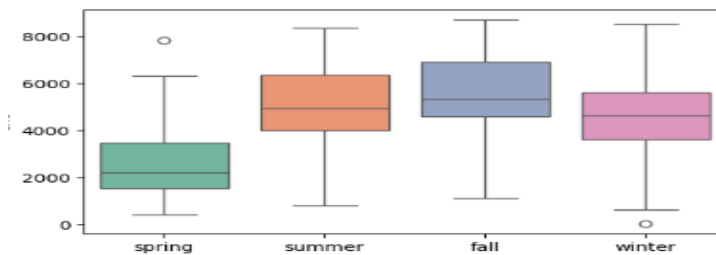


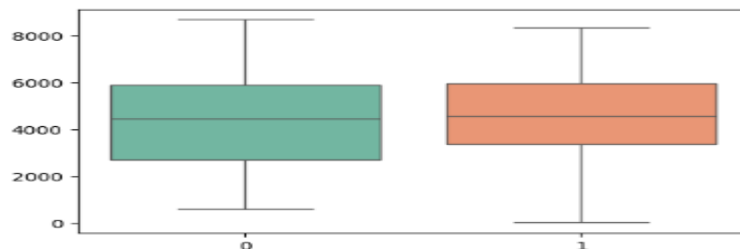
Linear Regression Subjective Questions

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

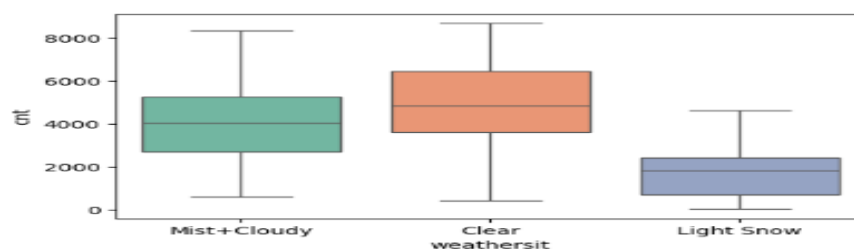
- The categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.
- “season” –
 - As we show you in graph, the most favourable seasons for biking are summer and fall. We should plan advance strategy to take advantage of peak season.
 - Spring has significant low consumption ratio.



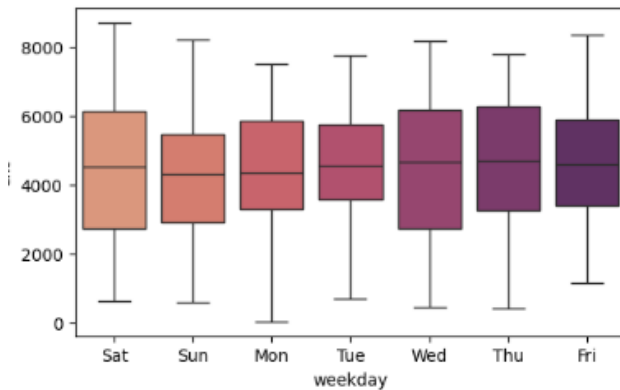
- “workingday” –
 - Working day represents weekday and weekend/holiday information.
 - The registered users are renting bikes on working days whereas casual users prefer the bikes on non-working days. This effect is nullified when we look at the total count because of the contradictory behaviour of registered and casual users.
 - Registered and casual users’ identity and relevant strategy for working and not working days shall help to increase the numbers.



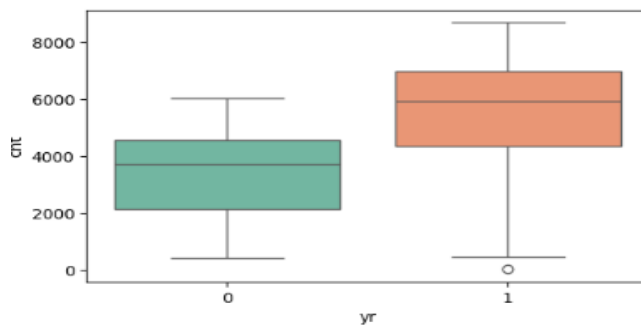
- “weathersit” –
 - Most favourable weather condition is the clean/few clouds days.
 - Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace.



