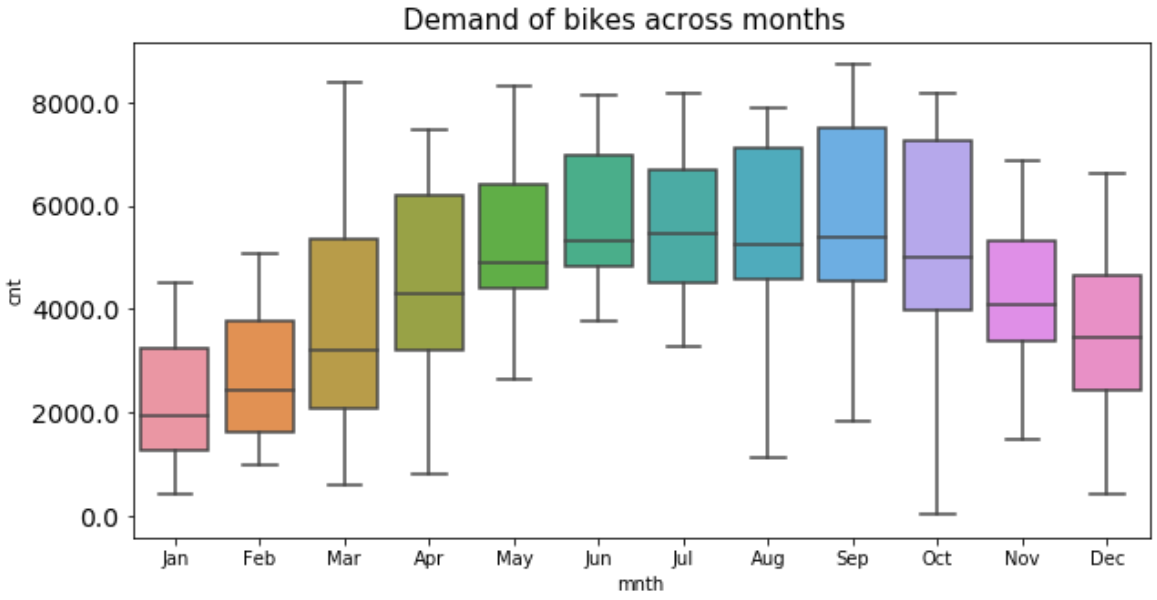
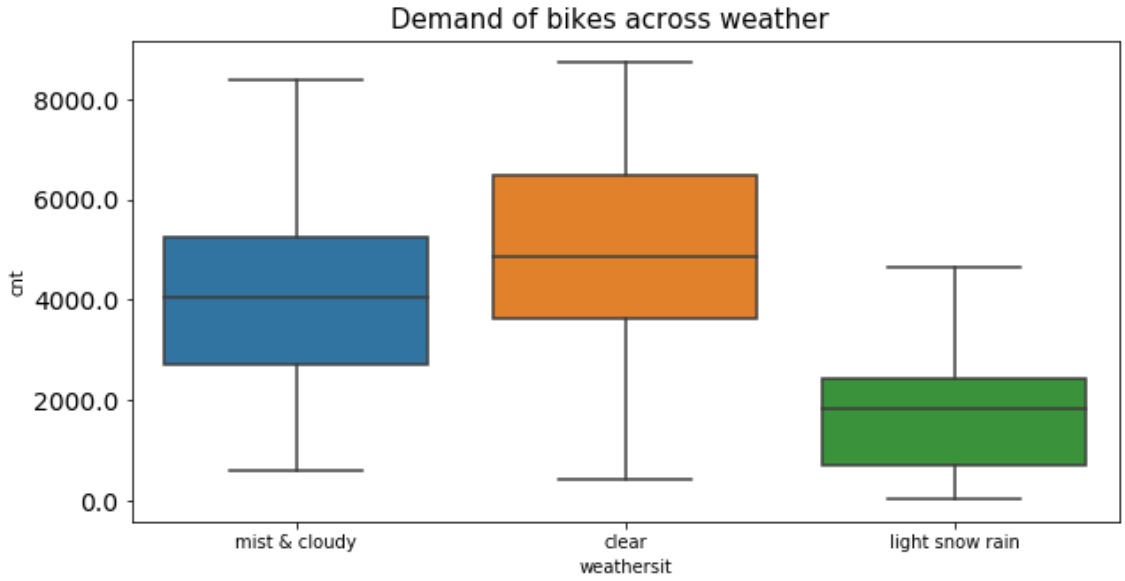
Assignment-based Subjective Questions

