Linear Regression Subjective Questions

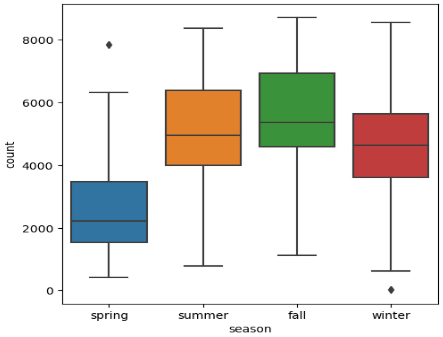
Assignment-based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

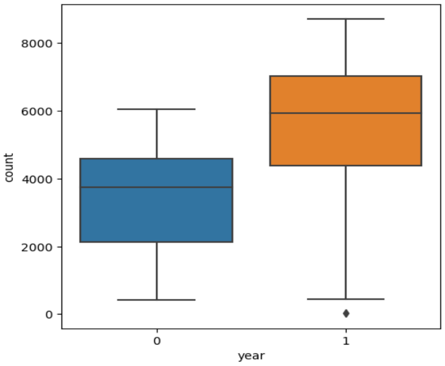
Answer:

Based on the analysis of the categorical variables, inference about their effect on the dependent variable:

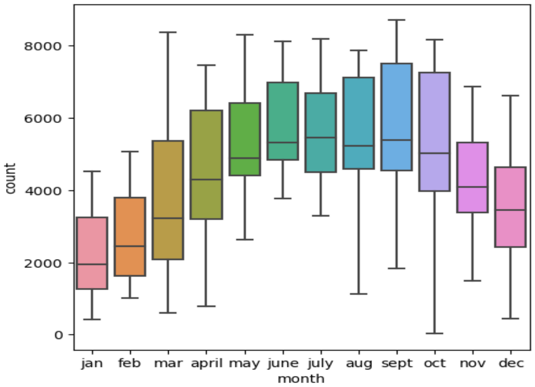
- **Season:** The plots indicate that bike rentals are higher during the fall season and summer months compared to other seasons. This suggests that seasonality has an impact on bike demand, with increased rentals during these periods.



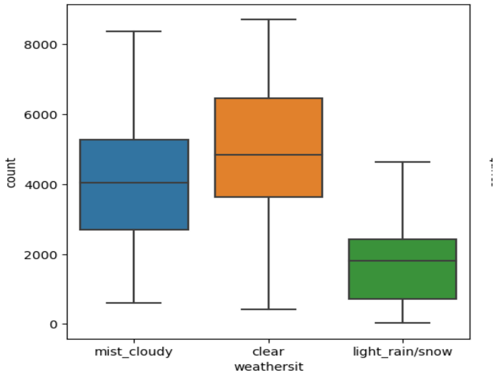
- **Year**: The plot shows higher bike rentals in 2019 compared to 2018, it suggests that bike rentals increased from 2018 to 2019. The year variable indicates a temporal trend, and it can be inferred that the demand for shared bikes increased over time.



- **Month**: The plots show that bike rentals are more in the month of September and October. This implies that there might be certain factors associated with these months that lead to higher bike demand, such as favorable weather conditions or specific events.



- **Weather**: The analysis suggests that bike rentals are more likely to occur when the weather is clear. This implies that weather conditions have an influence on bike demand, with better weather leading to increased rentals.



Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

Using drop_first=True during dummy variable creation is important for several reasons:

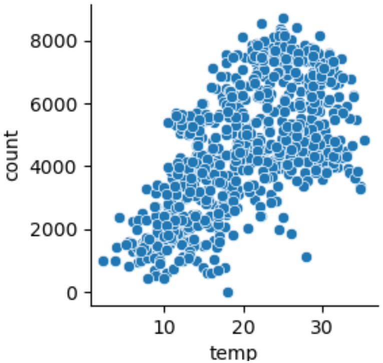
- **Avoiding_multicollinearity**: Including dummy variables for all categories of a categorical variable can lead to multicollinearity. It occurs when there is a high correlation between predictor variables, making it difficult for the model to estimate the true effect of each variable. By dropping the first category and excluding it from the model, we create a baseline reference category, avoiding perfect multicollinearity.

