==

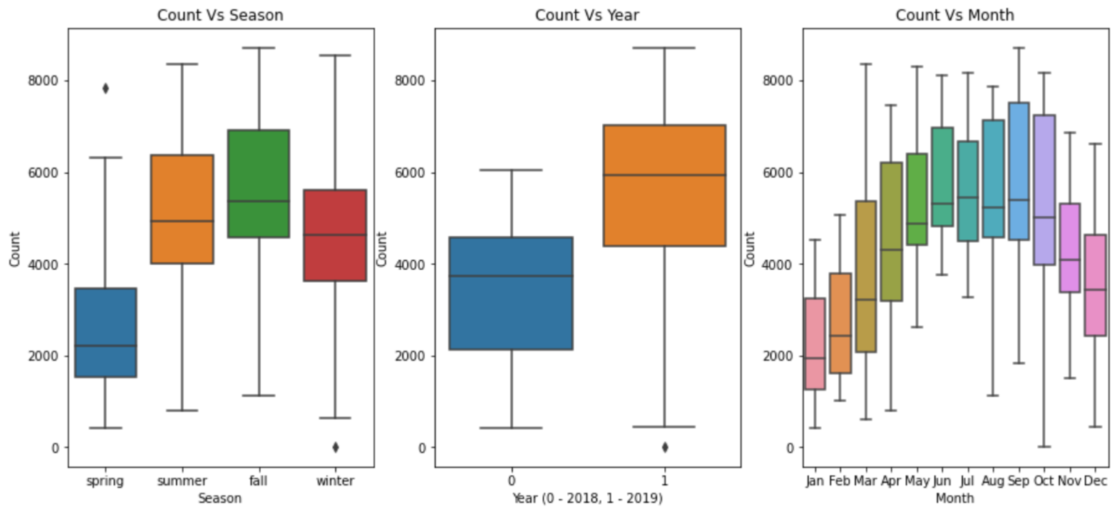
Assignment-based Subjective Questions

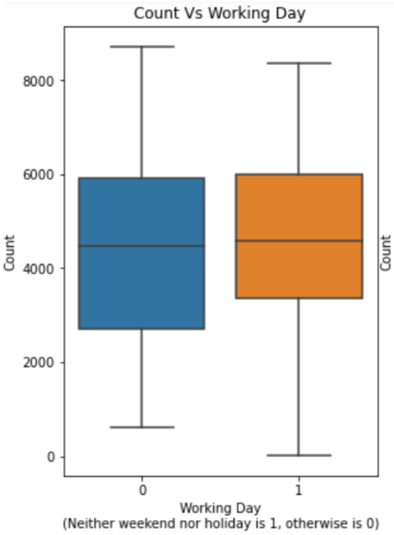
==

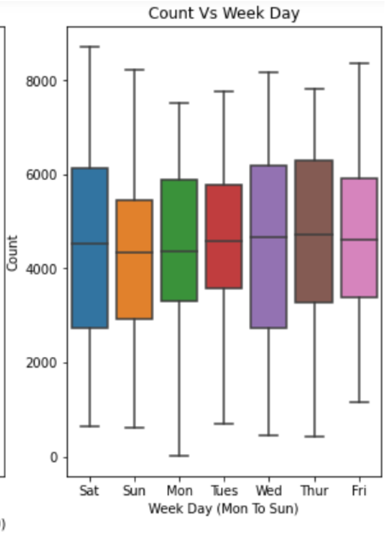
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Answer 1:
In this assignment, based on analysis and data dictionary provide we came across “season”, “yr”, “holiday”, weathersit”, “mnth” and “weekday”. So best way to infer categorical variable is visualize using Boxplot. All these variables have some effect on target or dependent variable “cnt” as below:

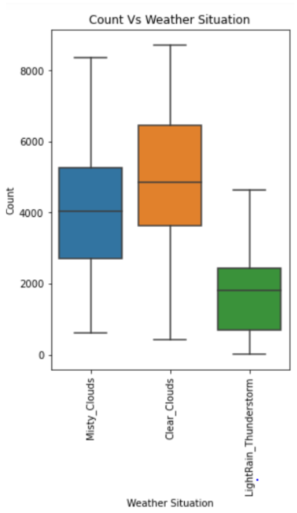
- **Season**: The count of bike sharing is least in spring season while maximum in fall season. In summer and winter bike sharing count is satisfactory or intermediate but far better than spring.
- **Yr**: Count of bike sharing increased in 2019 as compare to 2018.
- **Holiday:** Count of bike sharing is less on holidays.
- **Weathersit:** Count of bike sharing is more and favorable when weather is “Clear, Few Clouds and Partly Clouds” while it is totally opposite or unfavorable (i.e., No customers) when weather is “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog".
- **Mnth:** Bike sharing count increase during between April to October while the highest bike sharing is in the month of September.
- **Weekday:** Daily on an average 4000 – 5000 booking took place.
- **Workingday:** Overall total around 68% count of bike sharing booking is more as compare to non-working day.

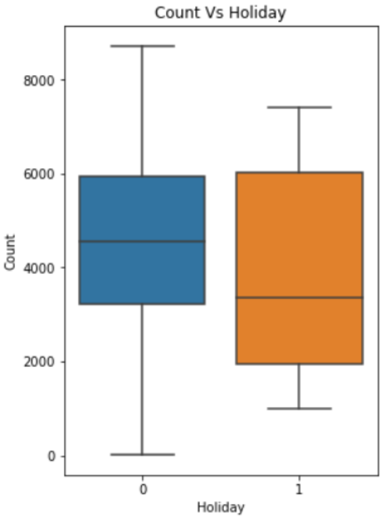
Below is the graphical representation of all categorical variable with respective to count.











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

Answer 2:

Yes, it is very important to use **drop_first=True** option during dummy variable creation. This will avoid to create redundant features. If we don't do that then it will impact our model creation and end result. More dummy variable may lead to multi-collinearity between independent variables.

Let's take an example like we have a categorical feature "XYZ" whose values are "yes" or "no" which is equivalent to "1" or "0". So once dummy variable is created for "XYZ", it will end up in below table

