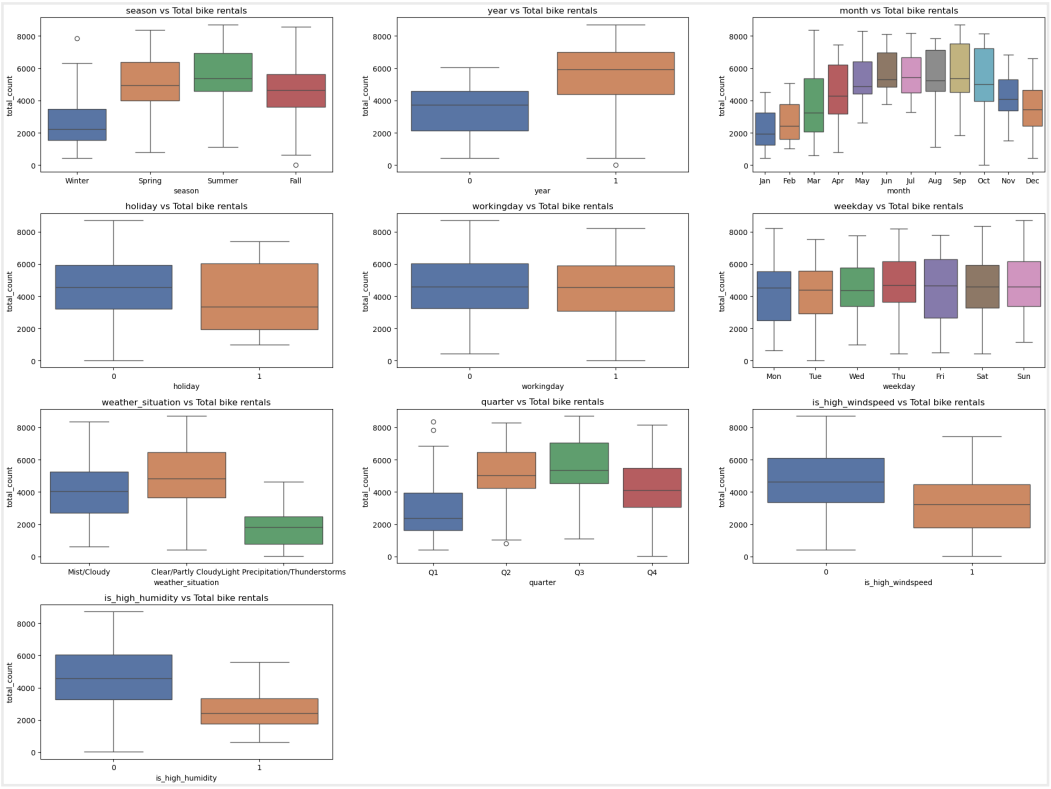
Assignment-based Subjective Questions

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

Ans)



Seasonal Trends:

- **Summer and Spring** have the highest bike rentals, while **Winter** sees the lowest, showing that warmer temperatures lead to more rentals.

Yearly Trends:

- Rentals increased from 2018 to 2019, suggesting growing popularity of bike rentals.

Monthly and Quarterly Trends:

- Rentals peak between **May and October (Month has Normal Distribution)**, with **Q2 (April to June)** and **Q3 (July to September)** being the most popular. **Quarter 1 and Quarter 4** see fewer rentals, particularly during colder months.

Day and Holiday Trends:

- Rentals on **holidays** are slightly lower than on non-holidays, indicating more rentals for commuting on regular days. Weekends tend to have higher rentals compared to holidays. Rentals are consistent across weekdays, with a slight peak on **Thursdays**.

Weather Impact:

- **Clear or partly cloudy days** see the most rentals, while **light precipitation or thunderstorms** reduce them, indicating that extreme weather conditions reduce bike rentals.

Other Factors:

- **High wind speeds** (> 20 m/s) and **high humidity** (> 90%) correlate with fewer rentals.

———————————————————————————————

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

Ans)

Short Answer: To avoid Multicollinearity and Keep model simple.