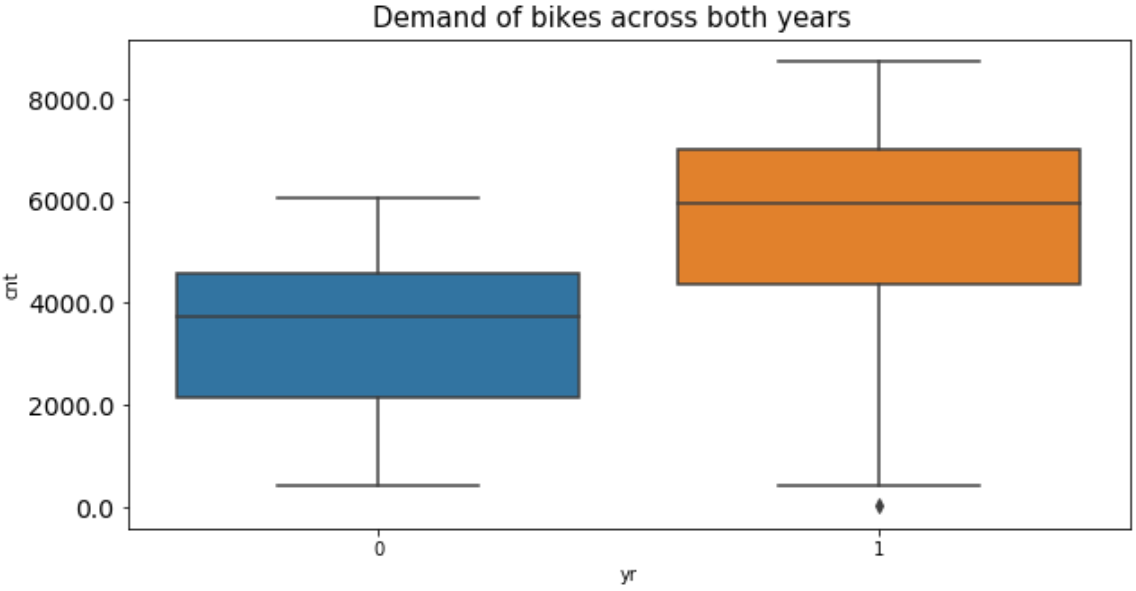
Q.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

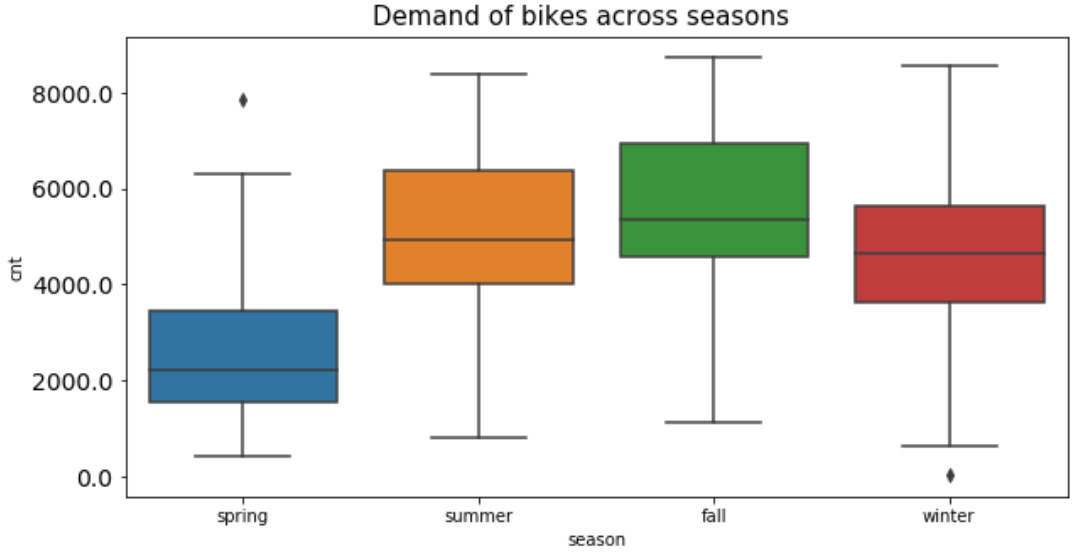
Ans. Let us look at the categorical variables we have.

