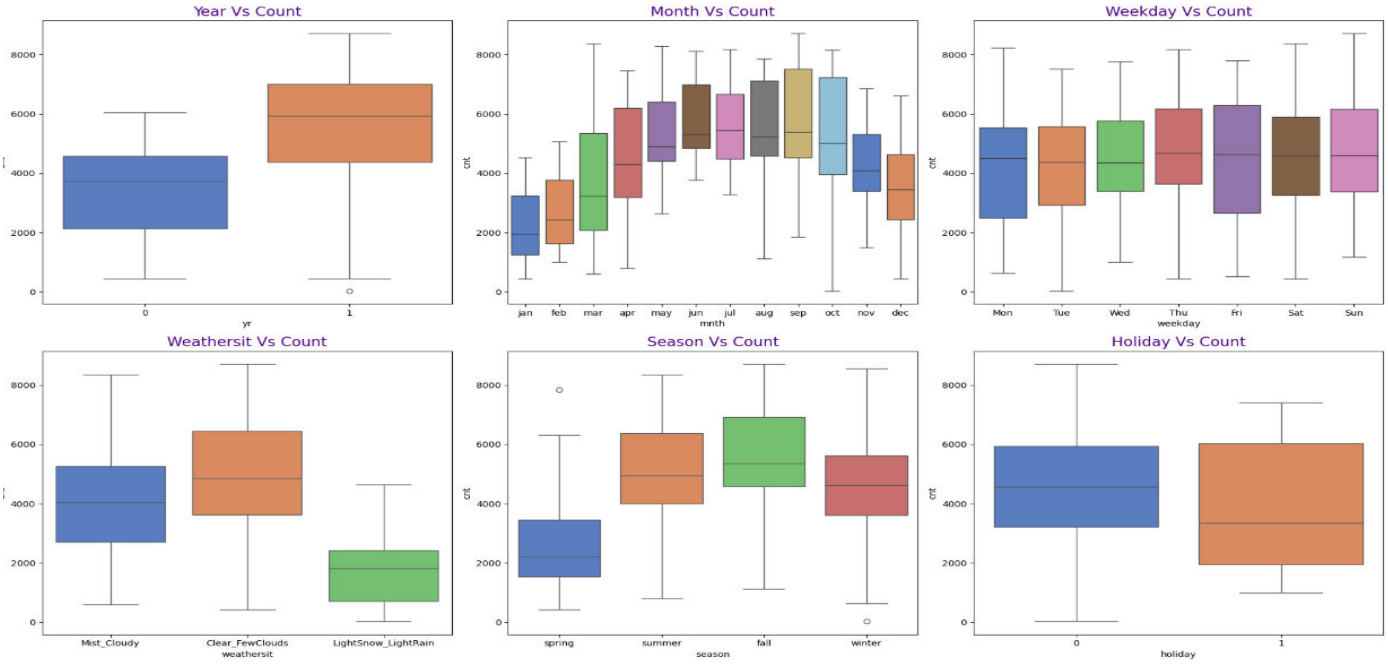
Assignment based Subjective Questions

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

Answer 1. My Inferences from categorical variables are as follows:



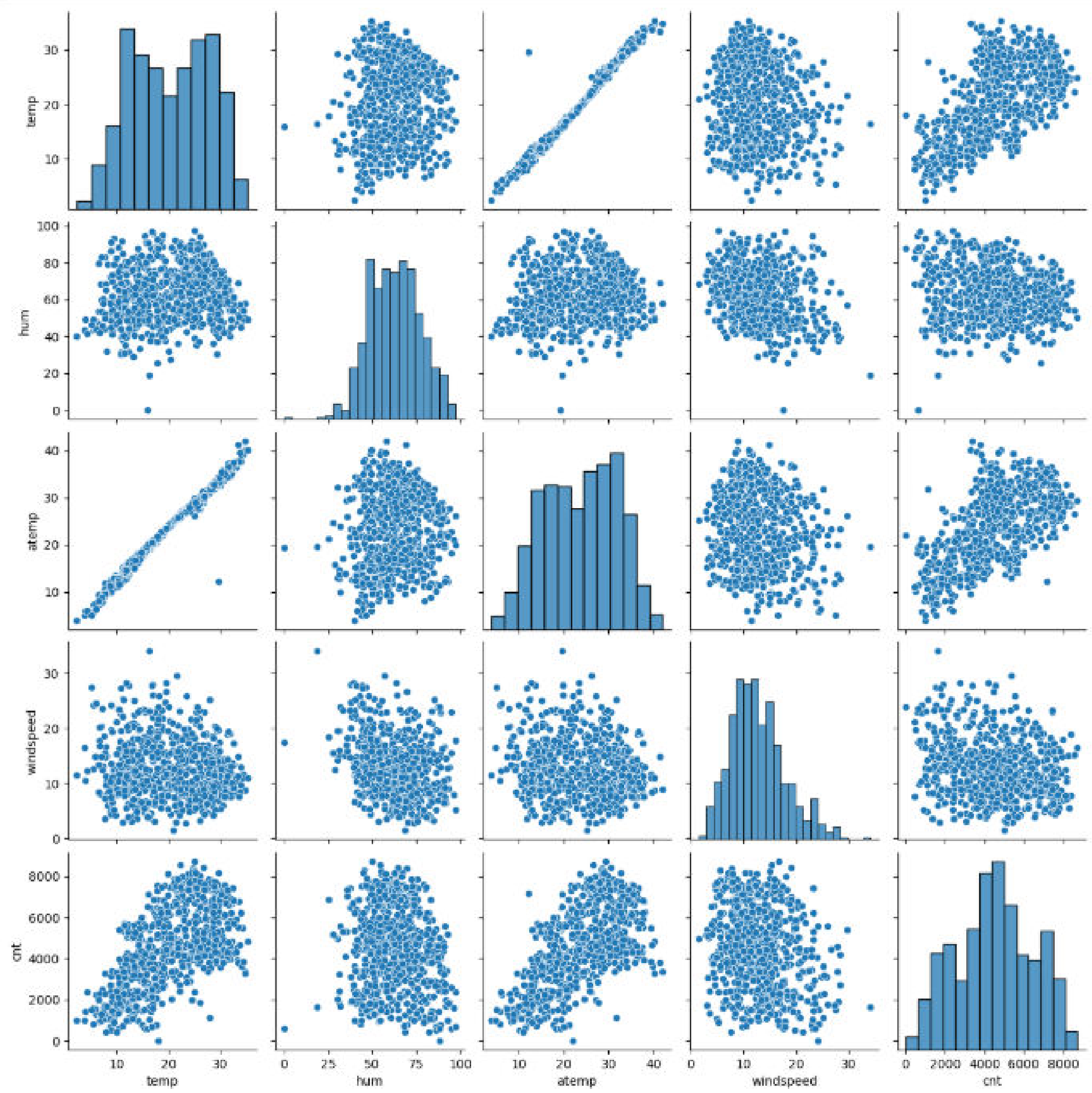
- **Increase in Customers**: The count of customers has increased significantly in 2019 compared to 2018.
- **Monthly Trends**: The count of bookings by customers is relatively higher in the mid-year months.
- **Weekly Distribution**: The count of bookings is spread similarly across all seven days of the week.
- **Weather Impact**:
  a)-If the weather is clear, the count of bookings is relatively higher.
  b)-The count decreases in misty weather.
  c)-The bookings fall significantly in rainy weather.
- **Seasonal Trends:** The count of bookings increased in summer and fall seasons and lowest in spring season.
- The count of bookings is higher on holidays as compare to working days.

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

Answer 2. We use it because it helps in preventing the multicollinearity, also it helps in streamlining the model by removing an extra column that is introduced during dummy variable creation. We also know that a variable with n levels can be represented by n-1 dummy variable. So, for creating a robust model we should use drop_first=True when introducing the dummy variable.

