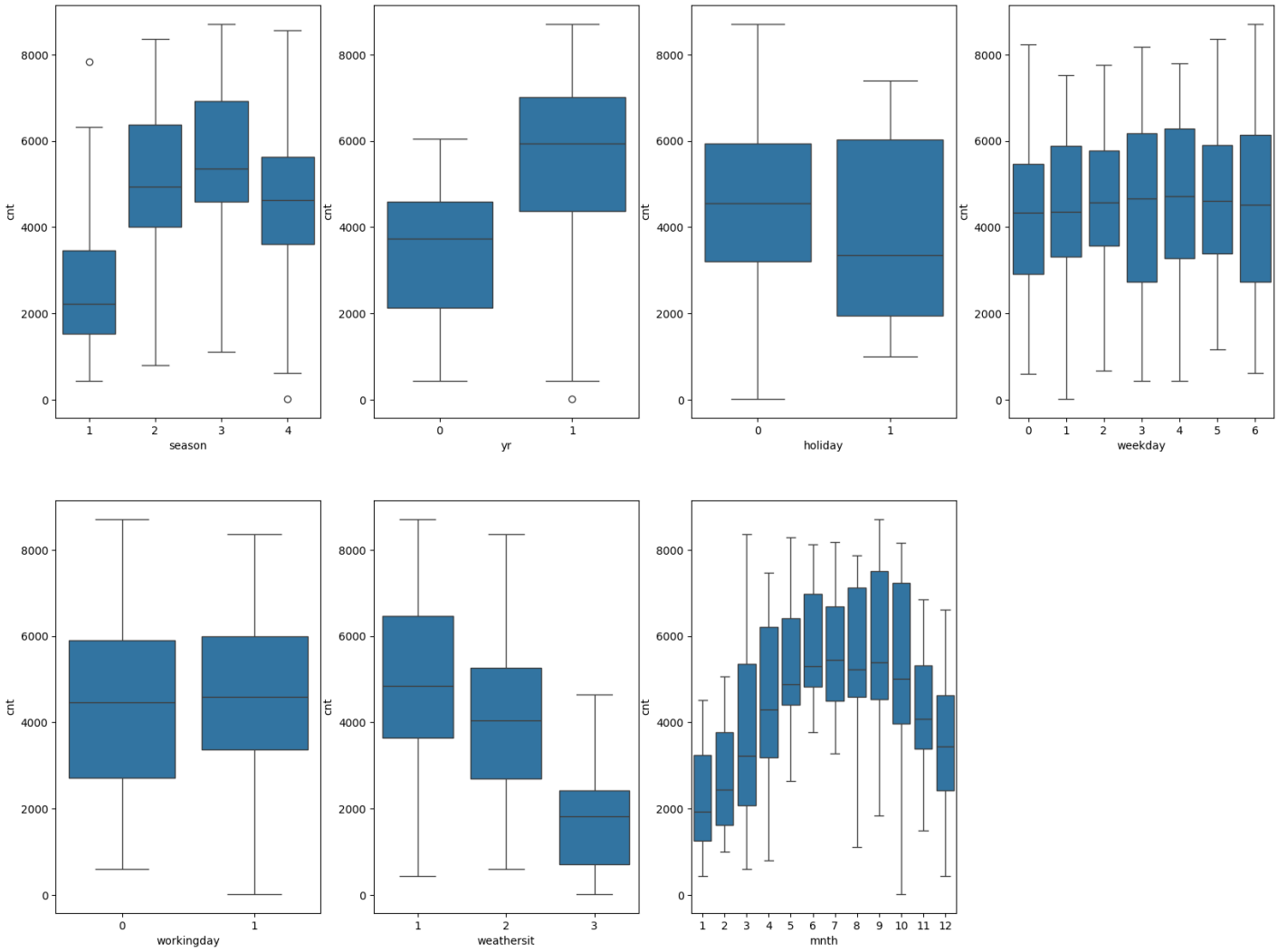
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer:



