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)

From analyzing the categorical variables in the dataset, several clear patterns emerge in how they affect bike rentals:

Yearly Pattern:
There's significant growth in rentals from 2018 to 2019, with counts increasing from around 3,400 to 5,600, showing growing popularity in bike-sharing.

Seasonal Impact:
- Summer leads with highest rentals
- Fall follows with good usage
- Spring shows moderate activity
- Winter has the lowest rentals (2,000-3,000)

Weather Conditions:
- Clear days show the highest rentals (6,000-8,000)
- Misty conditions have moderate usage
- Light snow/rain sees lowest rentals (around 2,000)

Time-Based Patterns:
- Holidays show higher rentals (around 6,000) compared to regular days
- Weekends (Saturday/Sunday) see peak activity
- Mid-week days have lower counts
- Monthly trend shows a steady increase from January, peaking in July-August (6,000+)

These patterns suggest that bike rentals are heavily influenced by:
1. Weather conditions (people prefer clear days)
2. Seasonal temperatures (warm weather encourages more rides)
3. Day type (holidays and weekends are preferred)
4. Time of year (summer months are most popular)

Understanding these relationships helps predict demand patterns and could be valuable for:
- Planning maintenance schedules
- Adjusting bike availability
- Developing targeted marketing strategies
- Optimizing operations based on weather forecasts

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)

Using `drop_first=True` when creating dummy variables is crucial to avoid the dummy variable trap, which occurs when dummy variables are highly correlated. When you create dummies from a categorical variable with ‘n’ categories, you generate ‘n’ new columns. However, these columns sum up to 1 for any observation since each belongs to one category. This redundancy can lead to multicollinearity, adversely affecting the stability of regression coefficients and making it difficult to interpret the model.

By dropping the first category, you retain only ‘n-1’ dummy variables, which provides sufficient information while preventing multicollinearity. This practice simplifies the model and improves its interpretability, ensuring that the omitted category serves as a reference point. Overall, it helps in making more reliable predictions and clearer insights from your analysis.

For example:
Consider a categorical variable "furrnishstatus" with three categories: Furnished, Semi-furnished, and Unfurnished.
If you create dummy variables for all three furnishstatus without dropping one, you'll have:

- furrnishstatus_Furnished
- furrnishstatus_Semi-furnished
- furrnishstatus_Unfurnished