	XYZ_yes	**XYZ_no**
0	0	1
1	1	0

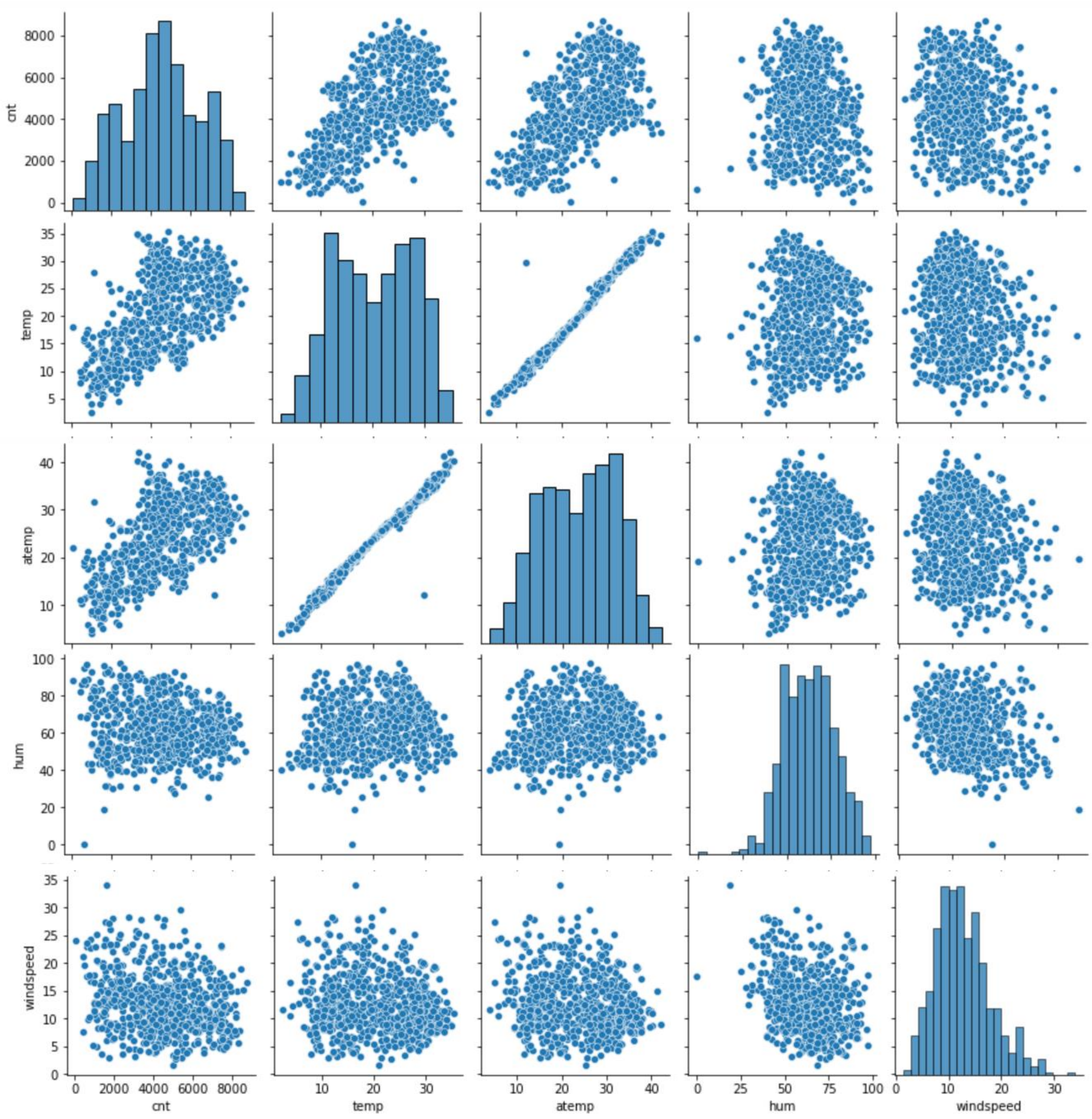
As you can see in above able both columns are giving same result i.e.

IF XYZ_yes is equal to “0” then XYZ_no is automatically “1” and vice-versa. So rather than keeping both columns which gives same result we have to drop first column and after which only “XYZ_no” will left with two values which is “1” or “0”.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer 3:

Based on **pair-plot** on all numerical variables, "**temp**" and "**atemp**" is highly correlated with target variable "**cnt**"



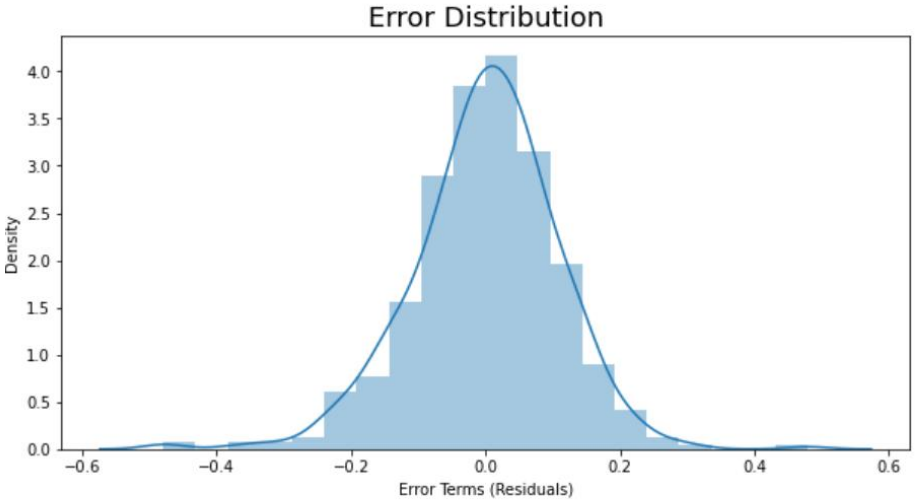
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer 4:

Once model is prepared on training set, we validate using below assumption:

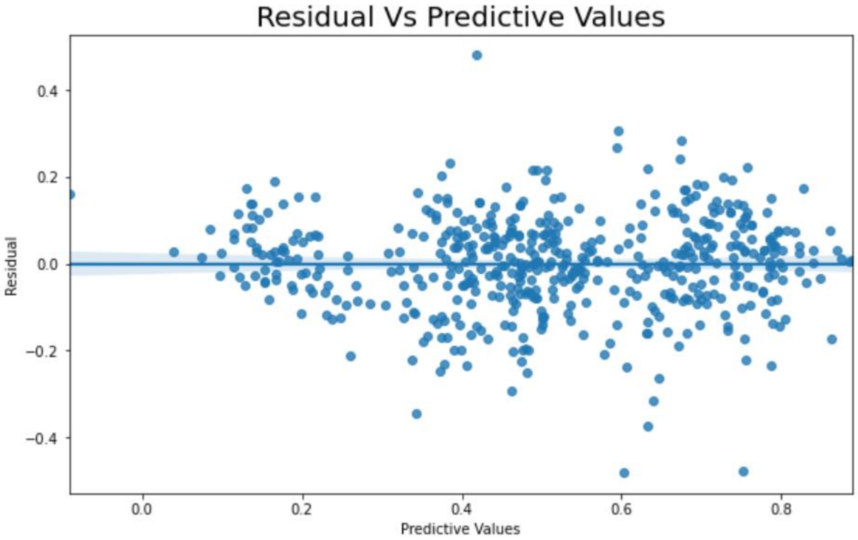
- **Normal Distribution of the error terms**

  Validate if the error terms (or residual distribution) are normally distributed i.e. around 0 or not (which is infact, one of the major assumptions of linear regression), this is proven by plotting those using distplot.



- **Error terms should be independent**

  Based on below plot, we can see that there is no relation and specific pattern between residual and predicted values which is good sign for model.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer 5:
Top 3 feature which contributes significantly in explaining demand of shared bikes is below

- **Year (yr)** is the most significant with 0.2476 coefficient value
- **Spring Season (season_spring)** with -0.2988 coefficient value signifies that the demand of bike is less in Spring season.
- **Weather condition (weathersit_LightRain_Thunderstorm)** with -0.2964 coefficient value significantly portray that during light rain, snow, thunderstorm and scattered clouds demand is less.

==

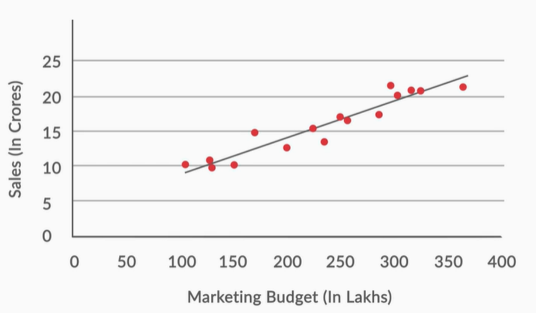
General Subjective Questions

==

Question 1: Explain the linear regression algorithm in detail. (4 marks)
Answer 1:

Linear regression is one of the machine learning algorithms widely used using Supervised Learning method. It is used to predict linear relationship between dependent and independent variable. Regression is performed on dependent variable is continuous or numeric in nature while predictive or independent variable can be of any form like continuous or categorical.

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.



Linear regression can be classified into two types depending upon the number of independent variables:

1. **Simple Linear Regression (SLR)**: SLR is used when the number of independent variables is 1 to predict dependent variable.
   Mathematical equation is:

$$\boxed{\begin{array}{l}\mathbf{Y = c + mx}\\ \quad\textbf{OR}\\ \mathbf{Y = \beta_o + \beta_1 X}\end{array}}$$

   Here,
   Y is predictive or dependent variable (Y-axis)
   X is independent variable (X-axis)
   β_o is Interception on Y when X = 0
   β_1 is slope or gradient or Coefficient of **X** (Change in Y / change in X)

