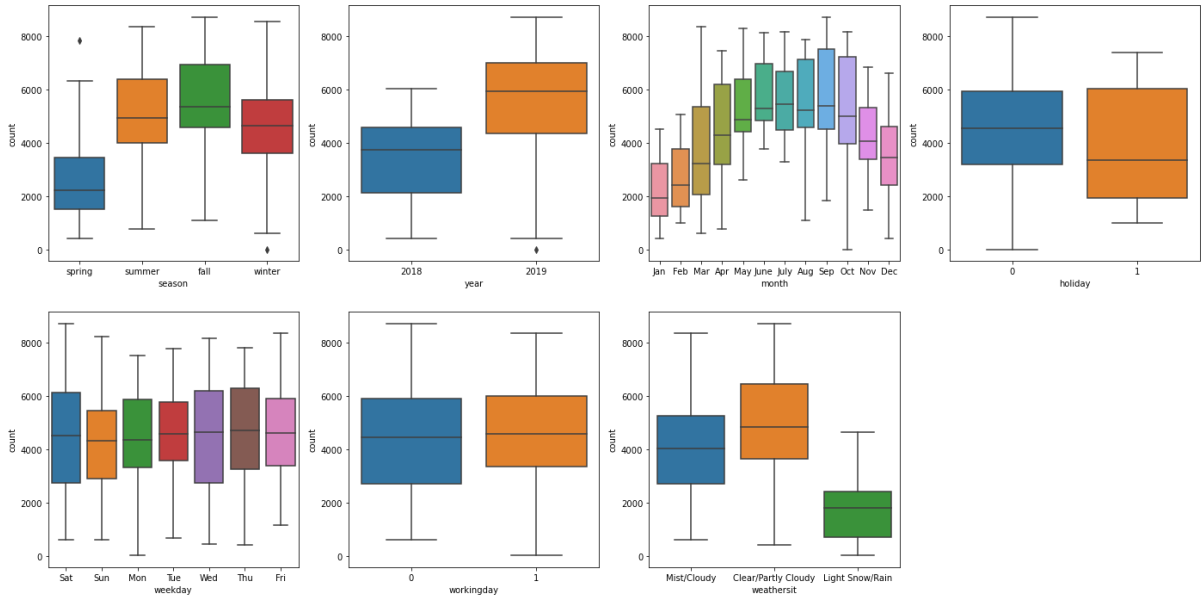
Assignment-based Subjective

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

Ans. The categorical variables identified are

- season
- year
- month
- holiday
- weekday
- workingday
- weathersit

Below is the box plot of count for all categorical variables-



Inferences drawn are stated below-

- Demand of shared bikes is more during FALL(highest median) followed by SUMMER as the bike riding conditions are favourable during these seasons.
- Demand of shared bikes is clearly more in 2019 than in 2018. This may be due to an increase in awareness with every passing year.
- Clearly, the demand of bikes is more in the months from May to October which are again FALL and SUMMER months.
- Over non-Holidays, demand is more. This may be because people might be spending more family time at home or preferring car rentals for family commutation.
- weekday/workingday hardly affects the count
- Most of the shared bikes are rented on 'Clear, Few clouds, partly cloudy' day.
- There is not even a single record of any bike rental on a 'Heavy Snow/Rain/Hail/Fog' day.

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

Ans. This is known as one hot encoding. This is a technique where a categorical data representation is done in a form so that our machine learning model can interpret it. Generally, if a categorical variable has n levels, then it can be represented by n-1 dummy variables. This not only reduces an extra column but also reduces correlation between newly introduced dummy variables.

