

# Linear Regression – Graded Assignment

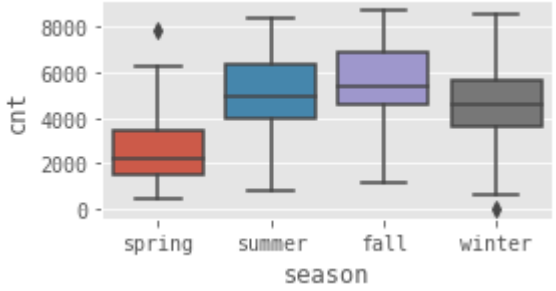
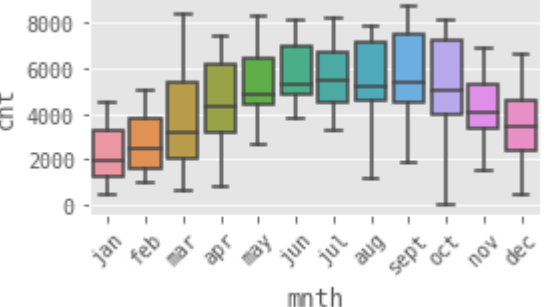

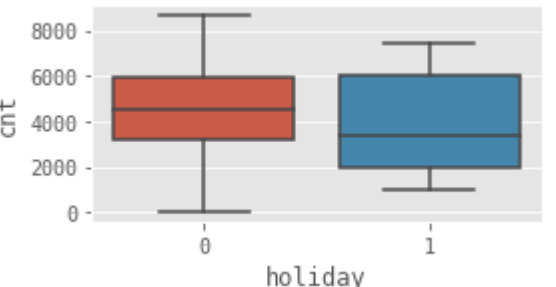
BIKE SHARING - MULTI LINEAR REGRESSION -  
ASSIGNMENT

**BHAVESHKUMAR THAKER**

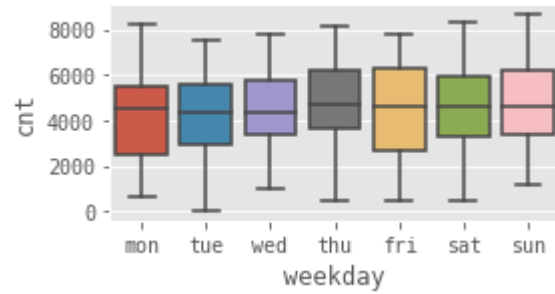
# Assignment based Subjective Questions

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

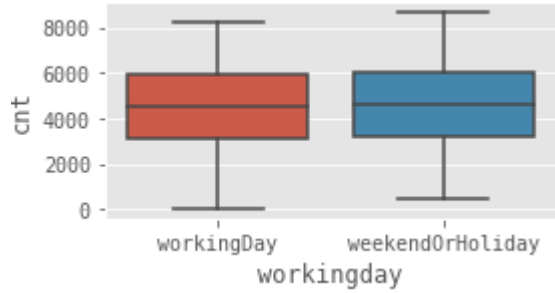
**Answer 1:**

The demand of bike(s) is less in the month of <b>spring</b> when compared with other seasons	
<b>June to September</b> month is the period when bike the demand is high. In the January month the demand is low.	
The demand for bike has increased in the year <b>2019</b> when compared with year 2018.	
Bike demand is less in holidays in comparison to non-holiday.	

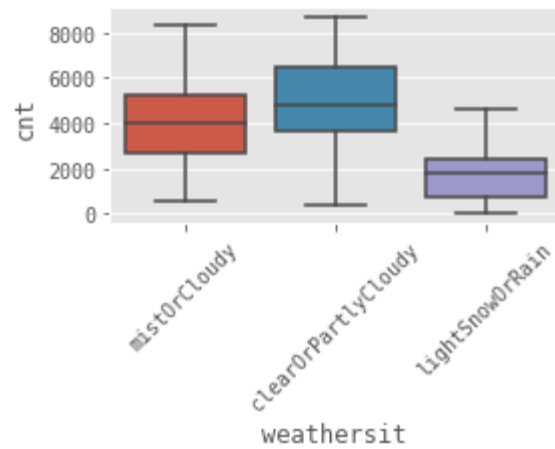
The demand of bike is almost similar throughout the weekdays.