- **Enhancing_model_interpretability**: Dropping the first category allows us to interpret the coefficients of the remaining dummy variables more easily. The coefficients represent the change in the response variable relative to the baseline category. It provides a clear reference point for understanding the effect of each category.

- **Efficiency and simplicity**: Including dummy variables for all categories increases the dimensionality of the dataset. This can lead to computational inefficiencies, especially when dealing with large datasets. By dropping the first category, we reduce the number of variables and simplify the model without sacrificing information.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:



Based on the analysis of the pair-plot among the numerical variables, the variable with the highest correlation with the target variable (count) is temperature.

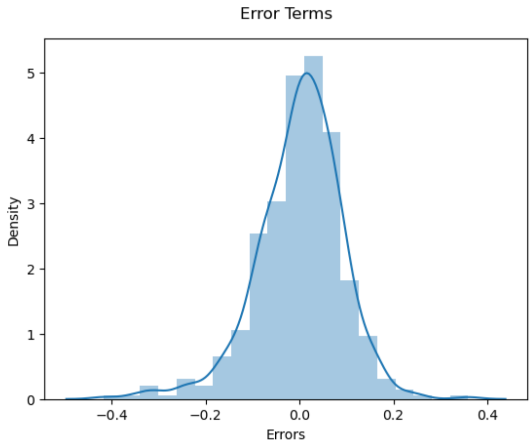
The positive correlation suggests that bike rentals tend to increase as the temperature rises. This indicates that people are more likely to rent bikes on warmer days.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

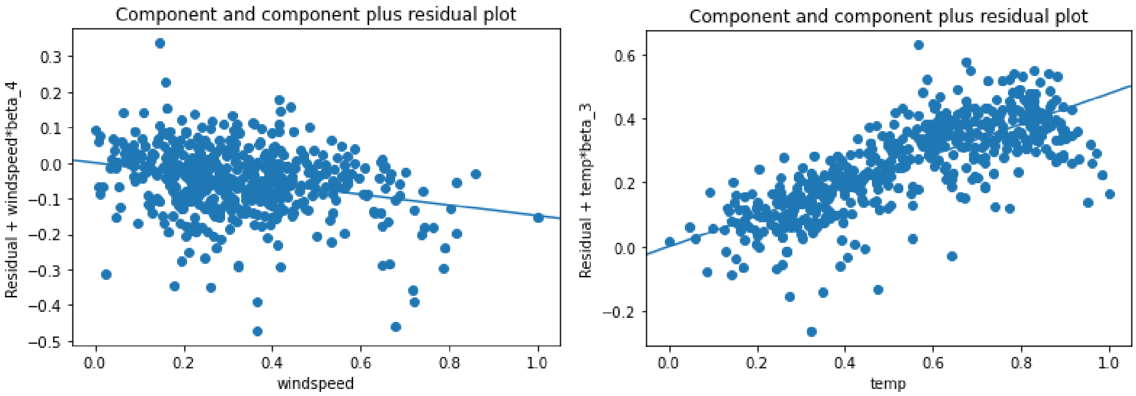
Answer:

Validating the assumptions of linear regression is an important step to ensure that the model is reliable and accurate.

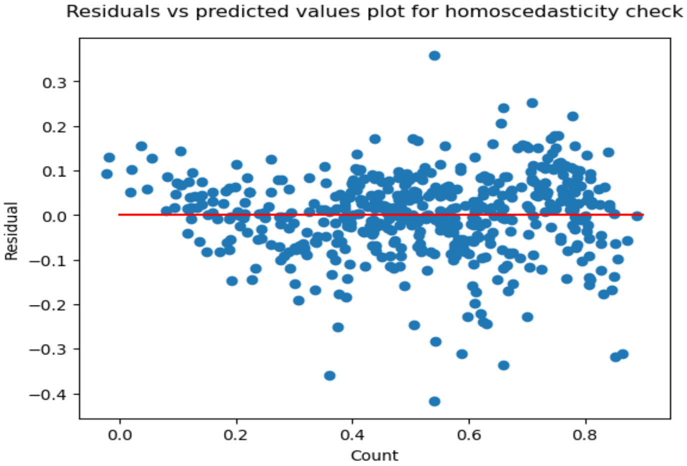
- **Normality of Residuals:** To validate this assumption, we plot a distplot of the residuals and visually inspect whether they follow a roughly normal distribution.



