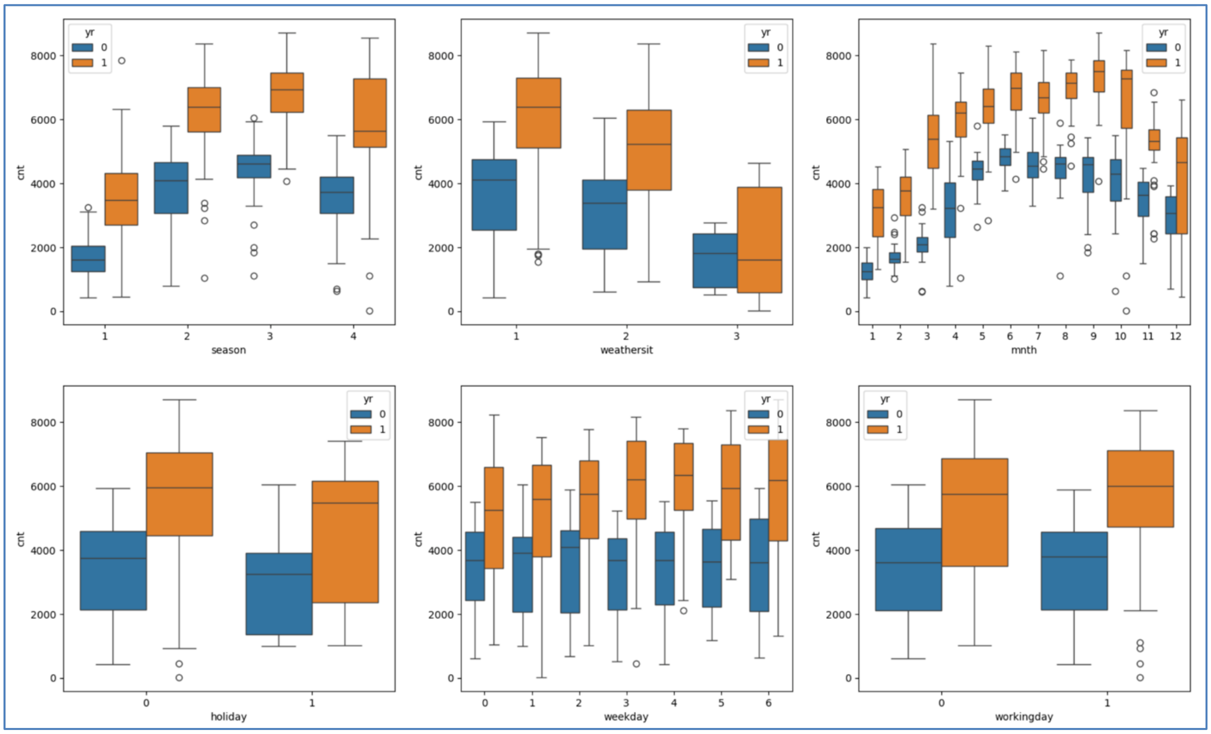
Assignment-based Subjective Questions

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: <Your answer for Question 1 goes below this line> (Do not edit)

After reading & analyzing Bike sharing dataset, I arrive at the following 7 categorical variables: “season”, “yr”, “mnth”, “holiday”, “weekday”, “workingday”, “weathersit”. These were visualized using a boxplot as shown below.



These categorical variables had the following effect on the dependent variables:

- ✓ season (1:spring, 2:summer, 3:fall, 4:winter) –
  - ○ Spring season has the least demand for shared bikes, where as fall season has the highest demand, followed by winter and summer.
  - ○ In all the seasons the demand has increased significantly from year 2018 to year 2019.
- ✓ weathersit (1: Clear, 2: Mist, 3: Light Snow, 4: Heavy Rain) –
  - ○ There are no users in Heavy rain/snow, clearly indicating the weather is extremely unfavorable.
  - ○ Highest count of Bike sharing seen when the weather is pleasant (clear & partly cloudy).
  - ○ Again Y-o-Y, The demand has gone up significantly from 2018 to 2019 in case of Clear weather conditions, and considerably in Misty & cloudy conditions.
- ✓ mnth –
  - ○ September month saw the highest demand followed by August, October and June.
  - ○ For all the months demand has increased significantly from year 2018 to year 2019.
- ✓ holiday –
  - ○ Rentals are relatively lower on holidays then on working days, and the demand increased from 2018 to 2019 in the similar pattern.