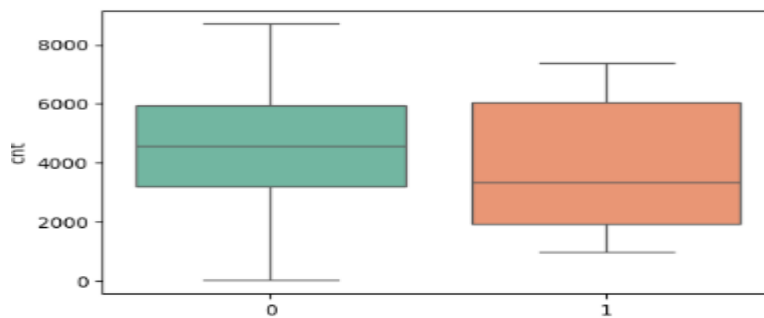
- “Weekday” – We observe , Fir, Sat and Sun have more number of bookings as compared to the start of the week.



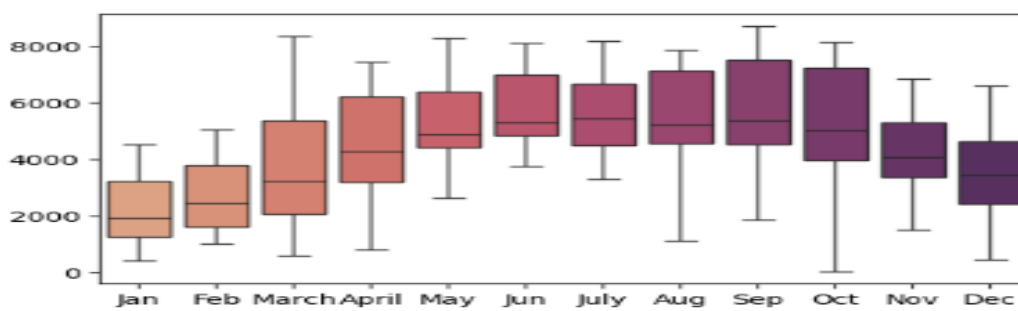
- “yr” –
 - 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
 - As we can see in graph 0 – 2018 , 1 – 2019, it can show significant growth in 2019 than 2018



- “holiday” –
 - Holiday consumption of bikes if compared within “registered” and “casual” users then the observation is “casual” users are using bikes more on holiday.



- “mnth” –
 - The bike rental ratio is higher for June, July, August, September and October months.
 - 75 quantile grows in the months mentioned in point 1.



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

The `drop first = True` is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category. Since one of the columns can be generated completely from the others, and hence retaining this extra column does not add any new information for the modelling process, would it be good practice to always drop the first column after performing One Hot encoding, regardless of the algorithm of choice.

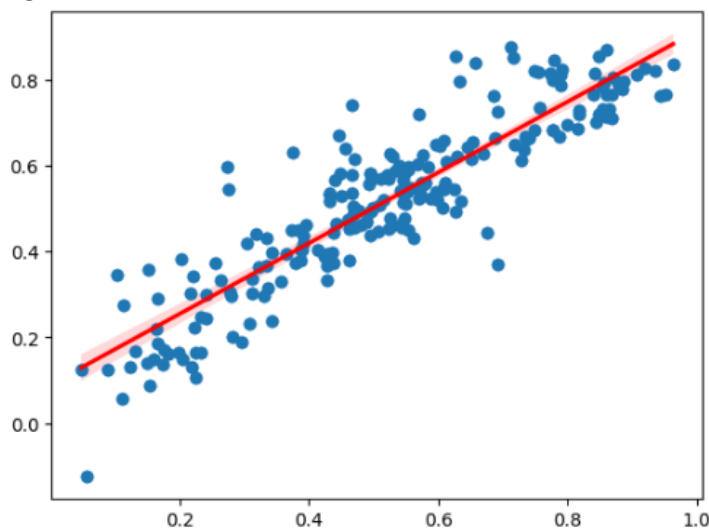
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- “temp” is the variable which has the highest correlation with target variable i.e. 0.63.
- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- “atemp” is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation to avoid multi-collinearity.

