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:
After conducting an analysis on the categorical columns using box plots and bar plots, I have observed several significant insights that can be inferred from the visualizations. –

1. The year 2019 witnessed a higher number of bookings compared to the previous year, indicating positive progress in terms of business.
2. On non-holiday days, the number of bookings appears to be lower, which is reasonable since people may prefer to spend time at home and enjoy with family during holidays.
3. Thursdays, Fridays, Saturdays, and Sundays have a higher number of bookings compared to the start of the week.
4. The fall season attracted more bookings, and the booking count increased significantly in each season from 2018 to 2019.
5. Most of the bookings were made during the months of May, June, July, August, September, and October. The trend increased from the beginning of the year until the middle of the year and then decreased towards the end of the year.
6. Clear weather attracted more bookings, as people are more likely to come out and travel during such conditions.

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

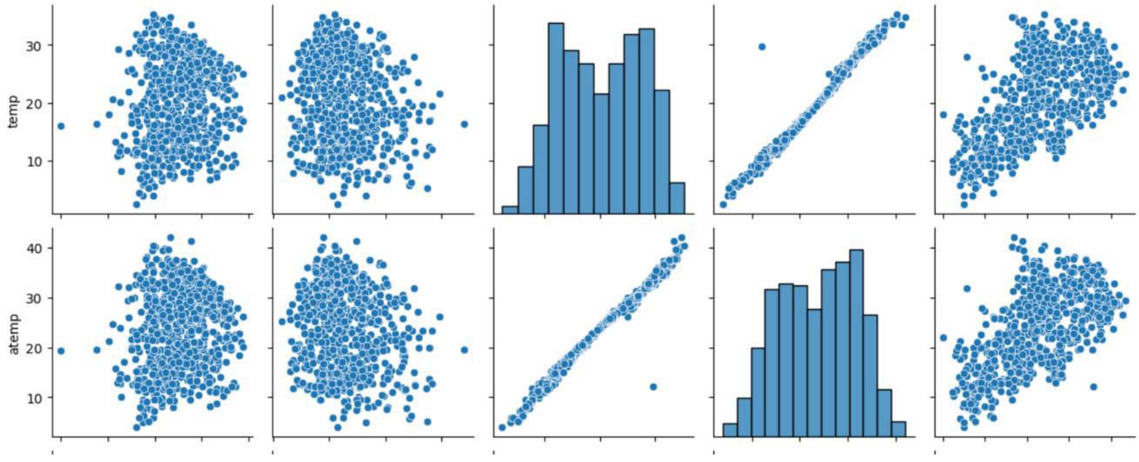
Answer:
The drop_first=True parameter is crucial to use during the creation of dummy variables, as it helps in mitigating the issue of multicollinearity. When set to True, it ensures that we get k-1 dummies out of k categorical levels by removing the first level.

Consider a scenario where we have a categorical column with three types of values: A, B, and C. If we create dummy variables without dropping the first level, we will end up with three variables: A, B, and C. However, we can infer that if a data point is not classified as A or B, then it must be C. Therefore, we do not need a third variable to identify C, as its presence is implicit from the absence of A and B.

By using drop_first=True, we avoid introducing redundant information, reducing the chances of creating correlations among dummy variables. This optimization helps in obtaining a more efficient representation of categorical data without compromising the integrity of the information.

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

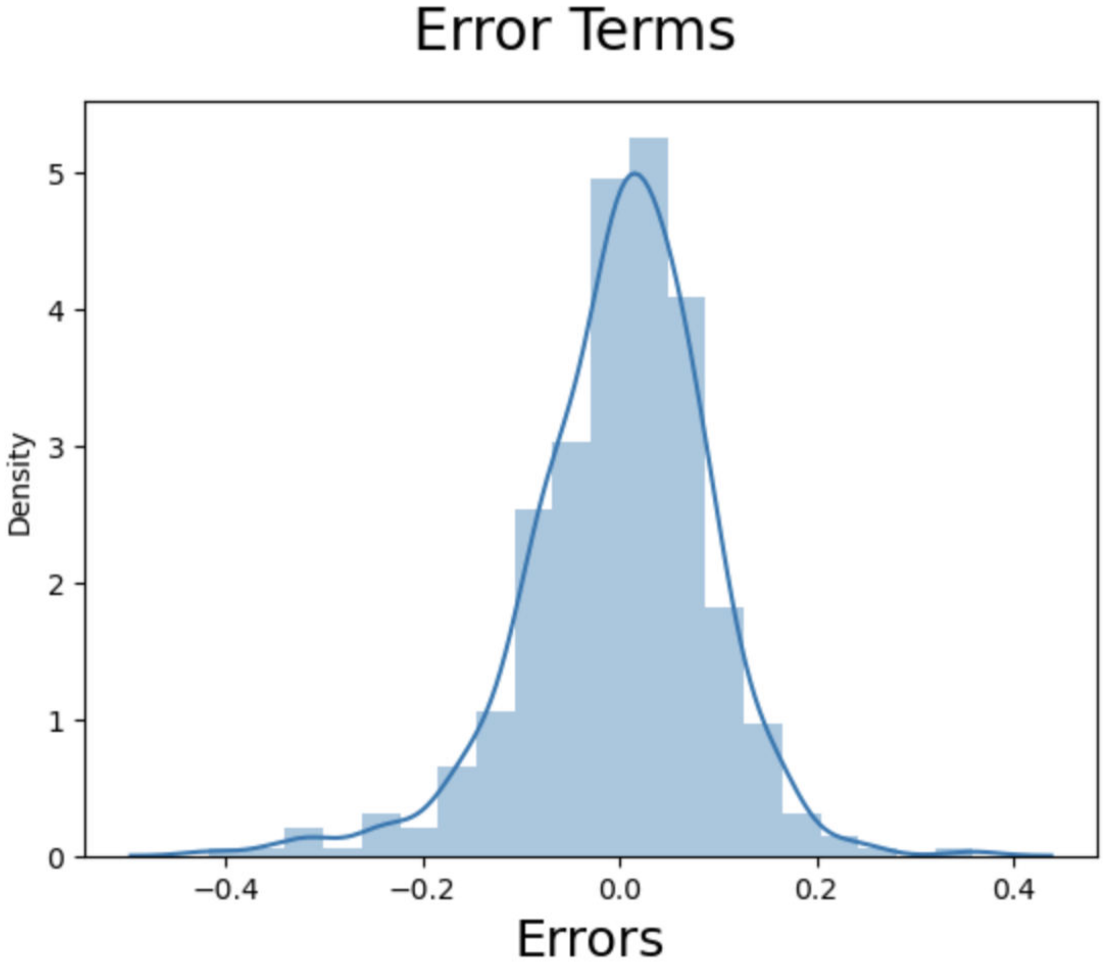
Answer:
The 'temp' and 'atemp' variable exhibits the strongest correlation with the target variable. Graphs screenshots are mentioned below :-