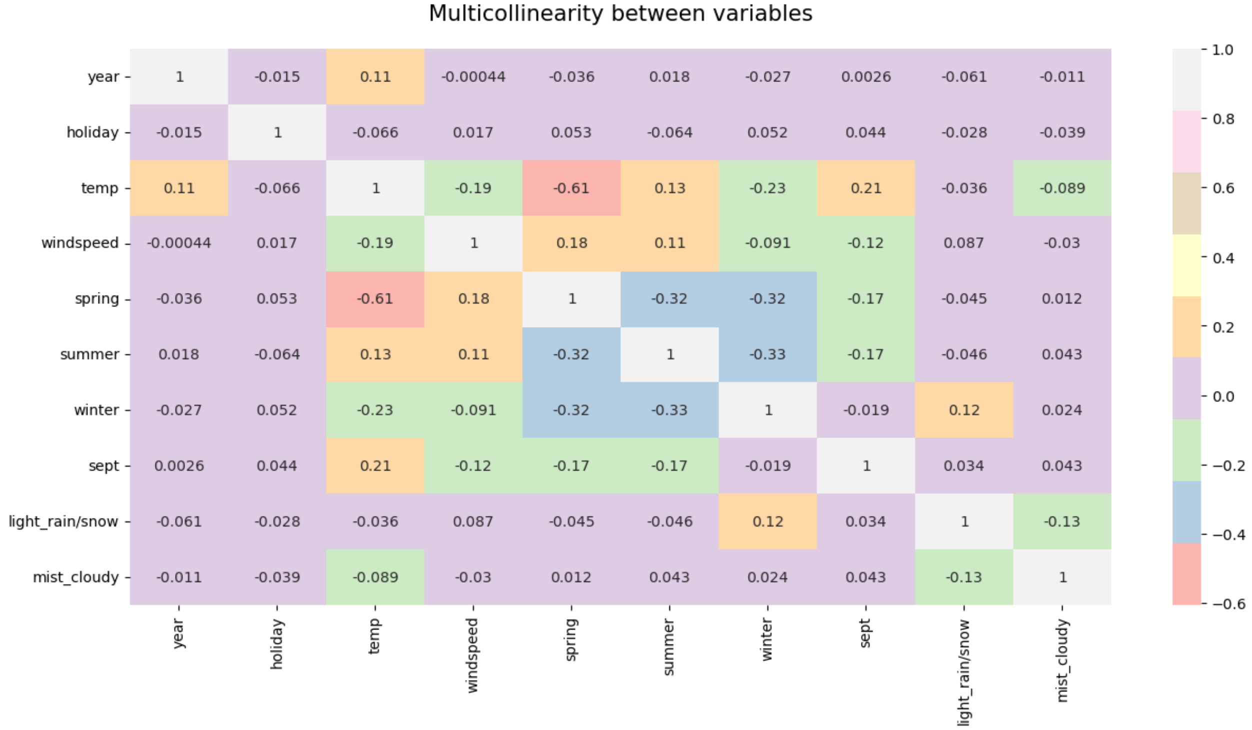
- **Linear Relationship:** To assess the linearity assumption, we plot a CCPR (Component-Component plus Residual) plot to visualize the relationship between an independent variable and the response variable, while controlling for the effects of other independent variables.



- **Homoscedasticity:** To check for homoscedasticity, we plot a scatter plot of the residuals against the predicted values. The spread of the residuals appears roughly constant across different ranges of the predicted values.

