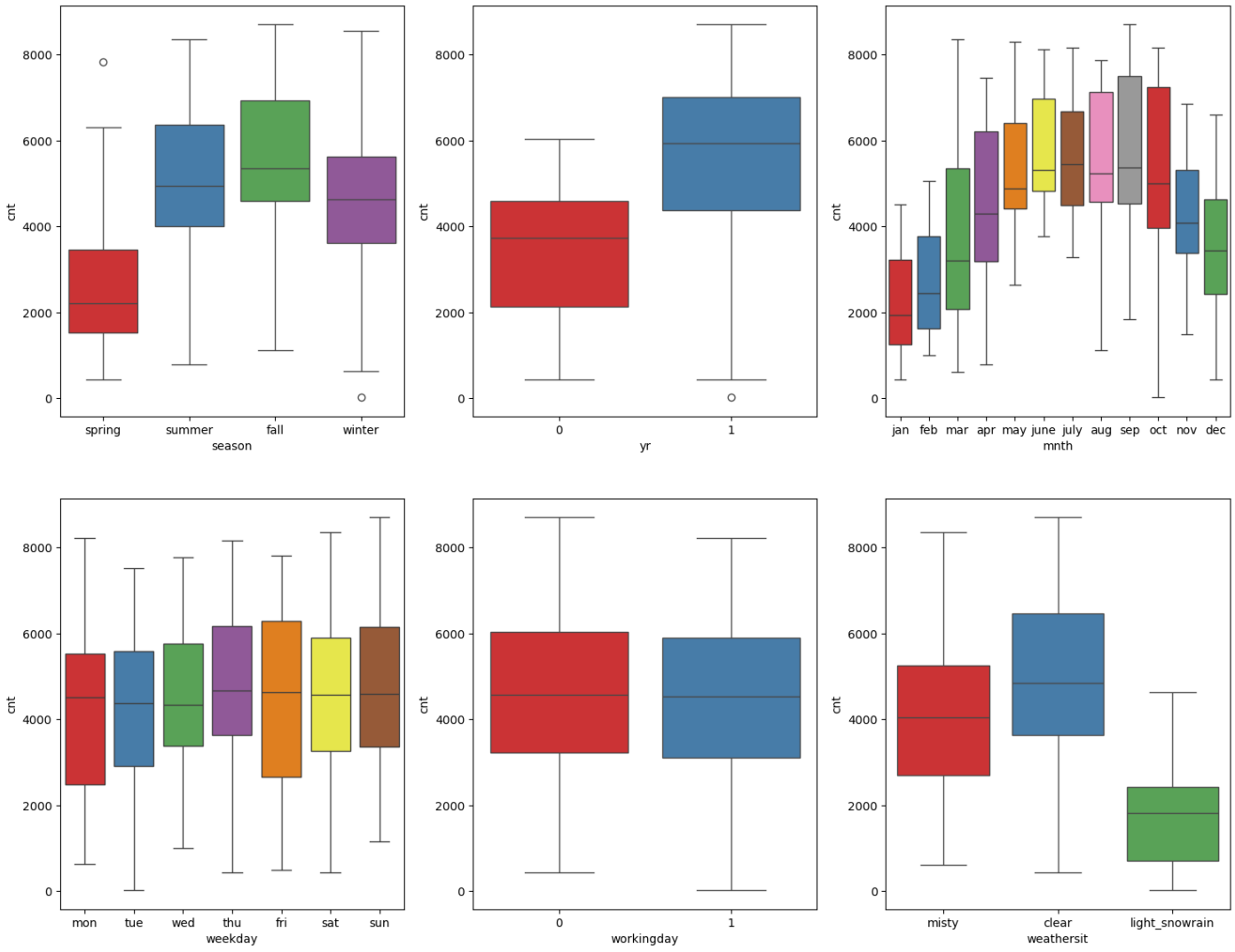
Assignment-based Subjective Questions

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

Ans:

After doing analysis on categorical variables from the dataset, we can infer that variable like 'season', 'yr', 'mnth', 'weekday', 'workingday' and 'weathersit' has major effect on the dependent variable '**cnt**'.

Please find below boxplot showing the correlation of variables with respect to 'cnt'.



Below are the some points we inferred from the analysis :

- Bike bookings are seen more in Fall season and becomes less while spring season.
- In every season bookings were increased drastically from year 2018 to 2019.
- Most of the bookings were made in June, July, Aug, Sep, Oct months, we can see bookings started to increase from year start till peak point in sep and then decreased at year end.
- People are more likely to book bikes on Thursday, Friday, Saturday and Sunday, as they seem to have holidays for these week days.
- It doesn't matter whether its working day or non-working day, bookings are done almost for either of the days.
- People would like to have majority of their bookings done when the weather is clear. And very less bookings when weather is likely to be snowy or rainy.

