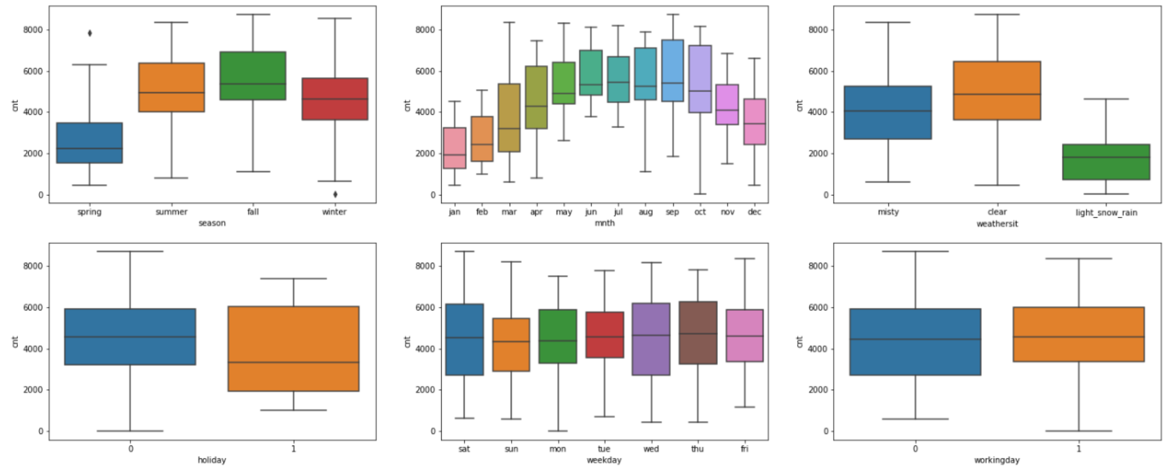
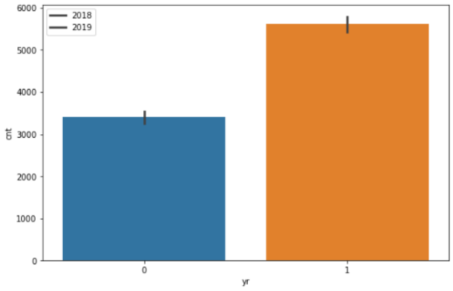
Assignment-based Subjective Questions

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



- For seasons variable, bike sharing is more in fall followed by summer while it is less in spring.
- From the month variable, demand is more from May to October. Least demand is seen in January.
- Demand is more during clear good weather.
- Bookings are more on days which are not holidays.
- There is not much trend observed on weekdays variables.
- Slightly more demand is observed on working days
- 2019 attracted more customer from the previous year. This shows good progress in terms of business.



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

For a categorical variable with n distinct categories, dummy variables are used to create n columns if drop_first = True is not used. However, it is ideal to use drop_first = True in order to create (n-1) columns as the presence of this first column has no impact whatsoever on the resultant model. Apart from this, dropping the first column also helps reduce the collinearity between the dummy variables.

To create a good model, we require features that are relevant and during creation of dummy variables, the first columns is irrelevant to our training and hence its important to use drop_first = True.

For example, lets consider a weather table for a particular city. There are 3 categories of weather for the city, namely, Sunny, Cloudy, Rainy. The below table is created as dummy variables before dropping the first column.

