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? (3 marks)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

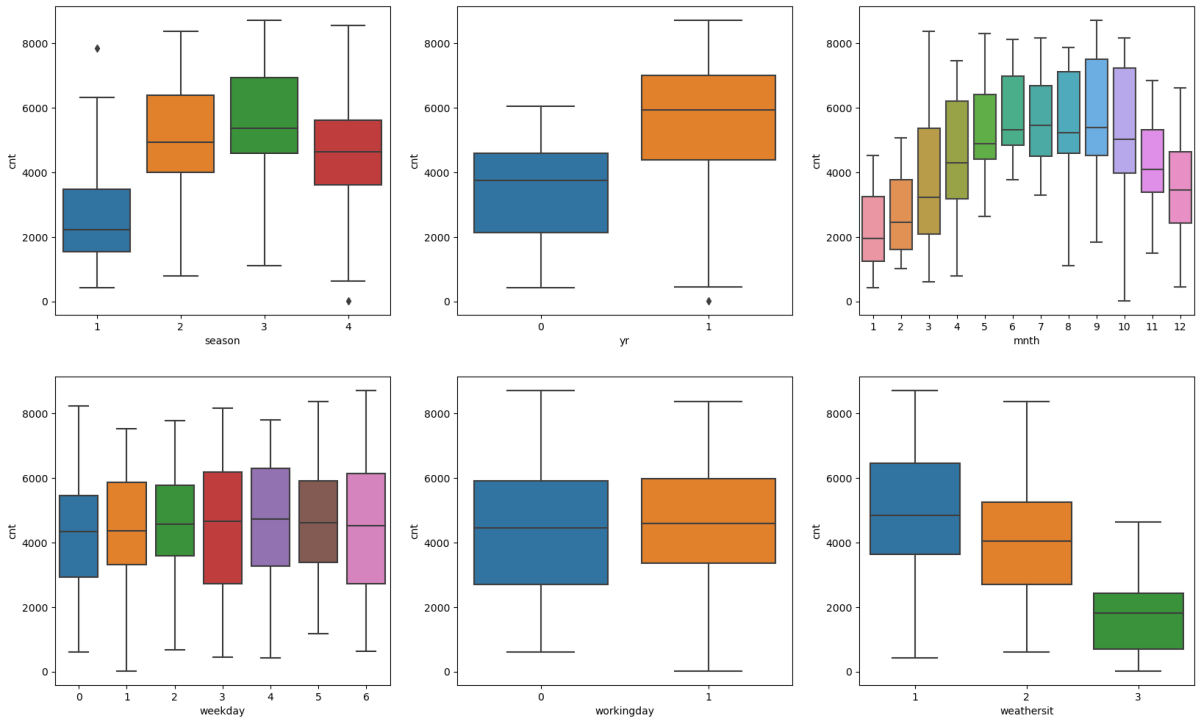
General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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?

The dataset has some of the variables like 'weathersit' and 'season' etc. that have values as 1, 2, 3, 4 which have specific labels associated with them. Despite them having numerical variables, they should be considered as categorical variables. Following image indicates the association of individual categorical variables with target variable (cnt)



- **season** – Season 1 (spring) has relatively less rentals. Other 3 seasons have similar trend. Fall (Season 3) season has the highest demand.
- **yr** - Bike-sharing systems are gaining popularity Y-o-Y. It indicates that demand will only increase with supply.
- **mnth** – Rentals a less in months of Dec, Jan and Feb due to (probably) environmental conditions.
- **weekday** – Rentals are mostly uniform over weekdays.
- **workingday** – Workingday factor has slight impact on rentals (0.092)
- **weathersit** – No rentals on days with value 4 (extreme weather conditions) and relatively few rentals on days with value 3 (light rain etc.). Overall good weather (1 & 2) increases chances of bike rentals.