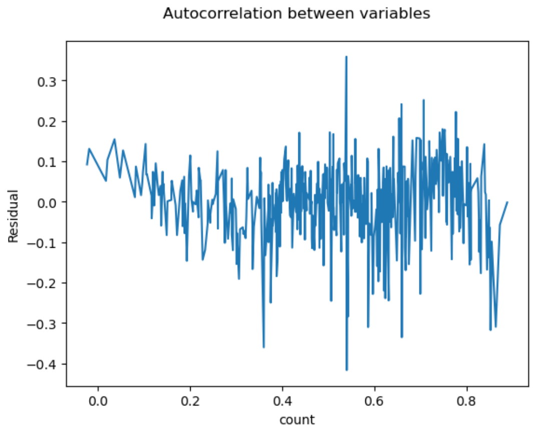


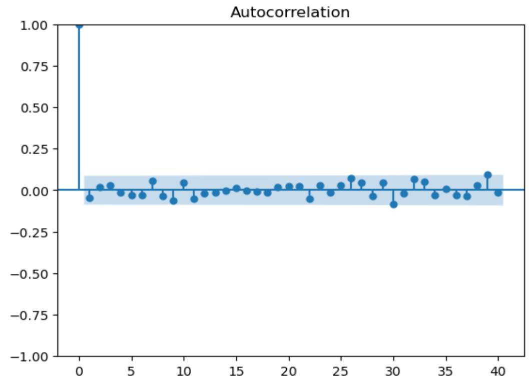
- **Multicollinearity:** The VIF (Variance Inflation Factor) values of our final model are less than 5, indicating no multicollinearity. Additionally, to cross check the same we plot a heatmap to assess presence of multicollinearity.



	year	holiday	temp	windspeed	spring	summer	winter	sept	light_rain/snow	mist_cloudy
year	1	-0.015	0.11	-0.00044	-0.036	0.018	-0.027	0.0026	-0.061	-0.011
holiday	-0.015	1	-0.066	0.017	0.053	-0.064	0.052	0.044	-0.028	-0.039
temp	0.11	-0.066	1	-0.19	-0.61	0.13	-0.23	0.21	-0.036	-0.089
windspeed	-0.00044	0.017	-0.19	1	0.18	0.11	-0.091	-0.12	0.087	-0.03
spring	-0.036	0.053	-0.61	0.18	1	-0.32	-0.32	-0.17	-0.045	0.012
summer	0.018	-0.064	0.13	0.11	-0.32	1	-0.33	-0.17	-0.046	0.043
winter	-0.027	0.052	-0.23	-0.091	-0.32	-0.33	1	-0.019	0.12	0.024
sept	0.0026	0.044	0.21	-0.12	-0.17	-0.17	-0.019	1	0.034	0.043
light_rain/snow	-0.061	-0.028	-0.036	0.087	-0.045	-0.046	0.12	0.034	1	-0.13
mist_cloudy	-0.011	-0.039	-0.089	-0.03	0.012	0.043	0.024	0.043	-0.13	1

- **No Autocorrelation:** To check autocorrelation we plot a line plot of the residuals against the predicted values and the Durbin-Watson statistic (value of 2.085) suggest no positive autocorrelation. Additionally, we also plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to further validate the absence of autocorrelation.





Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

Based on the coefficients (coef) and the p-values (P>|t|) of the final model, we can determine the top three features that contribute significantly towards explaining the demand of the shared bikes.

1. **Temp**: The coefficient for temperature is 0.4777, and it has a p-value of 0.000, indicating a highly significant relationship with the demand for shared bikes. An increase in temperature leads to an increase in bike demand.

2. **Windspeed**: The coefficient for windspeed is -0.1481, and it also has a p-value of 0.000, indicating a significant negative relationship with the demand for shared bikes. Higher windspeeds tend to decrease the demand for bikes.

3. **Light_rain/snow**: The coefficient for light_rain/snow is -0.2850, and it has a p-value of 0.000, indicating a significant negative relationship with the demand for shared bikes. Presence of light rain or snow reduces the demand for bikes.

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a popular and widely used algorithm in statistics and machine learning that aims to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable.

The algorithm starts by defining a linear equation that represents the relationship between the independent variables (also known as features or predictors) and the dependent variable (also known as the target or response variable).

The equation takes the form $y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn + \varepsilon$
where:

- y represents the dependent variable,
- x1, x2, ..., xn represent the independent variables,
- β0, β1, β2, ..., βn represent the coefficients or weights associated with each independent variable,
- ε represents the error term or residual, which accounts for unexplained variation in the dependent variable.