SUNNY	CLOUDY	RAINY
1	0	0
0	1	0
0	0	1

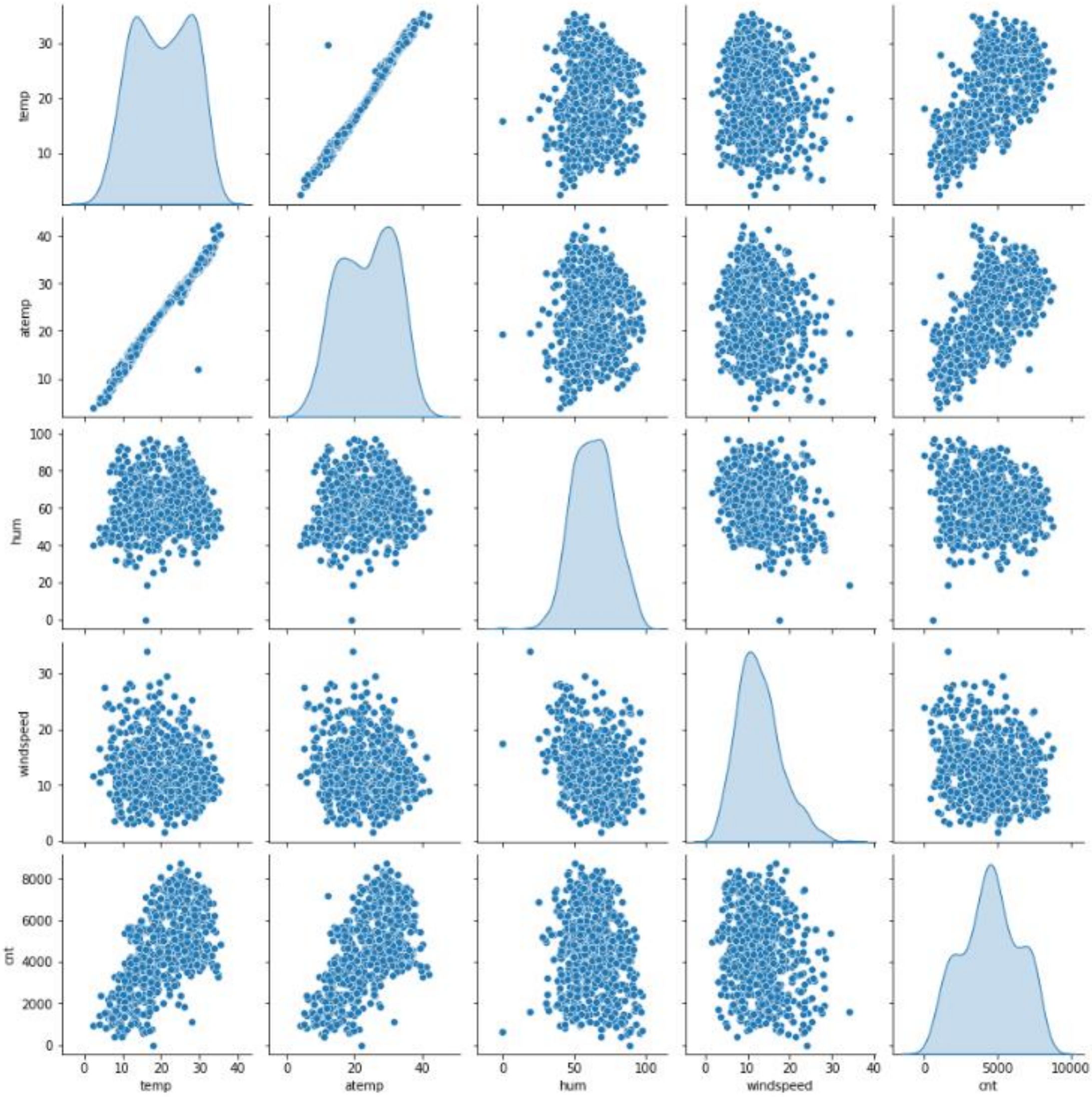
Lets check the impact after removing the first column.

CLOUDY	RAINY
0	0
1	0
0	1

We notice that when cloudy and rainy are both zeros, it indicates sunny even without the presence of the sunny column.

Hence, we only require (n-1) columns during dummy creation.

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

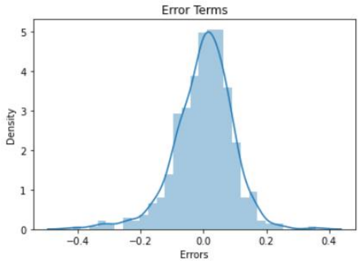


From the above pair plots, we can clearly see that *'temp'* and '*atemp*' has highest correlation with the target variable '*cnt'*.

