Linear Regression Subjective Questions

Assignment based subjective questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable.**

   Based on boxplot following are the interpretation about the effect of categorical variable on the dependent variable (cnt) from the dataset

   1. **Season** – Bike rentals are more during **summer and fall**.
   2. **Month** – From **March till September** the bike rentals are higher than the other months (**Oct-Feb**)
   3. **Weathersit** – There is no bike rental during '**Heavy snow rain**'. Bike rental is high when it is '**Clear'** and it is slowly getting reduced once the weather changes to '**Cloudy'** and '**Light Snow Rain**'
   4. **Weekday** – The spread of data is high on '**Friday'** compared to other days.
   5. **Holiday** – The spread of data is high on '**Holiday'** compared to '**workingday'**

2. **Why it is important to use drop_first = True during dummy variable creation?**

   Yes, it is important to use **drop_first = True** when creating dummy variable.

   Consider if a college offers a Data Science course in full-time and part-time. If you have a feature called full-time, you do not need a column part-time because if full-time = 1 denotes the student is doing the

course on full-time basis. If it is ‘0’ it means the student is pursuing the course in part time. So, we don’t need a separate feature (part-time) to identify it. It can be identified using the single features ‘full-time’.

In our assignment, for ‘season’ rather having four columns spring, summer, fall and winter we can have three columns(spring,summer,fall). If spring =0, summer=0 and fall = 0, it means the season is ‘winter’. There is no need to have a separate column ‘winter’. The same applies for all the other categorical variables (mnth, weekday, weathersit). So using **drop_first = True** (drops the first dummy column), we are dropping the column which is not necessary.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   From the pairplot, we can say that the variable ‘temp’ has the highest correlation with the target variable ‘cnt’.

4. **How did you validate the assumptions of Linear regression after building the model on the training set?**

   Assumption: The error terms should be normally distributed with mean ‘0’.
   We have to find the error terms which is the difference between the actual y-value of the train data and the predicted y-value of the train data (calculate the predicted y-value from the model which we have created).

Plot the error terms in a distplot. The error terms should be normally distributed with mean ‘0’.

Predicting the y-value: (from the assignment)

```
y_train_pred = lr_4.predict(X_rfe_lm_4)
```

y_train_pred – Predicted y-value

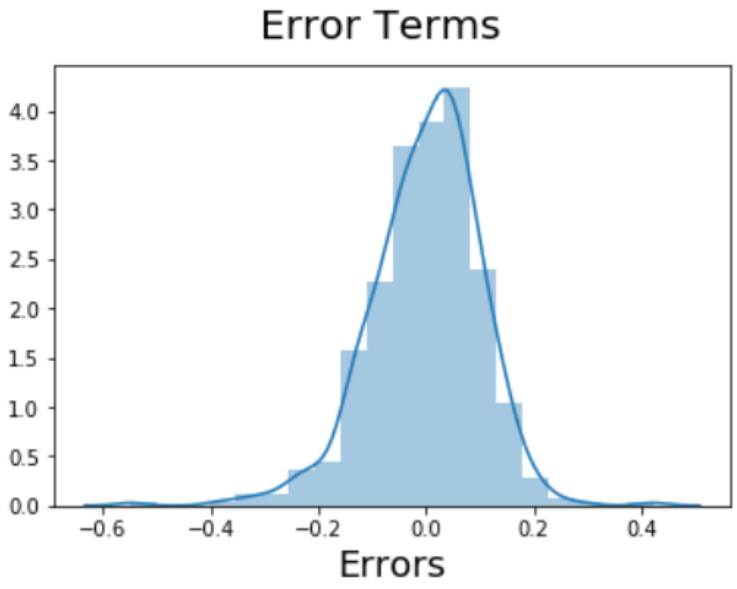
lr_4 – linear regression model

X_rfe_lm_4 – X-values

Plotting the distplot: (from the assignment)

Error term = y_train – y_train_pred

```
sns.distplot((y_train - y_train_pred), bins = 20)
```



Error terms in the above figure is normally distributed with mean '0'. This is how we validate the assumption of Linear regression after our model is built.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

The top 3 features (with their **dummy variables**(explained below) for **Weathersit** and **Season**) that contribute significantly in explaining the demand of the shared bikes are Season, Weathersit and Windspeed.

1. **Weathersit:** The Dummy variable **'Light snow rain'** shows negative correlation with the bike rental count which says, if there is Light snow rain the '**cnt'** of bike sharing decreases and vice -versa. In this case, if Weathersit is 3 and above the count of bike rental gets affected.