The goal of linear regression is to estimate the coefficients β0, β1, β2, ..., βn that best fit the observed data, minimizing the sum of squared errors between the predicted values (obtained from the linear equation) and the actual values of the dependent variable.

To estimate the coefficients, the algorithm typically employs the method of least squares. This method aims to find the coefficients that minimize the sum of squared differences between the observed dependent variable values and the predicted values from the linear equation.

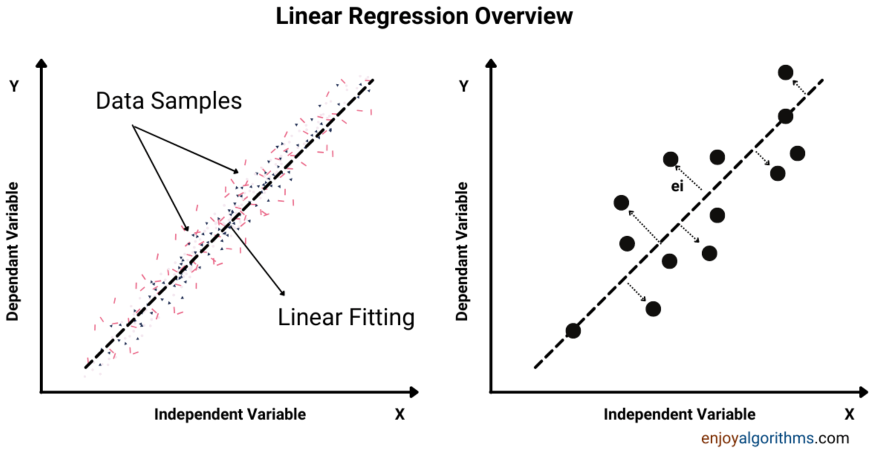
The algorithm uses various mathematical techniques to estimate the coefficients. One common approach is the ordinary least squares (OLS) method, which involves solving a set of equations to find the values of the coefficients that minimize the sum of squared errors.

Once the coefficients are estimated, they can be used to predict the values of the dependent variable for new or unseen data. By plugging in the values of the independent variables into the linear equation, the algorithm can generate predictions.

The performance of the linear regression model can be evaluated using various metrics, such as the mean squared error (MSE), mean absolute error (MAE), or coefficient of determination (R-squared). These metrics assess the accuracy and goodness of fit of the model.

Linear regression assumes several key assumptions, including normality of errors, linearity, multicollinearity, homoscedasticity (constant variance of errors), and autocorrelation. It is important to validate these assumptions to ensure the reliability and validity of the linear regression model.

Overall, linear regression is a versatile algorithm that provides a straightforward and interpretable way to model and understand the relationship between variables in many real-world problems.



Q2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe's quartet is a collection of four datasets that were introduced by the statistician Francis Anscombe in 1973. The quartet is widely cited and studied in statistics and data analysis to illustrate the importance of graphical exploration and visualization in understanding data.

The unique aspect of Anscombe's quartet is that all four datasets have nearly identical statistical properties, including the same mean values, regression coefficients, and coefficients of determination (R-squared).

Despite these similarities, when the datasets are graphically represented, they exhibit distinct patterns and relationships, highlighting the limitations of relying solely on numerical summaries.