2. **Multiple Linear Regression (MLR):** MLR is used when the number of independent variables is more than 1 to predict dependent variable. Mathematical equation is:

$$\boxed{\mathbf{Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_n X_n}}$$

   Here,
   Y is predictive or dependent variable (Y-axis)
   X_1, X_2, X_3…………X_n independent variables (X-axis)
   β_o is Interception on Y when X = 0
   β_1 is Coefficient of **X_1**
   β_2 is Coefficient of **X_2**
   β_3 is Coefficient of **X_3** and so on…

Question 2. Explain the Anscombe's quartet in detail. (3 marks)

Answer 2:

While going through various description on various search engines, I can conclude that Anscombe's quartet was constructed by statistician Francis Anscombe in 1793 to illustrate the importance of graphical representation before analyzing and model building.

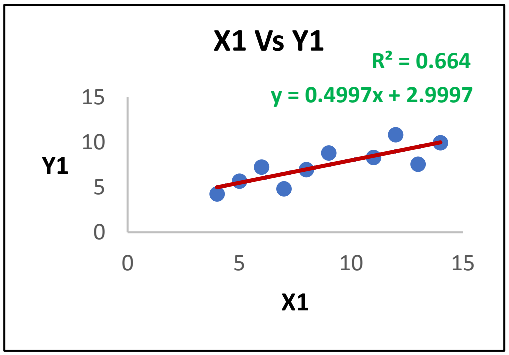
Anscombe's quartet build on 4 Data Set which is symmetrical in terms of statistical value of below:

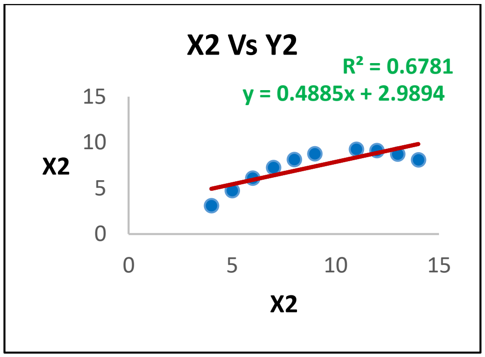
- Mean of X and Y values
- Standard deviation of X and Y values
- Correlation between X and Y will be same for all 4 Data Set
- Best Fit line equation will also be same

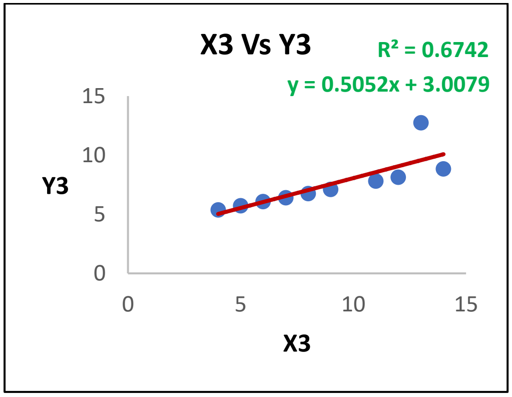
Though statistical values for all 4 data set are same but it's graphical representation will give you totally different result. So due to only statistical value we cannot predict which regression model is good.

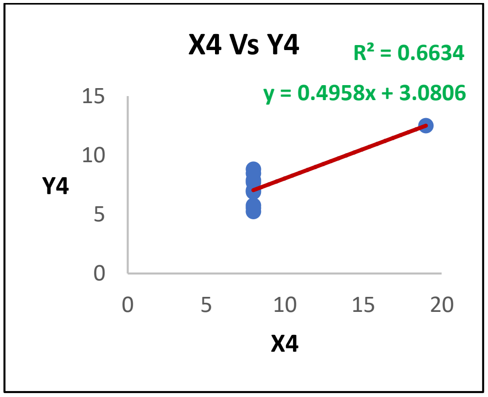
Let’s take an example of below 4 Data Set.