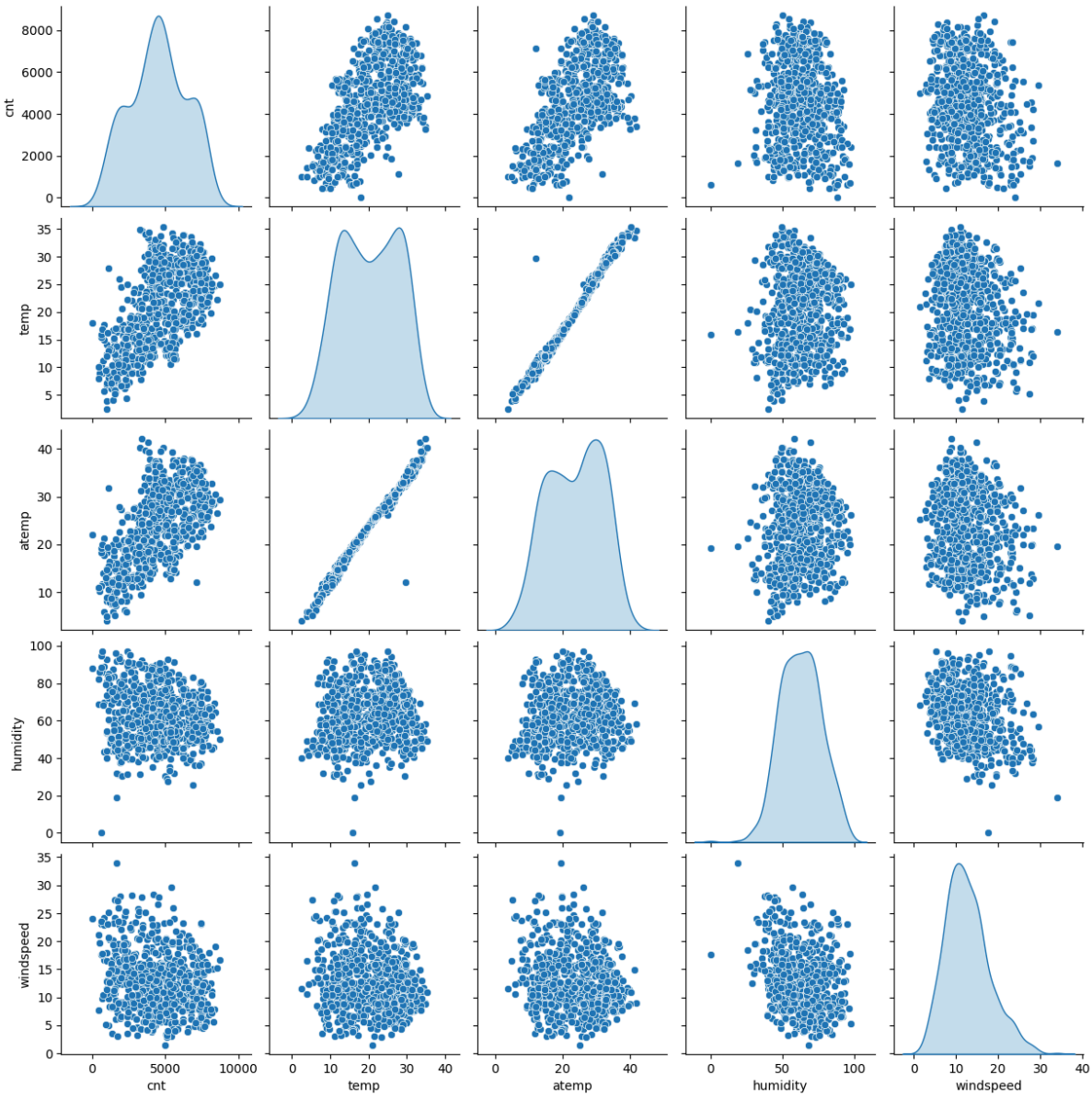
Now, if an observation is "furrnishstatus_Furnished," it will have values (1, 0, 0), while for "furrnishstatus_Semi-furnished," it will be (0, 1, 0), and for "furrnishstatus_Unfurnished," (0, 0, 1). This creates a perfect multicollinearity situation, as the values of furrnishstatus_Furnished , furrnishstatus_Semi-furnished, furrnishstatus_Unfurnished will always sum up to 1.

By using `drop_first=True`, you drop "furrnishstatus_Furnished," resulting in:

- furrnishstatus_Semi-furnished
- furrnishstatus_Unfurnished

Now, Furnished is the reference category. If an observation has (0, 0), it means it's Furnished, (1, 0) means Semi-furnished, and (0, 1) means Unfurnished. This prevents multicollinearity while maintaining all necessary information.

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)