There is no significant change in bike demand between working day and weekend or holiday day.



The bike demand is high when weather is clear or partly cloudy.



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

**Answer 2:** "`drop_first=True`" is a technique of one hot encoding. One hot encoding is a technique which converts categorical data into a form which is understandable by ml model. `drop_first=True` reduces the correlations created among dummy variables.

"`drop_first=True`" meaning that k-1 dummies created out of k categorical levels

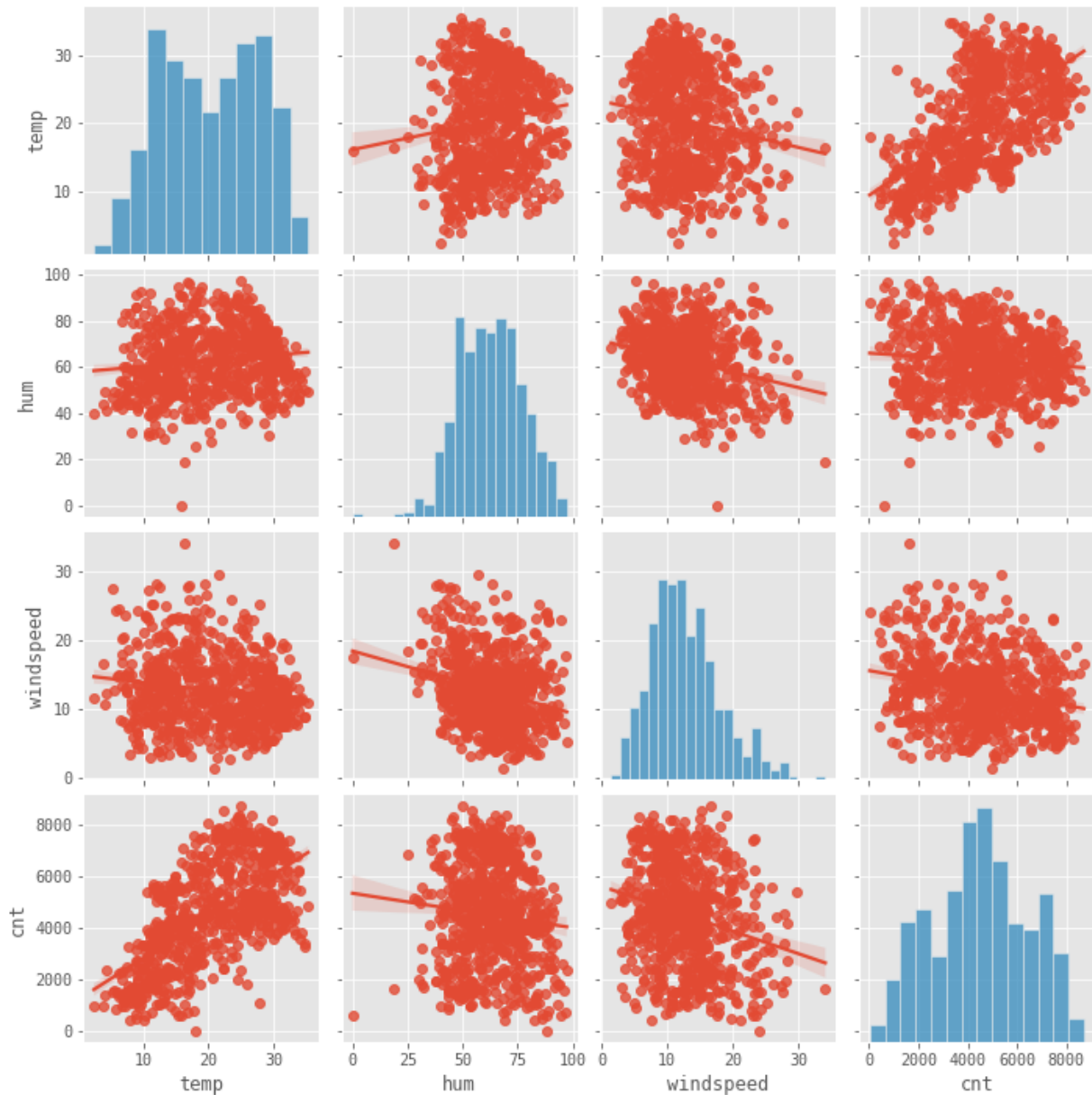
"`drop_first=False`" meaning that the reference is not dropped and k dummies created out of k categorical levels.