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

Ans :

It is important to use **drop_first = True** while creating dummy variables because it removes the first column which is created for the first unique value of a column.

Hence, it helps to avoid multicollinearity by reducing the extra column made while dummy variable creation.

For Example: You have categorical variable named '*Relationship Status*' with 'n'=3 levels, namely, '*Single*', '*Divorced*', and '*Married*' as shown in below table.

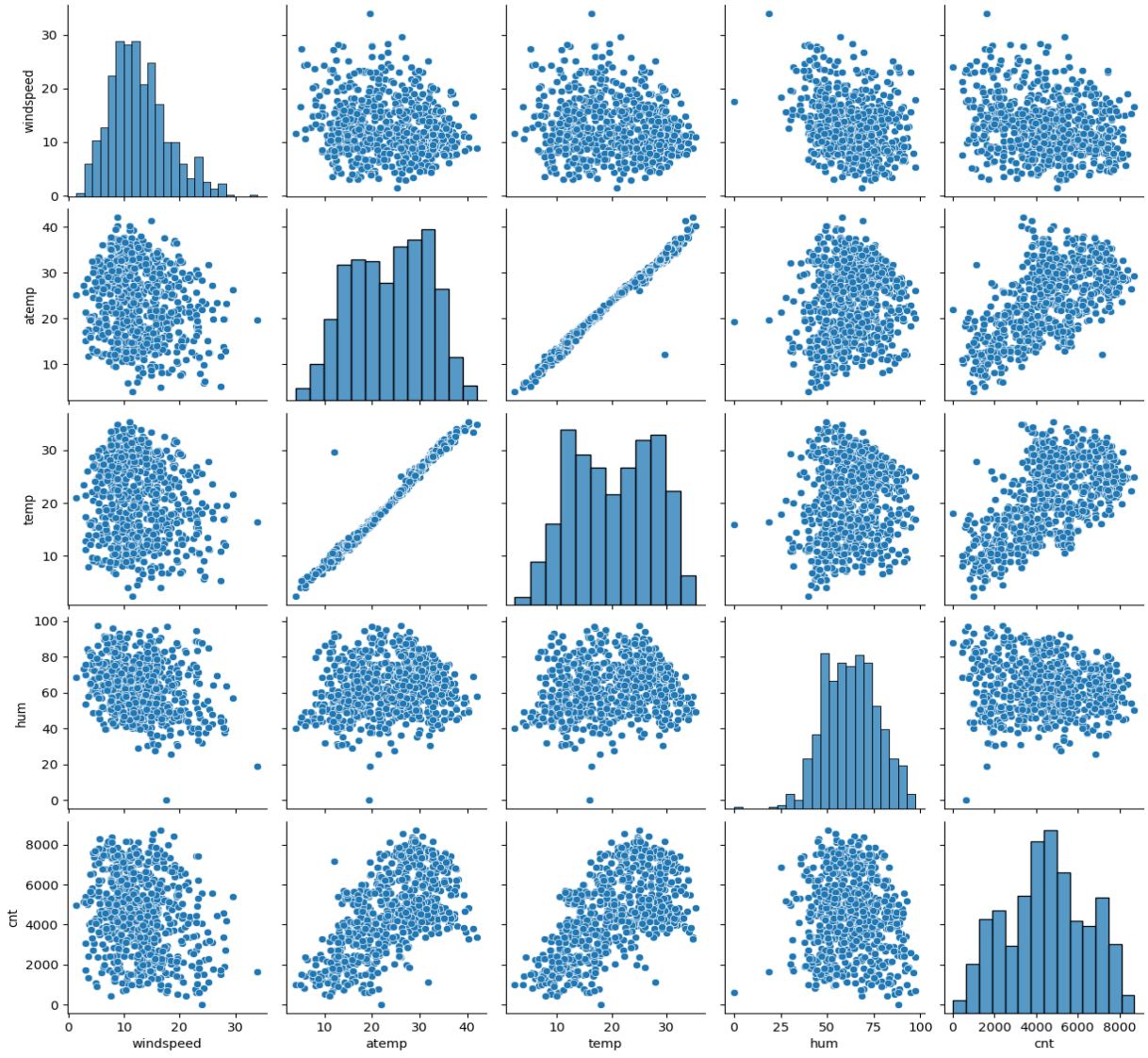
Relationship Status	**Single**	**Divorced**	**Married**
Single	1	0	0
Divorced	0	1	0
Married	0	0	1