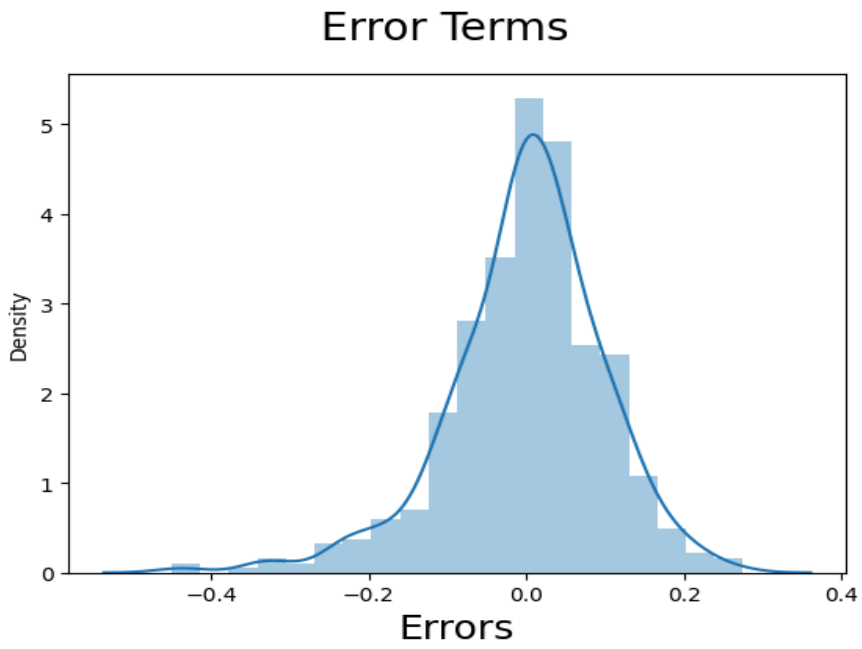


"temp" and "atemp" are the most correlated to the target variable "cnt" in numerical variables, as shown in the pair plot above.

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)

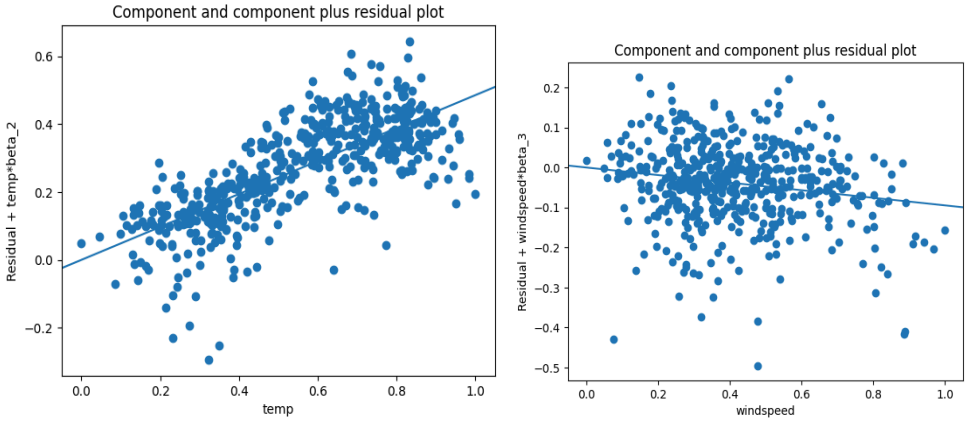
1. **Normality of Error Terms**:

    - **Assumption**: The residuals (errors) should follow a normal distribution.
    - **Validation**: I checked the residuals using a distplot, and they were found to be normally distributed.



2. **Multicollinearity Check**:
    - **Assumption**: There should be no significant multicollinearity among predictors.
    - **Validation**: I calculated the **Variance Inflation Factor (VIF)**. Variables with VIF greater than 5 indicate multicollinearity. The VIF values for each predictor were less than 5, indicating no multicollinearity.

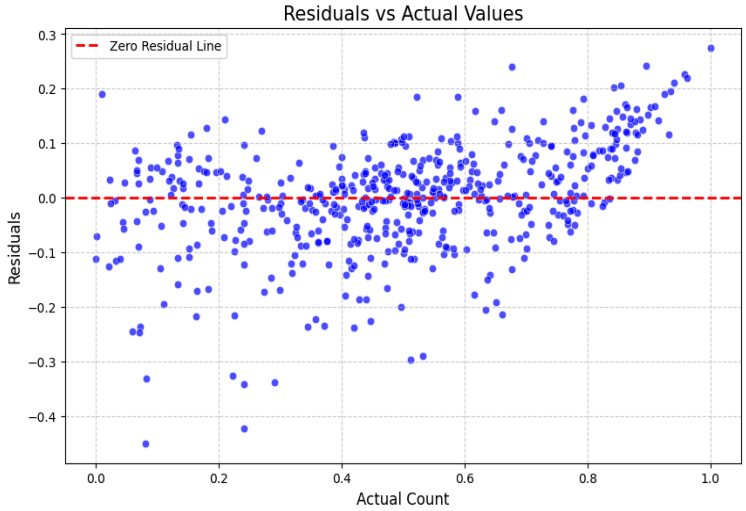
3. **Linear Relationship**:
    - **Assumption**: A linear relationship should exist between independent and dependent variables.
    - **Validation**: I validated this assumption by plotting scatter plots of predictors against the target, revealing linear relationships.