- ✓ weekday –
    - o count of rentals is comparable on all days, and the demand increased from 2018 to 2019 in the similar pattern.
- ✓ workingday –
    - o median count is almost same whether it's a workingday or not, and the demand increased from 2018 to 2019 in the similar pattern.

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)

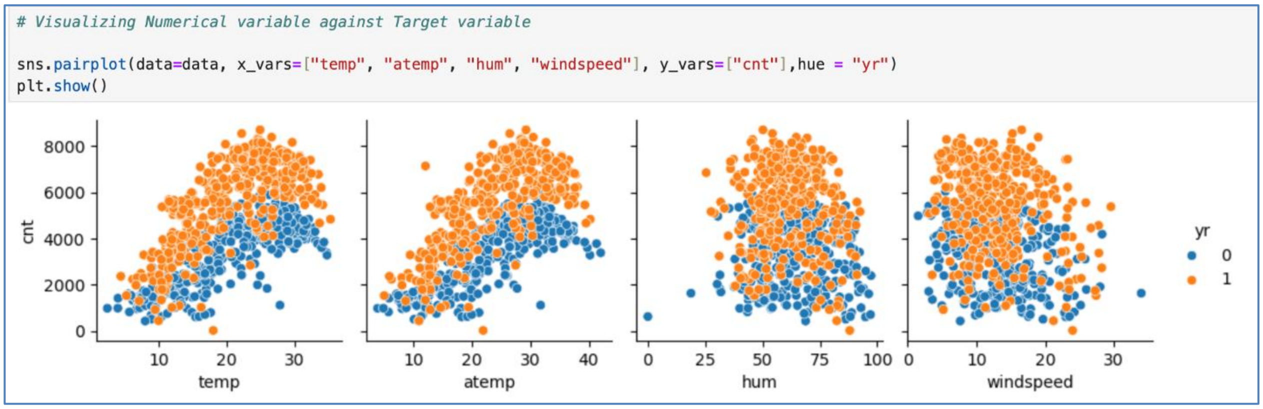
Syntax: dummies = pd.get_dummies(df['Category'], drop_first=True)

Using drop_first=True during dummy variable creation is important for several reasons, particularly when working with regression models.

Avoiding Multicollinearity: When you create dummy variables for a categorical variable with (k) categories, you end up with (k) dummy variables. However, one of these dummy variables can be perfectly predicted by the others, leading to multicollinearity.
By setting drop_first=True, you drop the first dummy variable, which helps avoid this issue and ensures that the model remains stable and interpretable.

Model Efficiency: Reducing the number of dummy variables by one can make the model more efficient, especially when dealing with large datasets with many categorical variables. This can lead to faster computation and less memory usage.

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)



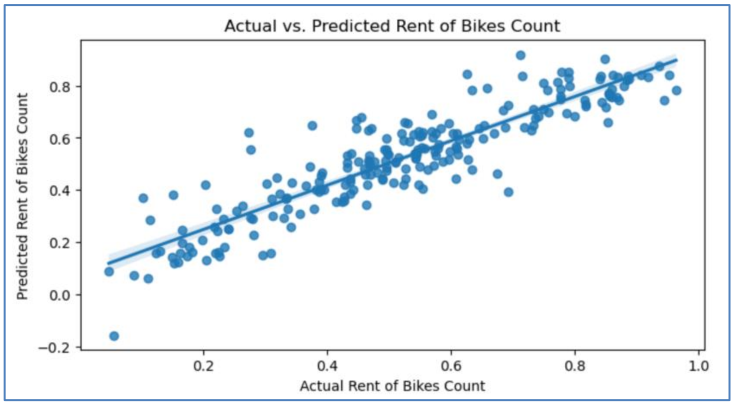
```
# Visualizing Numerical variable against Target variable

sns.pairplot(data=data, x_vars=["temp", "atemp", "hum", "windspeed"], y_vars=["cnt"],hue = "yr")
plt.show()
```

As per the above pair plot, "temp" & "atemp" are the two numeric variables, having the highest correlation with the target variable "cnt".
As "atemp" is highly correlated with "temp" approx. 99%, to remove redundancy, we further removed it from the regression model numeric variables.