By creating dummy variable will reduce the levels to 'n-1' = 2 levels. So, after dropping the first column, say, '*Single*' from columns, we can see the final table looks like:

Relationship Status	Divorced	Married
Single	0	0
In a Relationship	1	0
Married	0	1

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

Ans: Looking at the pair-plot, the '**temp**' and '**atemp**' variable has the highest correlation with the target variable '**cnt**'. *(As shown in below figure)*

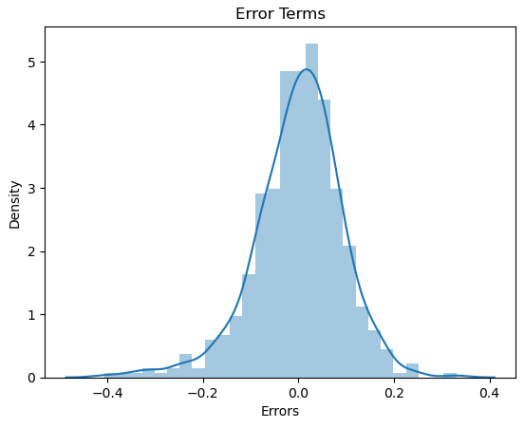


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

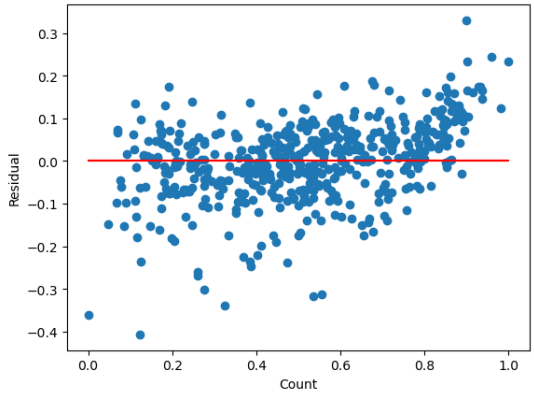
Ans :

We can validate the assumptions of Linear regression models by considering following factors:

- Normality of error terms
  - ➢ There should be Normal Distribution of error terms.



- Multicollinearity
  - ➢ There should be an insignificant multicollinearity among all variables.
- Independence of Residual
  - ➢ There should not be any autocorrelation.
- Homoscedasticity
  - ➢ There should not be any visible pattern in residual values.



- Linear Relationship
  - ➢ There should be linear relation among variables.

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

Ans :

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- 'yr'
- 'temp'
- 'winter'

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.827
Method:	Least Squares	F-statistic:	271.9
Date:	Sun, 29 Dec 2024	Prob (F-statistic):	2.83e-186
Time:	19:43:17	Log-Likelihood:	491.27
No. Observations:	510	AIC:	-962.5
Df Residuals:	500	BIC:	-920.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1259	0.017	7.508	0.000	0.093	0.159
yr	0.2329	0.008	27.962	0.000	0.216	0.249
holiday	-0.0987	0.026	-3.738	0.000	-0.151	-0.047
temp	0.5480	0.020	27.360	0.000	0.509	0.587
windspeed	-0.1532	0.025	-6.039	0.000	-0.203	-0.103
summer	0.0881	0.010	8.437	0.000	0.068	0.109
winter	0.1293	0.011	12.314	0.000	0.109	0.150
sep	0.1012	0.016	6.330	0.000	0.070	0.133
light_snowrain	-0.2829	0.025	-11.295	0.000	-0.332	-0.234
misty	-0.0784	0.009	-8.844	0.000	-0.096	-0.061

Omnibus:	57.077	Durbin-Watson:	2.097
Prob(Omnibus):	0.000	Jarque-Bera (JB):	114.844
Skew:	-0.648	Prob(JB):	1.15e-25
Kurtosis:	4.930	Cond. No.	10.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

General Subjective Questions

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

Ans :

Linear Regression : -

- ✓ Linear regression is a type of *supervised machine learning* algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.
- ✓ It is a data analysis technique which is used to predict continuous values.
- ✓ This technique attempts to plot a line graph between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. Independent variables are also called explanatory variables or predictor variables.

Types of Linear Regression :-

There are two types of linear regression :

- ➢ **Simple Linear Regression :**

  The linear regression involves only one independent variable and one dependent variable is known as *Simple Linear Regression*. It is the simplest form of linear regression.
  Example: The relationship between pollution levels and rising temperatures.