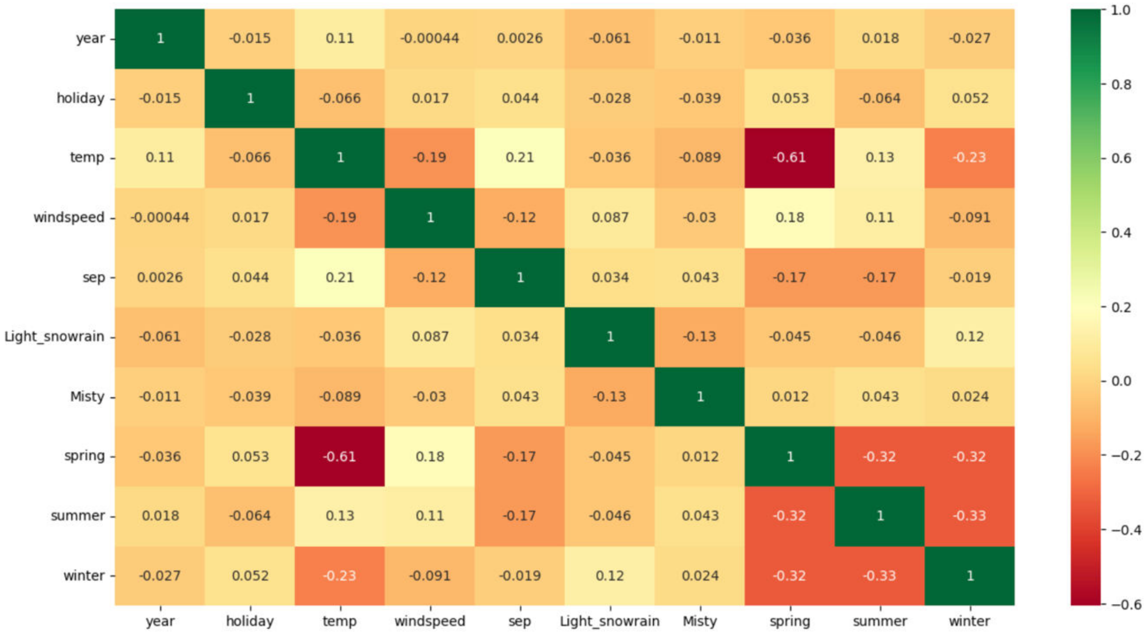
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:

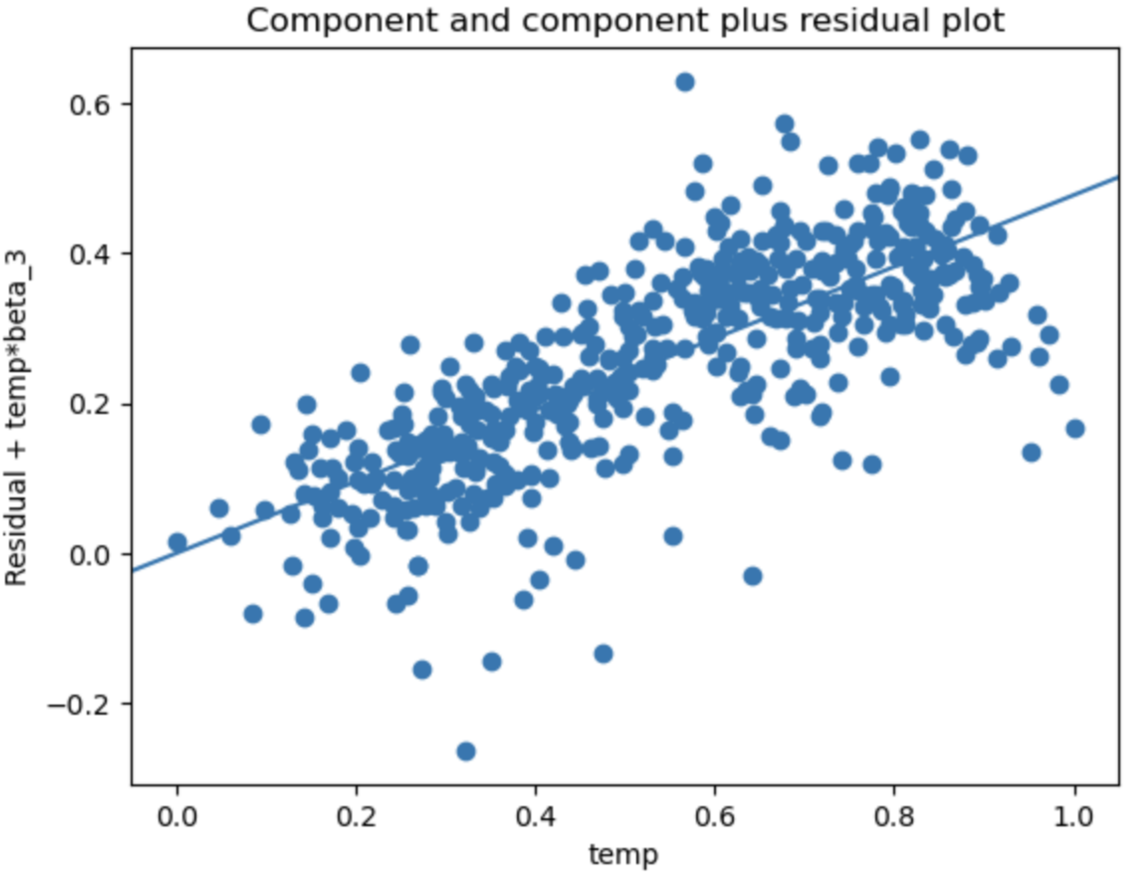
The assumption of Linear Regression Model were validated based on mentioned assumptions -
Normality of error terms: The error terms were found to be normally distributed.

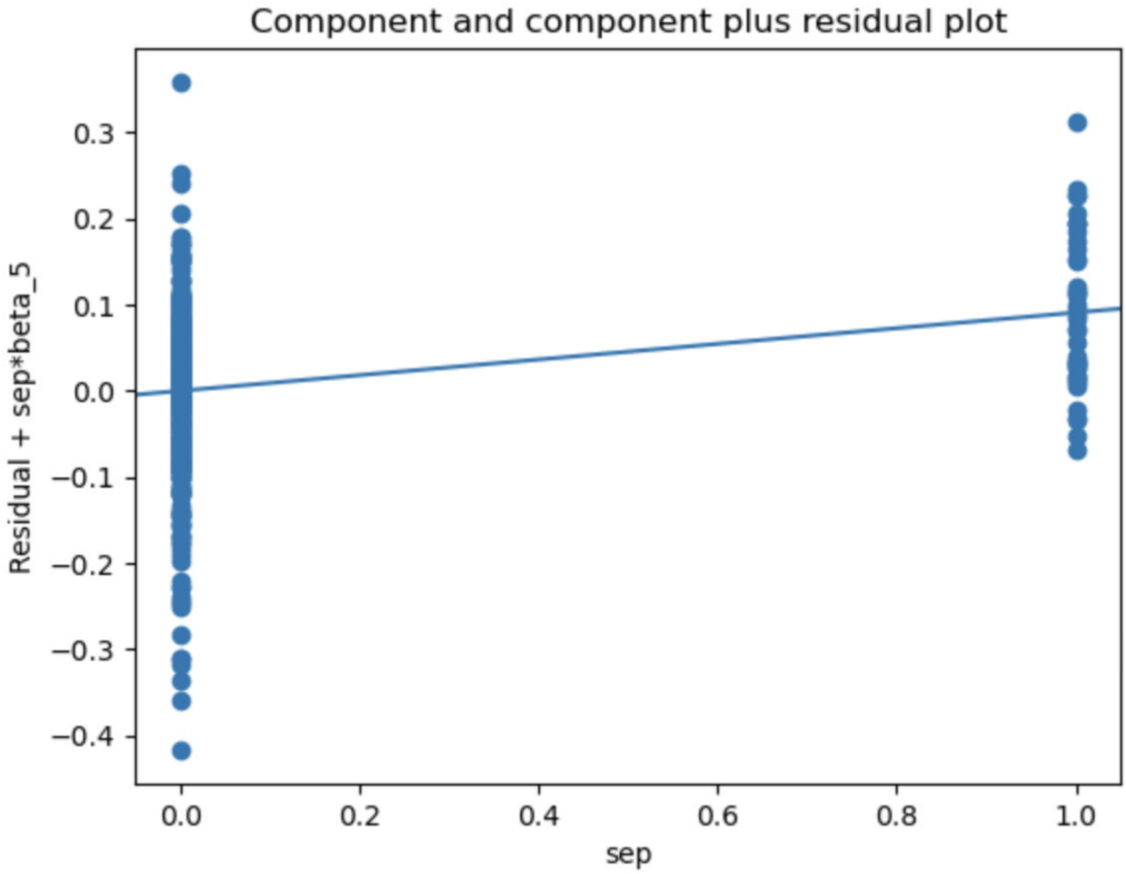


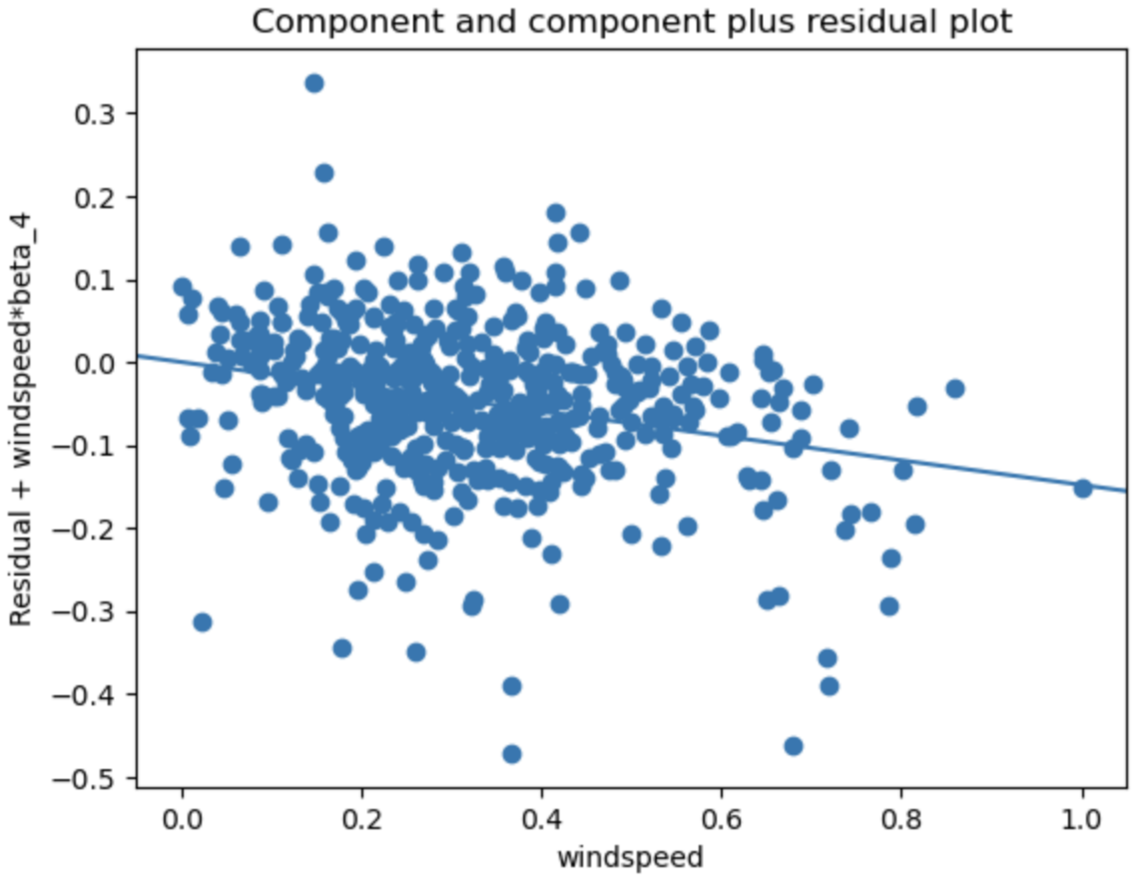
Multicollinearity check: There was no significant multicollinearity among variables.



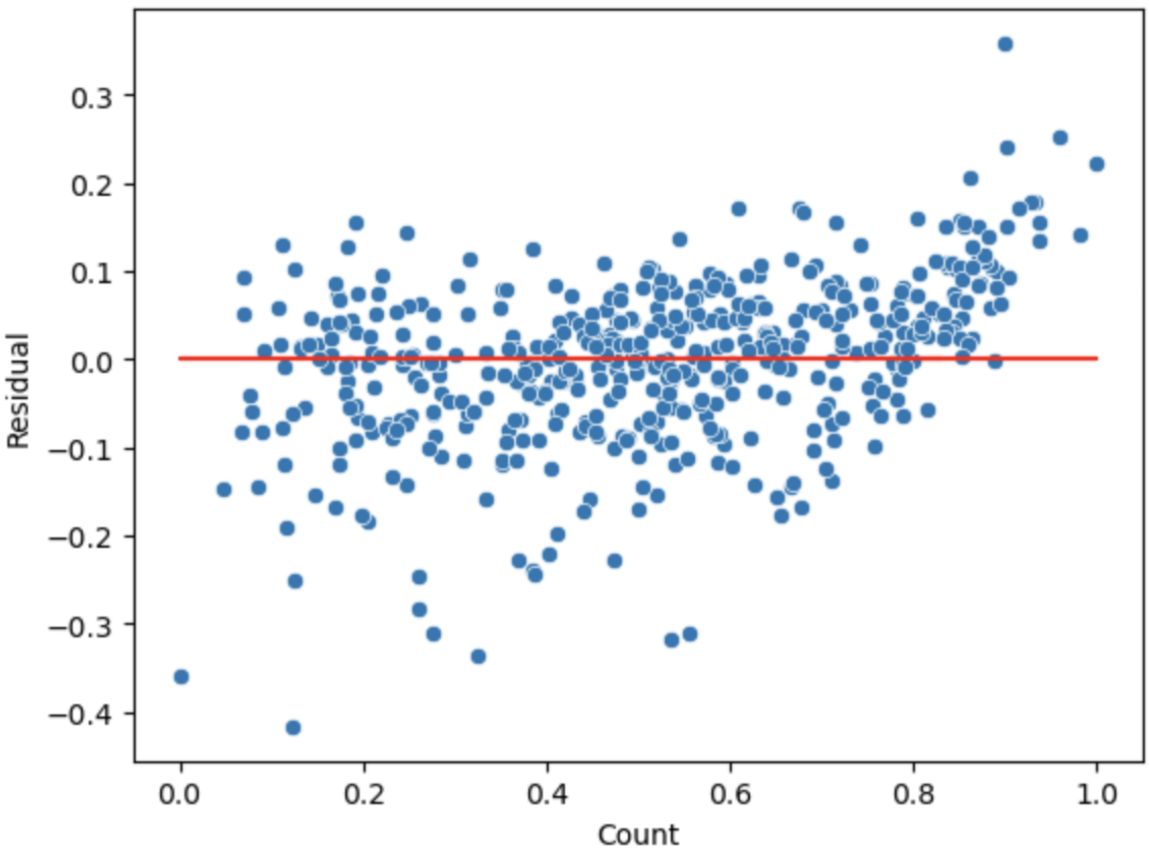
Linear relationship validation: Linearity was evident among variables.







Homoscedasticity: There was no visible pattern in the residual values.



Independence of residuals: No auto-correlation was observed in the residuals.
Durbin-Watson value of final model lr_6 is 2.085, which signifies there is no autocorrelation.

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

Answer:
Top three predictors variables that influence booking of bike are:

1) Temperature (temp)
2) Weather (weathersit)
3) Year (yr)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:
Linear regression can be defined as a statistical model that examines the linear association between a dependent variable and a given set of independent variables. This linear relationship implies that when the values of one or more independent variables change (increase or decrease), the value of the dependent variable will also change accordingly (increase or decrease).

Mathematically, this relationship can be represented by the equation:
$Y = mX + c$

In this equation:

Y represents the dependent variable we are trying to predict.

X denotes the independent variable used for making predictions.
m is the slope of the regression line, representing the effect of X on Y.
c is a constant known as the Y-intercept. If X equals 0, Y would be equal to c.

Linear regression can be positive or negative depending on the slope of the line.