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)

All the 4 Assumptions of the Linear regression were validated, by plotting relevant charts as below:

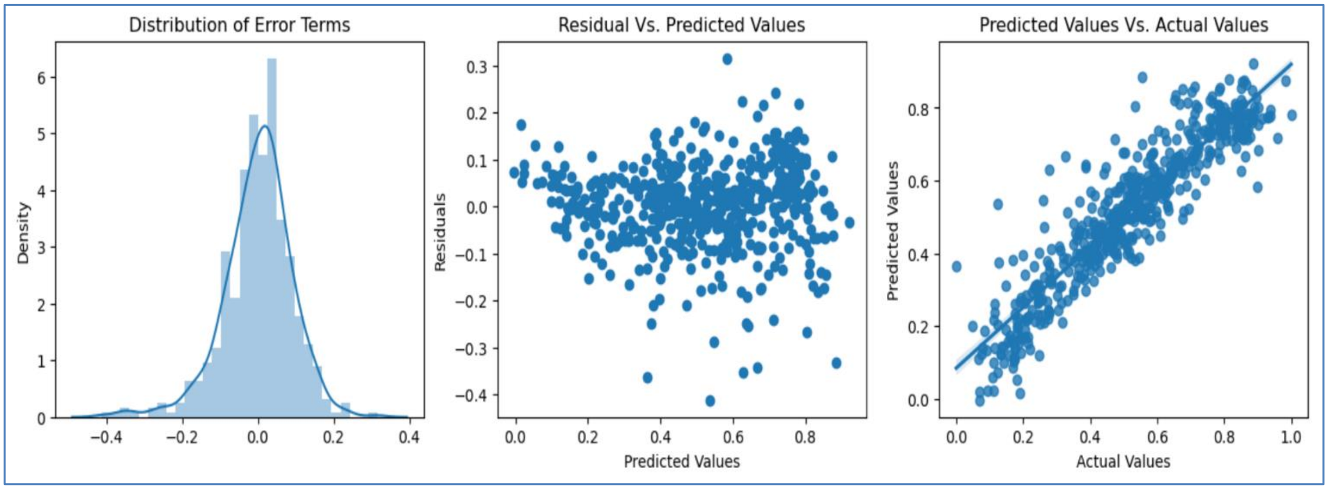
1. **Linear relationship between independent and dependent variables** – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the count of actual vs predicted Rent for bikes plot as shown in the below figure.



2. **Error terms are normally distributed**: Histogram and distribution plot helps to understand the normal distribution of error terms with the mean of 0.

3. **Error terms are independent of each other** – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction.

4. **Error terms have constant variance (homoscedasticity)**: We can see Error Terms have approx. a Constant Variance, hence it follows the Assumption of Homoscedasticity



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)

Top 3 contributors are:

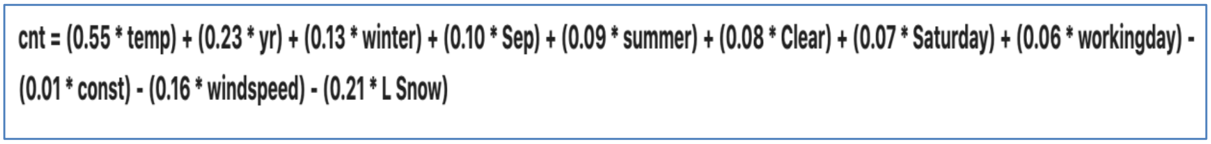
- **temp** (***coefficient:*** ***0.55***) → most significant feature affecting the business positively.

- **Yr** (***coefficient:*** ***0.23***) → Year over Year (Y-o-Y) growth seems to be promising too.

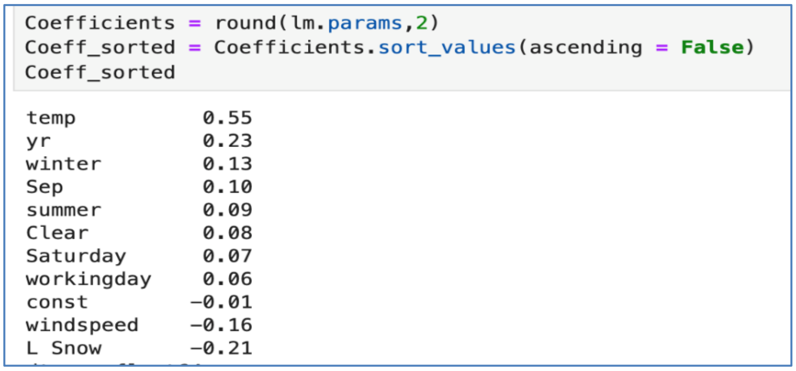
- **weathersit- L Snow** (***coefficient:*** ***-0.21***) → This feature negatively impacts the business.