  SLR Equation :

  The equation derived for simple linear regression is :

  $y = \beta 0 + \beta 1X$

  where,

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

- **Multiple Linear Regression :**

The linear regression involves more than one independent variable and one dependent variable is known as *Multiple Linear Regression*.

Example: "*House Prices Prediction*"- The price of a house depends on factors like its size, location, number of rooms, type and age. Linear regression helps us analyze how one of these factors, such as the size, type, rooms and age of the house, influences the price.

MLR Equation :

The equation derived for simple linear regression is :

$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots\ldots \beta nXn$

where:

- Y is the dependent variable
- X1, X2, ..., Xn are the independent variables
- β0 is the intercept
- β1, β2, ..., βn are the slopes

Goal of Linear Regression :

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Assumptions of Linear Regression :

- Normality of error terms
- Multicollinearity
- Linearity
- Homoscedasticity
- Independence of Residual

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

Ans :

Anscombe's quartet :-

Anscombe's quartet is collection of a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Each dataset consists of eleven (*x*, *y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

Here are the data sets from Anscombe's Quartet – both as raw data, and plotted on a chart. The *x* values are the same for the first three datasets.

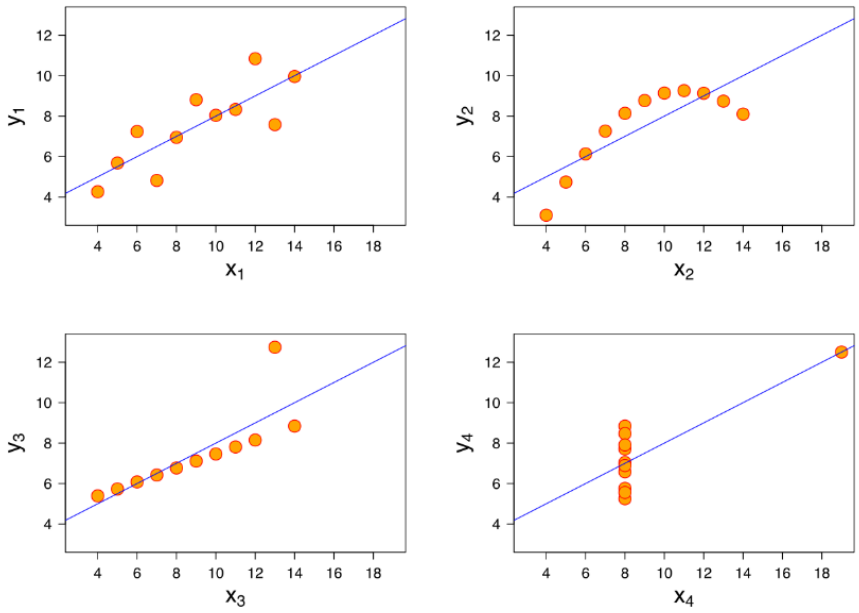
Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

You can tell just by looking that these data sets are very different. However, each data set is practically identical when you calculate the following summary metrics:

- Mean of $X = 9$
- Standard deviation of $X = 3.16$
- Mean of $Y = 7.5$
- Standard deviation of $Y = 1.94$
- Correlation between X & $Y = 0.816$
- The linear regression (the line of best fit) is also the same

The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different:



The four datasets can be described as:

- **Dataset 1:** this fits the linear regression model pretty 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.

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

Ans :

Pearson's R :-

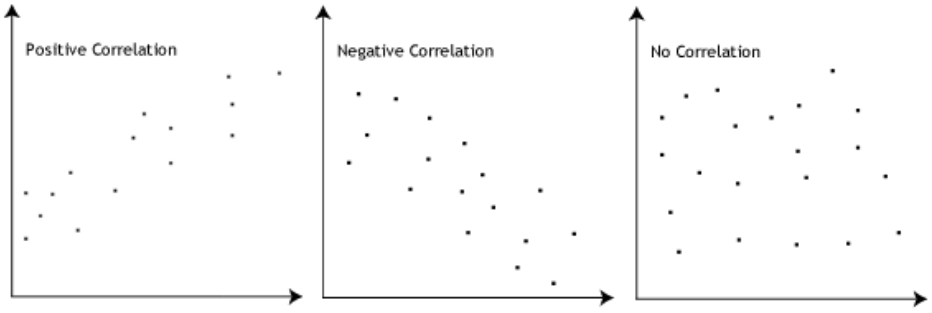
Pearson's correlation coefficient is a statistical measure that not only evaluates the strength but also direction of the relationship between two continuous variables.