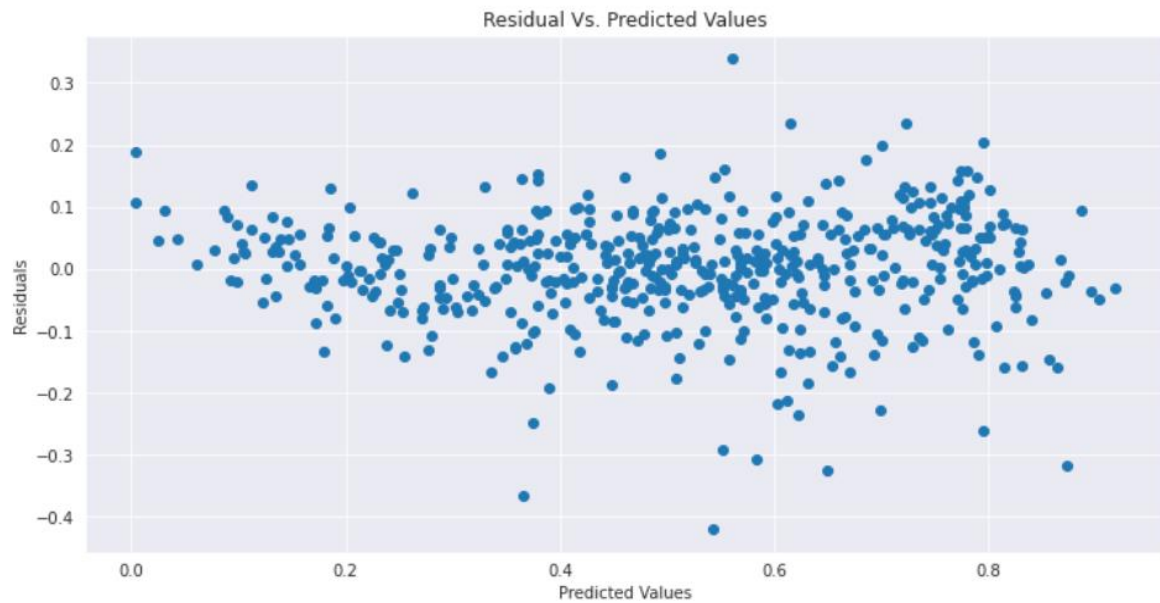
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Linear regression is useful for finding out a linear relationship between the target and one or more predictors. In statistics, there are two types of linear regression, simple linear regression, and multiple linear regression. let's study the assumption :