To understand it better, lets take the example of Season categorical variable in the BOOMBIKES dataset. So, there are 4 seasons

- SPRING
- SUMMER
- FALL
- WINTER

The number of levels is 4. So, lets first encode them in the below manner

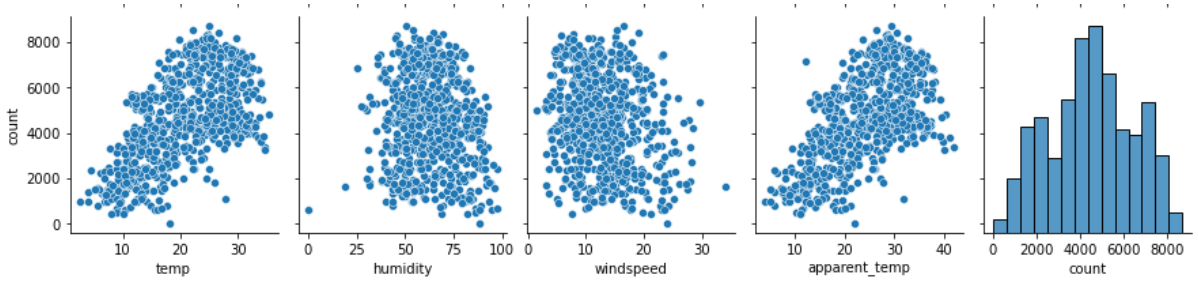
FALL	SPRING	SUMMER	WINTER
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

Now, we can clearly see that if all SPRING, SUMMER and WINTER dummy variables have value 0 then obviously it's a FALL season. Hence, we drop the first column and still have a clear interpretation of FALL from remaining 3 dummy variables

SEASON	D_0	D_1	D_2
WINTER ->	0	0	1
SUMMER ->	0	1	0
SPRING ->	1	0	0
FALL ->	0	0	0

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. 'temp' and 'apparent_temp' have a very strong correlation of 0.63 with the 'count' variable(target).