Long Answer (using Example of Season):
When we create dummy variables for a categorical variable like "Season" with categories ['Fall', 'Spring', 'Summer', 'Winter'], we might end up with something like this:

Season_Fall	Season_Spring	Season_Summer	Season_Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

If we know three of these values, we can always figure out the fourth. So, having all the values leads to redundancy / multicollinearity that can confuse our model.
Using **drop_first=True** drops one of the dummy columns, say "Season_Winter":

Season_Spring	Season_Summer	Season_Winter
1	0	0
0	1	0
0	0	1
0	0	0

Now, if all values are 0, the model knows it's Fall. This prevents multicollinearity, making the model simpler and more reliable.

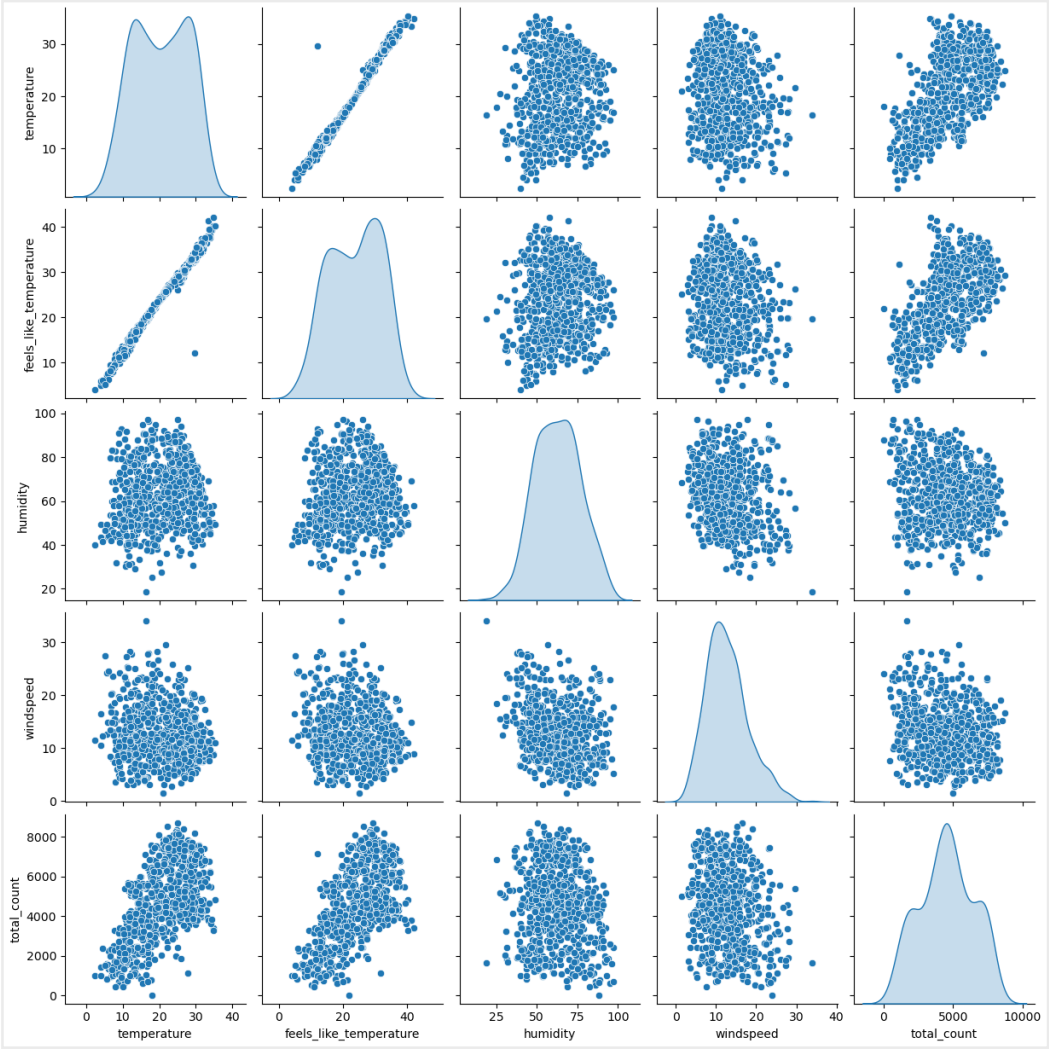
———————————————————————————————

3. Looking at the pair-plot among the numerical variables, which

one has the highest correlation with the target variable?