1. Weather-related variables (season, weather sit) have strong effects on usage
2. Temporal patterns (monthly variation) are significant
3. The system shows year-over-year growth
4. Work/holiday status has less impact than might be expected
5. Usage is relatively stable across different days of the week

These patterns suggest us that this might be data from a bike-sharing or similar transportation system where

Weather and seasonal conditions are major factors. Also, during model building on inclusion of categorical features such as year (yr), season etc we saw a significant change in the value of

R-squared and adjusted R-squared. This implies that the Categorical features were helpful in explaining a greater proportion of variances in the data sets.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables (to avoid a problem called **multicollinearity**)
if we have categorical variable with n-levels, then we need to use n-1 columns to represent the **Dummy variables.**

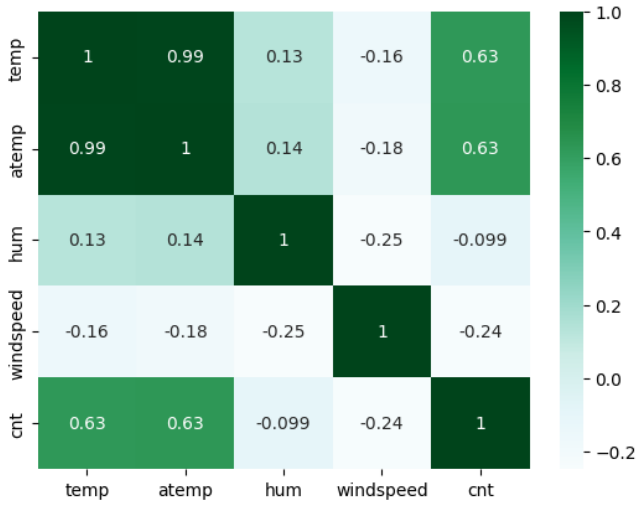
Example: Suppose we have a feature "city" with values "London," "Paris," and "Tokyo."

- Without **drop_first=True**, you'd have three dummy variables: "city_London," "city_Paris," and "city_Tokyo."
- With drop_first=True, you'd have two: "city_Paris" and "city_Tokyo." "London" becomes the reference.
    - If "city_Paris" is 1 and "city_Tokyo" is 0, it means the city is Paris.
    - If "city_Paris" is 0 and "city_Tokyo" is 1, it means the city is Tokyo.
    - If both "city_Paris" and "city_Tokyo" are 0, it means the city is London (the reference).

In essence, **drop_first=True** ensures that your model remains interpretable and avoids the statistical pitfalls of **multicollinearity** when dealing with categorical data.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