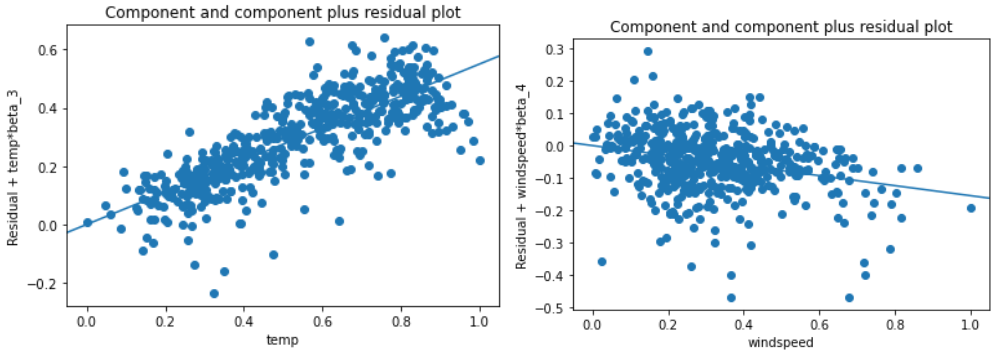


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

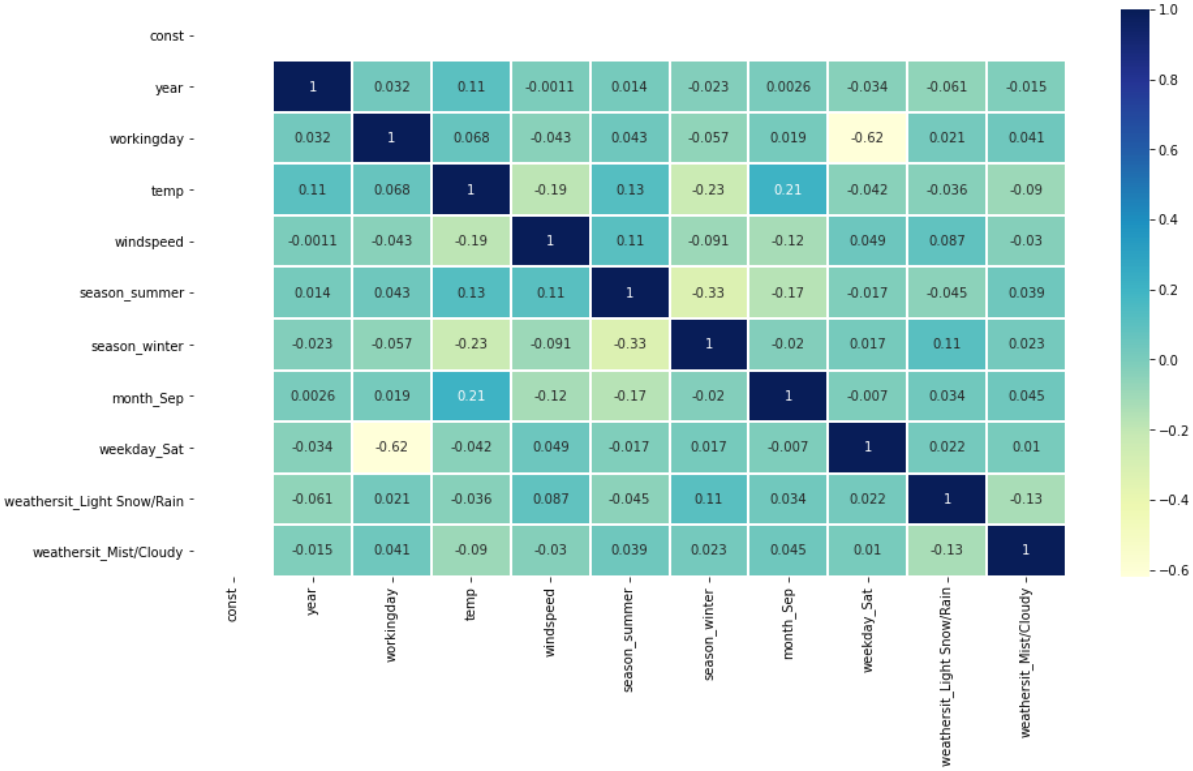
Ans. There are basically 5 assumptions of linear regression-

- Linearity Assumption
- Little or no Multicollinearity between the features
- Homoscedasticity Assumption
- Normal distribution of error terms
- Little or No autocorrelation in the residuals

1. **Linearity Assumption –** Linearity Assumption validation is done through scatter plots between dependent and independent variables. After the model was build, we plotted ccpr regression plot and could see that there was a linear relationship between predictor and target variables. We did this on 'temp' and 'windspeed'.

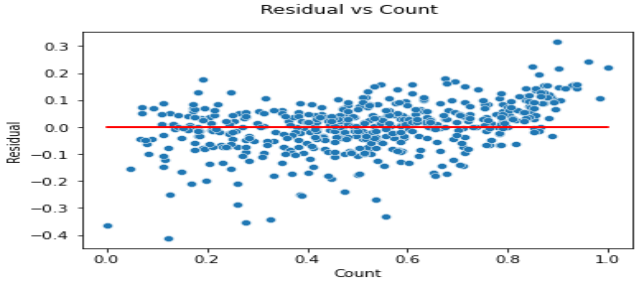


2. **Little or no Multicollinearity between the features –** There should not be any or very little collinearity between predictor variables. This can be checked by Pearson's correlation coefficient between predictor variables or by checking VIF, Since, all VIF are less than 5, there is very little multicollinearity observed.