Ans)

Temperature (dataset name: temp) and feels-like temperature (dataset name: atemp) show the highest and positive correlation with total bike rentals (dataset name: cnt). The pairplot's last column or row highlights a nearly direct and linear relationship between these variables. As temperature and feels-like temperature rise, bike rentals increase. Similarly, bike rentals decrease as temperatures drop.



---

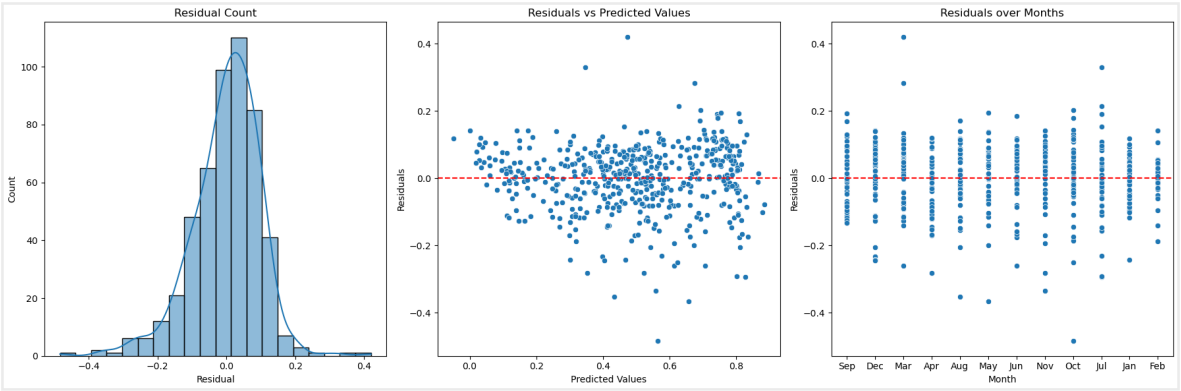
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans)

- **Residual Analysis:**
  - **Normality of Residuals/Errors:** The histogram with KDE (Kernel Density Estimate) indicates that the residuals are approximately

normally distributed, centered around zero. This suggests the normality assumption is satisfied.

- **Linearity and Homoscedasticity:** The residuals vs predicted values plot shows that residuals are fairly randomly scattered around zero with no clear funnel shape, indicating that the linearity and homoscedasticity assumptions (Constant Variance of Errors) are met.
- **Residual Independence/No auto-correlation:** The residuals over months plot does not show a strong pattern, indicating residuals to be independent of each other (no correlation between errors).



- **No Multicollinearity:**
  - All the predictors in final model have **low VIF (less than 5)**, indicating multicollinearity is not a problem in this model.

Features	VIF
temperature	4.54
season_Summer	3.13
year	2.06
month_Jul	1.62
weather_situation_Mist/Cloudy	1.50
month_Sep	1.40
season_Winter	1.25
month_Oct	1.18
holiday	1.03
weather_situation_Light Precipitation/Thunders...	1.03

---

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

Ans)

Below are the top 3 features that contribute most significantly to explaining the demand for shared bikes in your model.

- **temperature** (**Coefficient:** 0.4878 indicating that as temperature increases, the demand for shared bikes significantly increases.
- **year** (**Coefficient:** 0.2433) indicating an increasing trend in bike-sharing usage over time.
- **weather_situation_Light_Precipitation_Thunderstorms** (**Coefficient:** -0.2777) indicating that when the weather involves light precipitation or thunderstorms, the demand for shared bikes decreases significantly. Adverse weather conditions are likely to reduce people from using bikes.

Multiple Linear Regression Equation:

total_count = 0.1822 + (0.4878 * temperature) + (0.2433 * year) - (0.2777 * weather_situation_Light_Precipitation_Thunderstorms) - (0.1476 * season_Winter) + (0.0925 * month_Sep) - (0.0802 * weather_situation_Mist_Cloudy) - (0.0743 * holiday) + (0.0609 * month_Oct) - (0.0449 * season_Summer) - (0.0418 * month_Jul)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans)

Linear Regression is used to model the relationship between a dependent variable/target variable, and one or more independent variables/features/ predictors. The goal is to find the best-fitting linear equation that can predict the dependent variable based on the values of the independent variables.