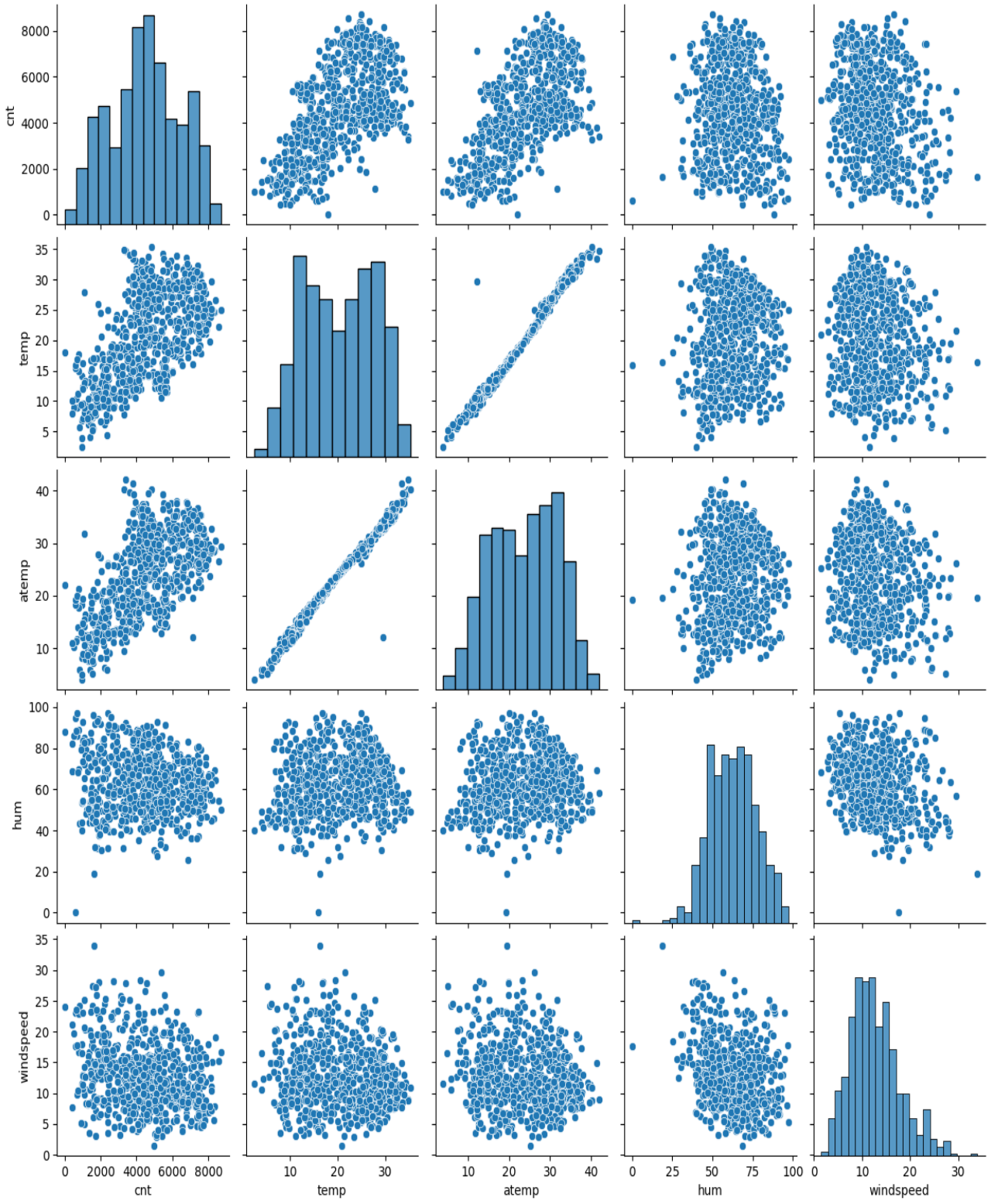
The numerical variable ‘registered (0.99)’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multi collinearity the numerical variable ‘atemp (0.63)’ has the highest correlation with the target variable ‘cnt’.

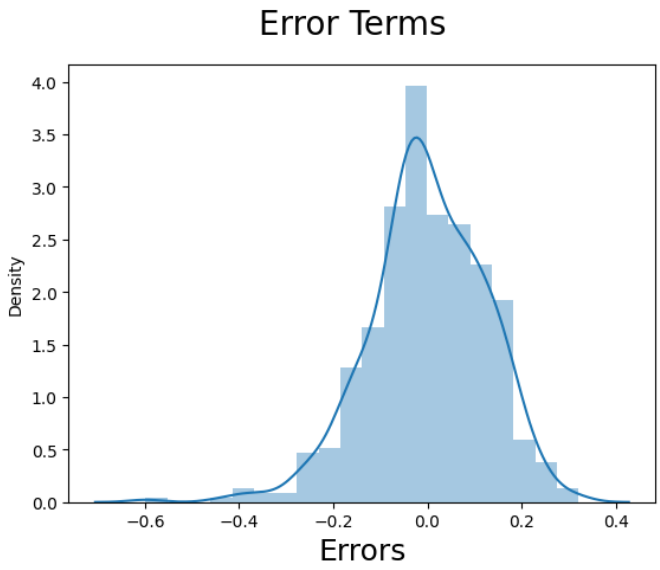


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I have done following tests to validate assumptions of Linear Regression:

- There should be linear relationship between independent and dependent variables. We visualized the numeric variables using a pairplot to see if the variables are linearly related or not.
- Residuals distribution should follow normal distribution and centred around 0 (mean = 0).We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.
- linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.
- For more information


cnt
temp
atemp
hum
windspeed
0
2000
4000
6000
8000
5
10
15
20
25
30
35
40
0
20
40
60
80
100
0
5
10
15
20
25
30
35
0
2000
4000
6000
8000
10
20
30
10
20
30
40
0
25
50
75
100
0
10
20
30



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 significant features are:
1. temp - coefficient : 0.480127
2. yr - coefficient : 0.233612
3. weathersit_Light Snow & Rain - coefficient : -0.291342

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable (the outcome we want to predict) and one or more independent variables (the predictors). It assumes a linear relationship, meaning the change in the dependent variable is proportional to the change in the independent variable(s).

Types:

- **Simple Linear Regression:** One independent variable.
- **Multiple Linear Regressions:** Two or more independent variables.

The Equation:

The core is a linear equation:

- **Simple:** $y = B0 + B1x + E$
  - y: Dependent variable.
  - x: Independent variable.
  - B0: Y-intercept (value of y when x is 0).
  - B1: Slope (change in y for a one-unit change in x).
  - E: Error term (difference between observed and predicted values).
- **Multiple:** $y = B_0 + B1x1 + B2x2 + ... + Bnxn + E$
  - y: Dependent variable.
  - x1, x2, ..., xn: Independent variables.
  - B0: Y-intercept.
  - B1, B2,... Bn: Coefficients for each independent variable.
  - E: Error term.

Goal:

Find the "best-fit" line (simple) or hyper plane (multiple) that minimizes the difference between predicted and actual values of the dependent variable. This is often done by minimizing the sum of squared errors (residual sum of squares).

Finding the Best Fit:

1. **Ordinary Least Squares (OLS):** Finds coefficients (B0, B1, etc.) that minimize the sum of squared errors. Has a direct formula for calculation.
2. **Gradient Descent:** An iterative optimization algorithm used for large datasets or when OLS is computationally expensive. Adjusts coefficients step-by-step to reduce error.

Assumptions:

- **Linearity:** Linear relationship between variables.
- **Independence:** Errors are independent.
- **Homoscedasticity:** Errors have constant variance.
- **Normality:** Errors are normally distributed.
- **No/Little Multicollinearity (Multiple Regression):** Independent variables are not highly correlated.

Evaluation:

- **R-squared:** Proportion of variance in the dependent variable explained by the model.
- **Mean Squared Error (MSE):** Average squared difference between predicted and actual values.

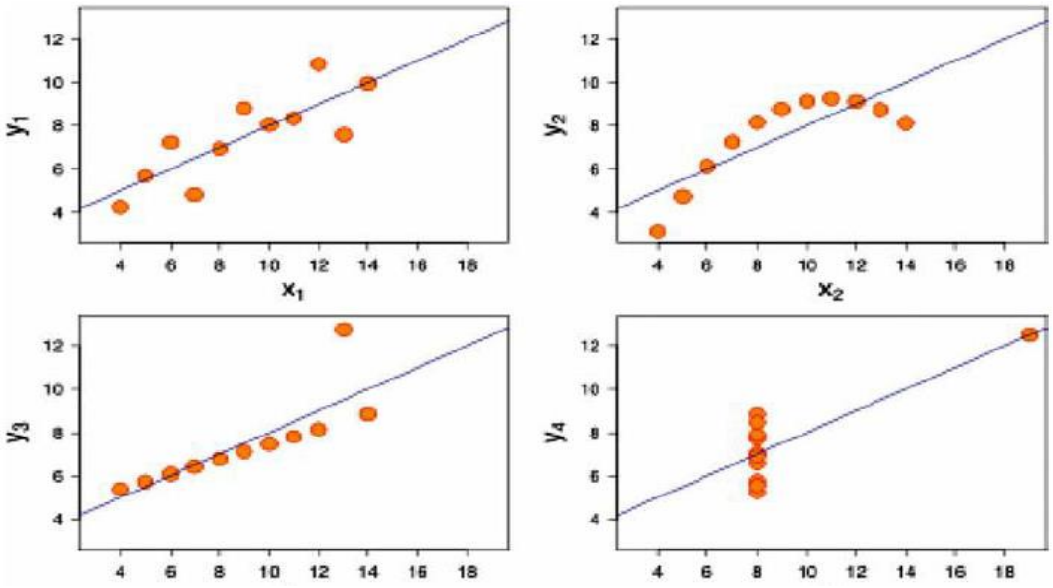
- **Root Mean Squared Error (RMSE):** Square root of MSE, [1] in the same units as the dependent variable.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it
and the effect of outliers and other influential observations on statistical properties.