- **Season:** There is a strong relationship between season (categorical variable) & the demand (output). Fall has maximum demand followed by Summer, and then Winter and spring.
    - **Fall** has maximum demand.
    - **Spring** has minimum demand.
- **Month:** Demand in general increases steadily from **Feb** to **Sept** and then starts dropping.
    - **Dec & Jan** are months with considerably lower demands.
- **Year:** Demand of later year (with value 1) is considerably greater than the prior year (with value 0).
- **Weekdays:** There isn't a strong relationship between demand & weekdays. The demand is on average similar across days.
- **Weather:** The demand in **clear** weather is significantly higher than others.
    - **Light snow rain** has least demand among other weather.
- **Holiday:** Demand on holiday is more varying than the demand on non-holiday.
    - The median of demand lies higher in case of non-holiday than the holiday, so on average there is more consistent demand on a non-holiday.
- **Working day:** There isn't a strong relationship between demand & working/non-working day. The demand seems same on working as well as non-working day.




Demand of bikes across weather
cnt
8000.0
6000.0
4000.0
2000.0
0.0
mist & cloudy
clear
light snow rain
weathersit


Demand of bikes across both years
cnt
8000.0
6000.0
4000.0
2000.0
0.0
0
1
yr


Demand of bikes across seasons
cnt
8000.0
6000.0
4000.0
2000.0
0.0
spring
summer
fall
winter
season