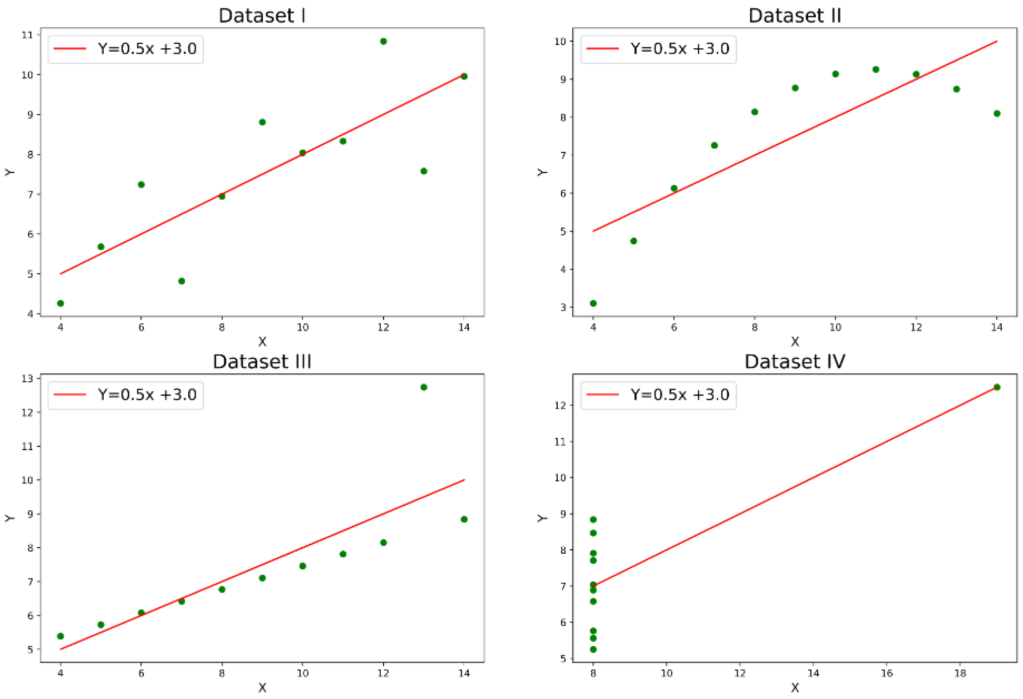
Datasets in Anscombe's quartet:

- **Dataset I**: It consists of a simple linear relationship between the x and y variables. When plotted, the data points closely follow a linear trend, suggesting a strong linear association.

- **Dataset II**: It also exhibits a linear relationship between x and y, but with one outlier. The presence of this outlier significantly impacts the regression line and correlation coefficient, demonstrating the influence of individual data points on statistical analysis.

- **Dataset III**: It challenges the assumption of linearity. It showcases a curvilinear relationship between x and y. While the overall relationship is not linear, it can be better approximated using a quadratic model, illustrating the need to consider non-linear relationships in data analysis.

- **Dataset IV**: It appears to have no apparent relationship between x and y when viewed as a whole. However, upon closer inspection, it reveals an interesting relationship when divided into two subsets. In each subset, there is a strong linear relationship between x and y. This dataset highlights the importance of investigating data in different segments or groups to uncover underlying patterns.



The purpose of Anscombe's quartet is to emphasize that numerical summaries alone, such as means, variances, and correlation coefficients, are insufficient for understanding data. Visualizations, on the other hand, provide valuable insights into the structure and relationships within the data.

Q3. What is Pearson's R?

Answer:

Pearson's R, or Pearson correlation coefficient, is a statistical measure that assesses the strength and direction of the linear relationship between two continuous variables. It quantifies how closely the data points in a scatterplot adhere to a straight line.

To calculate Pearson's R, you need a set of paired observations for the two variables of interest. Let's consider an example to illustrate this:

Suppose we have data on the hours of study and the corresponding exam scores for a group of students. Here's a hypothetical dataset:

Hours of Study (X): [5, 10, 8, 3, 7]
Exam Scores (Y): [65, 85, 75, 50, 80]

To calculate Pearson's R, follow these steps:

1. Calculate the means of X and Y. In this case, the mean of X is 6.6 and the mean of Y is 71.

2. Calculate the deviations from the means for each pair of observations. These are the differences between each value and its respective mean. For example, the deviations for the first pair are (5-6.6) and (65-71).

   Deviations for X: [-1.6, 3.4, 1.4, -3.6, 0.4]
   Deviations for Y: [-6, 14, 4, -21, 9]

3. Multiply the deviations for X and Y for each pair and calculate their sum. For example, the product of the first pair's deviations is (-1.6 * -6).

   Sum of (X deviation * Y deviation): 273.8

4. Square each deviation for X and Y, and calculate their sum.

   Sum of squared X deviations: 28.8
   Sum of squared Y deviations: 940

5. Calculate the square root of the product of the sums of squared deviations for X and Y. In this case, the square root is approximately 30.67.