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

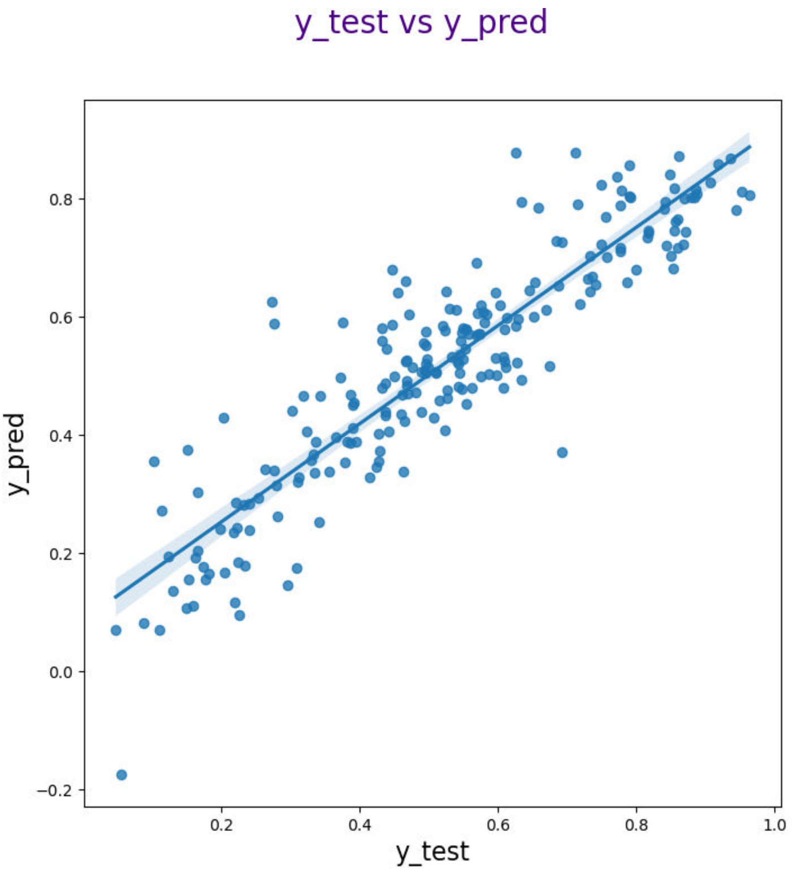
Answer 3.



“temp” and “atemp” variable has the highest correlation with the target “cnt” variable. And also temp and atemp variables are highly correlated with each other so we would drop the atemp variable to avoid multicollinearity. So that the “temp” variable is the one which has the highest correlation with the target variable.

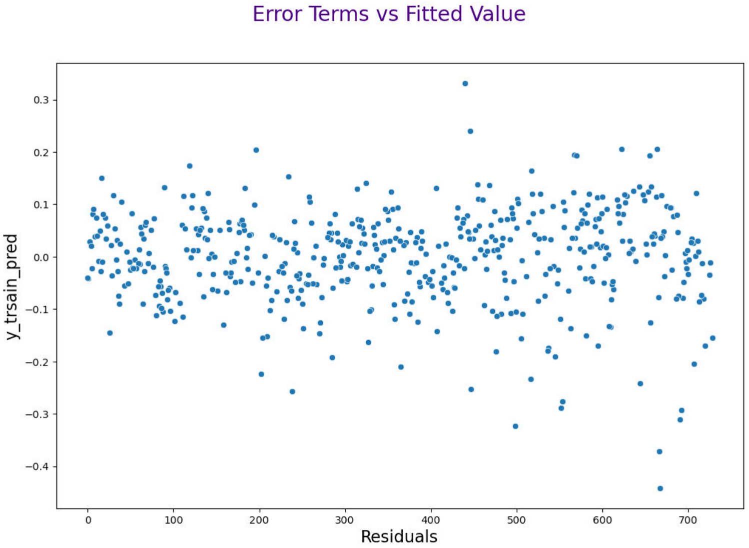
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer 4. 1-**Linearity:** By visualizing the linear relationship between test and predicted variables, showing a clear linear trend indicates a good fit.