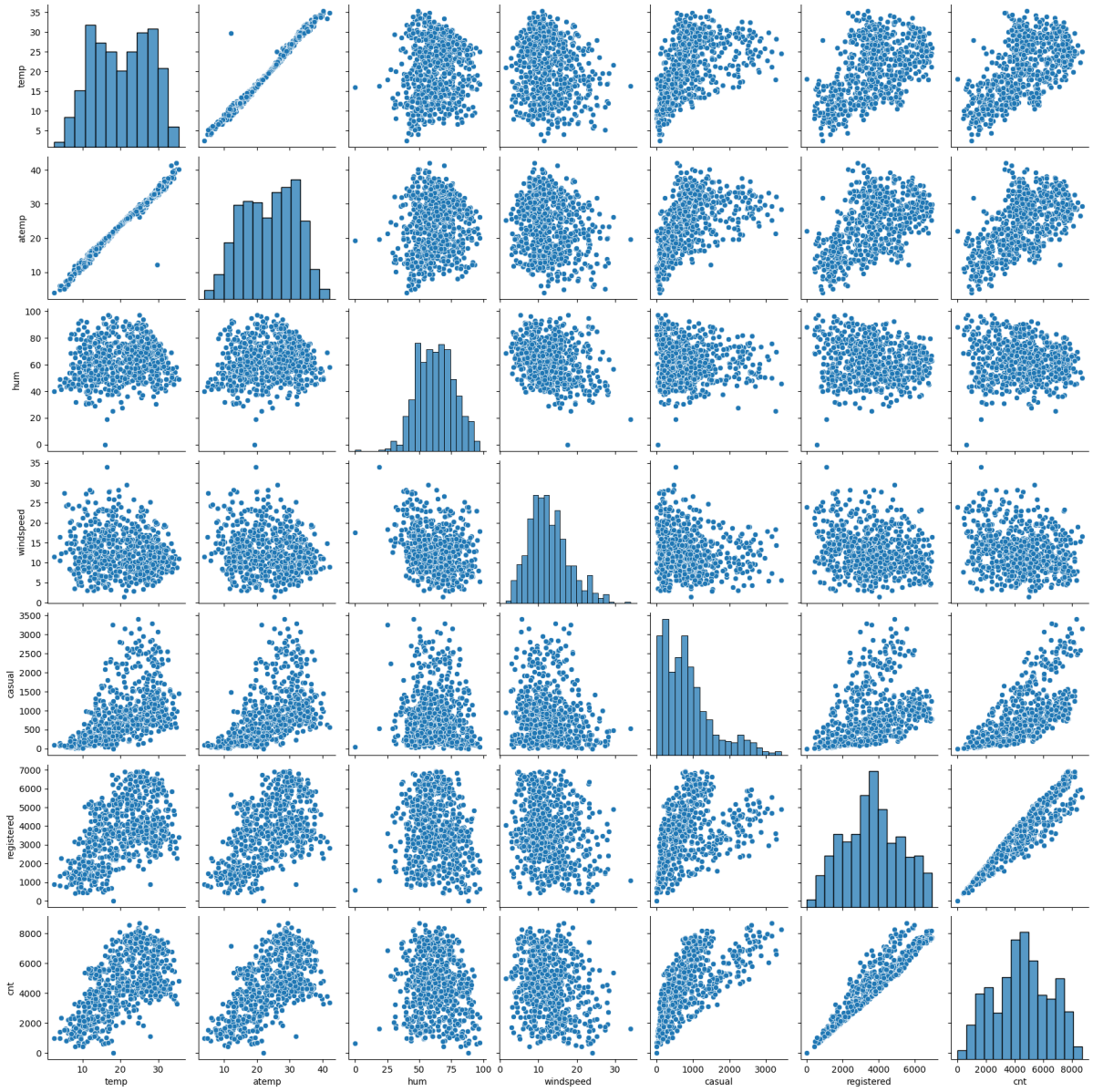
Q2: Why is it important to use drop_first=True during dummy variable creation?

Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity, which is when independent variables are highly correlated with each other. This can cause problems in the interpretation of the coefficients in regression analysis because it becomes difficult to isolate the individual effect of each variable.

By dropping the first category, you create a reference category against which the other categories are compared, thus reducing multicollinearity and ensuring a more stable and interpretable model.

If a categorical variable has n levels of categorization, it can be analyzed with n-1 levels. drop_first = True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummyvariables.

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



Order of highest correlation (top-3) with the target variable among numeric variables

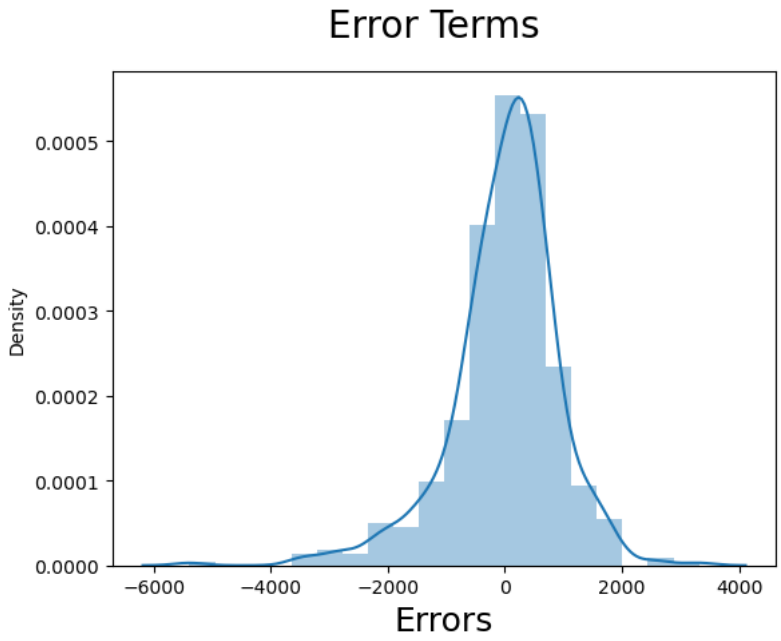
1. registered
2. casual
3. temp ~= atemp