Square root of (Sum of squared X deviations * Sum of squared Y deviations): 30.67

6. Divide the sum of the (X deviation * Y deviation) by the square root calculated in step 5.

Pearson's R = Sum of (X deviation * Y deviation) / Square root of (Sum of squared X deviations * Sum of squared Y deviations)
= 273.8 / 30.67
≈ 8.92

Therefore, the Pearson correlation coefficient (R) for this dataset is approximately 8.92. Since R is positive, we can infer a positive linear relationship between hours of study and exam scores.

Pearson's R only measures the linear relationship between variables and assumes that the relationship is roughly linear. It may not capture nonlinear or complex relationships. Additionally, correlation does not imply causation; it merely indicates the strength and direction of association between the variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling, in the context of data analysis and machine learning, refers to the process of transforming variables to a common scale. It involves adjusting the range or distribution of the data to make it more suitable for analysis or modeling.

Scaling is performed for several reasons:

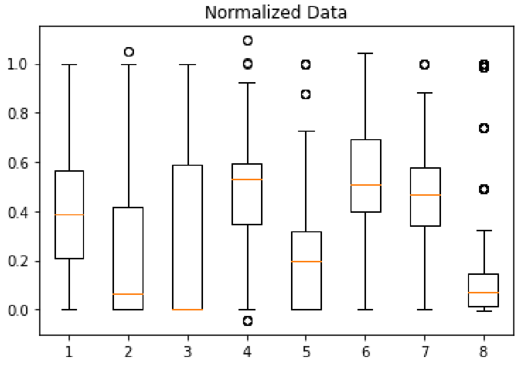
- **To ensure fairness and equal treatment**: When variables have different scales or units, it can lead to biased results. Scaling helps eliminate such biases, allowing for fair comparisons between variables.

- **To improve convergence and performance of algorithms**: Many machine learning algorithms, such as gradient descent-based optimization, perform better when variables are on a similar scale. Scaling facilitates faster convergence and prevents certain variables from dominating the learning process due to their larger values.

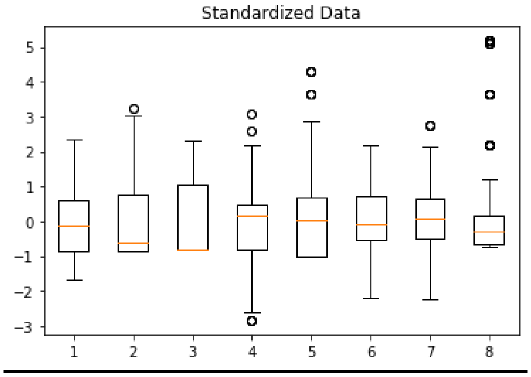
- **To enhance interpretability**: Scaling makes it easier to interpret the effect of different variables. When variables are on a similar scale, the coefficients or weights associated with them can be compared more directly.

There are two common methods of scaling:

- **Normalized scaling (or feature scaling)**: It involves transforming the values of a variable to a common range, usually between 0 and 1. This is achieved by subtracting the minimum value from each observation and dividing by the range (maximum value minus minimum value). Normalized scaling preserves the relative relationships between the values of a variable.



- **Standardized scaling (or z-score normalization)**: It involves transforming the values of a variable to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean from each observation and dividing by the standard deviation. Standardized scaling not only adjusts the range but also centers the data around the mean.



The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the original distribution and range of the data, while standardized scaling standardizes the data to have a mean of 0 and a standard deviation of 1. The choice between these methods depends on the specific requirements of the analysis or modeling task at hand.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

When the Variance Inflation Factor (VIF) is infinite, it typically 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 using a linear combination of the other independent variables.

The VIF is a measure used to assess multicollinearity between independent variables in a regression model. It quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity. VIF values are calculated for each independent variable, and a high VIF suggests a high degree of multicollinearity.

If the VIF for a particular variable is infinite, it means that this variable is a perfect linear combination of the other independent variables in the model. In other words, it can be precisely predicted using a linear equation involving the other variables. As a result, the estimated regression coefficients become indeterminate or infinite, leading to an infinite VIF.