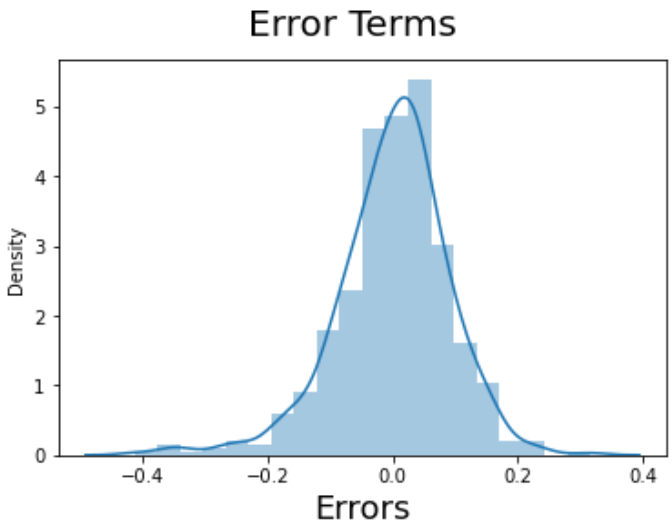


	Features	VIF
2	temp	4.76
1	workingday	4.04
3	windspeed	3.43
0	year	2.02
7	weekday_Sat	1.69
4	season_summer	1.57
9	weathersit_Mist/Cloudy	1.53
5	season_winter	1.40
6	month_Sep	1.20
8	weathersit_Light Snow/Rain	1.08

3. **Homoscedasticity Assumption –** Homoscedasticity in a model means that the error is constant along the values of the dependent variable. We plotted a residual v/s count scatter plot to check this. There was a constant deviation of points from the zero line and there was no discernible pattern, preserving Homoscedasticity.



4. **Normal distribution of error terms –** On plotting a histogram of residuals, we could clearly see that the error terms are normally distributed and were centred around 0.



5. **Little or No autocorrelation in the residuals-** We did Durbin-Watson test for checking the degree of correlation of each residual error with the 'previous' residual error. The Durbin Watson test reports a test statistic, with a value from 0 to 4; where if the value is close to 2, the less auto-correlation there is between the various variables.

Durbin-Watson value for the Final Model is 2.08

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

Ans. The top 3 features contributing significantly towards explaining the demand of shared bikes are-

- **temp** with the coefficient of 0.5499
- **weathersit_Light Snow/Rain (weathersit = 3)** with the coefficient of -0.2880
- **year (yr)** with a coefficient of 0.2331

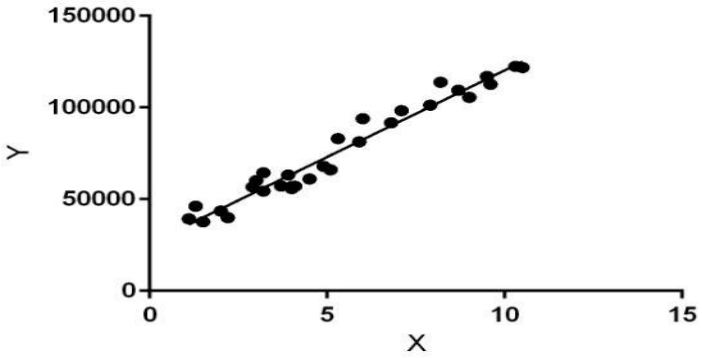
General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Ans. There are two types of Machine Learning algorithms basically- Supervised and Unsupervised. **Linear Regression** comes under **Supervised Machine Learning** algorithm. And it is done for prediction of a continuous target variable. It performs a regression task to predict a target variable value based on some independent variables. In Linear Regression model, the relationship between dependent and independent variables is linear and so is the name. Linear Regression is used for prediction and forecasting.

There are two types of Linear Regression models-

- **Simple Linear Regression(SLR)** – is the relationship between a Dependent variable(Y) and One Independent variable(X)