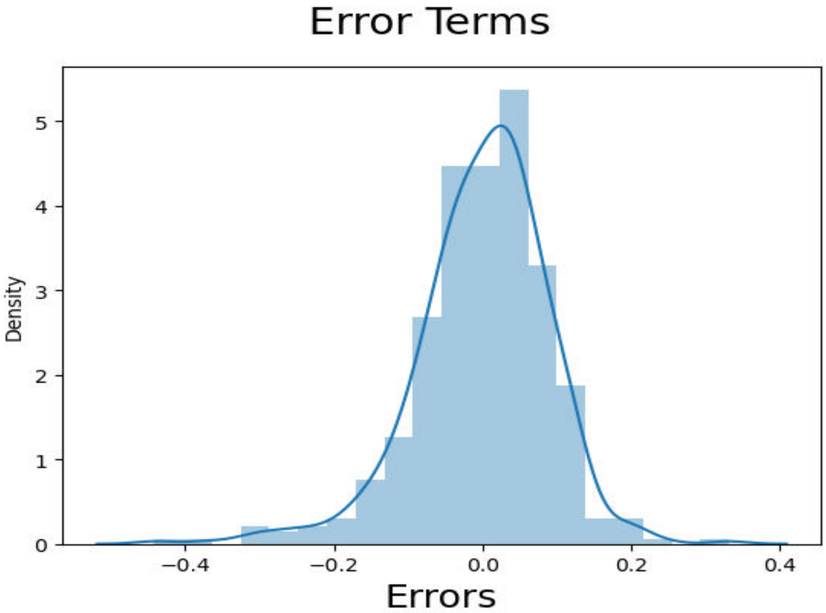


2-Homoscedasticity: Example: Plot of residuals vs. predicted values shows residuals spread evenly around zero.

Method: Create a scatter plot of residuals vs. predicted values; check for a random spread without funnel-like shapes.



3-Normality: By verify that the error terms follow a normal distribution, by plotting a histogram of the error terms and examine its shape.



4-Multicollinearity: Example: VIF values for all independent variables are below 10, indicating low multicollinearity.

Method: Calculate the Variance Inflation Factor (VIF) for each independent variable; VIF values above 10 indicate high multicollinearity.

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

Answer 5. Top 3 features contributing significantly the demand of shared bikes are: **1-temp (Temperature)**

2- weathersit_LightSnow_LightRain

3- yr (Year)

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer 1. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps us to find out the relationship between Input and the target variable. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as we can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically,

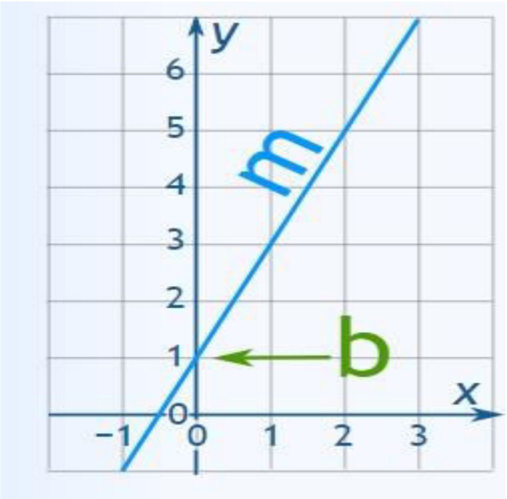
$y = mx + b$ (Equation of a straight line)

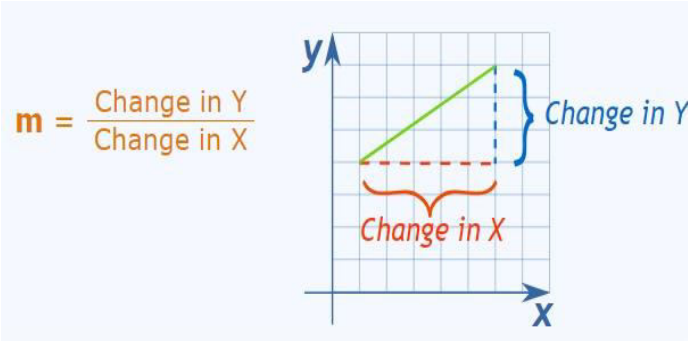
y = how far up

x = how far along

m = Slope or Gradient (how steep the line is)