Perfect multicollinearity can arise for various reasons, such as data errors, redundant variables, or incorrect model specification. It is important to identify and address multicollinearity issues as it can affect the interpretation and stability of the regression model.

Let's consider a regression model with two independent variables, X1 and X2, and a dependent variable Y. The dataset is as follows:

X1: [1, 2, 3, 4, 5]
X2: [2, 4, 6, 8, 10]
Y: [3, 6, 9, 12, 15]

To calculate the VIF, you perform separate regressions for each independent variable, using the other variable(s) as predictors. However, in this case, we can see that X2 is a perfect linear combination of X1: X2 = 2 * X1. This means that X2 can be perfectly predicted using X1.

When you attempt to calculate the VIF for X2, it involves regressing X2 against X1. However, since X2 can be perfectly predicted by X1, the regression results become indeterminate. The estimated coefficients become infinite, resulting in an infinite VIF for X2.

This scenario occurs because the presence of perfect multicollinearity violates the assumptions of the regression model. It creates an unstable and ill-posed regression problem, making it impossible to estimate reliable coefficients and resulting in an infinite VIF.

Q6. What is Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Answer:

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the similarity between the quantiles of a dataset and the quantiles of a theoretical distribution. It helps determine if the data follows a particular distribution or if it deviates significantly from it.

In the context of linear regression, a Q-Q plot is often used to evaluate the assumption of normality for the residuals. Residuals are the differences between the observed values and the predicted values from a regression model. The Q-Q plot provides a visual comparison between the observed residuals and the expected residuals assuming a normal distribution.

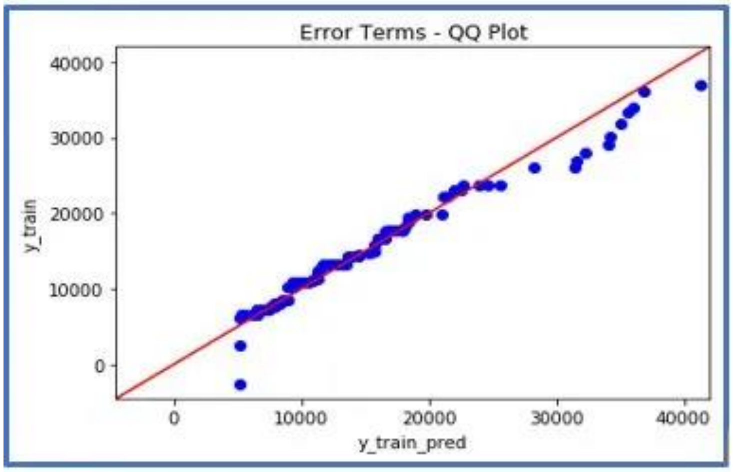
How a Q-Q plot is constructed and its importance in linear regression:

1. **Residuals are calculated**: After fitting a linear regression model, the residuals are obtained by subtracting the predicted values from the observed values.

2. **Quantiles are computed**: The quantiles of the residuals are determined, indicating the relative standing of each residual within the distribution.

3. **Theoretical quantiles are obtained**: Assuming the residuals follow a normal distribution, the expected quantiles are calculated based on the mean and standard deviation of the residuals.

4. **Plotting the Q-Q plot**: The observed quantiles (residuals) are plotted against the expected quantiles (theoretical quantiles). If the residuals closely align with the expected quantiles, it suggests that the residuals follow a normal distribution.



The use and importance of a Q-Q plot in linear regression are as follows:

- **Assessing normality assumption**: Linear regression assumes that the residuals are normally distributed. Departure from normality can impact the validity of statistical tests, confidence intervals, and prediction intervals. The Q-Q plot allows you to visually inspect if the residuals deviate from normality. If the points on the plot deviate significantly from the expected line, it suggests non-normality.

- **Identifying outliers**: Outliers, which are extreme values in the data, can distort the regression analysis. A Q-Q plot can help identify outliers by showing data points that deviate significantly from the expected quantiles. Outliers may appear as points that deviate from the straight line on the Q-Q plot.

- **Model refinement:** If the Q-Q plot reveals deviations from normality, it indicates potential issues with the model assumptions. This insight can guide model refinement strategies, such as transforming variables or considering alternative regression techniques.