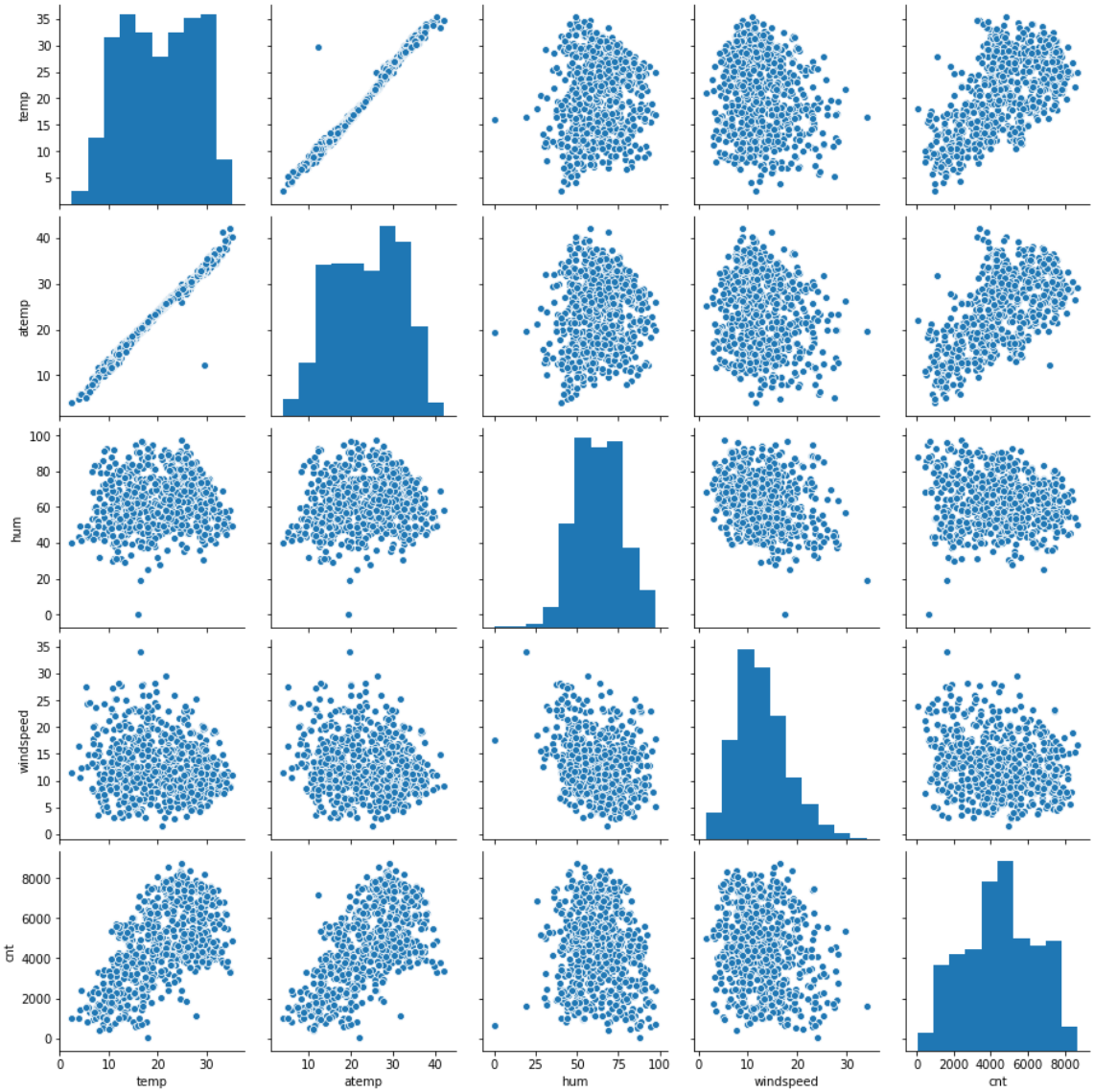
Q.2 Why is it important to use drop_first=True during dummy variable creation?

Ans. We need **k-1** dummy variables to denote **k** unique terms.

So, we can simply drop the first column, which will bring down the count from k to k-1.
In general, we always try to represent our model in lesser variables. Since, more the number of variables, more the model be subjected to changes if one variable is altered.
So, we always use **k-1** variables instead of **k** variables.

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

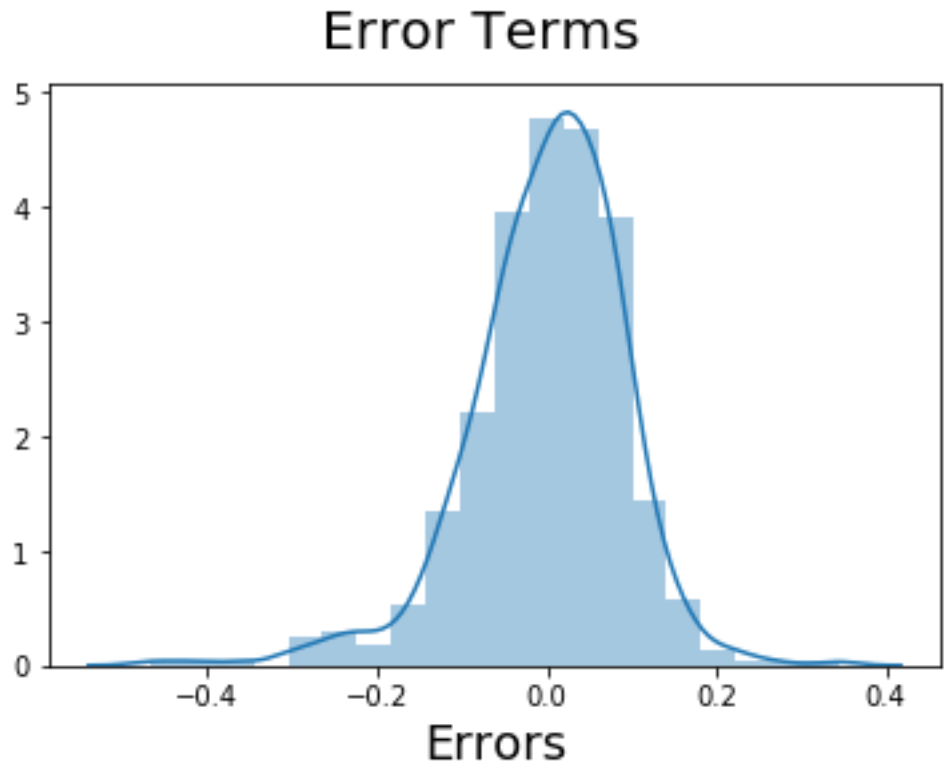
Ans. **temp** & **atemp** has the highest correlation with the target variable.
Also, among themselves their plot is almost symmetric, i.e., high level of multi-collinearity is present in them.