Linearity can be observed from above visualizations.

4. **Homoscedasticity**:
    - **Assumption**: The variance of residuals should remain constant across all levels of independent variables.

- **Validation**: I checked residuals for patterns or varying spread. The residual plot shows no visible patterns, confirming homoscedasticity.



5. **Independence of Residuals**:
   - **Assumption**: Residuals should be independent (no autocorrelation).
   - **Validation**: I used the **Durbin-Watson test** for autocorrelation, which returned a value of 2.099, indicating no autocorrelation.
   - `Durbin-Watson:  2.099`

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of 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 Features Contributing to Bike Demand

Based on the **OLS Regression Results**, the **top three significant features** influencing bike demand (**cnt**) are:

1. **Year (coef = 0.2414, p-value = 0.000)**
   - A positive coefficient means that bike demand **increases over the years**.

2. **Temperature (temp) (coef = 0.4082, p-value = 0.000)**
   - Higher temperatures lead to an **increase in bike demand**.

3. **Weather Situation – Light Snow/Rain (coef = -0.2650, p-value = 0.000)**
   - **Bad weather (light snow or rain)** significantly reduces bike demand.

These features have the **strongest impact** based on their **coefficient values** and **statistical significance (p-value < 0.05)**.

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 algorithm in machine learning and statistics used for predicting a **continuous** target variable based on one or more independent variables. It assumes a **linear** relationship between the dependent and independent variables.

Mathematically the relationship can be represented with the help of following equation –

Y = mX + c

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept.

If X = 0, Y would be equal to c. Furthermore, the linear relationship can be positive or negative in nature as explained below– o Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variables increases.

Negative Linear relationship: A linear relationship will be called positive if the independent increases and the dependent variable decreases.

Types of Linear Regression

Linear regression can be categorized into two types:

-**Simple Linear Regression**

This involves only **one** independent variable (XX) to predict the dependent variable (YY).

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

Example: Predicting **bike demand** based on **temperature**.

- **Multiple Linear Regression**

This involves **multiple** independent variables (X1,X2,X3,...,XnX_1, X_2, X_3, ..., X_n) to predict the dependent variable (YY).

$Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn$

Example: Predicting **bike demand** based on **temperature, wind speed, season, and weather conditions**.

Assumptions of Linear Regression

For linear regression to work effectively, the following assumptions must hold:

-Linearity

The relationship between the independent variable(s) and the dependent variable is **linear**. This means a straight-line relationship should exist.

Check: Scatter plots should show a linear trend.

-Independence of Errors (No Autocorrelation)

The residuals (prediction errors) should not be correlated with each other. This is especially important in **time-series** data.

Check: Durbin-Watson test (values close to 2 indicate no autocorrelation).

-Homoscedasticity (Constant Variance of Errors)

The variance of residuals should remain **constant** across all values of independent variables.

Check: Residual vs. fitted value plot should show **no clear pattern**.

Violation: Heteroscedasticity (when errors have **unequal variance**), often seen as a funnel shape in residual plots.

-Normality of Residuals

The errors (residuals) should follow a **normal distribution** for valid hypothesis testing.

Check: Histogram of residuals or Q-Q plot should be roughly **bell-shaped**.

Violation: If residuals are skewed, the model's predictions may be biased.

-No Perfect Multicollinearity

Independent variables should not be **highly correlated** with each other.

Check: Variance Inflation Factor (**VIF**) should be **<5**.

Violation: If VIF is high, **remove** or **combine** highly correlated features.

Advantages of Linear Regression

- **Easy to interpret** – Simple and widely used.

- **Computationally efficient** – Works well on large datasets.
- **Works well when assumptions hold**.

Limitations of Linear Regression

-**Assumes linear relationships** – Doesn't work for non-linear data.
-**Sensitive to outliers** – Extreme values can distort predictions.
-**Multicollinearity issues** – Highly correlated features can mislead the model.

Applications of Linear Regression

Finance: Stock price prediction.
Real Estate: House price estimation.
Marketing: Sales forecasting.
Bike Sharing: Demand prediction.