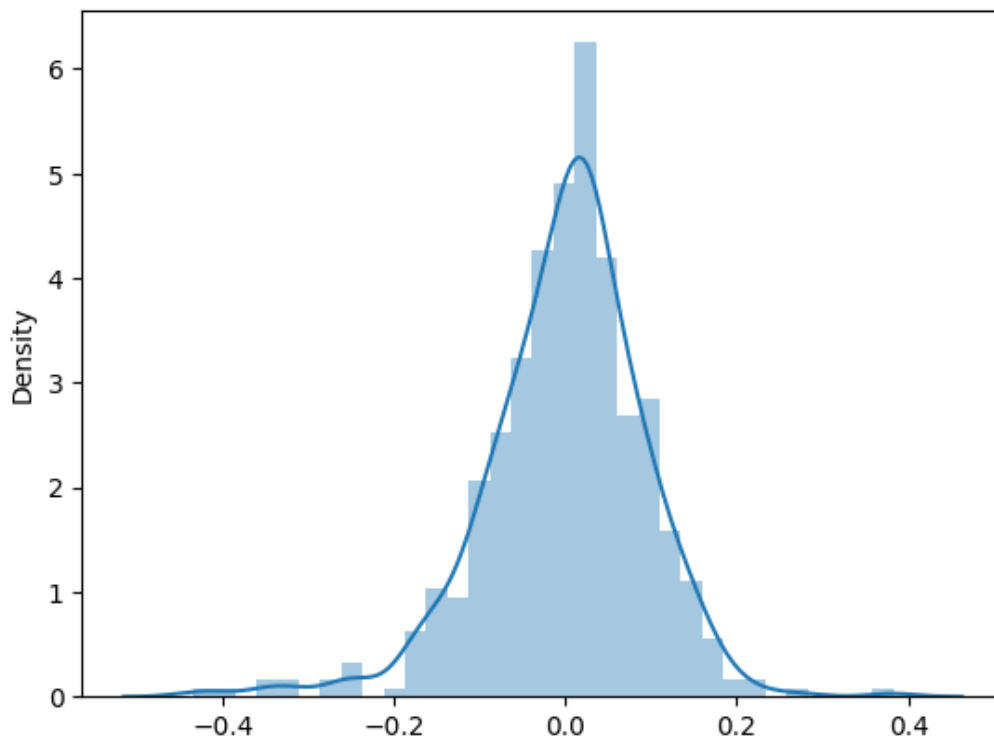
1. **The Two Variables Should be in a Linear Relationship**– The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



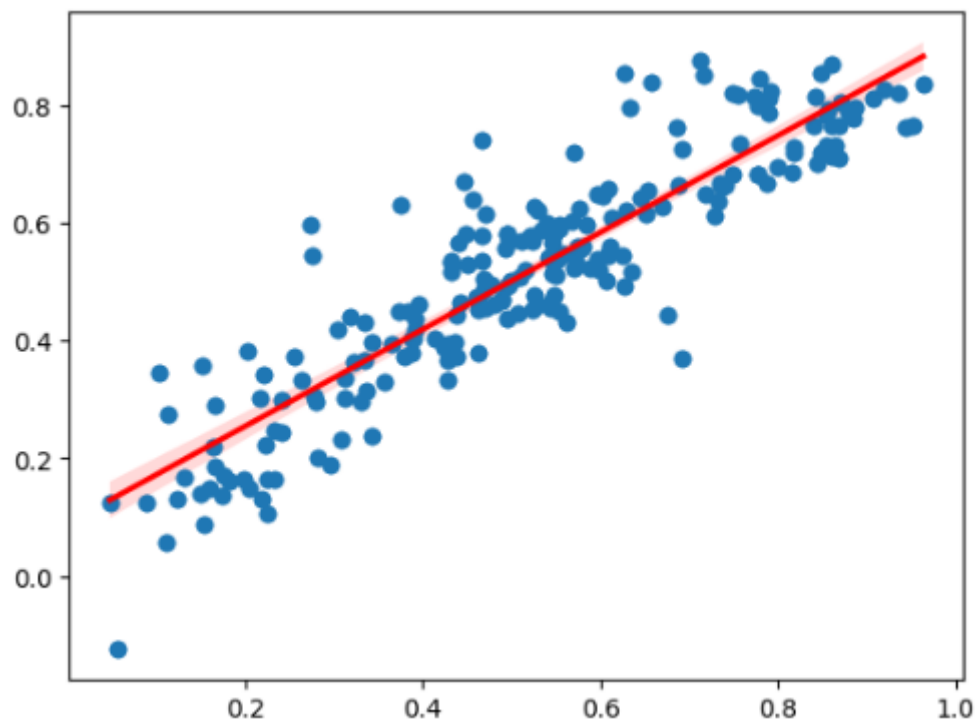
2. **Error terms are independent of each other** – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other.



- 3. Error terms are normally distributed:** Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



- 4. Error terms have constant variance (homoscedasticity):** We can see Error Terms have approximately a Constant Variance, hence it follows the Assumption of Homoscedasticity.



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 variables are:

weathersit :

Temperature is the Most Significant Feature which affects the Business positively.

Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively.

'Yr':

The growth year on year seems organic given the geological attributes.