Example: Let's say we have 3 values in Categorical column "unfurnished", "furnished", "semi\_furnished". If one variable is not "furnished" and "semi\_furnished", then It is obvious "unfurnished". So, we do not need 3rd variable to identify the "unfurnished".

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

**Answer 3:** By looking at the pair-plot, we can understand that the numerical variable **"temp"** has the highest correlation with the target variable **"cnt"**.

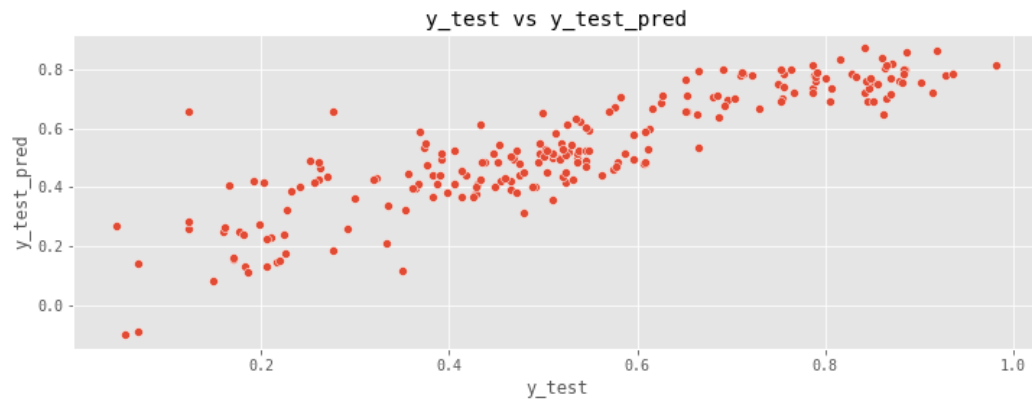
From the graph below, we can see the linear relationship between **"temp"** and **"cnt"** variable.