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)

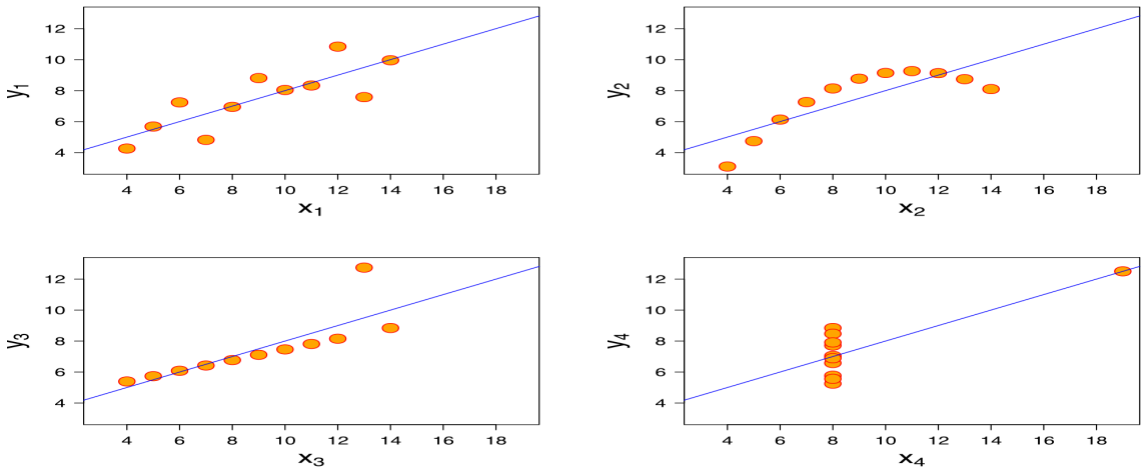
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

I		II		III		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

The summary statistics show that the means and the variances were identical for x and y across the

groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



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 (Pearson Correlation Coefficient) **:** Pearson's R is a statistical measure that calculates the strength and direction of the **linear** relationship between two continuous variables. It ranges from **-1 to +1**:

- **+1** ⟶ Perfect positive correlation (as one variable increases, the other also increases).
- **-1** ⟶ Perfect negative correlation (as one variable increases, the other decreases).
- **0** ⟶ No correlation (the variables are not related linearly).

Formula for Pearson's R:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Example Calculation:

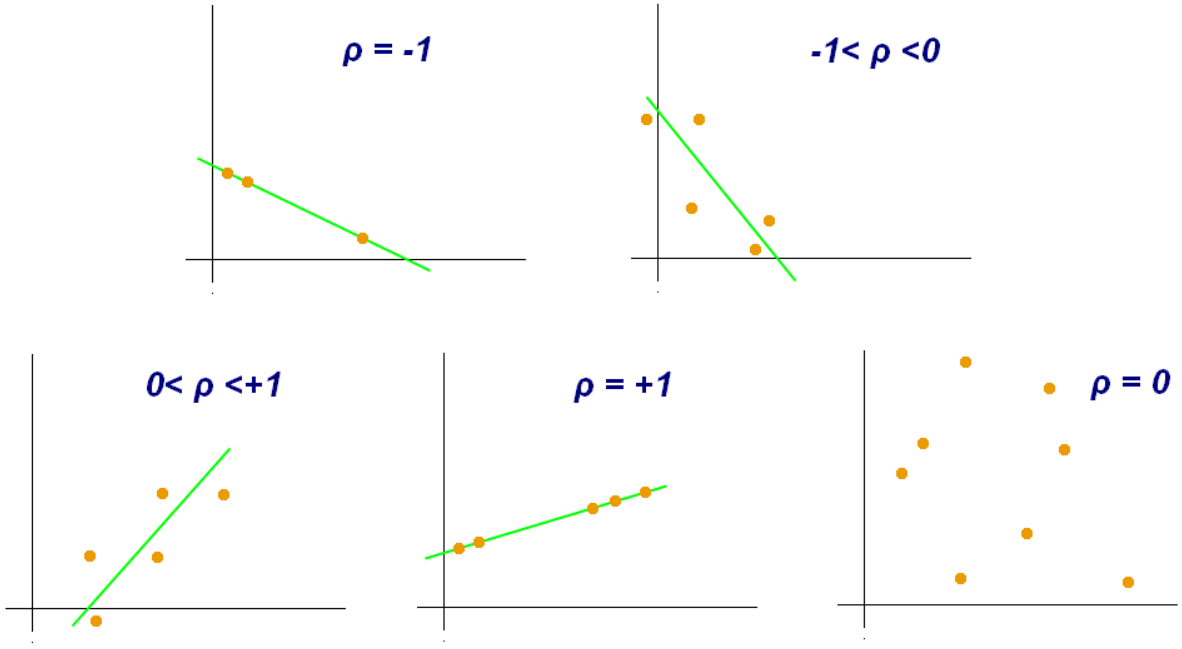
Let's say we have two variables:

Hours Studied (X)	Test Score (Y)
2	50
3	55
5	70
7	85
9	95

If we calculate Pearson's R for this dataset, we might get **r = 0.98**, indicating a **strong positive correlation** (as study hours increase, test scores increase).

Use Cases of Pearson's R:

- **Finance:** Correlation between stock prices and market indices.
- **Healthcare:** Relationship between smoking and lung cancer.
- **Marketing:** Correlation between advertising spend and sales revenue.



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 refers to the process of transforming the features of a dataset to a specific range or distribution, making them easier to work with in machine learning algorithms. Scaling is crucial because many algorithms, including gradient descent-based methods (like linear regression and neural networks), are sensitive to the scale of the input data. Features with larger ranges can disproportionately influence the model, potentially leading to longer convergence times or suboptimal solutions.

Reasons for Scaling:
1. Equal Weightage: Ensures that all features contribute equally to the distance calculations, preventing features with larger ranges from dominating the results.
2. Better Performance: Many machine learning models perform better when input features are scaled.

Types of Scaling:
1. Normalized Scaling: This technique scales the features to a specific range, usually [0, 1]. It is computed as:

X` = X - X_min / X_max - X_min

Example: If a feature value is 50, with a minimum of 0 and a maximum of 100, the normalized value would be:

X` = 50 - 0 / 100 -0 = 0.5

2. Standardized Scaling: This technique transforms the features to have a mean of 0 and a standard deviation of 1. It is calculated as:

X` = X - mean / standard deviation

Example: If a feature has a mean of 10 and a standard deviation of 2, a value of 12 would be standardized as:

X` = 12 - 10/ 2 = 1

In summary, scaling is essential for effective machine learning, and the choice between normalization and standardization depends on the specific requirements and characteristics of the data.

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)

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. A VIF value becomes infinite in the following scenarios:

1. Perfect Multicollinearity: This occurs when one independent variable is a perfect linear combination of one or more other independent variables. For example, if you have two variables where one is exactly twice the other, the VIF for one of those variables will be infinite.

2. Zero Variance: If an independent variable has zero variance (i.e., it is constant and does not change), the calculation of VIF will also lead to an infinite value. This is because the regression model cannot estimate the relationship between the dependent variable and a variable that does not vary.

3. Deterministic Relationships: Any situation where there is a deterministic relationship among the predictors can cause VIF to be infinite. This means that knowing the value of one predictor allows you to perfectly predict the value of another.

In summary, infinite VIF values indicate severe multicollinearity issues that can compromise the reliability of the regression coefficients.

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)

A Q-Q plot (quantile-quantile plot) is a visual tool used to compare two sets of data to see if they come from the same type of distribution (like whether both are normally distributed). Here's how it works in simple terms:

How the Q-Q Plot Works

1. Understanding Quantiles:
- A quantile is a way to understand the data. For instance, the 30th percentile (or 0.3 quantile) is the value below which 30% of the data points lie.

2. Creating the Plot:
- You plot the quantiles of one data set against the quantiles of another data set. If both sets are from the same distribution, the points will follow a straight diagonal line at a 45-degree angle (the reference line).

3. Interpreting the Plot:
- If points fall along the line, it suggests the two data sets are similar.
- If points deviate widely from the line, it indicates the distributions might differ.

Why Q-Q Plots Matter

- Identifying Distribution: They help us decide whether we can treat two samples as coming from the same distribution, which is important for further analysis.
- Understanding Differences: If the distributions differ, Q-Q plots can provide insights into how they differ, which can be more informative than complex statistical tests like the chi-square test.