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

To validate the assumptions of Linear Regression after building the model on the training set, several diagnostic checks and tests are typically performed:

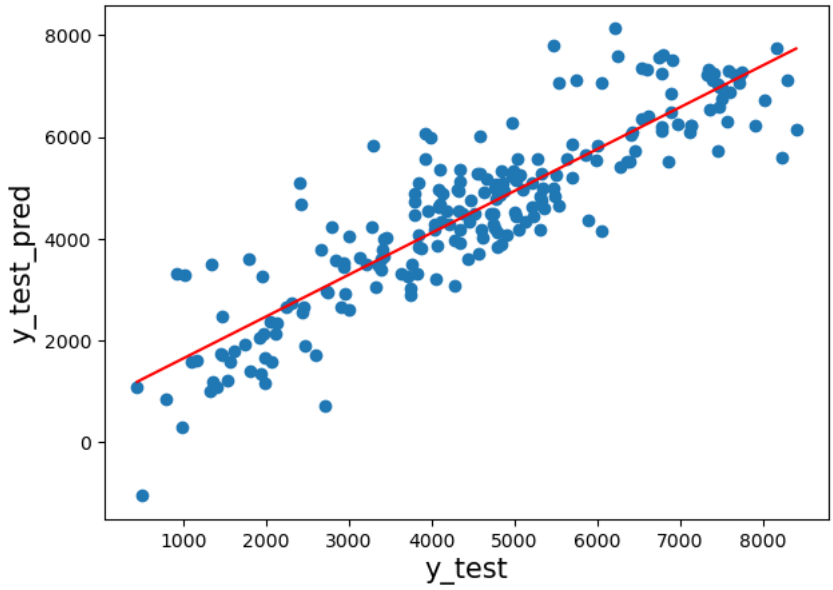
Residual Analysis

Checked if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), by plotting the histogram of the error terms.



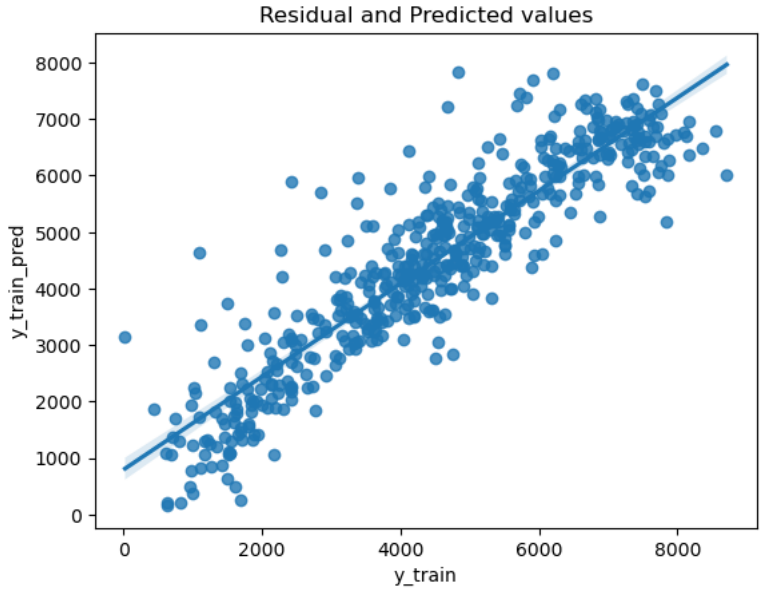
Linearity

Check for the linearity assumption by plotting the predicted values against the actual values. The points should ideally form a straight line. Additionally, scatter plots of residuals versus fitted values can help detect any patterns; the residuals should be randomly scattered.