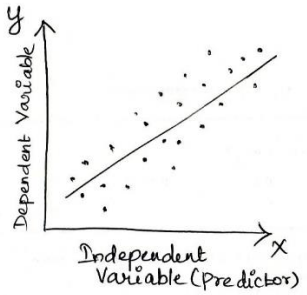
2. **Season:** Comparing all the four seasons, the count of bike rental is more during **fall(September)** and **summer** than in **winter** and **spring**.

3. **Windspeed:** The speed of the wind plays a role in deciding the count of bike rental. If the speed of the wind is less, the bike rental count is more and vice versa.

General Subjective Questions

1. **Explain Linear Regression Algorithm in detail.**

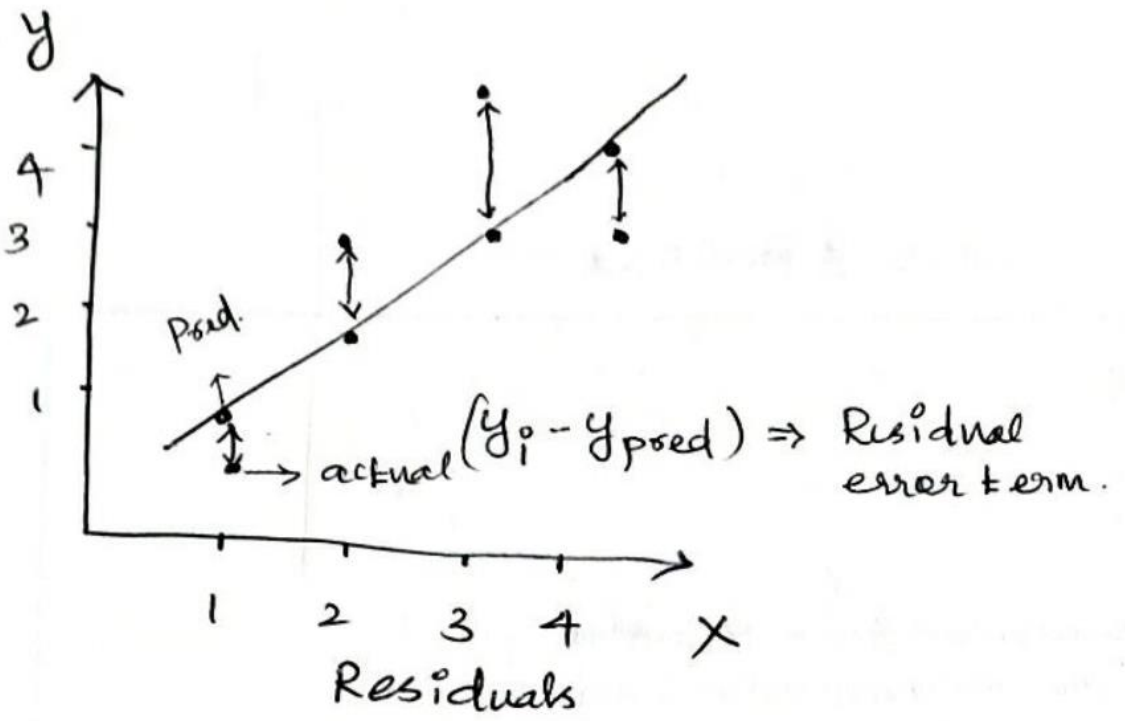
Linear Regression is a **Machine Learning Algorithm** which is based on **"Supervised Learning"** (The output variable to be predicted is either continuous or categorical and they have labels). This is a method of modelling the target variable based on the independent variable(predictors).
Simple linear regression (one independent variable)/Multiple linear regression (more than one independent variable) are the type of regression analysis where there should be a linear relationship between the independent variable (x) and dependent(y) variable. The line in the below fig is referred to as the **"best fit line".**
The equation of the linear regression is represented by



$$y = \beta_0 + \beta_1 x + \varepsilon$$

where $y \rightarrow$ dependent Variable
$x \rightarrow$ independent Variable
$\beta_0 \rightarrow$ intercept
$\beta_1 \rightarrow$ Slope
$\varepsilon \rightarrow$ Error terms.

An increase in the value of X by '1' unit, there will be an increase in y. We have to find the best line fit. For, that we need to find the Residuals.



It is denoted by $e_i = y_i - y_{pred}$

$e_i \rightarrow$ error

$y_i - y_{pred} \rightarrow$ diff between Original & predicted Value.

Now, **Ordinary Least Square** Method is used.

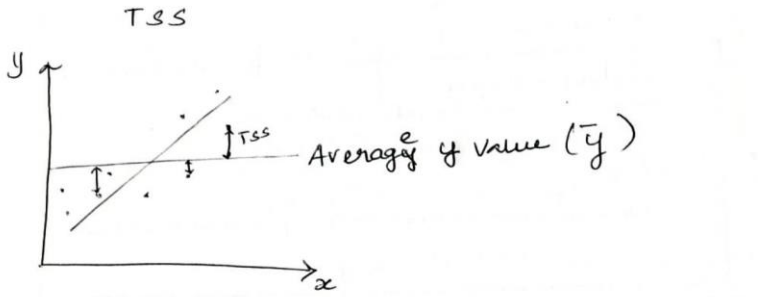
$e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2 = RSS$

This means minimising the Sum of Squares.

The best fit line is identified by the minimising residual sum of squares(RSS).

To find the best model, we need R-Squared value.

$$R^2 = 1 - (RSS/TSS)$$



TSS → difference between the Datapoint and the average value.

$$TSS = \Sigma (y_p - \bar{y})^2$$

$$R^2 = 1 - (RSS/TSS)$$

If the value of R^2 is $\simeq 1$ then, we can say that the line is a best fit line.

The value of R^2 depends on β_0 and β_1.

At the backend, the cost function is used to find the best possible values for b0(beta 0) and b1(beta 1). The significant thing behind this is the 'Gradient Descent method' which is used find the best line fit. The idea is that it starts with some values for b0(beta_0) and b1(beta1) and then we change these values iteratively to reduce the cost.

The entire data will be divided into train and test set. Build the model using the train dataset. Test the model using the test data set. The R-squared value of the train and test dataset should be approximately equal or there should be minimal difference between them. If so, it is considered as best model else we need to redesign our model.

(**Note**: In case of Multiple Linear Regression, the number independent variables should be less (should not be one) so that our model is more reliable and simpler model.)

2. **Explain the Anscombe's Quartet in detail**

   Before running a linear regression analysis, we need to analyse the complete dataset. Sometimes, the statistical summary of the data is not enough to come to a conclusion. We need to analyse the dataset completely. This is best explained by 'Anscombe's Quartet'.

   In 'Anscombe's Quartet', four datasets have been taken. The statistical summary of those datasets looks similar.

Anscombe's quartet

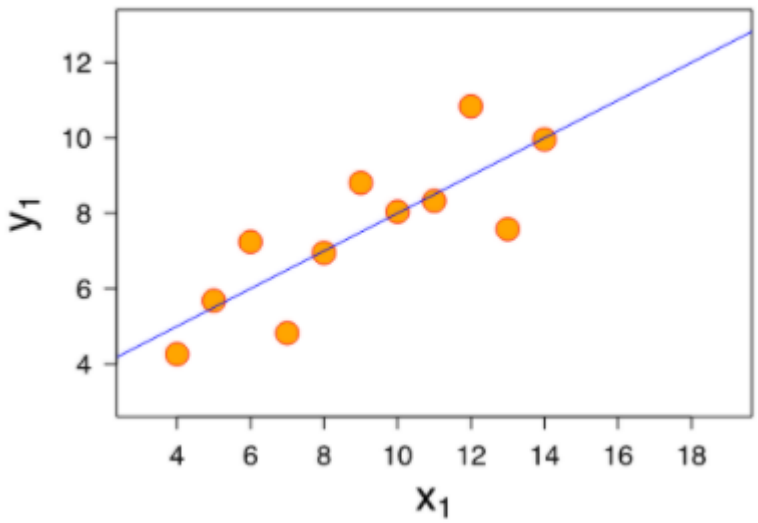
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The four datasets are I,II,III and IV. Let us now look at the statistical summary (mean value, variance, correlation and line of best fit) of each datasets.

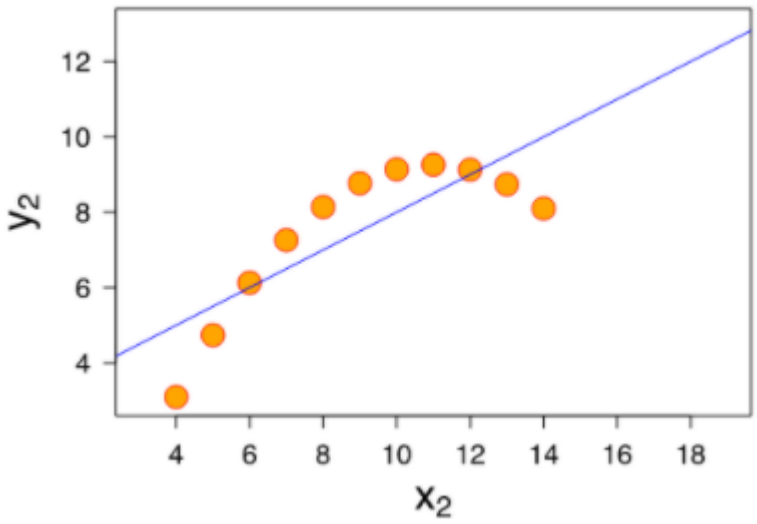
Statistical summary of each dataset

1. The mean value of x = 9 (each dataset)
2. The average value of y = 7.50 (each dataset)
3. The variance for x = 11 (each dataset)
4. The variance for y = 4.12 (each dataset)
5. The correlation between x and y is 0.816(each dataset)
6. Line of best fit = 0.5x + 3(each dataset)

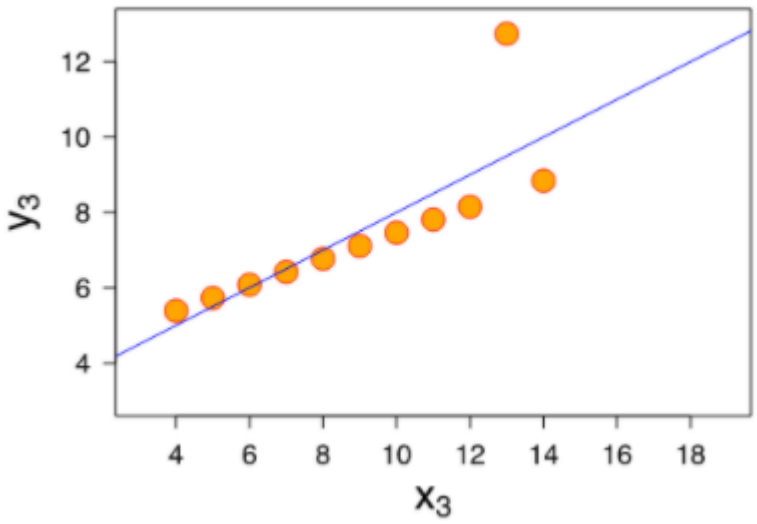
On seeing, the above statistics, we feel that all the datasets look similar. But when we plot these four data sets on a graph it is not the same. Look into the below diagram (plot for the four datasets)



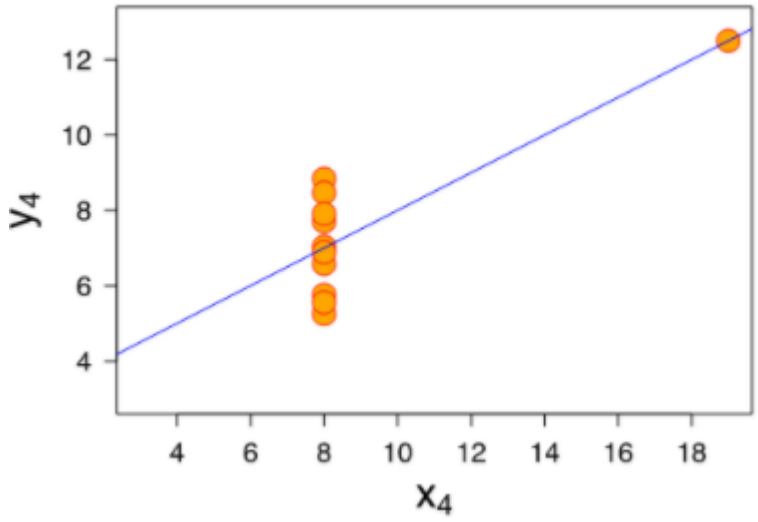
Above plot is for the dataset I. This shows that there is a linear relationship with some variance.



Above plot represents dataset II. There is no linear relationship between x and y (Linear relationship – If the value of x increases the value of y increases or vice versa. Relationship can be either positive or negative).



Above plot represents dataset III. There is linear relationship between x and y, but there is an outlier.