There are two types of Linear Regression –

- Simple Linear Regression
- Multiple Linear Regression

2. **Explain the Anscombe's quartet in detail.**

Answer:

Anscombe's quartet is a set of four datasets that was introduced by the statistician Francis Anscombe in 1973. Anscombe's quartet comprises four datasets with remarkably similar simple descriptive statistics. However, these datasets exhibit distinct distributions and appear significantly dissimilar when graphed. Each dataset is composed of 11 (x, y) data points.

The four datasets are as follows:

Data set 1: y = x + 3
Data set 2: y = 0.5x + 2.5
Data set 3: y = x^2 + 1
Data set 4: y = 3x^2 + 2

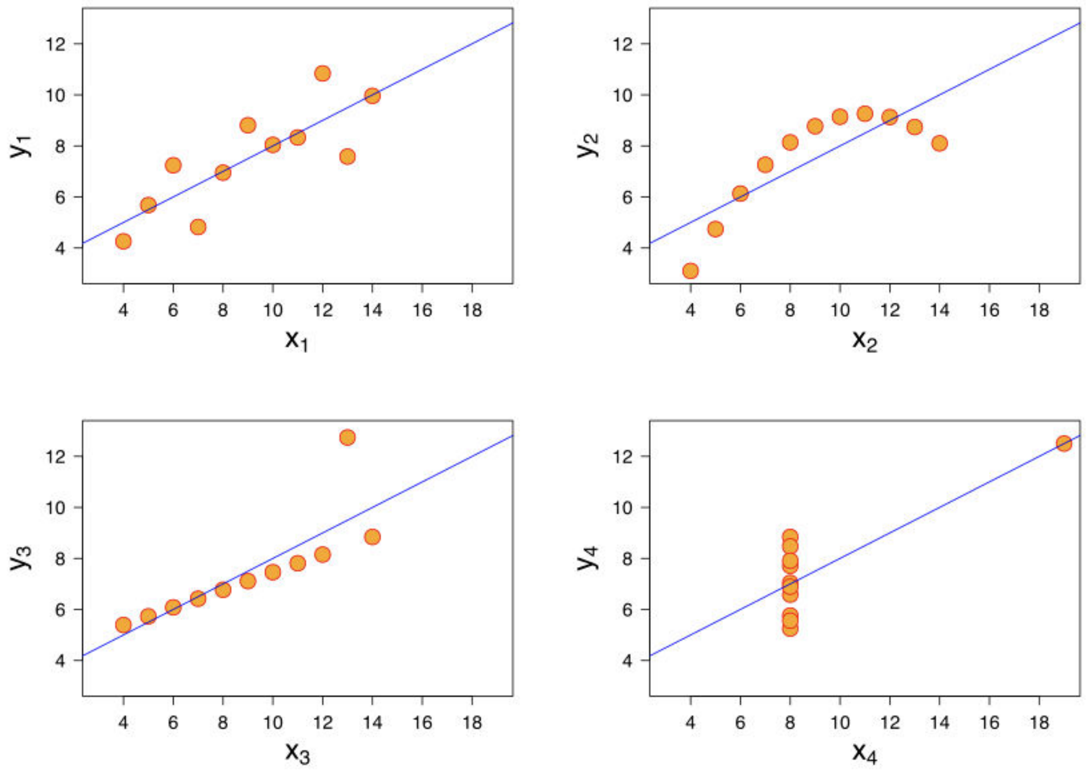


Diagram reference – Wikipedia

3. **What is Pearson's R?**

Answer: (Reference google.com)
Pearson's R is a statistical measure that is used to quantify the strength and direction of the linear relationship between two variables. It is a number between -1 and 1, where:

-1 indicates a perfect negative correlation
0 indicates no correlation
1 indicates a perfect positive correlation
A positive correlation means that as one variable increases, the other variable also increases. A negative correlation means that as one variable increases, the other variable decreases.

Here are some of the things to keep in mind when using Pearson's R:
• It is a measure of linear correlation. This means that it only measures the strength of a linear relationship between two variables. If the relationship between the variables is not linear, then Pearson's R may not be an accurate measure of the strength of the relationship.
• It is sensitive to outliers. Outliers are data points that are far away from the rest of the data. Outliers can have a significant impact on the value of Pearson's R, so it is important to check for outliers before using Pearson's R.
Overall, Pearson's R is a valuable statistical measure that can be used to quantify the strength and direction of the linear relationship between two variables. However, it is important to keep in mind its limitations, such as its sensitivity to outliers and its inability to measure non-linear relationships.

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

Answer: (Reference google.com)

Scaling is a data preprocessing technique used in machine learning to bring all the features or variables to a similar scale or range. It involves transforming the numerical values of the features so that they fall within a specific range or have a similar magnitude. Scaling is performed to ensure that the features do not dominate each other during model training, especially when using algorithms that rely on distances or gradients.

The main reasons for performing scaling are:

- Improved Model Performance: Scaling helps algorithms that use distance-based metrics (e.g., k-nearest neighbors, support vector machines, etc.) to work effectively by ensuring that all features contribute equally to the model.
- Faster Convergence: Scaling can speed up the convergence of gradient-based optimization algorithms, leading to faster model training.
- Avoiding Numerical Instabilities: Scaling prevents numerical instabilities that may arise due to large differences in feature magnitudes.
- There are two common methods of scaling: normalized scaling and standardized scaling.

Normalized Scaling (Min-Max Scaling):

Normalized scaling transforms the data to a specific range, typically between 0 and 1.

The formula to perform Min-Max Scaling is:
X_scaled = (X - X_min) / (X_max - X_min)
Where X_scaled is the scaled value, X is the original value, X_min is the minimum value of the feature, and X_max is the maximum value of the feature.
Standardized Scaling (Z-Score Scaling):

Standardized scaling, also known as Z-Score Scaling or standardization, scales the data to have a mean of 0 and a standard deviation of 1.
The formula to perform Standardized Scaling is:
X_standardized = (X - X_mean) / X_std

Where X_standardized is the standardized value, X is the original value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.
The main difference between normalized scaling and standardized scaling lies in the range of values after scaling. Normalized scaling bounds the values between 0 and 1, while standardized scaling centres the values around 0 with a standard deviation of 1. The choice between these methods depends on the requirements of the specific machine learning algorithm and the characteristics of the data.

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

Answer (Reference google.com)

In some cases, the Variance Inflation Factor (VIF) can become infinite. This occurs due to perfect multicollinearity in the data.

Perfect multicollinearity happens when one or more independent variables in a regression model can be perfectly predicted by a linear combination of other independent variables. In other words, there is a perfect linear relationship among the independent variables. This situation leads to an inability to estimate the regression coefficients accurately, resulting in the VIF becoming infinite.

Mathematically, the formula for calculating VIF for a particular variable is:

$VIF = 1 / (1 - R^2)$

where R^2 is the coefficient of determination of the linear regression model when the variable in question is regressed against all the other independent variables.

When perfect multicollinearity exists for a variable, the R^2 value becomes equal to 1, and the denominator in the VIF formula becomes 0, resulting in VIF approaching infinity.

Perfect multicollinearity is a severe issue in regression analysis because it can lead to unstable estimates and inflated standard errors of the regression coefficients. To handle perfect multicollinearity, one or more of the correlated variables must be removed from the model or combined with other variables to create new features. By eliminating the collinear variables, the regression model can be stabilized, and accurate estimates can be obtained.

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

Answer

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective but useful in the understanding underlying distribution of a variables.

A Q-Q plot can be used to assess the following:

- Whether the data set is normally distributed. If the data set is normally distributed, the points on the Q-Q plot will fall along a straight line.
- Whether the data set is from the same distribution as another data set. If two data sets are from the same distribution, their Q-Q plots will be similar.
- Whether the data set is affected by outliers. Outliers are data points that are far away from the rest of the data. Outliers can cause the Q-Q plot to deviate from a straight line.