Steps in the Linear Regression Algorithm:

1. **Simple vs. Multiple Linear Regression**:
   Simple Linear Regression: Involves one independent variable.
   Equation: $y = \beta_0 + \beta_1 x + \epsilon$
   - **Multiple Linear Regression**: Involves two or more independent variables. Equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$

     where,
     y = dependent variable
     x = independent variable
     ϵ = Residual/Error Term (difference between the observed and predicted values of y)

coefficients:
β_0 = intercept (the value of y, when x is 0).
$\beta_{1..n}$ = slope (the change in y for a one-unit change in x)

The goal is to estimate the coefficients such that the sum of the squared differences between the observed values and the predicted values (the residuals) is minimized. This is known as **Ordinary Least Squares (OLS)**.

2. **Cost Function:**
   cost function = difference between the predicted values and actual values.
   Since, goal is to minimise the cost function, It can be achieved by:
   - **Differentiation**
   - **Gradient descent method**

3. The **strength of a simple linear regression model** is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
   - **RSS:** Residual Sum of Squares. Formula: Sum of (Square of Predicted - Actual values)
   - **TSS:** Total Sum of Squares. Formula: Sum of (Actual values - Mean)

4. In case of Multiple Linear Regression,
   Below steps are added:
   - **Dealing with Categorical Variables**: Convert them into dummy variables.
   - **Feature Scaling**: Standardization and MinMax scaling ensure consistency in data, making it suitable for modeling.
   - **Model Assessment**: Adjusted R-squared, AIC, and BIC are key metrics for evaluating model performance and selecting the best set of features.
   - **Feature Selection**: Feature selection can be manual (which can be tedious) or automated using methods like forward/backward selection or regularization. The best practice is to reduce features to a certain number through an automated process and then manually drop additional features if needed.

5. Assumptions of Linear Regression (Residual Analysis):
   - **Linearity**: The relationship between the dependent and independent variables is linear.
   - **Independence**: The residuals (errors) are independent.
   - **Homoscedasticity**: The residuals have constant variance.
   - **Normality**: The residuals of the model are normally distributed.
   - **No Multicollinearity**: The independent variables are not highly correlated with each other.

Summarized Steps in Multiple Linear Regression (After EDA):
Step 1] Dummy Variables for Non Binary Categorical Features.
Step 2] Splitting the Datasets into Training and Test Dataset.
Step 3] Scaling: Fit and Transform Training Dataset.
Step 4] Creating Models and Feature Selection using manual and automated approach (Like RFE, regularization, etc).
Step 5] Residual Analysis.
Step 6] Transforming Dataset.
Step 7] Making Prediction on Test Dataset.
Step 8] R Squared and Visual Analysis on Test Dataset.