Note: L snow refers to " Light snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" in Weathersit

Below is the final Equation of MLR Model & the Coefficients of different significant features contributing to.



cnt = (0.55 * temp) + (0.23 * yr) + (0.13 * winter) + (0.10 * Sep) + (0.09 * summer) + (0.08 * Clear) + (0.07 * Saturday) + (0.06 * workingday) - (0.01 * const) - (0.16 * windspeed) - (0.21 * L Snow)



```
Coefficients = round(lm.params,2)
Coeff_sorted = Coefficients.sort_values(ascending = False)
Coeff_sorted
```

```
temp          0.55
yr            0.23
winter        0.13
Sep           0.10
summer        0.09
Clear         0.08
Saturday      0.07
workingday    0.06
const        -0.01
windspeed    -0.16
L Snow       -0.21
```

General Subjective Questions

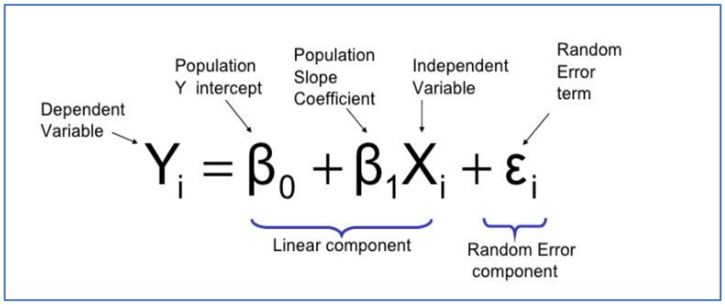
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a fundamental technique in machine learning and statistics used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting line that describes this relationship.

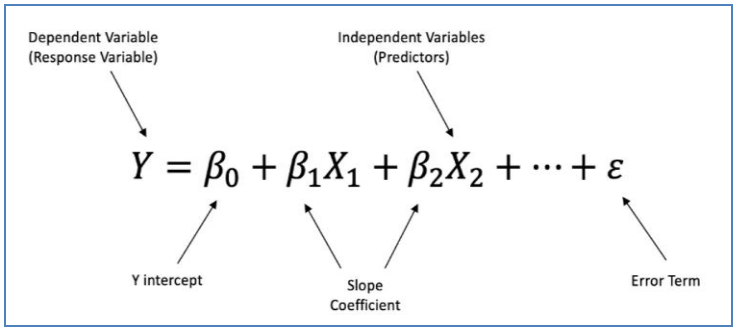
In Linear regression, the dependent variable is the function of the independent variable, respective coefficients and the error term.

Linear Regression is broadly classified into two:

1. Simple Linear Regression (SLR) – This is used when the dependent variable is predicted using only one independent variable. The equation for SLR is -



2. Multiple Linear Regression (MLR): This is used when the dependent variable is predicted using multiple independent variables. The equation for MLR is –


Dependent Variable (Response Variable)
Independent Variables (Predictors)
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$
Y intercept
Slope Coefficient
Error Term

Assumptions:

1. Linearity: The relationship between the dependent and independent variables is linear.
2. Independence: Error Terms are independent of each other.
3. Homoscedasticity: Error terms / residuals have constant variance.
4. Normality: Error Terms or Residuals are normally distributed with mean=0.

Steps in Linear Regression

Data Preparation:

Collect Data: Gather data with the dependent and independent variables.

Preprocess Data: Handle missing values, encode categorical variables, and normalize/standardize data if needed.

Model Training:

Split Data: Divide data into training and testing sets.

Fit the Model: Use the training data to estimate the coefficients β by minimizing the sum of squared residuals. The coefficients β are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared differences between observed and predicted values.