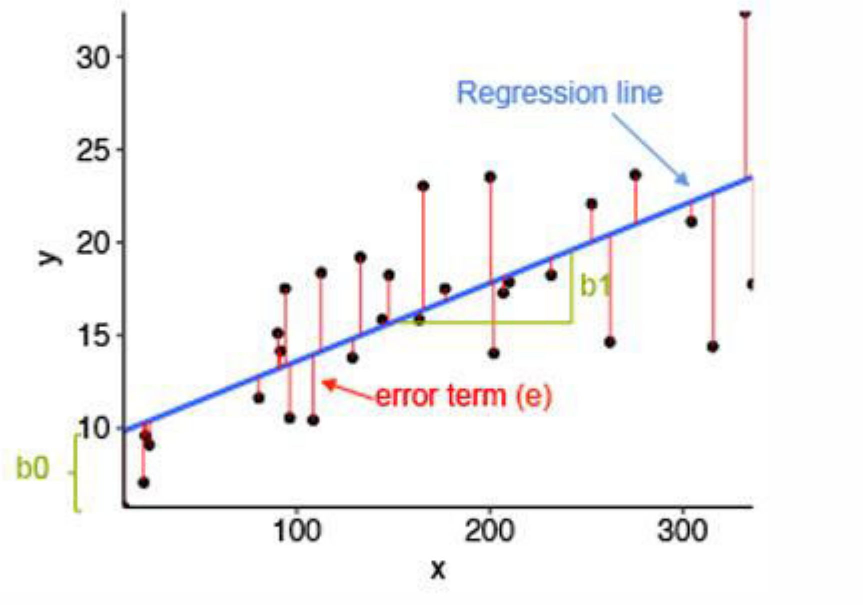
b = value of **y** when **x=0**





Broadly speaking, linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).

In mathematics, a basic linear regression model can be represented by the equation y = b0 + b1x. Here, y is the dependent variable that we aim to predict, x is the independent variable, b1 denotes the slope of the line, and b0 represents the intercept (a constant). The cost function helps us determine the most optimal values for the slope (b1) and intercept (b0), thereby providing the best-fitting line for the given data points.



Here, x and y are two variables on the regression line.

b1 = Slope of the line.

b0 = y-intercept of the line.

x = Independent variable from dataset

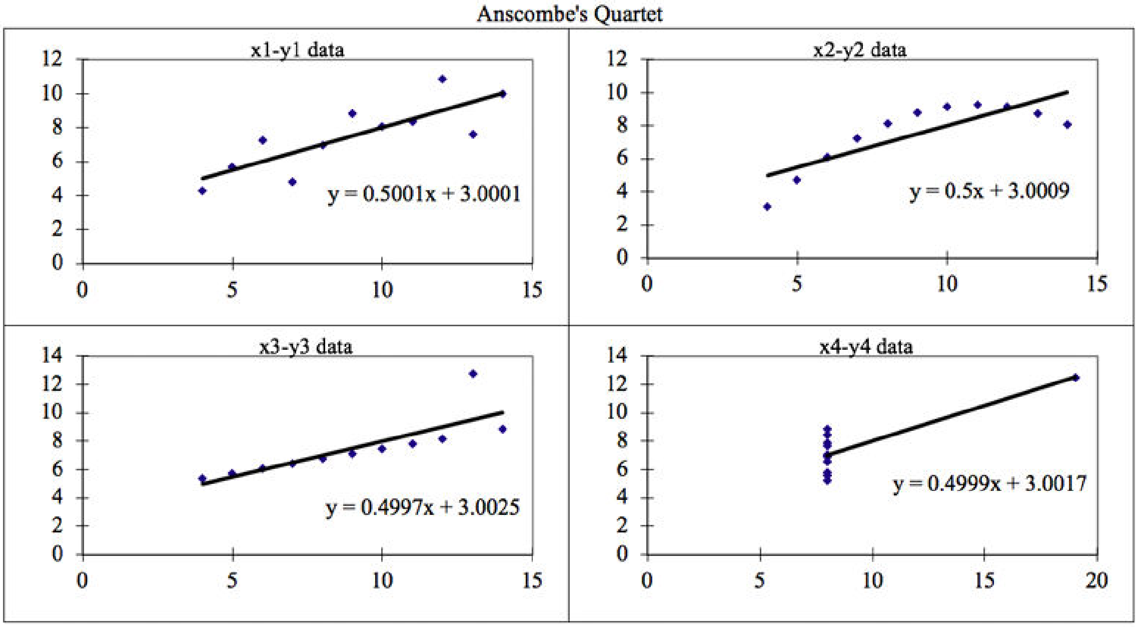
y = Dependent variable from dataset

There are two types of linear regression model:

- **Simple linear regression:** This is used when the number of independent variables is 1.
- **Multiple linear regression:** This is used when the number of independent variables is more than 1.