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

Ans. By performing **Residual Analysis**, we plot the histogram of errors.

- We can see that the errors are normally distributed.
- The plot is centred around zero.
- The plot is normally distributed.



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

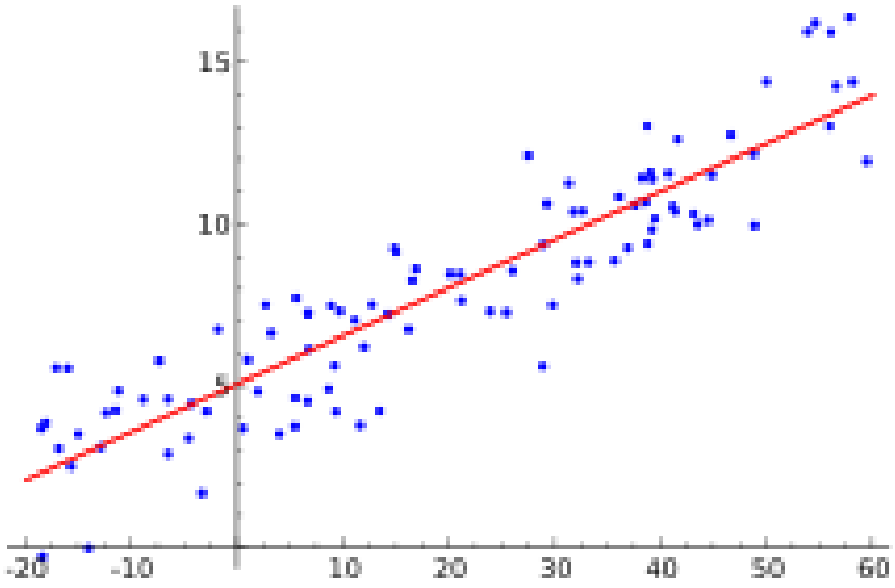
Ans. The top 3 contributors of the demands are as follows:

- **Temperature:** Temperature has a positive coefficient with a value of +0.358. So, it is positively affecting the target variable.
- **Light Snow Rain, i.e., weather:** It has a negative coefficient with a value of -0.26. So, it is negatively affecting to the target variable.
- **Year:** Year is also positively correlated with the coefficient of +0.229. Hence, It is also positively affecting the target variable.

General Subjective Questions

Q.1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a linear approach to model the relationship between a dependant variable and 1 or more independent variables.
In case where there is only one independent variable, it is referred to as Simple Linear Regression.
In case of more than one independent variables, it is referred to as Multiple Linear Regression.



In Laymen terms, Linear regression is method to find the best fit lines among the datapoints.

The linear regression fits one such line which has minimum sum of squares.

Where, Sum of squares are calculated as the sum of the squares of the differences between the predicted value and the actual value.

