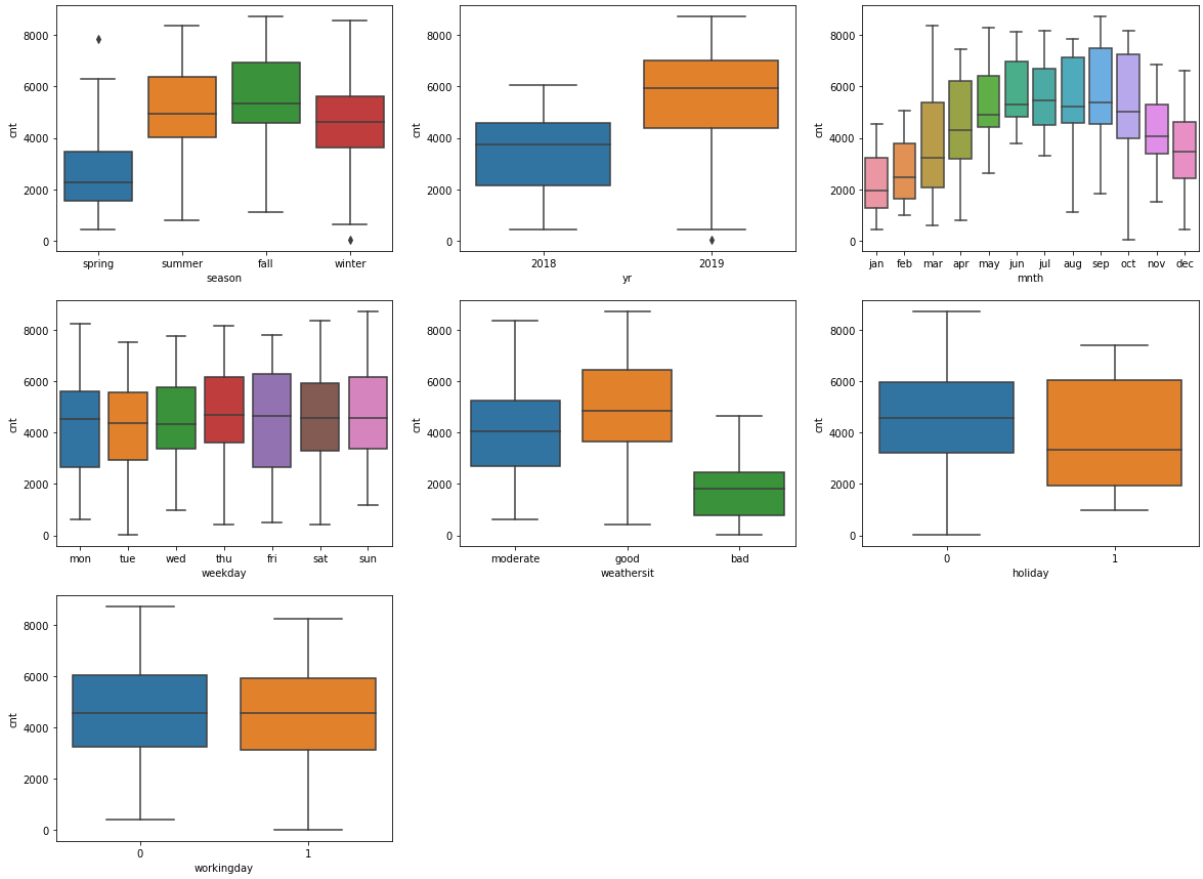
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A. Below are the plots on categorical vs response variable from dataset given, and inferences from these plots.



- Demand in bike sharing is highest in fall season and lowest in spring season.
- Demand for bikes in year 2019 is higher as compared to 2018.
- Demand for bikes is high in the months from May to October.
- No demand for bikes when the weather is severe (high rain/snow/fog), less demand when weather is bad (Light Snow/Rain/thunderstorm/Scattered clouds)
- The demand of bike is almost similar throughout the weekdays.
- Bike demand doesn't change whether day is working day or not.

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

A. To escape from so called Dummy Variable Trap, (or to reduce the collinearity between dummy variables.)

The Dummy Variable Trap occurs when two or more dummy variables created by one-hot encoding are highly correlated (multi-collinear). This means that one variable can be predicted from the others, making it difficult to interpret predicted coefficient variables in regression models. In other words, the individual effect of the dummy variables on the prediction model can not be interpreted well because of multicollinearity.