Anscombe's Data								
Observations	Data Set 1		Data Set 2		Data Set 3		Data Set 4	
	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	5.24
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
Summary Statistics								
N	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00
Mean	9.00	7.50	9.00	7.50	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-Square	0.67		0.67		0.67		0.67	


X1 Vs Y1


X2 Vs Y2


X3 Vs Y3


X4 Vs Y4

Outcome of above 4 datasets:

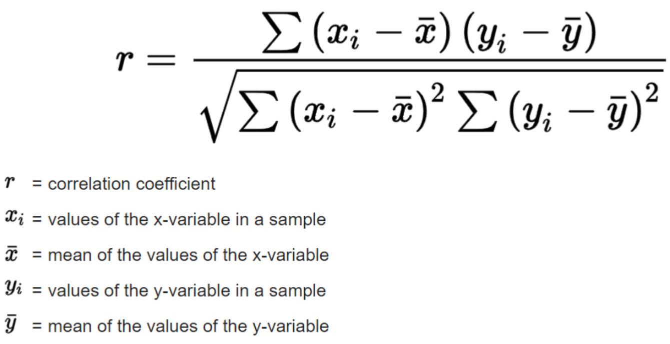
- **Dataset 1**: Best fits the linear regression model.
- **Dataset 2**: Not best fit linear regression model and non-linear data.
- **Dataset 3**: Outliers involved in the dataset cannot be handled by linear regression model
- **Dataset 4:** Outliers involved in the dataset cannot be handled by linear regression model

So, all the important features in the dataset must be visualized before implementing any machine learning algorithm which will help to make a good fit model.

Question 3. What is Pearson's R? (3 marks)
Answer 3:
Pearson's r is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. Basically, it tells us can we draw a line graph to represent the data?

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

Below, is the formula to calculate Pearson's R for a given dataset



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

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

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer 4:
Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data preparation to handle highly varying magnitudes or values or units.

It is extremely important to rescale the variables so that they have a comparable scale.
If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other

coefficients. It also will help us with interpretation and faster convergence of gradient descent.

The most popular methods for scaling:

- **Normalized** scaling means to scale a variable to have values between 0 and 1.
- **Standardized** scaling refers to transform data to have a mean of zero and a standard deviation of 1. It does not have a bounding range. So, even if we have any outliers in our data then they will not be affected by standardization

Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer 5:

As we know that VIF = 1 / (1 – R^2). So based on above question it means that there is a perfect correlation between the independent variables when R^2 value is "1"

VIF = 1 / (1 -1) = 1/0 => Result to infinity

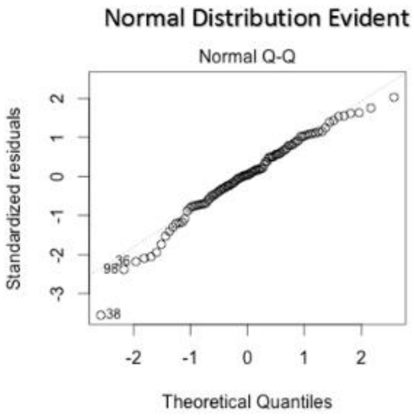
Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer 6:

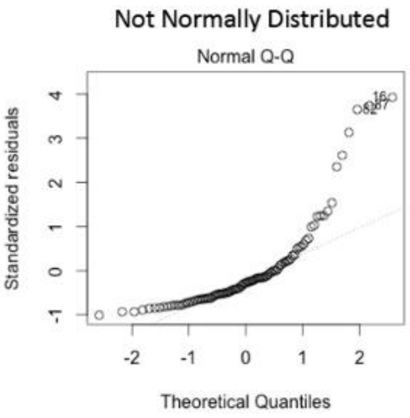
QQ-plot or quantile-quantile plot is formed when quantiles of two variables are plotted against each other. Generally, all points should lie on or close to the straight line at an angle of 45°. Scatter plot is used for visualization.

The q-q plot is used to:

- Check the assumption of normally distributed residuals.
- Compare the shapes of distributions and verify the properties such as location, scale, and skewness are similar or different in the two distributions
- Compare a data set to a theoretical model.



Normal Distribution Evident



Not Normally Distributed