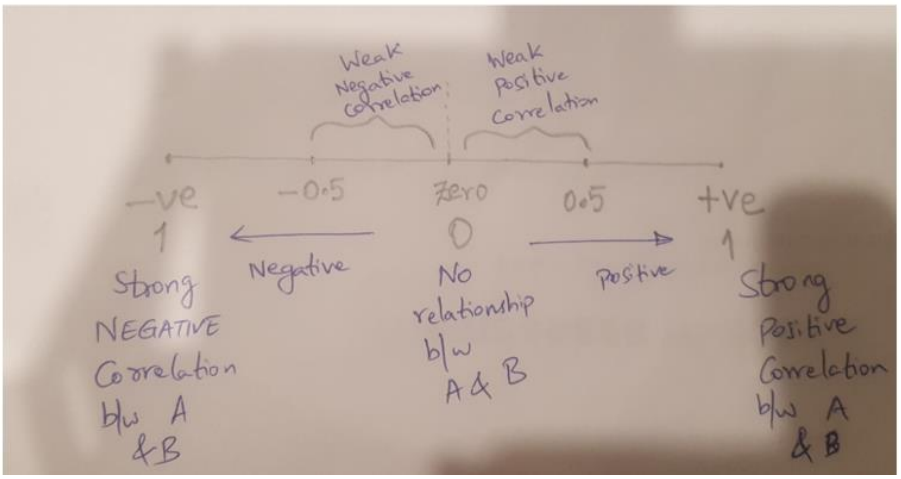
Above plot represents dataset IV. There is no linear relationship between x and y. There is an outlier and the value of x is constant.

Though the statistical summary of the four datasets look same their visual representation looks different.

So, Anscombe's Quartet says," We should not only rely on the summary statistics of the dataset. We need to completely visualise the dataset to get an idea about it and then we need to run a linear regression on it."

3. What is Pearson's R?

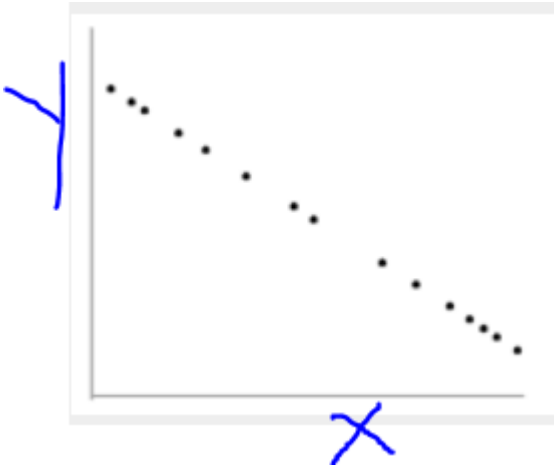
Pearson's R in other words it is referred to as Pearson's Correlation Coefficient. Correlation is analysing the relationship between the two variables X and Y Example - Price(X) and Sales(Y). Pearson's Correlation Coefficient measures the linear relationship between the two variables X and Y which has the value that range between -1 and +1.



The above figure represents the relationship between the two variables A and B.

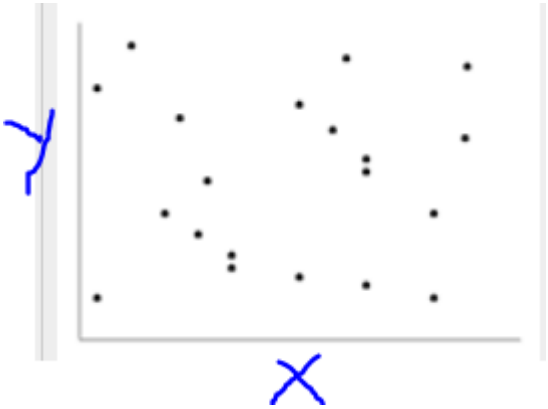
If the value (r) = -1, it has a strong negative correlation, if value (r) = 0 it means there is no relationship between those two variables and if the value (r) = 1 it means it has a strong positive correlation.

If the r = -1 -> Strong Negative Correlation, the plot looks as shown below.



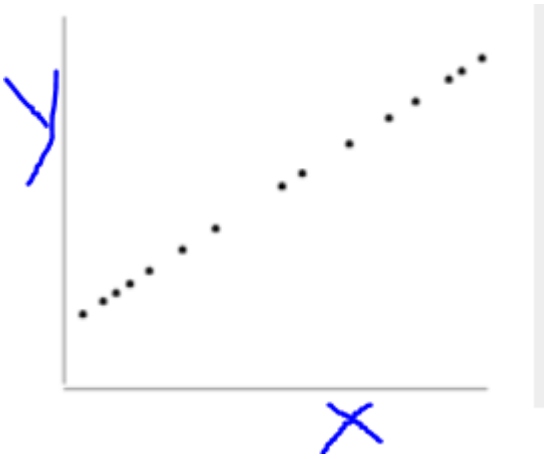
The above plot indicates the increase in the variable(X) causes decrease in the other variable.