Using the one-hot encoding method, a new dummy variable is created for each categorical variable to represent the presence (1) or absence (0) of the categorical variable. For example, if tree species is a categorical variable made up of the values pine or oak, then tree species can

be represented as a dummy variable by converting each variable to a one-hot vector. This means that a separate column is obtained for each category, where the first column represents if the tree is pine and the second column represents if the tree is oak. Each column will contain a 0 or 1 if the tree in question is of the column's species. These two columns are multi-collinear since if a tree is pine, then we know it's not oak and vice versa.

Further explanation

To demonstrate the dummy variable trap, consider that we have a categorical variable of tree species, and assume that we have 7 trees:

$$x_{species} = [pine, oak, oak, pine, pine, pine, oak]$$

f the tree species variable is converted to dummy variables, the two vectors obtained:

$$x_{pine} = [1, 0, 0, 1, 1, 1, 0] x_{oak} = [0, 1, 1, 1, 1, 0, 1]$$

Because a 1 in the pine column would mean a 0 in the oak column, we can say xpine=1–xoak. This results in two dummy variables that are multi-collinear, and so the dummy variable trap may occur in regression analysis.

To overcome the Dummy variable Trap, we drop one of the columns created when the categorical variable were converted to dummy variables by one-hot encoding. This can be done because the dummy variables include redundant information.

To see why this is the case, consider a multiple linear regression model for the given simple example as follows:

$$y=\beta0+\beta1xpine+\beta2xoak+\epsilon$$

where y is the response variable, xpine and xoak are the explanatory variables, β0 is the intercept, β1 and β2 are the regression coefficients, and ϵ is the error term. Since these two dummy variables are multi-collinear — hence we know if a tree is pine, then it's not oak — we can substitute xoak by (1–xpine) in the multiple linear regression equation.

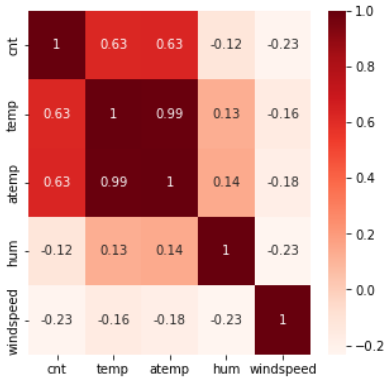
$$y=\beta0+\beta1xpine+(1-xpine)\beta2+\epsilon=(\beta0+\beta2)+(\beta1-\beta2)xpine+\epsilon$$

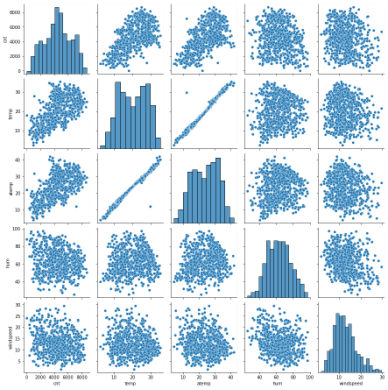
As you can see, we were able to rewrite the regression equation using only xpine, where the new coefficients to be predicted are (β0+β2) and (β1−β2). By dropping a dummy variable column, we can avoid this trap.

This example shows two categories, but this can be expanded to any number of categorical variables. In general, if we have p number of categories, we will use p−1 dummy variables. Dropping one dummy variable to protect from the dummy variable trap.

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

A. Both temp and atemp have the highest correlation of **0.63** with target variable (cnt).





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

A. Below are the assumptions and the approach followed to validate each of these:

Linear relationship: Created a scatter plot x vs y.

If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

No autocorrelation or independence: Conducted a Durbin-Watson (DW) statistic test.

Durbin-Watson (DW) statistic test: The values should fall between 0-4. If DW=2, no autocorrelation; if DW lies between 0 and 2, it means that there exists a positive correlation. If DW lies between 2 and 4, it means there is a negative correlation. Mine was 2.

No Multicollinearity: Determined the VIF (Variance Inflation Factor).

VIF<=4 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity.

Homoscedasticity: Created scatter plot that shows residual vs fitted value.

If the data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity).