———————————————————————————————————

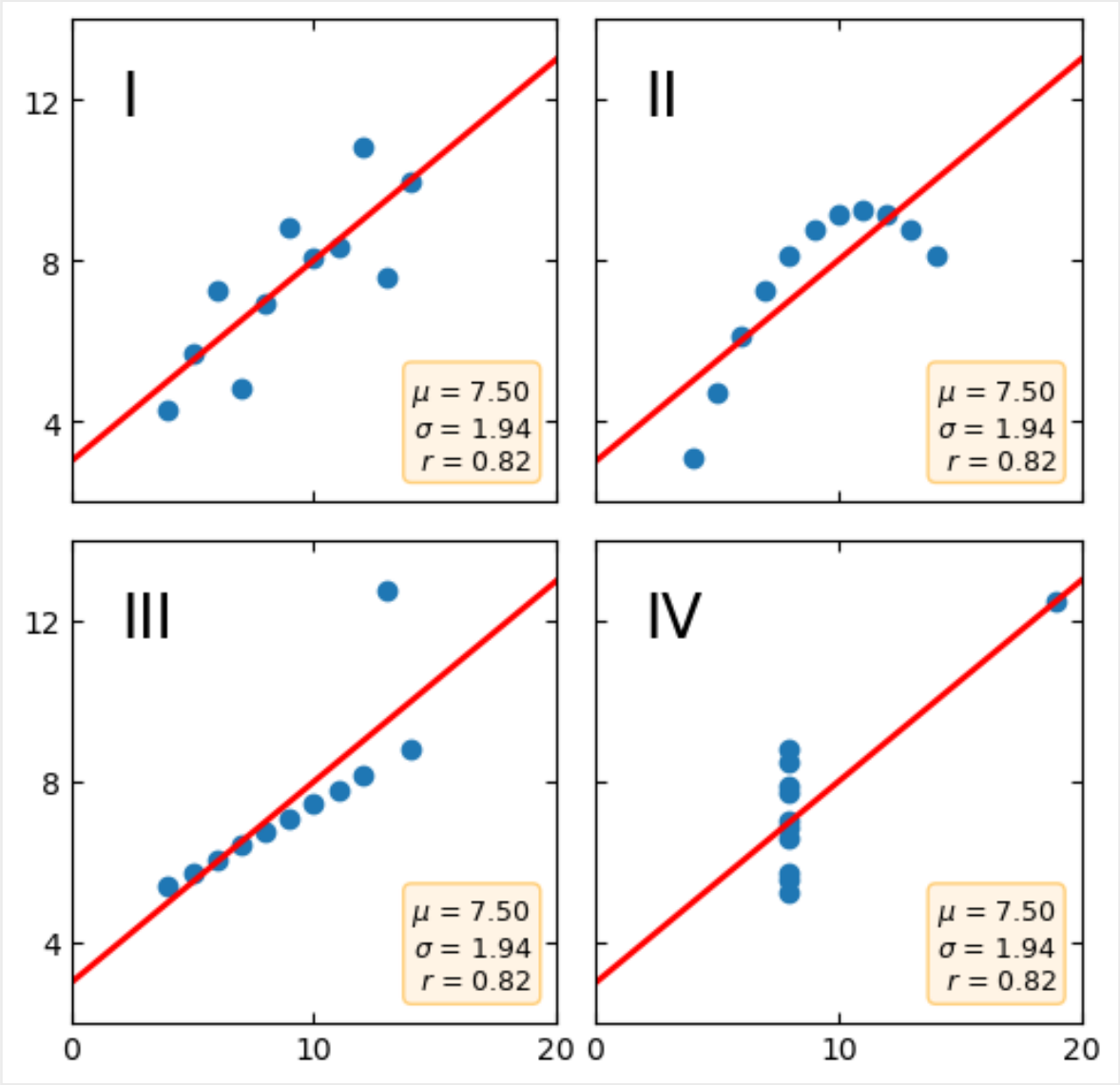
2. Explain the Anscombe’s quartet in detail.

Ans)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they are very different when graphed.

- **Anscombe's Quartet Importance:**
  The key idea behind Anscombe's Quartet is to show that different datasets can have the same statistical properties—such as mean, correlation, and regression line—but still be very different when you look at their distributions or plots.
- **Example:**
  Here are the four datasets:

Dataset	x Values	y Values
I	10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5	8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
II	10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5	9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
III	10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5	7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
IV	8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8	6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89



Visualization and Interpretation:
The difference becomes clear—by plotting these datasets:

1. **Dataset I**:
    - A typical linear relationship.
    - The data points are fairly evenly distributed around the regression line.
    - **Standard linear regression analysis.**

2. **Dataset II**:
    - **Non-linear relationship (curved).**
    - The regression line is not a good fit for this data.
    - High correlation doesn't necessarily mean a linear relationship.

3. **Dataset III**:
    - **Contains an outlier that affects the regression line.**
    - Removing the outlier would result in a very different regression model.

- Influence of outliers on statistical analysis.

4. **Dataset IV**:
   - Almost all data points have the **same x-value with one outlier that determines the slope of the regression line.**
   - **Correlation and regression statistics are misleading due to the lack of variability in x.**

Conclusion
Anscombe's Quartet is a powerful illustration of why it is essential to visualize your data and not rely solely on summary statistics.

---

3. What is Pearson’s R?

Ans)

Pearson's R (also known as the **Pearson correlation coefficient**) is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between the two variables.