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

Answer 2. Anscombe's Quartet is a collection of four data sets that are almost identical when viewed through simple descriptive statistics (like mean, variance, etc.), but exhibit some peculiar characteristics that can mislead regression models. These data sets have different distributions and display unique patterns when visualized using scatter plots. Each of the four data sets comprises eleven (x, y) points. The quartet illustrates the importance of graphing data before analysing it, as relying solely on statistics can overlook these underlying discrepancies.



Let's describe Anscombe's Quartet datasets with a bit of detail:

1. **Dataset 1:** This dataset fits the linear regression model quite well, following a relatively straightforward linear relationship between the variables x1 and y1.
2. **Dataset 2:** This dataset doesn't fit a linear regression model effectively because the relationship between x2 and y2 is non-linear, meaning that a straight line wouldn't accurately represent the data.
3. **Dataset 3:** This dataset features an outlier, a point that deviates significantly from the overall pattern. This outlier can distort the results of a linear regression model, making it less reliable.
4. **Dataset 4:** Similar to Dataset 3, this dataset also contains an outlier that disrupts the fit of a linear regression model, highlighting the model's limitations in dealing with anomalous data points.

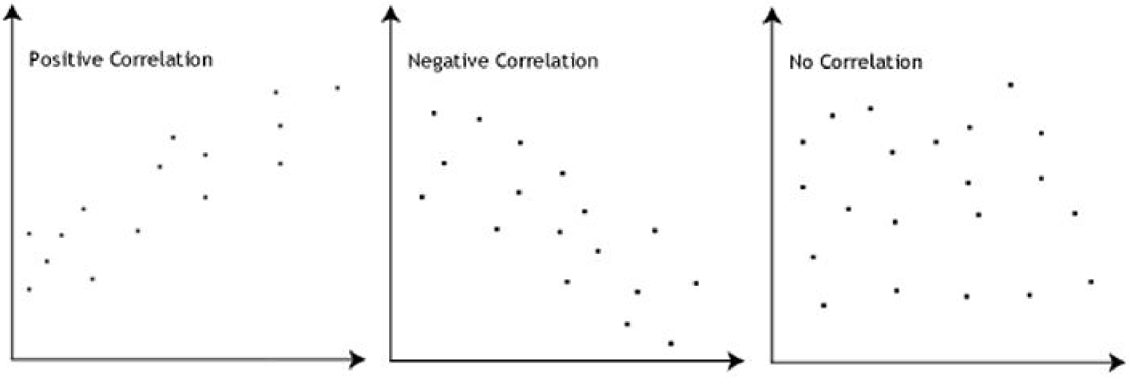
Question 3. What is Pearson's R?

Answer 3. Pearson's r is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. When the variables move in the same direction—both increasing or decreasing—the correlation coefficient (r) will be positive. It indicates how closely the data points follow a linear trend.

Pearson's r ranges between -1 and 1:

- A value of 1 means a perfect positive linear relationship.
- A value of -1 means a perfect negative linear relationship.
- A value of 0 means no linear relationship.

If the data points fall exactly on a straight line with a negative slope, then the correlation coefficient (r) will be -1.



Positive correlation means that both variables tend to increase or decrease together. In other words, as one variable goes up, the other also goes up, and as one goes down, the other also goes down.

On the other hand, a negative correlation indicates that as one variable increases, the other variable decreases, and vice versa. This means they move in opposite directions.

Both types of correlation help us understand the relationship between variables and can be important for making predictions or identifying trends in data.

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

Answer 4. Scaling is a technique used to normalize the range of independent variables in regression analysis. The goal is to bring all independent variables to the same scale. Without scaling, regression algorithms might incorrectly interpret greater numerical values as more significant, leading to inaccurate predictions.

It's essential to note that scaling affects only the coefficients and no other parameters like the t-statistic, F-statistic, p-values, R-squared, etc.

Example: Imagine the height of one person is 180 centimetres, and the height of another person is 1.8 meters. Without scaling, a machine learning algorithm might incorrectly interpret 180 centimetres as a much larger value compared to 1.8 meters, which is not the case. This could lead to inaccurate predictions.

Machine learning algorithms work with numbers, not units, so scaling is a crucial pre-processing step before performing regression on a dataset.