Normal distribution of error terms: Using a Q-Q (Quantile-Quantile) plot.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A. As per the final model, the top 3 predictors that are significant in predicting the demand for shared bikes are:

1. Temperature (temp) - coefficient value '0.4554'
2. Year (yr) - A coefficient value '0.2291'
3. Weather Situation (weathersit_bad) - A coefficient value '-0.2099'

General Subjective Questions

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

A. Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

Linear regression is of the 2 types:

i. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression: $Y=\beta0+\beta1X1+\epsilon$

ii. **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression: $Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by the following two methods:

· Differentiation

· Gradient descent

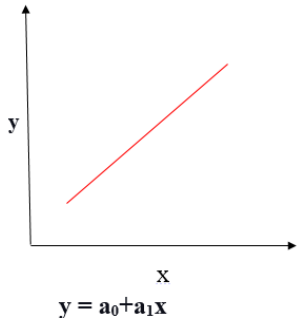
We can use statsmodels or SKLearn libraries in python for the linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables.

The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

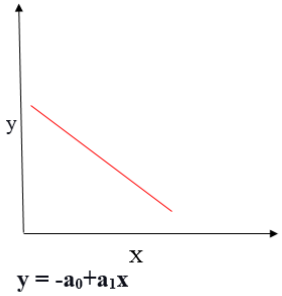
Positive Linear Relationship

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

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

A. Anscombe's Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

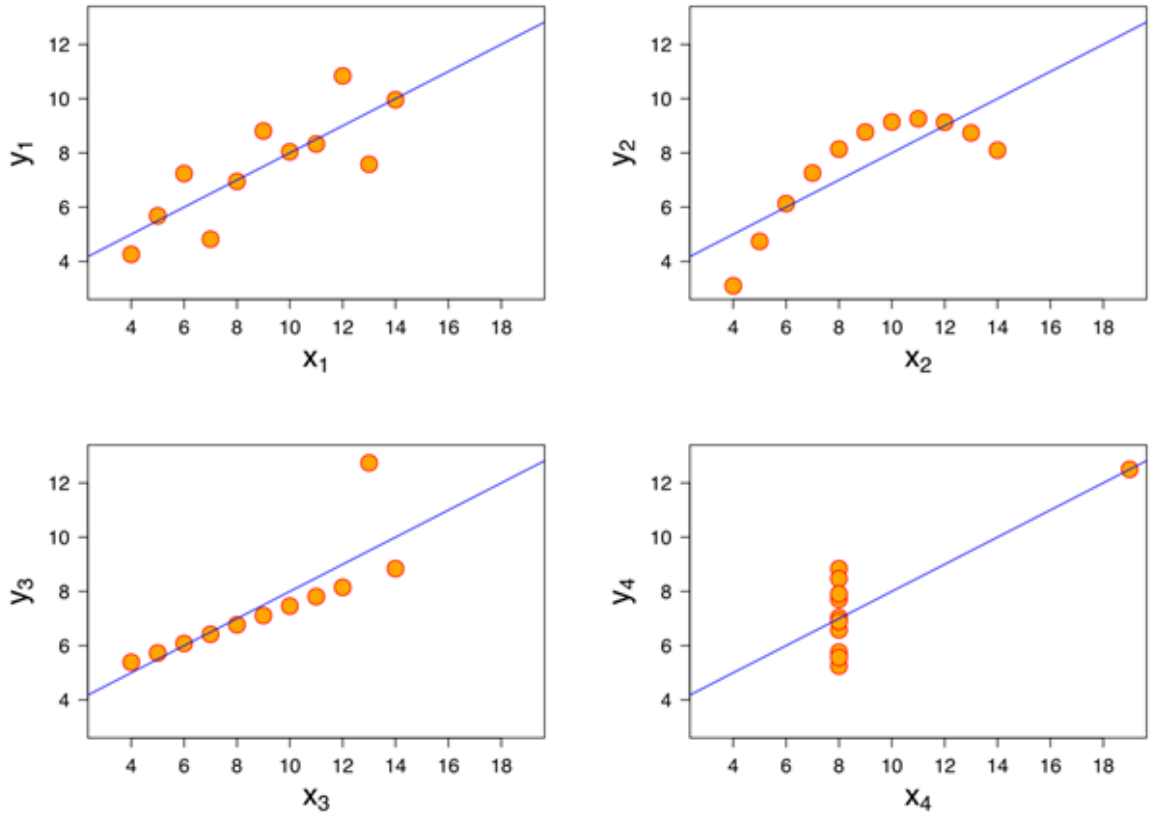
The summary statistics show that the means and the variances were identical for x and y across the groups:

· Mean of x is 9 and mean of y is 7.50 for each dataset.

· Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

· The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



· Dataset I appears to have clean and well-fitting linear models.

· Dataset II is not distributed normally.