Key Characteristics of Pearson's R:

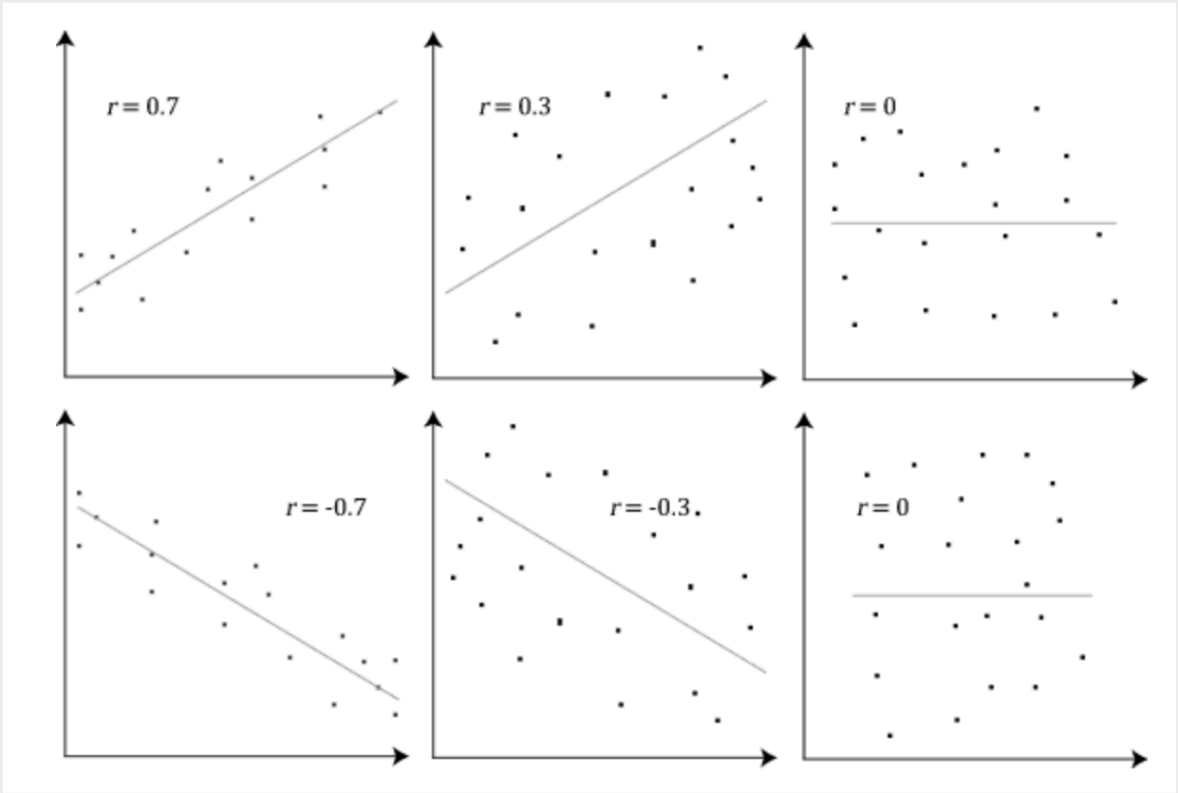
1. **Range**:
   - Pearson's R ranges from -1 to +1.
     - **Positive Correlation (R = +1)**: Perfect positive linear correlation (as one variable increases, the other increases proportionally).
     - **Negative Correlation (R = -1)**: Perfect negative linear correlation (as one variable increases, the other decreases proportionally).
     - **No Correlation (R = 0)**: No linear correlation (no linear relationship between the variables).

2. **Formula:**

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

3. **Example:**



- **Top Left (r = 0.7)** - **Positive Correlation**: The scatter plot shows a moderate to strong positive linear relationship. As one variable increases, the other also increases.
- **Top Middle (r = 0.3)** - **Weak Positive Correlation**: There is a weak positive linear relationship, with a slight upward trend. The points are more scattered compared to r = 0.7.
- **Top Right (r = 0)** - **No Correlation**: The scatter plot shows no linear relationship between the variables. The points are scattered.
- **Bottom Left (r = -0.7)** - **Negative Correlation**: A moderate to strong negative linear relationship. As one variable increases, the other decreases.
- **Bottom Middle (r = -0.3)** - **Weak Negative Correlation**: A weak negative linear relationship is present, with a slight downward trend. The points are more scattered compared to r = -0.7..
- **Bottom Right (r = 0)** - **No Correlation**: Similar to the top right, this plot shows no linear relationship, with points randomly dispersed.