Linear Regression has some assumptions. We need to know if the assumptions hold on our dataset, only then we can apply linear regression on the data. The assumptions are mentioned below:

- **Weak Exogeneity:** It means that the predictor variable x can be treated as a fixed value, rather than the random variables.
- **Linearity:** There exist a linear relationship between the independent and dependent variables. In other words, mean of response variable is a linear combination of the parameters and predictor variables.

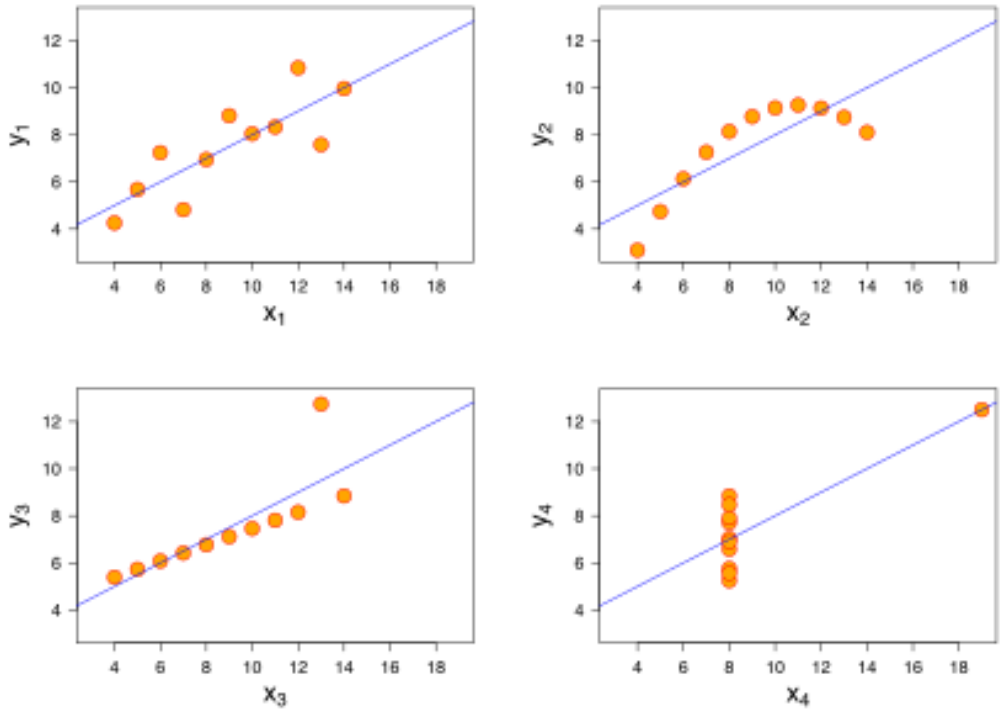
- **Constant Variance:** It means that the different values of response variable have same variance in their errors, regardless of the values of the predictor variables.
- **Multi-collinearity:** The errors of the response variables are uncorrelated with each other. In other words, there isn't a correlation between the error terms.

Usage of Linear Regression:

- Linear regression is widely used for prediction, forecasting or error reduction tasks.

Q.2 Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a group of four datasets that appear to be similar when using typical summary statistics yet tell four different stories when graphed.
In Anscombe's quartet, when each of the 4 data sets are graphed, they have very different distributions and appear very different.



They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.

Let's review all 4 datasets:

- The first scatter plot looks like a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph is not distributed normally. It is evident of visualising that the relationship between the two variables is not linear.
- In the third graph, the distribution is linear, but should have a different regression line.
- The fourth graph 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.

So, we can see how 4 datasets are very different from one another in Anscombe's quartet.

Q.3 What is Pearson's R?

Ans. The Pearson correlation coefficient is a statistic that measures linear correlation between two variables X and Y. Its values ranges from +1 and −1.

- **A value of +1:** means the data is perfectly linear with a positive slope
- **A value of 0:** means there is no linear association
- **A value of −1:** means the data is perfectly linear with a negative slope
- **A value between 0 to 5:** means there is a weak association
- **A value between 5 to 8:** means there is a moderate association
- **A value > 8:** means there is a strong association

By definition, Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

For the Pearson r correlation, there are few assumptions that needs to be their.