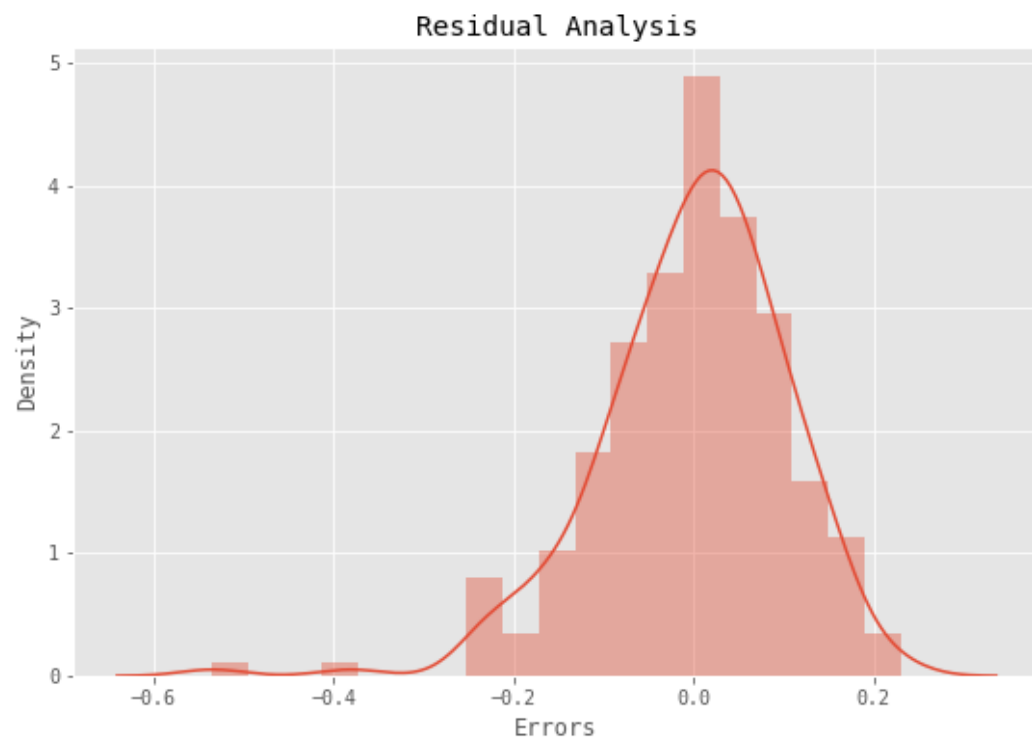
**Question 4:** How did you validate the assumptions of Linear Regression after building the model on the training set?

**Answer 4:** Following analysis performed to confirm the assumptions of the Linear Regression.

- Linearity check between dependent variables' test and predicted values.



- Residual Analysis: Residual error should follow normal distribution, mean should be 0.



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

**Answer 5:** Following are the top 3 features contributing significantly:

- Year
- Month\_September
- Month\_July

	model7
yr	0.247129
mnth_sept	0.078310
mnth_jul	0.021800
holiday	-0.045929
weathersit_mistOrCloudy	-0.084838
mnth_dec	-0.091556
mnth_nov	-0.102595
windspeed	-0.165688
season_spring	-0.263967
weathersit_lightSnowOrRain	-0.296458

# General Subjective Questions

---

**Question 1:** Explain the linear regression algorithm in detail.

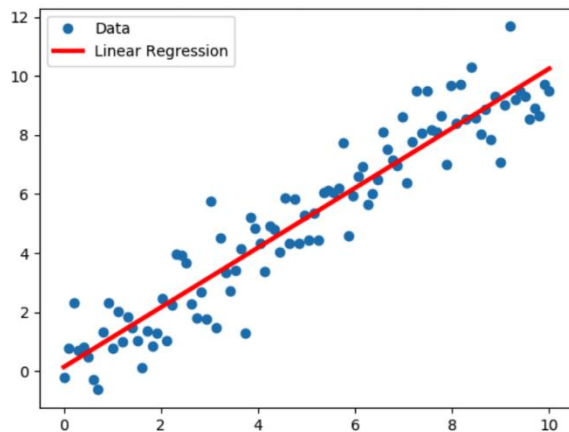
**Answer 1:**

## Regression

Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum. It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

## Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value, the dependent variable (Y-axis) based on independent variables, the independent variable (X-axis). If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables. It is mostly used for finding out the relationship between variables and forecasting.



## Hypothesis function for Linear Regression

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b are given by the formula:

$$b \text{ (slope)} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$



$$a \text{ (intercept)} = \frac{n \sum y - b(\sum x)}{n}$$

Here,

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

### **Cost Function**

Cost function of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

$$J = \frac{1}{n} \sum_{i=1}^n (pred_i - y_i)^2$$

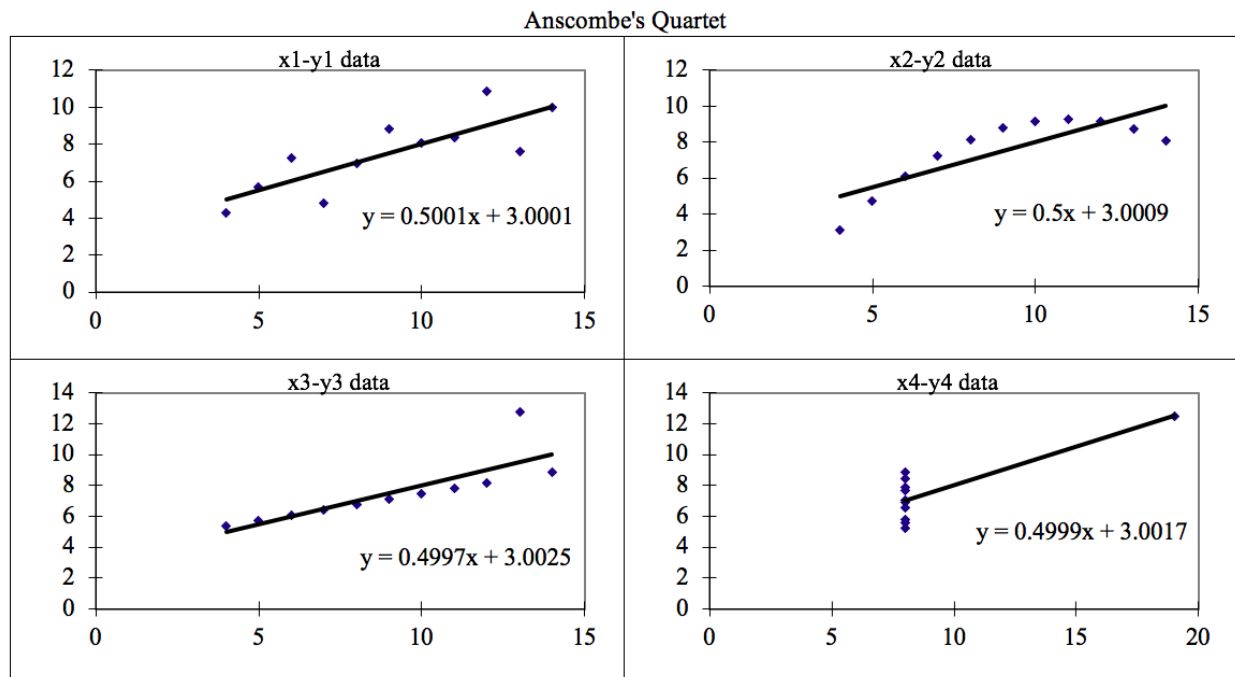
### **Advantage and disadvantage**

Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it's only limited to linear relationships.

Linear regression produces the best predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable.

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

**Answer 2:** Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, and y points in all four datasets.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				<u>Summary Statistics</u>							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The four datasets can be described as:

- Dataset 1: this fits the linear regression model well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

**Question 3:** What is Pearson's R?

**Answer 3:** Pearson's Correlation Coefficient is named after Karl Pearson. The Pearson's Correlation Coefficient is also referred to as Pearson's  $r$ , the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Pearson's correlation coefficient is the covariance of the two variables divided by 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

The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

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

**Answer 4:** Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. Scaling also helps in speeding up the calculations in an algorithm.

### Normalized Scaling

Normalized scaling brings the data in the range of 0 and 1. It is used when features are of different scales.

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

### Standardized Scaling

Standardized scaling replaces the values by their Z-scores. It brings all the data into a standard normal distribution which has a mean as zero and standard deviation as one. It is used when we want to ensure zero mean and unit standard deviation.

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

### Difference between Normalized Scaling and Standardized Scaling

Normalized Scaling	Standardize Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is really affected by outliers.	It is much less affected by outliers.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution.	It is useful when the feature distribution is Normal or Gaussian.

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

**Answer 5:** If there is perfect correlation, then  $VIF = \text{infinity}$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which lead to  $1/(1-R^2)$  infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

**Answer 6:** Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

It is used to check following scenarios:

- If two data sets
  - come from populations with a common distribution
  - have common location and scale
  - have similar distributional shapes
  - have similar tail behavior

### **Advantages**

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.
- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis