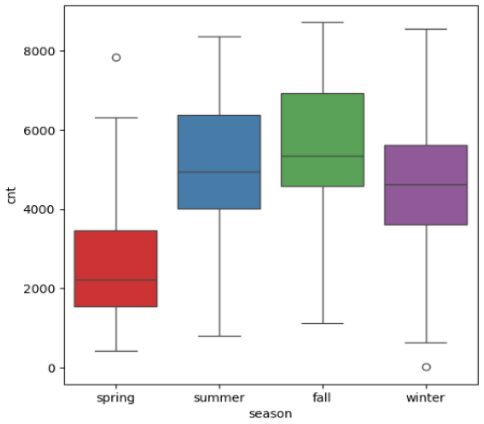
Assignment-based Subjective Questions

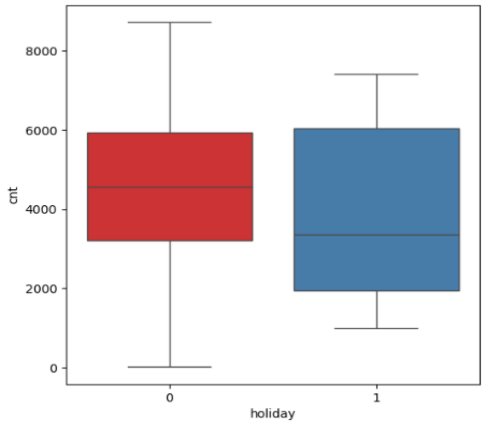
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

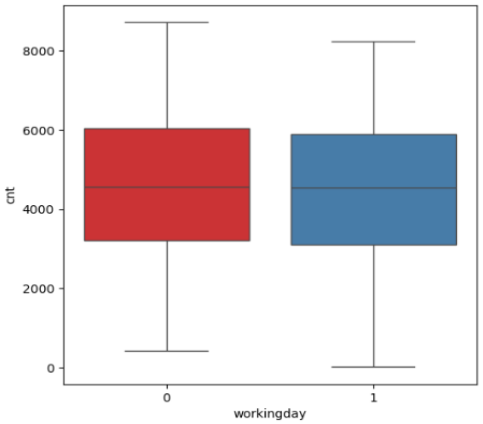
Answer:

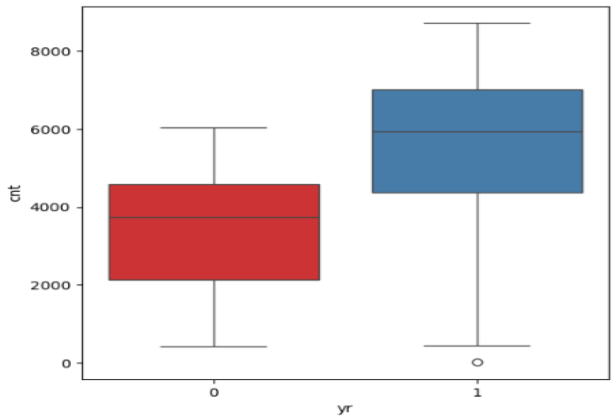
- **Seasonal Analysis:** Fall has the highest average rentals, followed closely by summer that means bike rental counts are higher in the Summer and Fall compared to Winter and Spring.
- **Year-wise Rentals:** The year 2019 (yr=1) shows a notable increase in demand compared to 2018 (yr=0), suggesting a rise in bike rentals over time.
- **Monthly Trend:** Bike demand varies greatly by month, with September showing the highest median counts, while January has the lowest.
- **Holiday vs. Working Days:** Non-working days (workingday=0) show a slightly higher demand compared to working days (workingday=1).
- **Weekday Analysis:** Median counts are relatively consistent throughout the week, with a slight peak on Sundays (weekday=0).

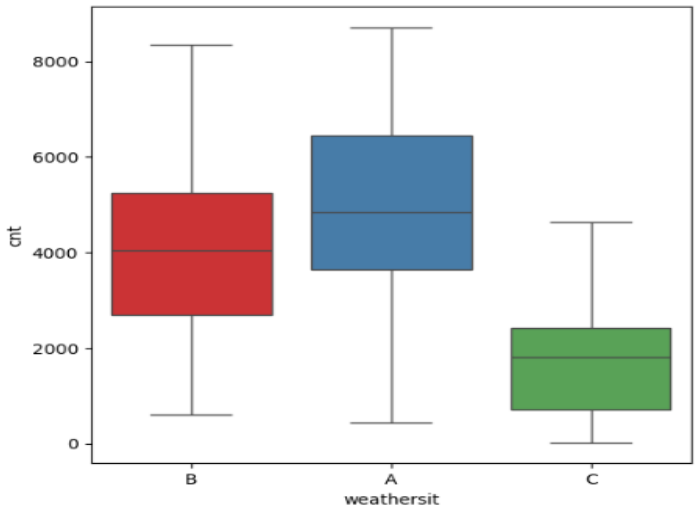
- **Weathersit trend :** Clear weather conditions (weathersit=1) result in the highest median bike count, followed by mist and cloudy weather (weathersit=2) and light snow weather (weathersit=3).

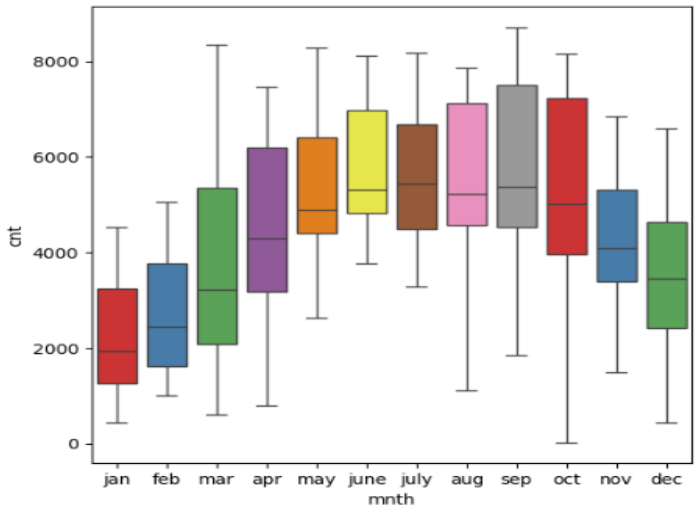


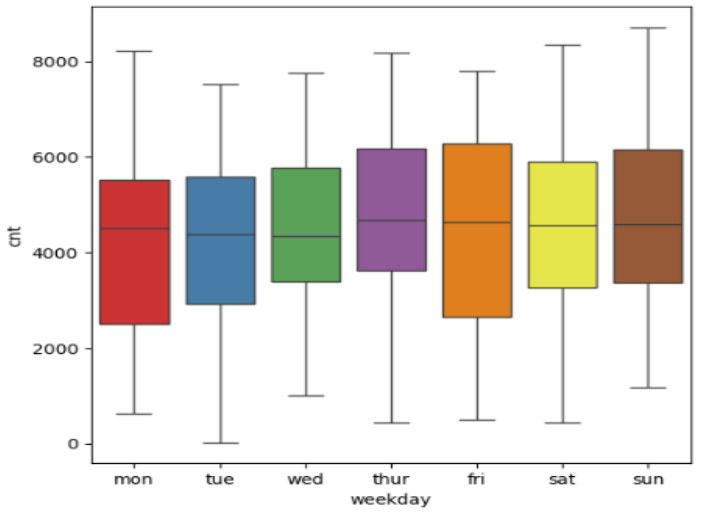












2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

Using drop_first=True during dummy variable creation is important to avoid the "dummy variable trap," which occurs when you have multicollinearity in your model.

- When you create dummy variables for a categorical feature with n categories, you end up with n binary columns, each representing one category. If you include all n dummy variables in your model, one of them can be perfectly predicted from the others. This creates perfect multicollinearity, which can lead to problems in regression model.

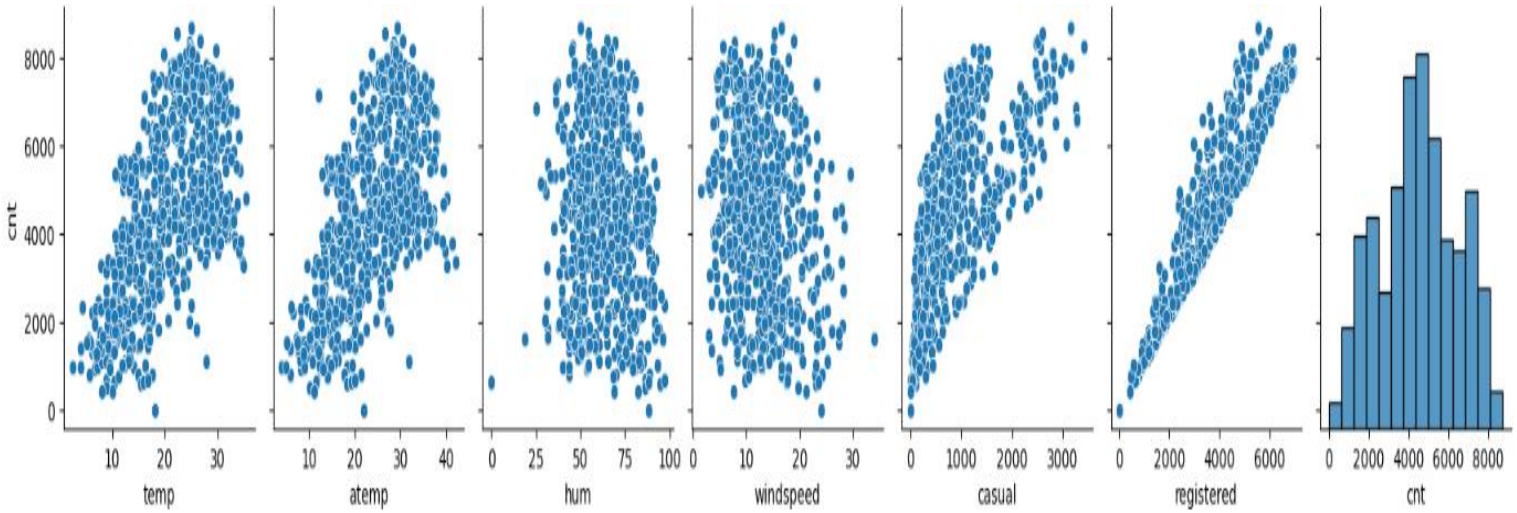
- Dropping the first dummy variable helps to eliminate this redundancy. This reduces the number of dummy variables by one, preventing multicollinearity while still allowing the model to capture the differences between the categories.

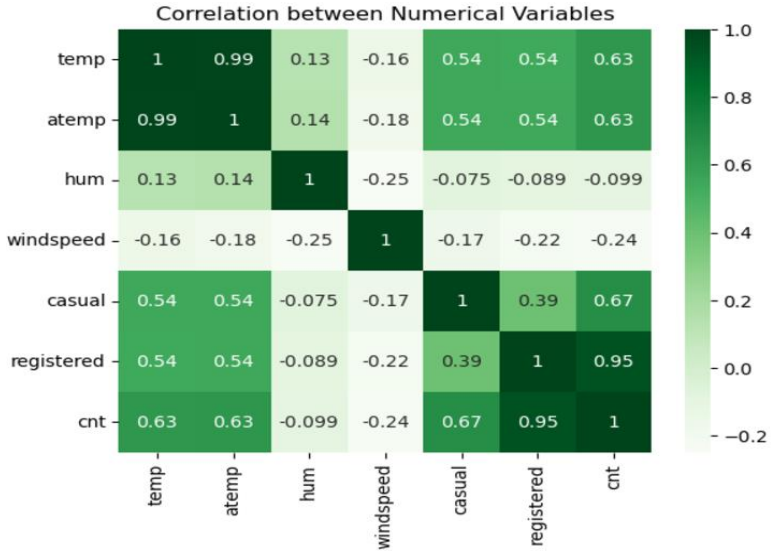
In essence, setting drop_first = True ensures that the dummy variables are linearly independent, which is key for reliable and interpretable model.

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

Answer :

- The variable **"temp" has the highest correlation with the target variable**.
- The "casual" and "registered" variables are components of the target variable, as their values combine to form the target, so their correlation is not considered.
- "atemp" is derived from "temp," humidity, and windspeed, and is excluded from the model preparation process, so it's not considered.



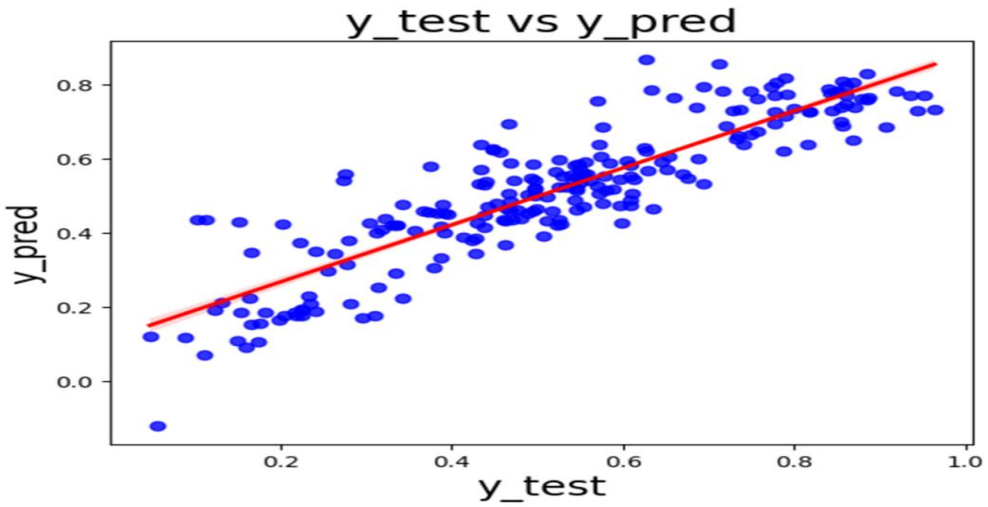


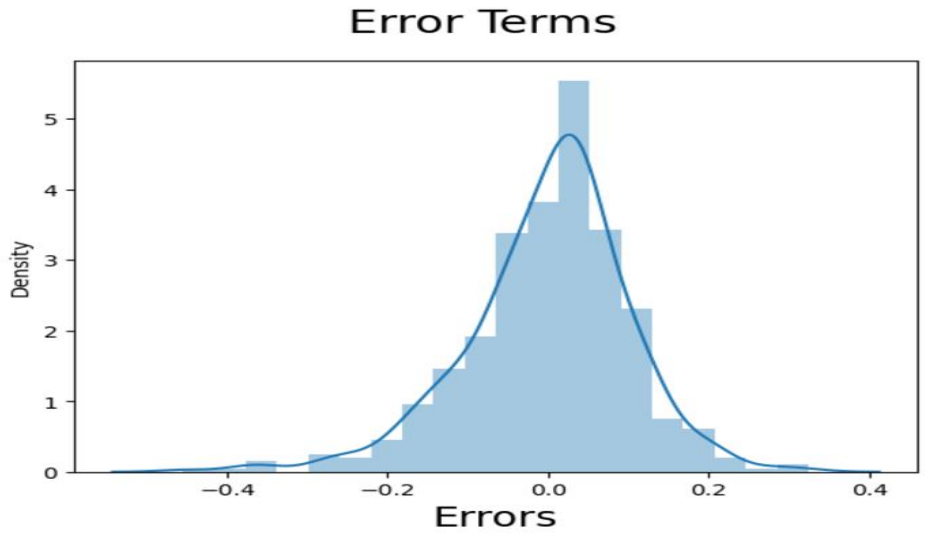
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer :

After building a linear regression model on the training set, we validate its assumptions to ensure the model's reliability and that the results are meaningful. The key assumptions of linear regression are:

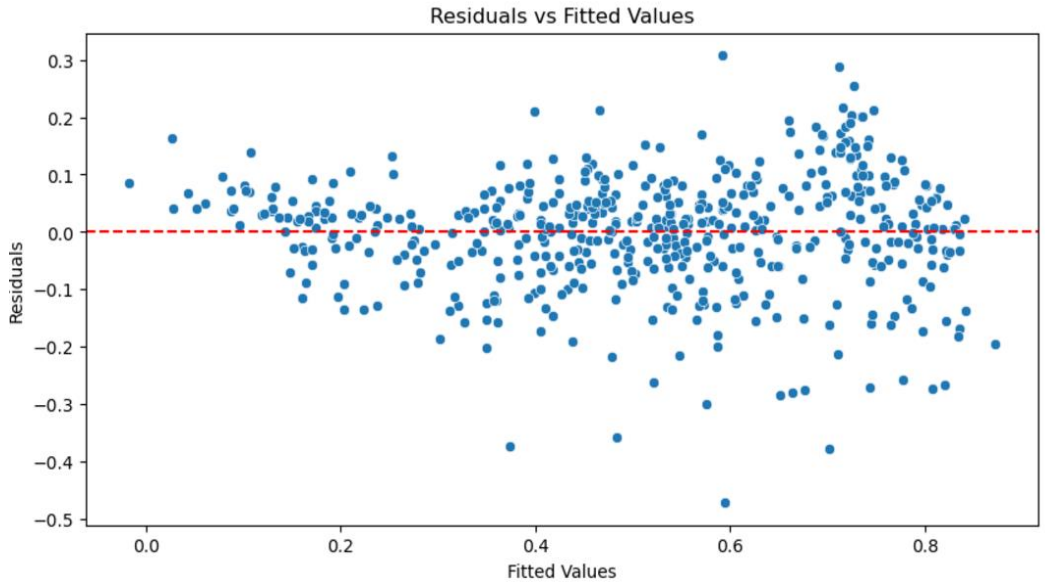
- **Linearity**: The relationship between the independent and dependent variables should be linear.
- **Independence**: The residuals (errors) should be independent of each other.
- **Homoscedasticity**: The residuals should have constant variance across all levels of the independent variables.

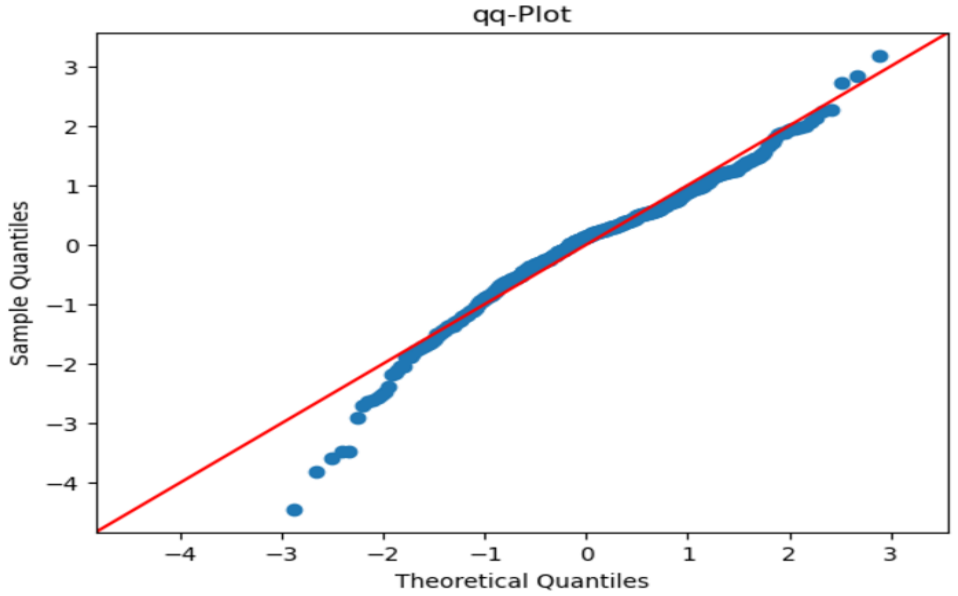




- **Normality of residuals**: The residuals should be normally distributed.
- **No multicollinearity**: The independent variables should not be highly correlated with each other.

For the above we can plot and check these assumptions.





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

Answer:

The top three factors that have the greatest impact on bike bookings on the basis of coefficient value are:

1. **temp** (cof:0.490781): This is the predictor with the highest absolute coefficient value, indicating that temperature has the strongest impact on the dependent variable. For each unit increase in **temp**, the dependent variable is expected to increase by **0.4908 units**.
2. **weathersit_C** (cof :-0.252097): This variable has the second highest absolute coefficient, indicating a strong negative effect. If the weather condition is classified as "C", the dependent variable is expected to decrease by **0.2521 units**.
3. **yr** (cof:0.234479): This variable has the third-highest absolute coefficient, indicating a positive effect. With each increase in **yr** (year), the dependent variable is expected to increase by **0.2345 units**.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: Simple and Multiple.

- Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

- Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

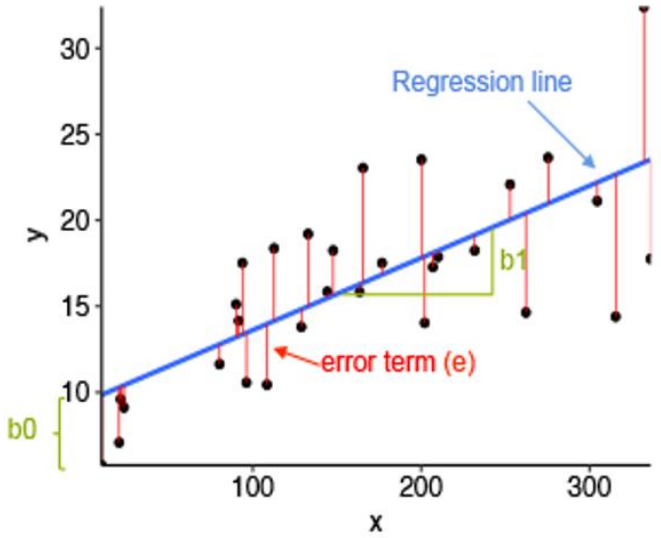
Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, b_1,b_2,b_3,b_4...,b_n are coefficients or slopes of the independent variables x_1,x_2,x_3,x_4...,x_n and y is the dependent variable.

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Let's understand this with the help of a diagram.



$$y = b_o + b_1x$$

- x is our independent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
- $e_i = y_i - y_{pred}$ provides the error for each of the data point.
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares (RSS).

$$\text{RSS} = = \sum_{i=1}^{n} \left(y_i - y_{pred}\right)^2$$

Assumptions:

- Linearity: Relationship between X and Y is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of error terms.
- No Multicollinearity: Independent variables are not highly correlated.
- Normality of Residuals: Errors are normally distributed

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet was created by statistician Francis Anscombe, consists of four datasets that share nearly identical statistical properties but exhibit very different distributions and appear quite distinct when plotted on a graph.

- **Objective**: The primary goal is to emphasize that visualizing data is essential to fully understand its distribution, as summary statistics alone can be insufficient or deceptive.
- **Core Characteristics**: All four datasets have the same mean, variance, correlation, and regression line, but look very different when visualized.
- **Interpretation**: It encourages the combined use of visual analysis along with summary statistics to gain a more accurate understanding of the data.

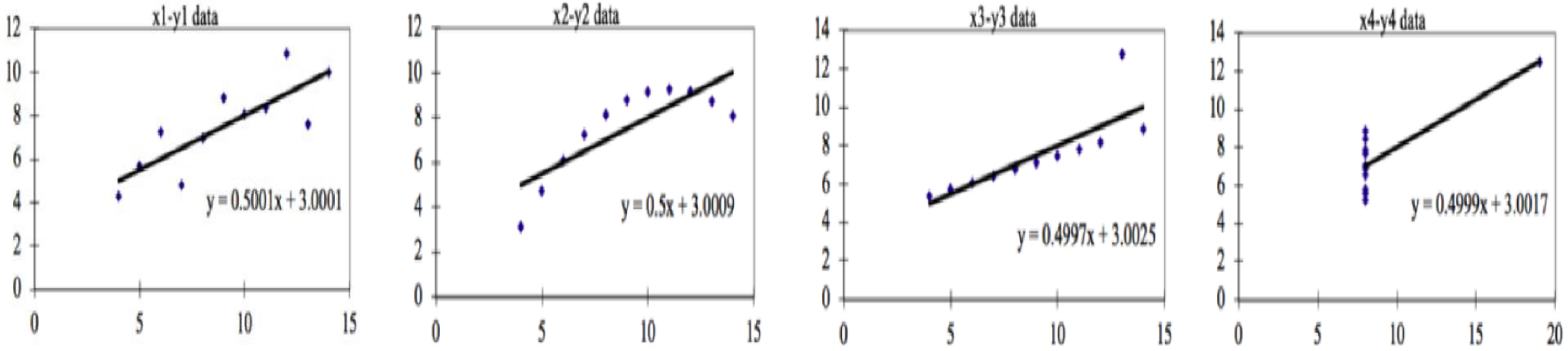
These four plots can be defined as follows →

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

- The statistical details for all four datasets are nearly identical and can be computed as follows:

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	

- When these datasets are plotted on a scatter plot, we see that they all share the same regression lines, yet each dataset conveys a different message. Now let's check scatter plot .



The four datasets can be described as:

- The first scatter plot shows a clear linear relationship between the variables.
- The second plot does not follow a normal distribution; although there is a relationship, it is not linear.
- In the third scatter plot , the distribution appears linear at first glance. However, the regression line is heavily influenced by one **outlier** that skews the results. This outlier distorts the overall linearity of the data, leading to a significant shift in the regression line.
- The fourth scatter plot illustrates how one high-leverage point can create a strong correlation coefficient, even though the other data points do not show any relationship between the variables.

3. What is Pearson's R?

Answer :

Pearson's r is a numerical measure that summarizes the strength of the linear relationship between two variables. If the variables increase and decrease together, the correlation coefficient will be positive. If one variable tends to increase while the other decreases, the correlation will be negative.

- The Pearson correlation coefficient, denoted as r, ranges from +1 to -1:
- A value of **+1** indicates a perfect positive linear relationship, where both variables increase together.
- A value of **-1** indicates a perfect negative linear relationship, where one variable increases as the other decreases.
- A value of **0** suggests no linear relationship between the variables.

Values greater than 0 indicate a positive relationship, meaning that as one variable increases, the other also tends to increase. Conversely, values less than 0 indicate a negative relationship, meaning that as one variable increases, the other tends to decrease.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

Answer : Scaling is the process of adjusting the values of variables so that they are on a similar level. This is important because if variables have different units or ranges, some might have more influence on the analysis than others. By scaling the data, we make sure all variables are treated equally, which helps improve the performance of machine learning models.

Why : Scaling is important because it ensures all features contribute equally to the model, improves the performance of algorithms like gradient descent, makes comparisons easier, and helps distance-based algorithms like KNN or SVM avoid being dominated by larger values.

Example: Let's take a dataset with two features:

- **Age** (in years) ranging from 18 to 100.
- **Income** (in rupees) ranging from 25,000 to 100,000.

Without Scaling:

- **Age**: Values like 18 and 100 are small compared to income.
- **Income**: Values like 20,000 and 100,000 are much larger.

If you use this data without scaling, the income feature will dominate the model because its values are much larger than the age feature. The model will focus more on income than on age, leading to biased predictions.

After Scaling:

- Both features are transformed to the same scale, such as between 0 and 1 (normalization) or with a mean of 0 and standard deviation of 1 (standardization).
- Now, both **age** and **income** contribute equally to the model, allowing the algorithm to consider both features fairly and avoid bias toward the larger scale of income.

Difference between normalized scaling and standardized scaling

S.no.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

Answer :

- **VIF (Variance Inflation Factor)** measures the extent to which the variance of a regression coefficient is increased because of multicollinearity.

Reason for Infinite VIF:

- **Perfect Multicollinearity**: This occurs when one independent variable is a perfect linear function of one or more other variables. In such cases, the correlation between the variables is either 1 or -1, meaning they are perfectly correlated.
- When perfect multicollinearity exists, it becomes impossible to separate the individual effects of the correlated variables on the dependent variable. The matrix used in the calculation of VIF cannot be inverted, leading to an infinite value for the VIF.

Cause: This occurs when one independent variable is an exact linear combination of the others. As a result, $1-R2=0$, which causes division by zero in the VIF formula.

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

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

Answer :

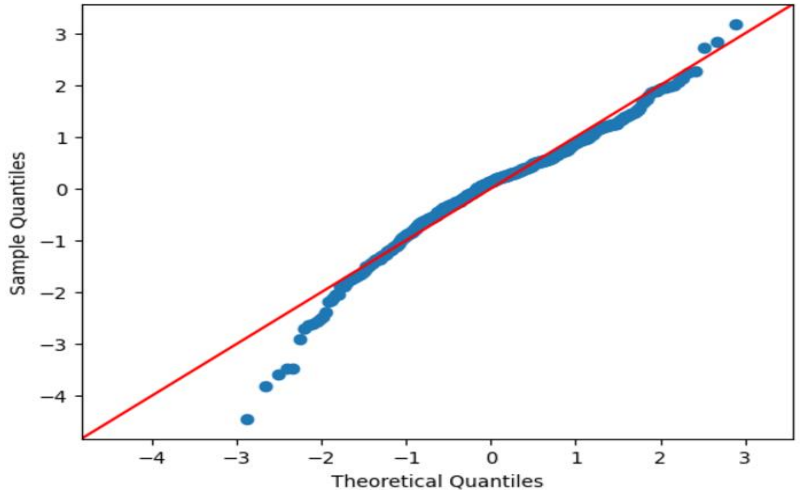
A Q-Q (Quantile-Quantile) plot is a graphical plot used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of a dataset with the quantiles of a specified theoretical distribution .

Use of Q-Q Plot:

In linear regression, a Q-Q plot is primarily used to check the normality of the residuals (the differences between the observed and predicted values). The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below:

Interpretations :

- Points on the line: Data is normally distributed.
- Points off the line: Indicates skewness or kurtosis.



Importance of Q-Q Plot :

A Q-Q plot in linear regression is important for:

1. It helps to visually check if residuals are normally distributed, a key assumption for valid regression results.

2. Ensures reliable hypothesis testing and confidence intervals.

3. Detects outliers that may distort model accuracy.

4. Provides a quick visual check for potential issues with the regression model.

Hence Q-Q plots are essential in linear regression for verifying normality of residuals, ensuring accurate inferences, detecting outliers, and diagnosing model issues.