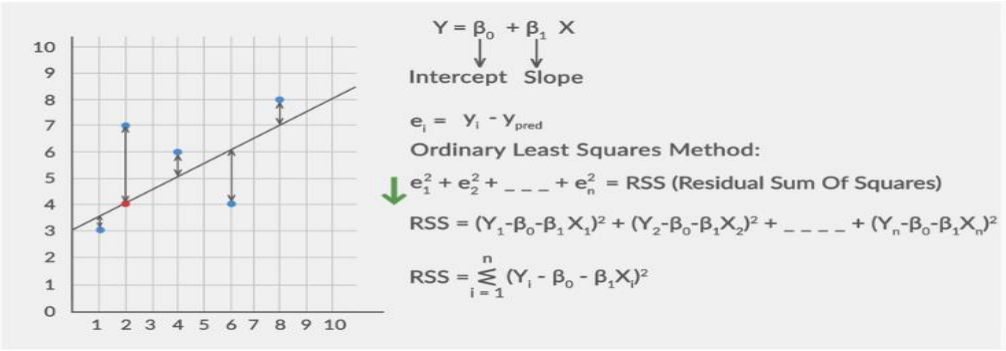


In SLR, a straight line is fitted on the scatter plot of these two variables which is called as regression line. The standard equation of this line is

$$Y = \beta_0 + \beta_1.X$$

where β_0 is the intercept and β_1 is the slope of regression line.

The **Best fit line** is found by minimising the Residual Sum of Squares (RSS)



Now, coming to the metrics associated with Linear Regression Model, we have mainly 2-

1. R-squared or Coefficient of Determination(R2)
   It explains what percentage of the given data variation is explained by the developed model. It takes value between 0 and 1 and is calculated as-

   $$R2 = 1 - (RSS / TSS)$$

   where, RSS is residual Sum of Squares
   TSS is Total Sum of Squares
2. Residual Standard Error (RSE)
   It is the average deviation between the actual outcome and the true regression line of the fitted model.

- **Multiple Linear Regression(MLR) -**– is the relationship between a Dependent variable(Y) and Many Independent variables (X1, X2, ...etc)
  Its equation can be represented as –

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 + \beta_4.X_4 + ...$$

  Here, also the equation of best fit 'hyperplane' is obtained.
  The Multiple Linear Regression metrics are mentioned below

1. R-squared and adjusted R-squared – Adjusted R-squared penalizes the value of R-squared for unnecessary addition of variables. It is a better metric than R-squared.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where
R^2 Sample R-Squared
N Total Sample Size
p Number of independent variable

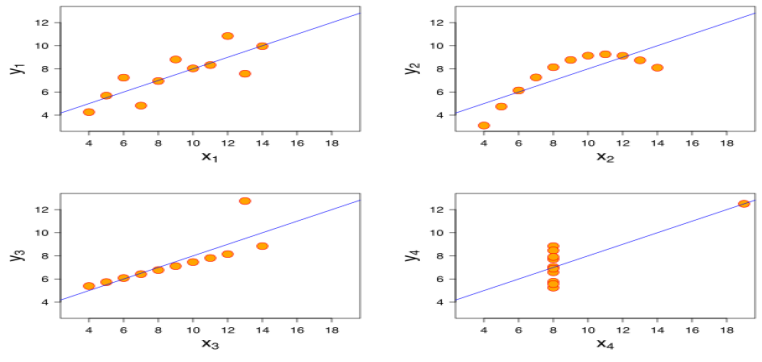
2. Variance inflation factor (VIF) - Multicollinearity is a problem in prediction model. If VIF value of a variable is high, it means that it is largely explained by other variables.

$$VIF = 1/(1-R2)$$

Question 2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet was constructed by Francis Anscombe to show the importance of plotting data before analysing it purely based on its statistic. And it also shows the effect of outliers on different statistical properties.

He created four datasets that had nearly identical simple statistical properties but when he graphed them, they were completely different.



- The first graph (y1 vs x1) represents a simple linear relationship between correlated variables
- The second graph (y2 vs x2) is not normally distributed but a non-linear relationship exists and Pearson's correlation coefficient is not relevant here.
- The third graph (y3 vs x3) represents a linear relationship but because of an outlier, the regression calculated is offset by some value.

- Finally, the fourth graph (y4 vs x4) shows that there is no relationship between variables but because of one high-leverage point, it is showing a high correlation between variables.