If the r = 0 -> No Correlation between the two variables, the plot looks as shown below.



The above plot (No Relationship) indicates the change in the values of X does not have any impact on the other variable Y

If the r = 1 -> strong positive correlation between the two variables, the plot looks as shown below.



The above plot indicates the increase in the variable(X) causes increase in the other variable(Y).

This is also referred to as “Bivariate correlation” as it involves two variables.

4. **What is Scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?**

   Scaling is a process applied on all the features (especially numerical features) of a dataset to **standardise them and bring all of them to same units**.

   Example: In simple terms, consider a dataset has values in metre, centimetre and kilometre. Converting it to same magnitude as Kilometre is called scaling.

   Need for scaling:

   Scaling has to be done during the data pre-processing.

   If the features are not scaled, our model will give a wrong prediction.

   Example: Consider a particular column has the value in metres as 3000. The other has the value in Km say 5. We have the figures only, which shows 3000>5. But actually, that's not the case. When we predict the model that has values in different unit, it will lead to mistakes and our model will become an unreliable one.

   So, it is must to scale all the features (numerical variables) before building the model.

   Normalised Scaling:

   This is one of the scaling techniques where the values of the variables (numerical variables) are scaled to values that range between **0(min) to 1(max)**. This is also referred to as **Min-Max Scaling.**

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where Xmax – maximum value of the feature

Xmin – minimum value of the feature

Standardised Scaling:

This scaling technique will bring the entire data to a standard normal distribution with the mean = 0 and standard deviation =1.

$$X' = \frac{X - \mu}{\sigma}$$

Where μ - mean value of the feature

σ - Standard deviation of the feature

5. **You might have observed that sometimes the value of VIF infinite. Why does this happen?**

   VIF is used to detect the multicollinearity between the independent variables. Multicollinearity occurs when independent variables in a regression model are correlated to each other. The cut off value of VIF is less than 5. But sometimes it can be less than 2 or 3.

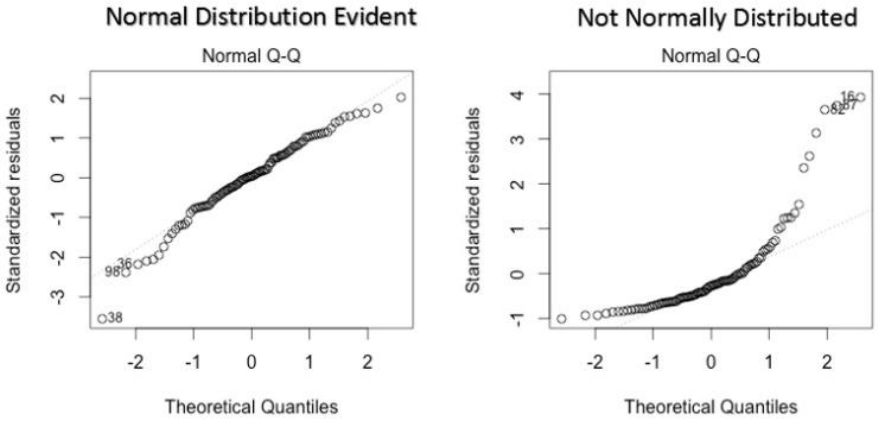
   If VIF is equal to infinite, it means that the feature (independent - X1) has a perfect correlation with the other features (independent – X2,X3,X4,….Xn). This shows that the variables (X1) can be expressed

exactly by the linear combination of all the other variables (independent – X2,X3,X4,….Xn).

6. **What is a Q-Q plot? Explain the use and importance of Q-Q plot in linear regression.**

   Q-Q plot (Quantile-Quantile) is a probability or graphical plot which is used to compare the two probability distributions (distributions of theoretical and practical events) by plotting the quantiles of those two against each other.

   If the data is normally distributed the plot will give a straight line. If it's not normally distributed there will be a deviation in the straight line.



Use and importance of Q-Q plot in linear regression

In our linear regression, we have training and test dataset. If we receive these datasets separately, we need to conform whether these two datasets are taken from a population that have same distributions. In this case, we can use a Q-Q plot to conform it.