- Both variables should be normally distributed.
- There should be no significant outliers.
- Each variable should be continuous variable.
- The two variables have a linear relationship.
- Observation are paired observations, i.e., for every observation of the independent variable, there must be a corresponding observation of the dependent variable.
- Homoscedascity – It refers to the equal variances.

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

Ans. Feature Scaling is a technique to standardize the independent features present in the data to bring it down to a fixed range. In other words, Feature scaling is a method used to normalize the range of independent variables or the features of data.

Why scaling is performed?

Scaling is performed during the data pre-processing to handle highly varying values in the dataset.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Some machine learning algorithms are sensitive to feature scaling while others are invariant to it.

- **Gradient Descent Based Algorithms**
    - Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled.
    - If we look at the formula for Gradient descent, we can see that is has a X term. This value of X will directly affect the step size of the gradient descent.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, the data needs to be scaled before feeding it to the model.

Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

- **Distance based algorithms**
    - Distance algorithms like KNN, K-means, SVMs are most affected by the range of features.
    - Feature scaling is also very important for these algorithms.

The two common scaling techniques are:

1. ***Min-Max Normalization***
2. ***Standardization***

Normalized scaling v/s standardized scaling

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Some of the differences b/w them are given in below table.

S. No	NORMALISATION	STANDARDISATION
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is affected by outliers.	It is much less affected by outliers.
5	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
6	It is also known as Scaling Normalization	It is also known as Z-Score Normalization.

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

Ans. VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
In other words, a high VIF value signifies high level of collinearity.

Why does VIF turns infinite?

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Let's look at the formula for VIF.

VIF = 1 / (1 – R2)

It is evident from the formula, when the value of **R2** reaches 1, the VIF will turns infinite.

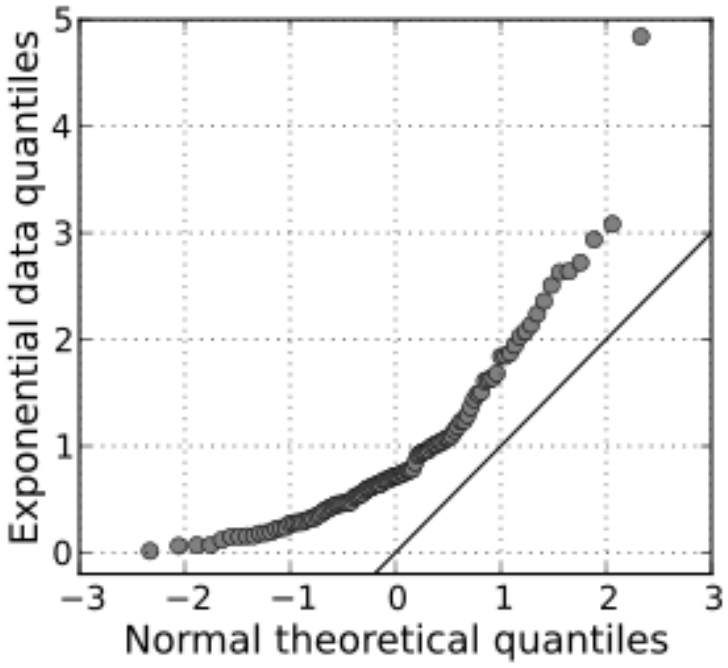
When the regression predictions perfectly fit the data, the R2 value is calculates to 1.

R2 = 1 indicates perfect fit that is you've explained all of the variance that there is to explain.

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

Ans. A Quantile-Quantile plot or Q-Q is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
- Whereas, If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.



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/different in the two distributions.

Importance & Use

- Q-Q plots is used to find out if two sets of data come from the same distribution.
- This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

- It can used with sample sizes.

- Distribution aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check If two data set –

1. Come from populations with a common distribution
2. Have common location and scale
3. Have similar distributional shapes
4. Have similar tail behaviour.