Model Evaluation:

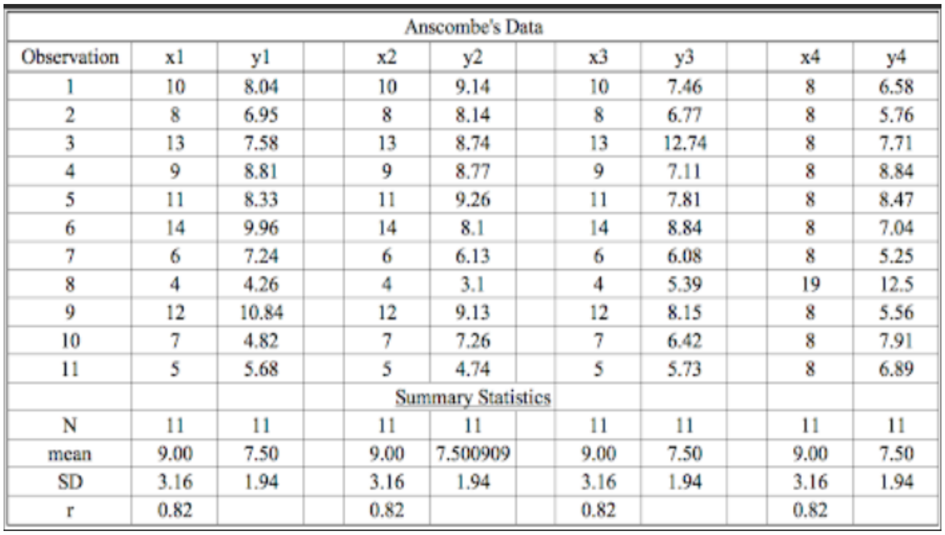
Predict: Use the model to make predictions on the test data.

Evaluate: Assess performance using metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

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)

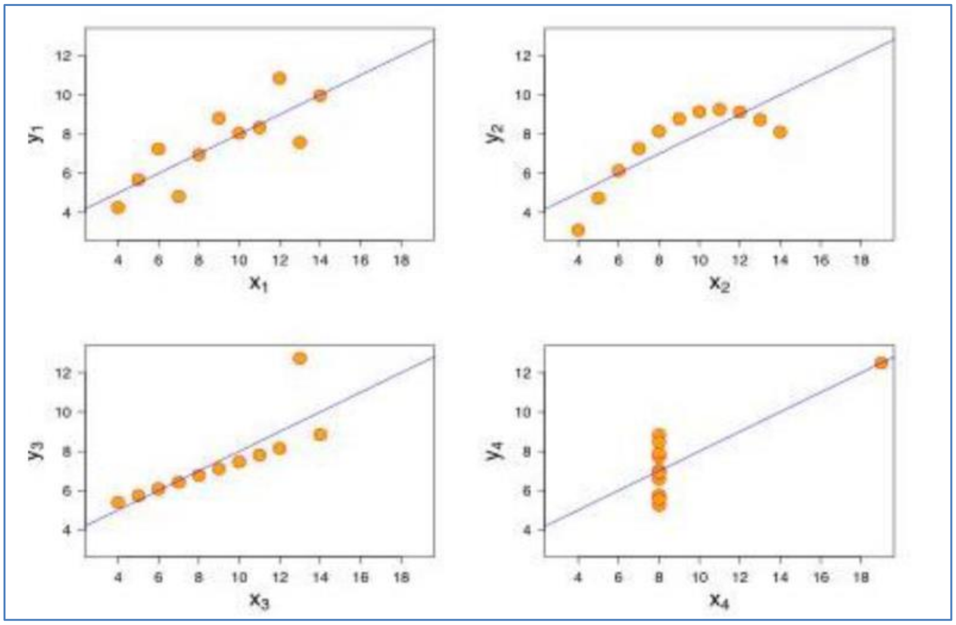
Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets are designed to have nearly identical simple descriptive statistics, such as mean, variance, and correlation, yet they appear very different when graphed. The purpose of Anscombe's quartet is to demonstrate the importance of visualizing data before analyzing it and to highlight how statistical properties alone can be misleading.

The statistical information for these four data sets are approximately similar:



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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm.



Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear

regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Conclusion: Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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)

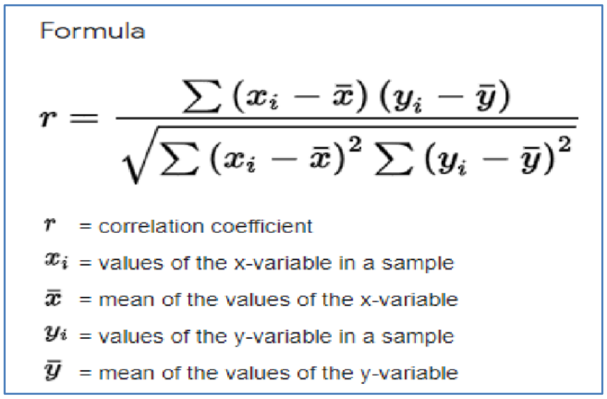
Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two quantitative variables. It quantifies the strength and direction of this relationship, providing a value between -1 and 1.

Value ranges:

- +1: Perfect positive linear relationship.
- -1: Perfect negative linear relationship.
- 0: No linear relationship.

Interpretation:

- Positive Values: Indicate that as one variable increases, the other variable also increases.
- Negative Values: Indicate that as one variable increases, the other variable decreases.
- Magnitude: The closer the value is to ±1, the stronger the linear relationship.
- The formula for r is:



Formula

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Applications:

1. Data Analysis: To determine the strength and direction of the relationship between two variables.
2. Hypothesis Testing: To test if there is a significant linear relationship between variables.

Visualization: A scatter plot can visually represent the relationship between two variables, with Pearson's R indicating how closely the data points fit a straight line.

Limitations of Pearson's R:

1. Linearity Assumption: It only measures the strength of a linear relationship. It does not accurately reflect the strength of non-linear relationships.
2. Sensitivity to Outliers: It is highly sensitive to outliers. A single outlier can significantly affect the correlation coefficient, potentially leading to misleading conclusions.
3. No Causation: It does not imply causation. It only indicates the strength and direction of a linear relationship, but it cannot determine which variable is the cause and which is the effect.
4. Homogeneity of Variance: It assumes that the variance of the variables is constant across the range of data. If this assumption is violated, the correlation coefficient may not be reliable.
5. Scale Sensitivity: It can be affected by the scale of measurement. Variables measured on different scales can lead to incorrect interpretations if not properly standardize.
6. Bivariate Normality: It assumes that the data for both variables are normally distributed. If this assumption is not met, the correlation coefficient may not accurately represent the relationship.

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)

Scaling is a data preprocessing technique used to adjust the range of independent variables or features of data. It ensures that all features contribute equally to the model, especially when they are measured on different scales.

Scaling is performed to:

1. Improves Model Performance: Many machine learning algorithms, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), are sensitive to the scale of the data. Scaling helps these algorithms perform better by ensuring that no single feature dominates due to its scale.

2. Speeds Up Convergence: For optimization algorithms like gradient descent, scaling can speed up convergence by making the cost function more symmetric.

3. Prevents Numerical Instability: Scaling helps prevent numerical instability in algorithms that involve matrix operations, such as linear regression and principal component analysis (PCA).

Difference between Normalized Scaling & Standardized Scaling:

S. No.	Normalized Scaling	Standardized Scaling
1	Normalization (or Min-Max Scaling) transforms the data to fit within a specific range, typically [0, 1].	Standardization (or Z-score Normalization) transforms the data to have a mean of 0 and a standard deviation of 1
2	MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$	Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
3	Useful when you want to ensure that all features are on the same scale, especially for algorithms that rely on distance calculations, like KNN and SVM.	Useful when the data follows a Gaussian distribution and you want to standardize the features to have the same scale but retain the original distribution shape.

4	Scikit Transformer used is – Minmax scaler
	Scaler = MinMaxScaler()
	X_train_norm = scaler.fit_transform(X_train)
	mod = sm.OLS(y_train, X_train_norm)
	res2 = mod.fit()
		Scikit Transformer used is – Standard Scaler
		Scaler = StandardScaler()
		df_standard = scaler.fit_transform(df)
		df_standard

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)

VIF (Variance Inflation Factor) – gives how much the variance of the coefficient estimate is being inflated by collinearity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is extremely important parameter to test Linear Regression model.

$$\boxed{VIF = \frac{1}{1 - R^2}}$$

Now, R^2 = 1, means there is a perfect correlation between 2 independent variables or the variance can be 100 % explained by the independent variables. This mathematically makes the denominator = 0, which results in VIF = Infinity.