We observe growth in 2019 is significantly more than the 2018

'season':

As we show you in graph, the most favourable seasons for biking are summer and fall. We should plan advance strategy to take advantage of peak season.

Spring has significant low consumption ratio.

Winter season is playing the crucial role in the demand of shared bikes.

1. Explain the linear regression algorithm in detail.

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
 - Simple Linear Regression – Single independent variable is used.
 - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
 - Multiple Linear Regression – Multiple independent variables are used.
 - $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$ is the line equation for MLR.
 - $\beta_0 = \text{value of the } Y \text{ when } X = 0 \text{ (} Y \text{ intercept)}$
 - $\beta_1, \beta_2, \dots, \beta_p = \text{Slope or the gradient.}$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained.**
 - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
 - The straight-line equation is $Y = \beta_0 + \beta_1 X$
 - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i .
 - Now the cost function will be $J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
 - The unconstrained minimization are solved using 2 methods
 - Closed form
 - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
 - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
 - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
 - $RSS = \sum_{i=1}^n (y_i - y_{pred})^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe's quartet in detail.

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.

Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

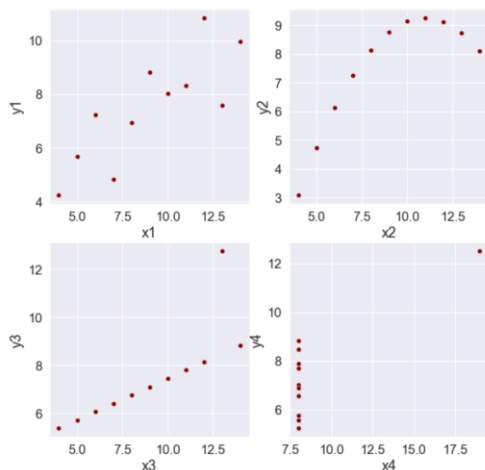
- Illustrations
 - One of the data set is as follows:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.980000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

- If the descriptive statistics are checked for above data set then all looks same:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

- However, when plotted these points, the relation looks completely different as depicted below.



- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points
 - Plotting the data is very important and a good practice before analysing the data.
 - Outliers should be removed while analysing the data.
 - Descriptive statistics do not fully depict the data set in its entirety.

3. What is Pearson's R?

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearson's R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$$r = \frac{n(\sum x * y) - (\sum x) * (\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] * [n\sum y^2 - (\sum y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx^2 = the sum of squared x scores

Σy^2 = the sum of squared y scores

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

- What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

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

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

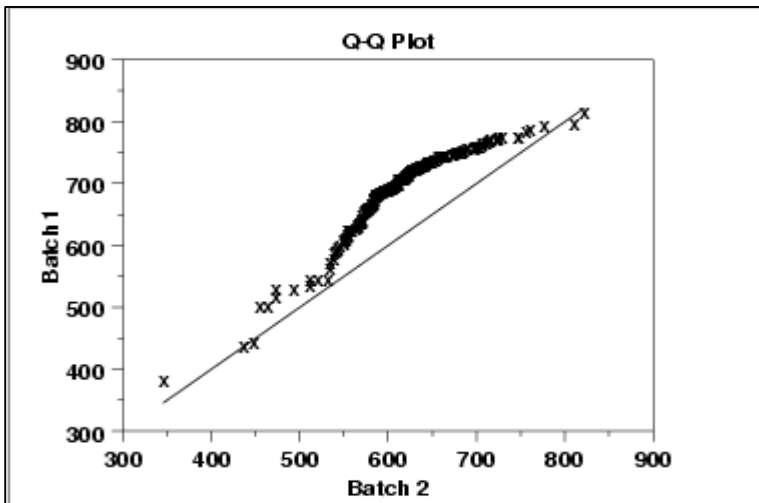
- Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.

- Y values < X values: If y-values quantiles are lower than x-values quantiles.
- X values < Y values: If x-values quantiles are lower than y-values quantiles.
- Different distributions – If all the data points are lying away from the straight line.

The advantages of the q-q plot are:-

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.



The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1

Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

□ □ □