Scaling can be performed in two ways:

1. **Normalization:** It scales a variable to a range between 0 and 1.
2. **Standardization:** It transforms data to have a mean of 0 and a standard deviation of 1.

To summarize, normalization scales data to a specific range, while standardization adjusts data to have a mean of 0 and standard deviation of 1. The choice between them depends on the specific requirements of your analysis or machine learning model.

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

Answer 5. Perfect Multicollinearity: When there is perfect collinearity, it means that the variables are perfectly correlated. For example, if variable X1 is an exact multiple of variable X2, the regression model cannot distinguish between the two variables, and it creates an issue in estimating their coefficients.

Infinite VIF: Since VIF is a measure of how much the variance of a coefficient is inflated due to multicollinearity, in cases of perfect multicollinearity, the calculation of VIF results in division by zero, leading to an infinite value.

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

Answer 6. A Quantile-Quantile (Q-Q) plot is a graphical tool that helps assess whether a set of data plausibly originates from a theoretical distribution (such as Normal, Exponential, or Uniform). It can also determine if two data sets come from populations with a common distribution. Essentially, a Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. This method can be used to compare various distributions, such as Gaussian, Uniform, Exponential distributions, etc.

Advantages:

a) It can be used with small sample sizes.

b) It can detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

Uses:

Q-Q plots can help check if two data sets:

i. Come from populations with a common distribution.

ii. Have a common location and scale.

iii. Have similar distributional shapes.

iv. Exhibit similar tail behaviour.

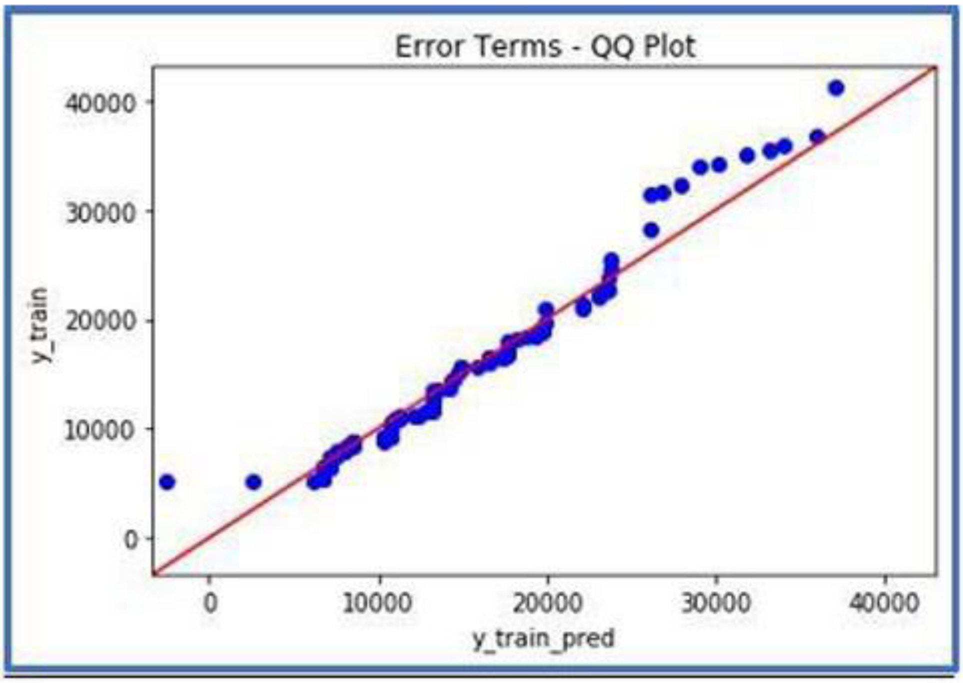
Interpretation:

A Q-Q plot is a plot of the quantiles of one data set against the quantiles of the other data set. Here are the possible interpretations for two data sets:

a) Similar distribution: If the quantiles lie on or close to a straight line at an angle of 45 degrees from the x-axis, it indicates that the two data sets have a similar distribution.

b) Y-values < X-values: If the y-quantiles are lower than the x-quantiles, it suggests a specific pattern or difference between the data sets.

c) X-values < Y-values: If the x-quantiles are lower than the y-quantiles, it indicates another specific pattern or difference between the data sets.



Overall, Q-Q plots are a valuable tool for assessing the distributional properties and relationships between data sets.