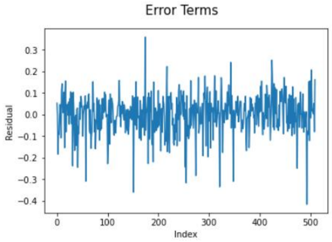
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Assumption of Linear Regression are as below:

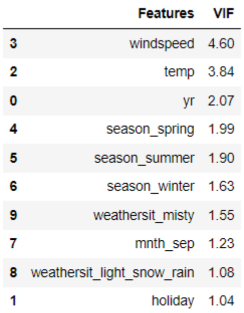
1. Error terms are normally distributed. This was verified by plotting the residual distplot.



2. Independence of residuals – there should not be any visible patterns in the spread of error terms which was validated using the below plot. No autocorrelation of errors.

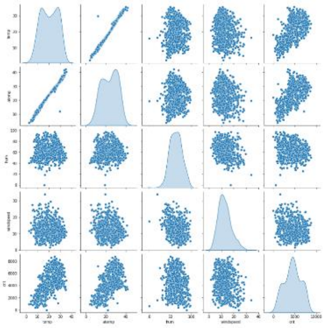


3. Multicollinearity check was performed by checking the VIF values of the variables. There should little to no multicollinearity between variables.

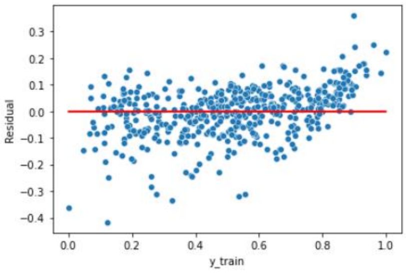


	Features	VIF
3	windspeed	4.60
2	temp	3.84
0	yr	2.07
4	season_spring	1.99
5	season_summer	1.90
6	season_winter	1.63
9	weathersit_misty	1.55
7	mnth_sep	1.23
8	weathersit_light_snow_rain	1.08
1	holiday	1.04

4. Linear relationship between variables was verified using the pairplots.



5. Homoscedasticity- Error terms have constant variance



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

Important variables from the model we have finalised are :

I. Temperature- it shows a linear relation with the demand. When temp increases, demand was also higher
II. Seasons- There is a better demand in fall followed by summer.
III. Weather Situation- Popularity for bike sharing is more when the weather is clear.

General Subjective Questions

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

Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, it finds how the value of the dependent variable is changing according to the value of the independent variable.

It is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables

Linear Regression is of two kinds:

I. Simple Linear Regression - uses a one single independent variable to predict the target variable
II. Multiple Linear Regression - uses multiple independent variables to predict the target variable

Equation for Simple linear regression:

y(x) = B0 + B1x

where, y = output variable. Variable y represents the continuous value that the model tries to predict.

x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.

B0 = y-axis intercept (or the bias term).

B1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.

Equation for Multiple linear regression:

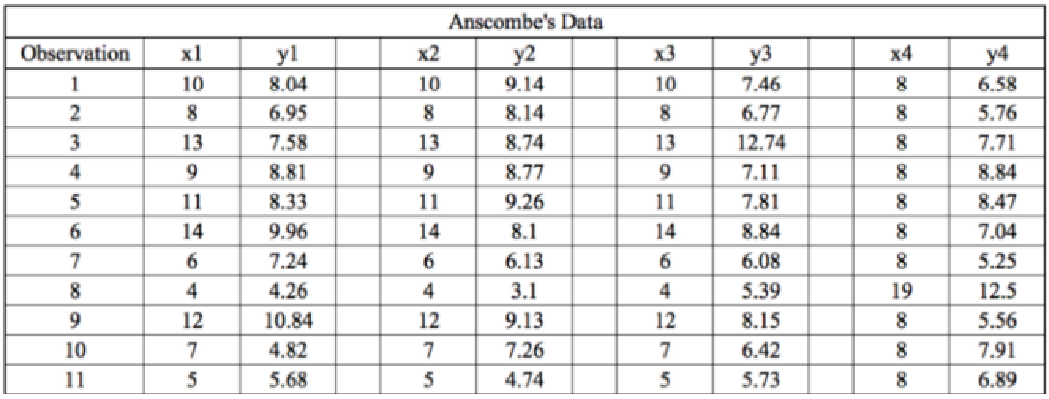
The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = B0 + B1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n).