———————————————————————————————————

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

Scaling is the process of adjusting the range of feature values in your data. This is important in machine learning and statistical modeling because many algorithms assume that all features are on a similar scale or perform better when they are.

Why is Scaling Performed?

1. **Improves Model Performance**
2. **Ensures Fair Comparison**: Without scaling, features with larger ranges can dominate the learning process, leading to biased models. For instance, in a dataset with height (in cm) and income (in dollars), income values might overshadow height values if not scaled.
3. **Convergence in Gradient Descent**: Scaling can help the optimization process converge faster by ensuring that the steps taken during the optimization process are more uniform.

Difference Between Normalized Scaling and Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling)**
   - This technique adjusts the values of each feature so that they fall within a specific range, typically between 0 and 1.
   - **When to Use**: Use normalization when you want to maintain the relationship between the original data points (e.g., ratios) and need to ensure all features are on a comparable scale.
   - **Example**: Suppose you have two features: Height (150-200 cm) and Weight (50-100 kg). After normalization, both features will have values between 0 and 1, making them comparable.
   - **Formula:** (x - min(x)) / (max(x) - min(x))

2. **Standardized Scaling (Z-Score Scaling)**
   - This technique centers the values of each feature around the mean and scales them according to the standard deviation. The result is a feature with a mean of 0 and a standard deviation of 1.
   - **When to Use**: Use standardization when the features have different units and distributions, especially when the algorithm assumes normally distributed data or when dealing with features that have outliers.
   - **Example**:Suppose you have Exam Scores (out of 100) and IQ Scores (mean 100, SD 15). Standardizing these features would convert them into z-scores with a mean of 0 and standard deviation of 1, making them comparable in the context of their distributions.
   - **Formula:** (x - mean(x)) / sd(x)

———————————————————————————————————

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

Ans)

Variance Inflation Factor (VIF) is a measure of multicollinearity for

independent variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, meaning that one variable can be predicted from the others with a significant degree of accuracy.
When the VIF for a variable becomes infinite, it indicates **perfect multicollinearity**. This happens when one independent variable is an exact linear combination of one or more of the other independent variables in the model. In other words, one variable can be perfectly predicted using the other variables, leading to infinite VIF.

Example:
Dataset with the following variables:

- **Income**
- **Savings**
- **Total Wealth = Income + Savings**

Here, **Total Wealth** is a perfect linear combination of **Income** and **Savings**:
If we add **Income**, **Savings**, and **Total Wealth** in a regression model, the VIF for these variables will be infinite because **Total Wealth** can be perfectly predicted by **Income** and **Savings**.

———————————————————————————————————

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

Ans)

A **Q-Q plot** (Quantile-Quantile plot) is a simple graphical tool that helps to see if your data looks like it comes from a specific distribution, most commonly a normal distribution. It’s like holding your data up against a "template" to see how well they match.