1. The first scatter plot (top left) appears to be a simple linear relationship.
2. The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
3. In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) 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.



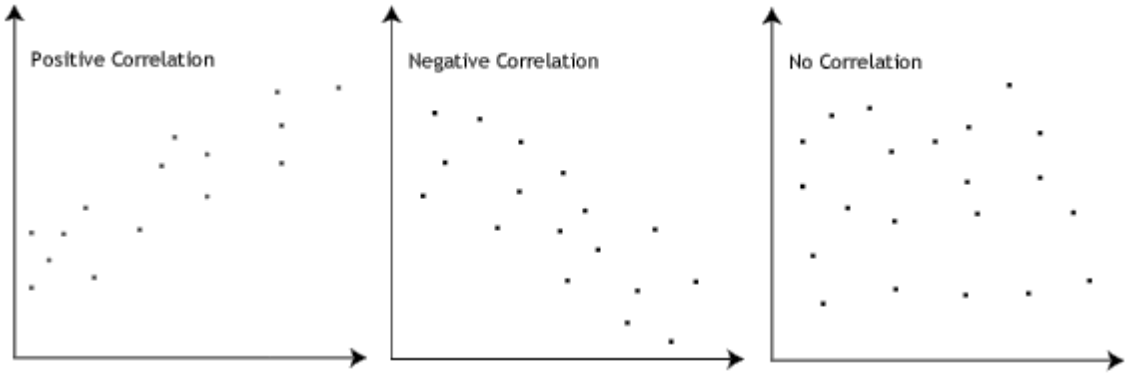
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

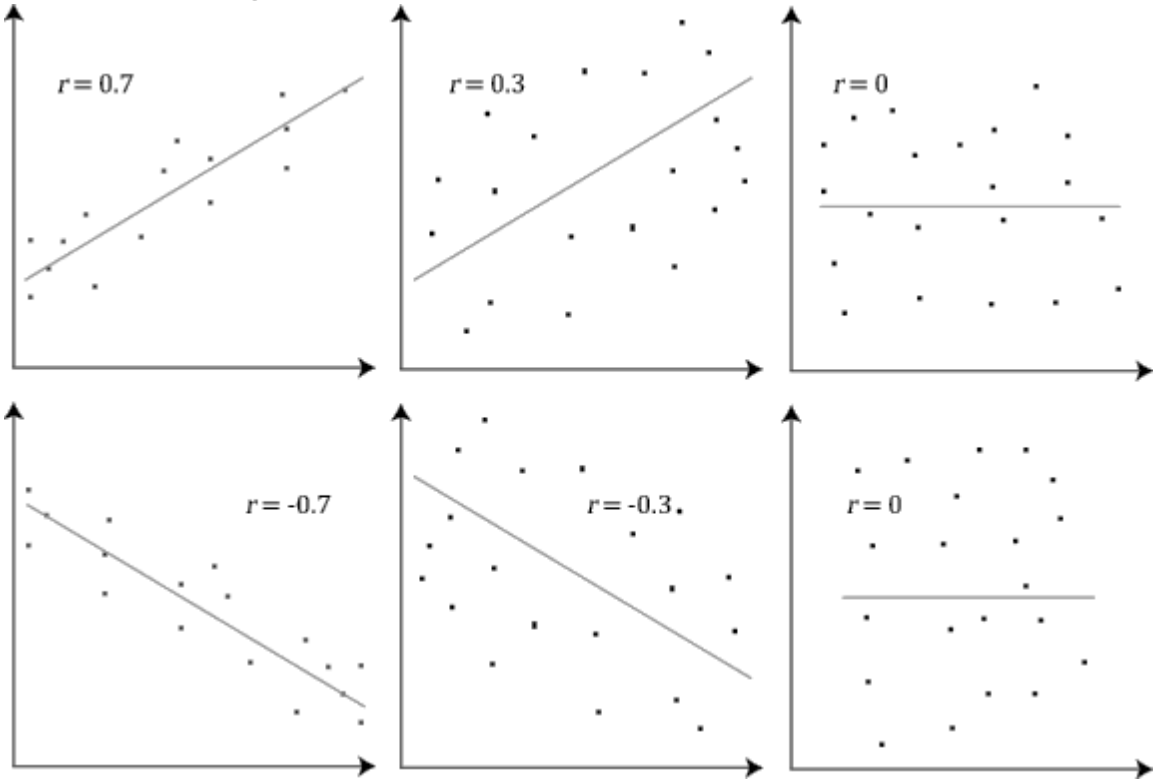
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