y(x) = B0 + B1x1 + B2x2 + … + B(n)x(n)

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

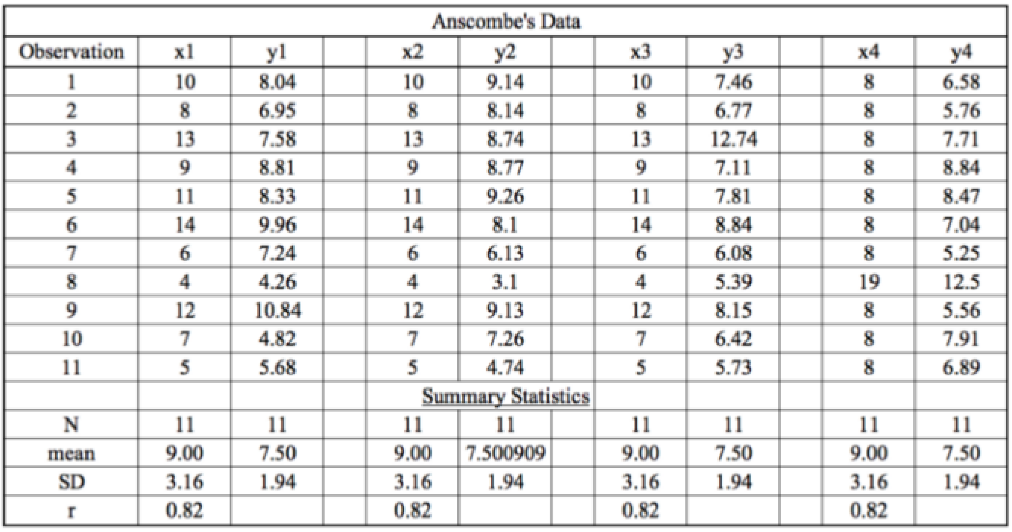
Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe's quartet intended to counter the impression among statisticians that "Numerical calculations are exact, but graphs are rough".

These four plots can be defined as follows:



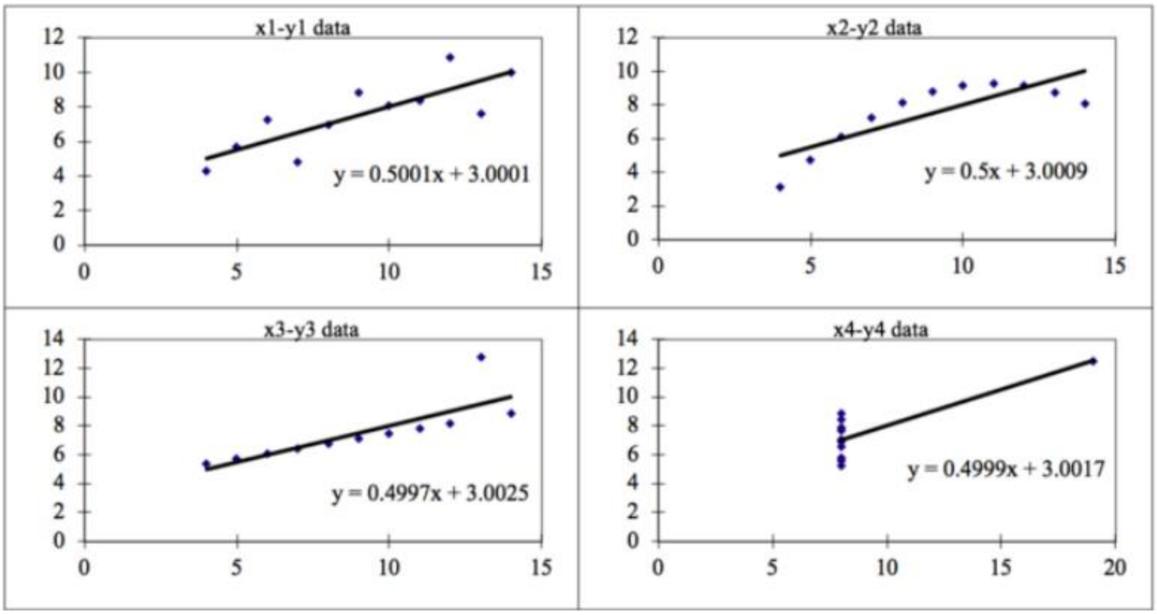
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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