Below are the 4 datasets and summary statistics-

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

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

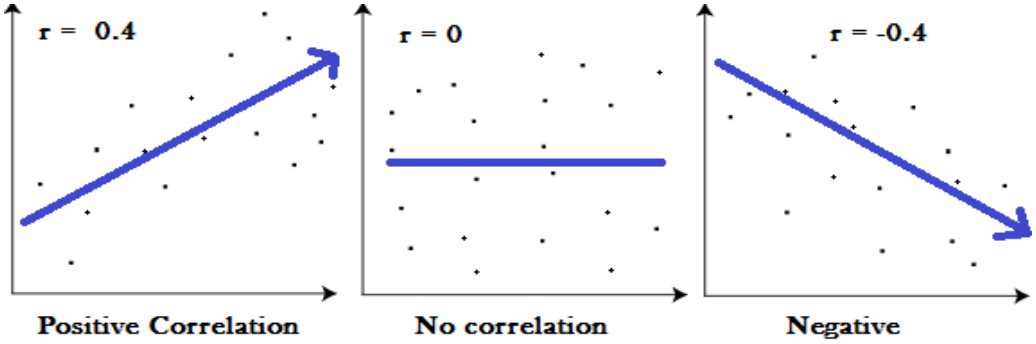
Question 3. What is Pearson's R?

Ans. The Pearson correlation coefficient, r, is to understand the strength of linear relationship between 2 variables. It is the ratio between the covariance of two variables and the product of their standard deviations

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Its value varies from -1 to 1. A negative value indicates a negative correlation , i.e., an increase in one variable will cause a decrease in the other and vice-versa. On the other hand, a positive value indicates a direct relationship or a positive correlation, which means an increase in one variable will result in the increase in other as well and vice-versa. A zero value indicates that there is no correlation between variables where as a unit value (+1 or -1) indicates a perfect relationship.



Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a pre-processing step, applied to independent variables before the model is build. It is use to normalize the data within a particular range.

In case of Multiple Linear Regression when there are lot many variables, many of them might be on very different scales. The model obtained will have varying coefficients which will be very difficult to interpret. So, scaling is needed for **ease of interpretation**. Another reason for Scaling is **faster convergence of Gradient Descent methods**

Difference between normalized scaling and standardized scaling-

1. **Normalized/MinMax Scaling:** The variables are scaled in such a way that all the values lie between 0 and 1 using the maximum and the minimum values in the data.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   Here, inliers are scaled to a very narrow range.
   It is done using Scikit-Learn **MinMaxScaler**

2. **Standardized Scaling:** The variables are scaled in such a way that their mean is zero and standard deviation is one. It is also called Z-score normalization

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

   It is done using Scikit-Learn **StandardScaler**

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

Ans. VIF indicates collinearity between independent variables. If there is a perfect correlation between two independent variables then we get R2 = 1

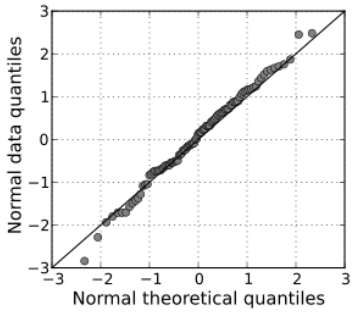
And if we calculate VIF which is 1/(1-R2) then it will come as infinity.

To solve this, we need to drop one of the variables which is causing Multicollinearity.

This could also happen if the variable with infinite VIF value is expressed perfectly with linear combination of other independent variables.

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

Ans. Q-Q Plot(Quantile-Quantile plot) is a plot of quantiles of 2 distribution against each other. The pattern of points in the plot is used to compare the distributions. If the 2 distributions are similar or linearly related then the points will lie on the line y=x, commonly called as 45-degree reference line. If the points are far from the reference line, the conclusion can be made that the datasets are from different distribution. Below is the q-q plot for Normal Distribution.



Use/Importance in LR – In Linear Regression, when Training and Test datasets are received from different source, we can use Q-Q plot to conclude that both of them are from Populations with same distribution.

Advantages –

- It can be used with varying sample sizes also
- We can use it to test many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.