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

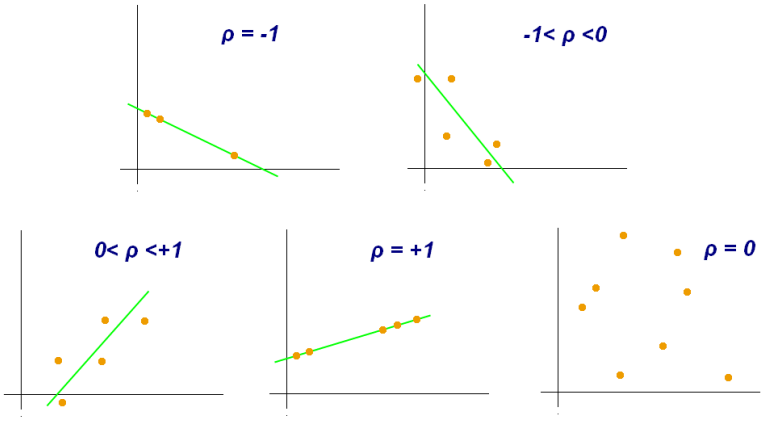
Pearson correlation coefficient values:-

- The Pearson correlation coefficient, *r*, can take a range of values from **+1** to **-1**.
- A value of 0 indicates that there is ***no association*** between the two variables.
- A value greater than 0 indicates a ***positive association***; that is, as the value of one variable increases, so does the value of the other variable.
- A value less than 0 indicates a ***negative association***; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



Below are examples of scatter diagrams with different values of correlation coefficient (ρ):



When *r* is 1 or –1, all the points fall exactly on the line of best fit.

When to use the Pearson correlation coefficient :-

Pearson's R correlation coefficient is used when:

- **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

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

Ans :

Scaling :-

Scaling, or feature scaling, refers to the process of transforming the values of variables to a specific range. This is often done to ensure that all variables have a comparable impact on the regression model.

Scaling can help prevent certain variables from dominating the model due to their larger magnitude. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

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.

Why Scaling used?

Scaling techniques aim to normalize the range, distribution, and magnitude of features, reducing potential biases and inconsistencies that may arise from variations in their values.

Types of Feature Scaling :

- **Normalization/Min-Max Scaling:**
  - *sklearn.preprocessing.MinMaxScaler* helps to implement normalization in python. It brings all of the data in the range of 0 and 1.
- **Standardization Scaling:**
  - *sklearn.preprocessing.scale* helps to implement standardization in python.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
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 really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	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.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

S.NO.	Normalization	Standardization
8.	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)

Ans :

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

formula for VIF is given as:

VIF = $1/(1-R^2)$

Now, when you're calculating the VIF for one independent variable using all the other independent variables, if the R^2 value comes out to be 1, the VIF will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables.

The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

Ans :

Q-Q (Quantile – Quantile) plot :-

A Q-Q (Quantile-Quantile) plot is used to see if a dataset follows a particular theoretical distribution. It works by comparing the quantiles of the observed data to the quantiles of this other distribution.

It is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. They are often used to describe the spread or distribution of a dataset.

The most common quantiles are:

1. **Median (50th percentile)**: The median is the middle value of a dataset when it is ordered from smallest to largest. It divides the dataset into two equal halves.
2. **Quartiles (25th, 50th, and 75th percentiles)**: Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.
3. **Percentiles**: Percentiles are similar to quartiles but divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

Uses of Q-Q plot :-

1. Assessing Distributional Assumptions:
   - Q-Q plots are frequently used to visually inspect whether a dataset follows a specific probability distribution, such as the normal distribution.
2. Detecting Outliers:
   - Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution.
3. Comparing Distributions:
   - Q-Q plots can be used to compare two datasets to see if they come from the same distribution.
4. Assessing Normality:
   - Q-Q plots are particularly useful for assessing the normality of a dataset.
5. Model Validation:
   - In fields like econometrics and machine learning, Q-Q plots are used to validate predictive models.

6. Quality Control:
    - Q-Q plots are employed in quality control processes to monitor the distribution of measured or observed values over time or across different batches.

Advantages of Q-Q plot :-

1. **Flexible Comparison**: Q-Q plots can compare datasets of different sizes without requiring equal sample sizes.
2. **Dimensionless Analysis**: They are dimensionless, making them suitable for comparing datasets with different units or scales.
3. **Visual Interpretation**: Provides a clear visual representation of data distribution compared to a theoretical distribution.
4. **Sensitive to Deviations**: Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
5. **Diagnostic Tool**: Helps in assessing distributional assumptions, identifying outliers, and understanding data patterns.