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:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with 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, irrespective of the units of the values.

- 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.
- 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.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.
Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no

multicollinearity — if there was no correlation with other predictors.
A rule of thumb for interpreting the variance inflation factor:
• 1 = not correlated.
• Between 1 and 5 = moderately correlated.
• Greater than 5 = highly correlated.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

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.

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.

Few advantages:

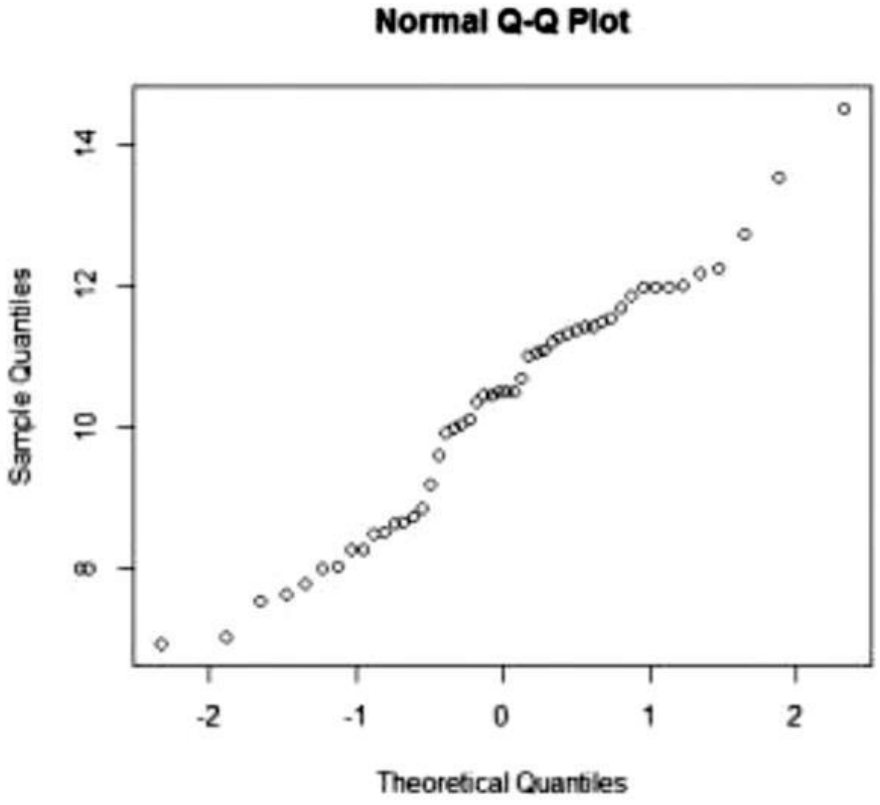
a) It can be used with sample sizes also

b) Many distributional 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 following scenarios: If two data sets.

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

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 i.e.


Normal Q-Q Plot
Sample Quantiles
8
10
12
14
-2
-1
0
1
2
Theoretical Quantiles