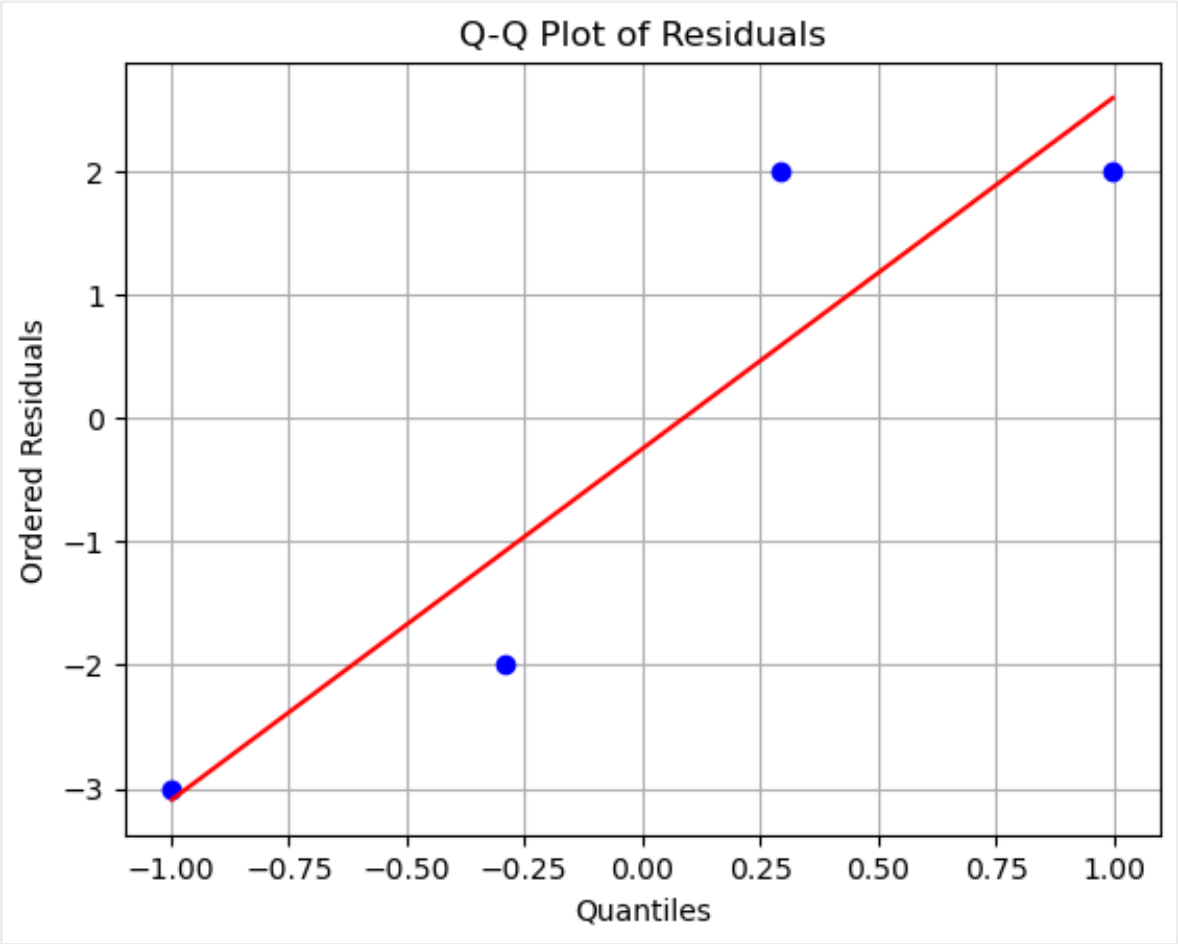
Use of Q-Q Plot in Linear Regression:
When we create a linear regression model, one of the important assumptions is that the residuals (predicted value - actual value) should be normally distributed. If they aren't, your model might not be reliable.

Example: Height of Students
Imagine we want to predict the height based on age. We built a simple linear regression model. After making predictions, we want to check if your model's residuals are normally distributed.

Data:

- **Student A**: Actual Height = 160 cm, Predicted Height = 162 cm, Residual = -2 cm
- **Student B**: Actual Height = 150 cm, Predicted Height = 148 cm, Residual = 2 cm
- **Student C**: Actual Height = 165 cm, Predicted Height = 163 cm,

Residual = 2 cm
- **Student D**: Actual Height = 155 cm, Predicted Height = 158 cm, Residual = -3 cm



Interpretation:

- **Points on the Red Line**: The plot shows that most of the points lie close to the red line, indicating that the residuals are approximately normally distributed.
- **No Extreme Deviations**: There are no extreme deviations from the line, suggesting that the normality assumption for these residuals is reasonable.

Steps to Use a Q-Q Plot:

1. **Calculate Residuals**
2. **Create a Q-Q Plot**
3. **Interpret the Plot**
    - **If the residuals are normally distributed**, they will line up along a straight diagonal line in the plot.
    - **If they aren't**, you'll see them curve away from the line. This might mean there are outliers, or the data is skewed in some way.

Importance:

- **Checking for Problems**: If your residuals don't line up well in the Q-Q

plot, it might mean your linear regression model isn’t fitting the data properly. You might have outliers, or maybe the relationship between your variables isn’t really linear.

- **Improving Your Model**: By spotting issues with the Q-Q plot, you can decide to maybe adjust your model, try a different approach, or transform your data to get a better fit.

Conclusion:
A Q-Q plot is a simple way to check if the residuals from linear regression model look like they should. If the points on the plot line up well with a straight line, your model is likely doing a good job. If not, need to investigate the model further. This check helps ensure that your model's predictions are reliable.