Coefficient r =0.82599 >0.85 so the strength is moderate
* Multiple R-squared: 0.6823<0.8 ; Adjusted R-squared: 0.6655
* So the difference between Multiple R-squared and Adjusted R-squared is less so sorting time is a good input
* variable to predict the output Delivery time.
* Linear model regression and P value is 3.98e -06<0.05;
* Correlation Coefficient pred =0.82599
* Square root Transformation:
* After transforming the data with sqrt of sorting time there is slight increase in the r value =0.83415
* Now transform the variables to check whether the predicted values are better using Sqrt:
* Multiple R-squared: 0.6958, Adjusted R-squared: 0.6758 Slight increase in the R square value.
* There is increase in the correlation coefficient value r =0.83415
* Log Transformation:
* Scatter plot seems to improve after the log transformation on sorting time.
* r = 0.8339 seems to increase very slightly and also the R-squared value =0.6954, P value= 2.642e-06
* after log transformation the value of R-squared = 0.7649 ; Residual value also decreased =0.1505

**Business Observation: Emp\_data -> Build a prediction model for Churn\_out\_rate**

* X = Salary\_hike Y = Churn\_out\_rate
* According to Boxplot there are no outliers in Salary\_hike and Churn\_out\_rate.
* Histogram: dataset in Salary\_hike is distributed on left ,positive Skewed and there seem to be an outlier; data is Churn\_out\_rate dataset is got positively Skewed. Concentrated towards one side also has an outlier.
* QQ plot : dataset of Salary\_hike is following normal distribution moderately ; dataset of Churn\_out\_rate is following normal distribution
* Scatter Plot: Direction is Negative, Strength is Moderate, Linear
* Correlation Coefficient |r| = -0.9117=0.9117 >0.85 so the strength is high
* Multiple R-squared: 0.8312, Adjusted R-squared: R Squared = 0.8312 >0.8
* So the difference between Multiple R-squared and Adjusted R-squared is less so sorting time is a good input
* Residual = 3.9975 which Is higher than normal P value = 0.00023 < 0.05
* Correlation coefficient r value using predicted value is 0.9117 >0.85.
* Scatter plot for predicted value is negative and linear.
* Now transform the variables to check whether the predicted values are better .
* Sqrt of regression improves the R Squared value = 0.84 > 0.8
* r value of Sqrt of salary\_hike = -0.916 ; Residual = 3.89
* r value of log of salary\_hike = -0.921 ; R Squared =0.8486 >0.8 ; Residual =3.786 => log transformation is improved the values above.
* Linear model regression and P value is 0.000136<0.05;
* R Squared value =0.853>0.8
* log of regression improves the R Squared value = 0.8735 > 0.8 ; P value=7.377e-05

**Business Observation: Salary\_hike -> Build a prediction model for Salary\_hike**

* X = Years of experience Y = Salary
* According to Boxplot there are no outliers in Salary and Years of experience.
* Histogram: dataset in Salary is distributed on left ,positive Skewed and there is no outlier; data is YearsExperience dataset is got positively Skewed. Concentrated towards one side also has no outlier
* QQ plot : dataset of Salary is following normal distribution moderately ; dataset of YearsExperience is following normal distribution
* Scatter Plot: Direction is Positive, Strength is Moderate, Linear
* Correlation Coefficient |r| = 0.9782 >0.85 so the strength is high
* Multiple R-squared: 0.957, Adjusted R-squared: R Squared = 0.955 >0.8 P value = 2.2e-16
* So the difference between Multiple R-squared and Adjusted R-squared is less so sorting time is a good input
* Residual = 5592.044 which Is higher than normal P value = 0.00023 < 0.05
* Correlation coefficient r value using predicted value is 1 >0.85.
* Scatter plot for predicted value is Positive and linear.
* Now transform the variables to check whether the predicted values are better .
* Sqrt of regression improves the R Squared value = 0.931> 0.8
* r value of Sqrt of YearsExperience = -0.916 ; Residual = 7080.096
* r value of log of YearsExperience = 1; R Squared =0.986>0.8 ; Residual =10302 => log transformation is improved the values above. P value is 2.2e-16<0.05;