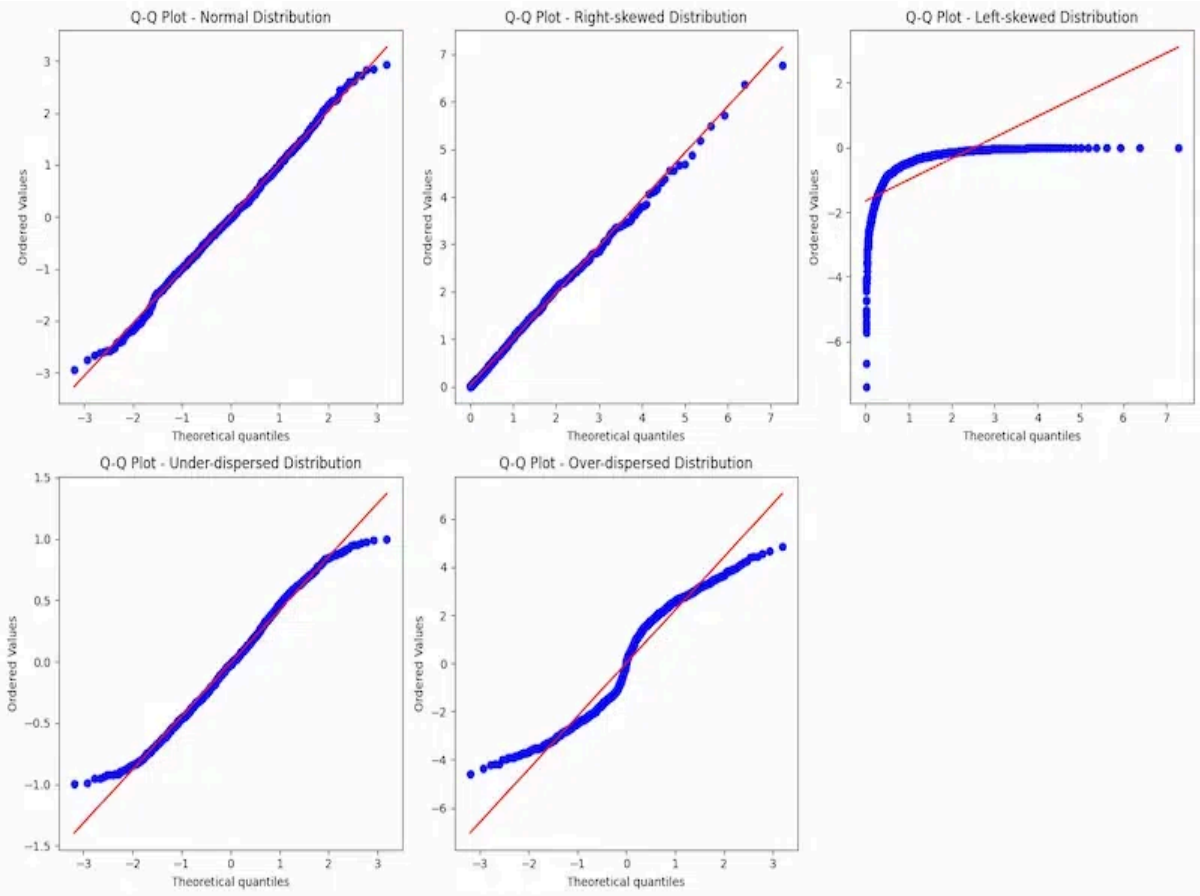
Example

Imagine you have two sets of test scores from different classes:

- Class A Scores: 75, 80, 85, 90, 95
- Class B Scores: 60, 70, 80, 90, 100

1. Calculate Quantiles: Determine the quantiles of both sets.
2. Plot: Create the Q-Q plot by plotting Class A's quantiles against Class B's quantiles.
3. Analyze: If the points align closely along the line, the classes might have similar score distributions. If not, they likely have different distributions.

Visualization



In summary, a Q-Q plot is a helpful tool in statistics for visually assessing whether two datasets share a common distribution.