The statistical information for all these four datasets are approximately similar and can be computed 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
					Summary Statistics						
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	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. The first scatter plot (top left) appears to be a simple linear relationship,

2. The second graph (top right); cannot fit the linear regression model because the data is non-linear

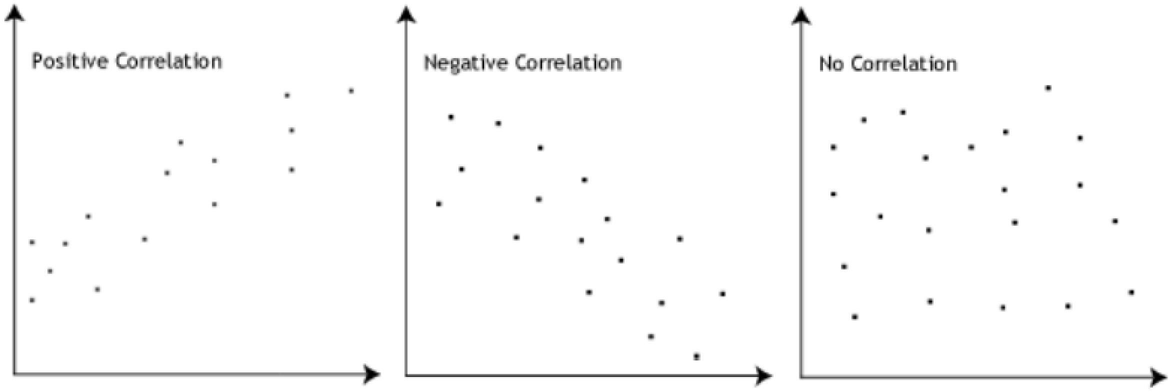
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Finally, the fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables. It shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that describes the linear relationship between two continuous variables. It is a value between-1 and 1, where-1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

Pearson's r measures the degree to which the variables are related by calculating the ratio of the covariance between the variables to the product of their standard deviations. In other words, it measures how much the variables vary together relative to how much they vary independently.



Pearson's R Formula is as follows:

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

Here,

I. r = correlation coefficient
II. xi = values of the x-variable in a sample
III. x̅= mean of the values of the x-variable
IV. yi = values of the y-variable in a sample
V. y̅ = mean of the values of the y-variable

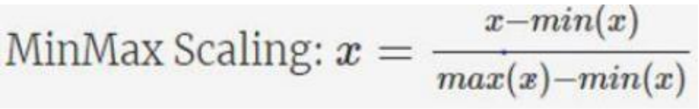
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. In simpler terms, in machine learning algorithms we need to bring all features in the same standing, so that one significant number doesn't impact the model just because of their large magnitude. This is called scaling or Feature scaling.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. 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 which results in an incorrect model. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

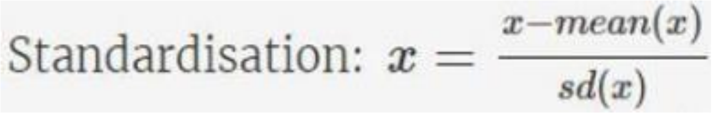
It is important to note that scaling just affects the coefficients and none of the other parameters like tstatistic, F-statistic, p-values, R-squared, etc.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. It brings all of the data in the range of 0 and 1



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

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. 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).



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

Normalization	Standardization
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is used when we want to ensure zero mean and unit standard deviation.	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is less affected by outliers.
It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of 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).

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

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. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. The power of Q-Q plots lies in their ability to summarize any distribution visually.

The advantages of the Q-Q plot are:

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

Q-Q plot is very useful to determine:

I. If two populations are of the same distribution
II. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
III. Skewness of distribution