· In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

· Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

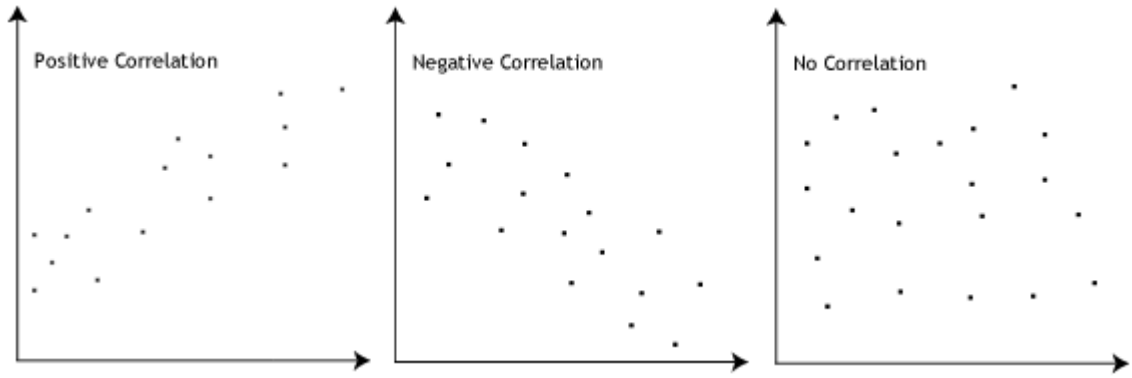
Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

3. What is Pearson's R? (3 marks)

A. In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association



Pearson r Formula

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

Here,
r=correlation coefficient
x_i=values of the x-variable in a sample
x_bar=mean of the values of the x-variable
y_i=values of the y-variable in a sample
y_bar=mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
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 have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

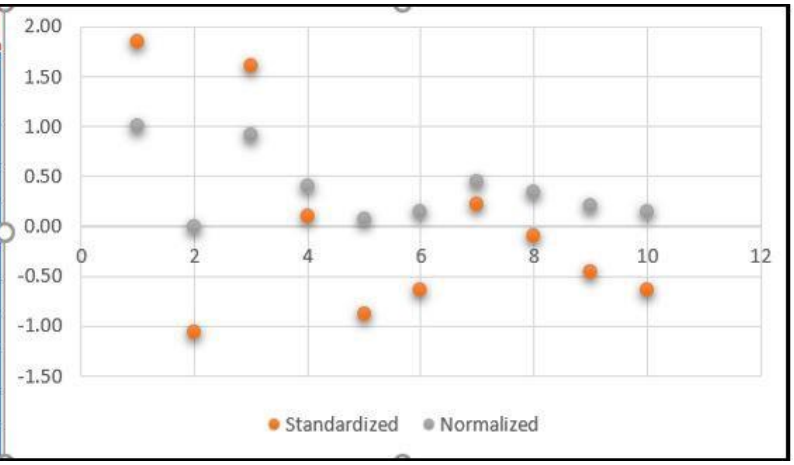
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- sklearn.preprocessing.scale helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Example: Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

A. The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

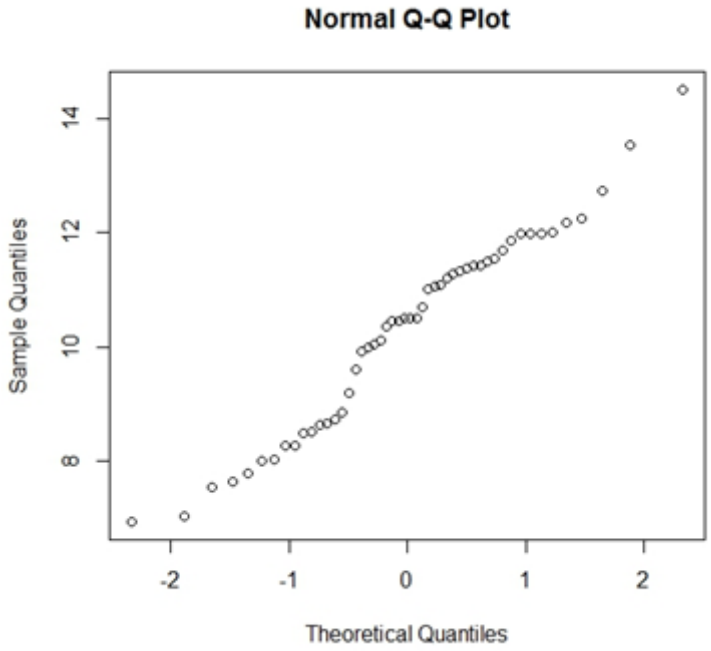
Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

A. The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.