Thumb Rule to interpret VIF:

- 1 = Not Correlated
- Between 1 & 5 = Moderately correlated
- Greater than 5 = Highly correlated

Now why VIF becomes infinite, could be more than 1 reason:

1. Perfect Collinearity (R2 = 1).
2. Dummy Variable Trap: This can occur when dummy variables are not properly handled. Example – If you include all categories of a categorical variable without dropping ones.
3. Redundant Variable: Including variables that are linear combination of each other, example weights in kgs and weights in pounds.

Implications:

1. Model Instability: Infinite VIF values indicate that the regression coefficients cannot be uniquely estimated, leading to instability in the model.
2. Interpretation Issues: High multicollinearity makes it difficult to determine the individual effect of each independent variable on the dependent variable.

Addressing Infinite VIF:

1. Remove Redundant Variables: Identify and remove or combine variables that are perfectly collinear.
2. Drop One Dummy Variable: When using dummy variables, always drop one category to avoid the dummy variable trap.

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)

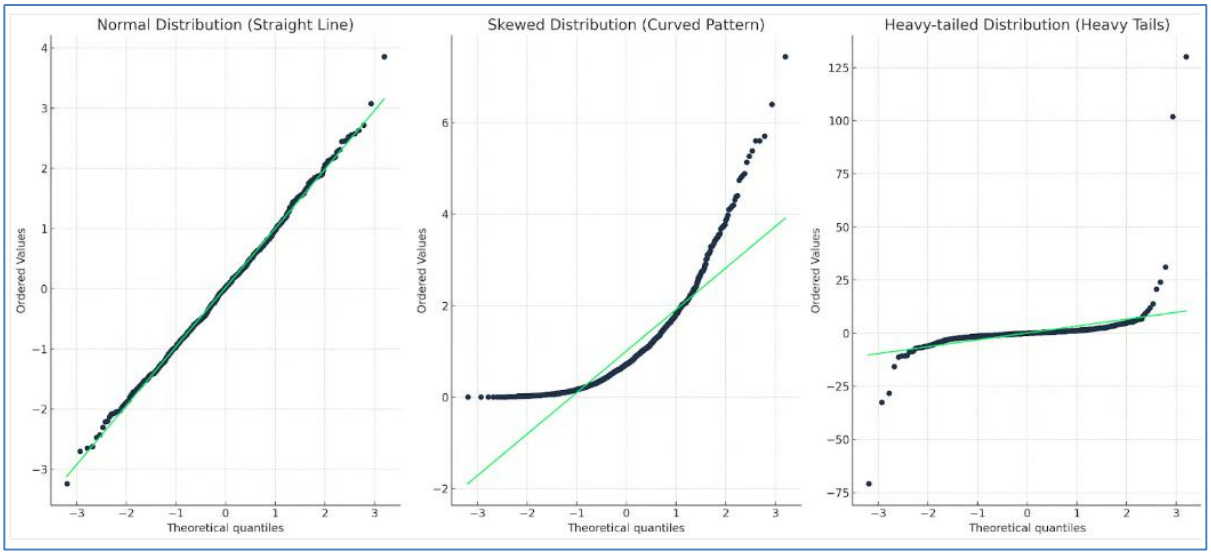
Answer: Please write your answer below this line. (Do not edit)

Quantile-Quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

QQ 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.

Q-Q plots are useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution.

How to Interpret a Q-Q Plot



1. Straight Line: If the points lie on or near the straight line, it indicates that the data follows the theoretical distribution.
2. Deviations from the Line: Deviations from the line suggest departures from the theoretical distribution. For example:
   - Upward Curve: Indicates a right-skewed distribution.
   - Downward Curve: Indicates a left-skewed distribution.
3. S-Shaped Curve: Indicates heavy tails (leptokurtic distribution)

Importance in Linear Regression

1. Assessing Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot of the residuals helps to visually check this assumption.
2. Model Diagnostics: By examining the Q-Q plot, you can identify if there are any significant deviations from normality, which might indicate issues with the model, such as the presence of outliers or the need for a different model.
3. Improving Model Fit: If the Q-Q plot shows deviations from normality, you might consider transforming the data or using a different regression technique to improve the model fit.

Conclusion

Q-Q plots are a valuable tool in linear regression for checking the normality of residuals and diagnosing potential issues with the model. They provide a visual method to assess whether the data conforms to a theoretical distribution, helping to ensure the validity of the regression model.