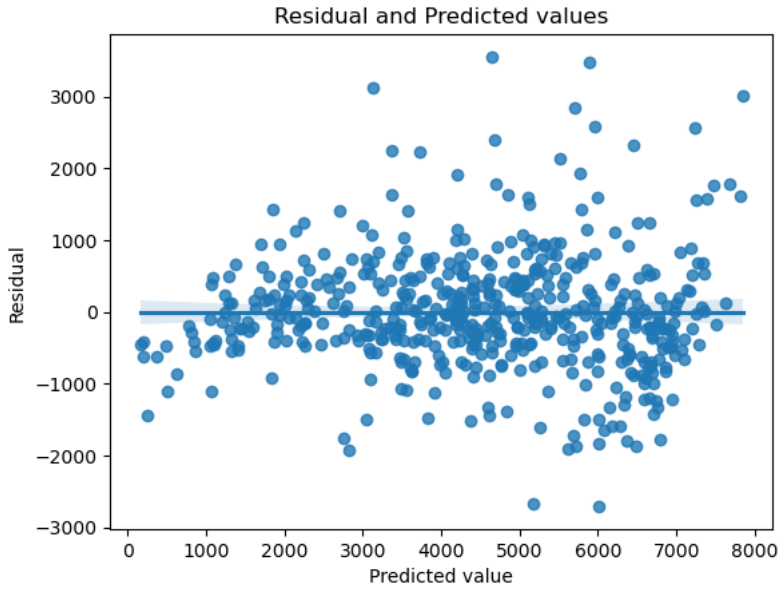
Homoscedasticity

Homoscedasticity means that the variance of the errors is constant across all levels of the independent variables. This can be validated by plotting the residuals against the predicted values. The residuals should display a random pattern without a funnel shape.



Independence

The independence of errors can be checked using below diagram.



Multicollinearity

Calculate the Variance Inflation Factor (VIF) for each predictor variable. None is > 5.

	Features	VIF
3	temp	4.35
4	windspeed	3.38
1	weekday	2.87
2	workingday	2.84
0	yr	2.01
5	weathersit_2	1.52
7	season_2	1.50
8	season_4	1.39
6	weathersit_3	1.08

Outliers and Leverage Points

Identify any potential outliers or leverage points that could disproportionately influence the model. This can be done using influence plots, Cook's distance, or leverage statistics. Removing or investigating these points can enhance model accuracy and reliability.

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

- 'temp'/'atemp' (0.64)
- 'yr' (0.59)
- 'season_2' (0.13)

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The key goal is to find the linear equation that best predicts the target variable based on the predictor variables. The algorithm can be broken down into the following steps:

1. **Types of Linear Regression**
   Simple Linear Regression
   Simple linear regression involves a single independent variable. The relationship between \(y \) and \(x \) is modeled as:

$$y = \beta_0 + (\beta_1 \times x) + \epsilon$$

   Multiple Linear Regression
   Multiple linear regression involves two or more independent variables. The relationship is modeled as:

$$y = \sum_{i=1}^{n} \beta_0 + (\beta_i \times x_i) + \epsilon$$

   Here, $\beta_0$is the intercept, β_i are the coefficients, and ϵ represents the error term.

2. **Estimation of Coefficients:**
   The goal of linear regression is to find the best-fitting line through the data points. This line is determined by estimating the regression. The most common method for estimating these coefficients is the Ordinary Least Squares (OLS) method.
   The coefficients (β_i) are estimated using the method of least squares. This can be expressed as:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

   $\hat{\beta}$ = ordinary least squares estimator

   X = matrix regressor variable X

   $\top$ = matrix transpose

   y = vector of the value of the response variable

3. **Assumptions**:
   Linear regression relies on several key assumptions:
   Linearity: The relationship between the predictors and the target is linear.

Independence: The residuals (errors) are independent.
Homoscedasticity: The residuals have constant variance.
Normality: The residuals are normally distributed.

4. **Fitting the Model:**
   Using the training data, the algorithm computes the best-fit line that minimizes the residual sum of squares.
5. **Making Predictions:**
   Once the model is trained, predictions for new data points can be made using the learned coefficients
6. **Model Evaluation:**
   Once the model is fitted, it is crucial to evaluate its performance. Common evaluation metrics include:

   R-Squared (r^2)
   The $\boldsymbol{r^2}$ statistic measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.

   Mean Squared Error (MSE)
   The MSE measures the average squared difference between the observed and predicted values. Lower values indicate a better fit.

   Root Mean Squared Error (RMSE)
   The RMSE is the square root of the MSE and provides a measure of the average magnitude of the prediction errors.

   Mean Absolute Error (MAE)
   The MAE measures the average absolute difference between the observed and predicted values. Like RMSE, lower values indicate a better fit.
7. **Interpretation:**
   The coefficients (β) indicate the change in the target variable for a one-unit change in the corresponding predictor variable, holding all other predictors constant. The intercept (β_0) represents the expected value of the target when all predictors are zero.

Q2: Explain the Anscombe's quartet in detail

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. This quartet was created to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if the data is not visualized.

The Four Datasets

Each of the four datasets in Anscombe's quartet consists of eleven (x, y) points. Despite having nearly identical statistical properties—such as mean, variance, correlation, and linear regression lines—the datasets have very different distributions and appearances when plotted. Below are the statistical properties that are identical (or nearly so) across the four datasets:

- Mean of x: 9
- Mean of y: 7.50
- Variance of x: 11
- Variance of y: 4.125
- Correlation between x and y: 0.816
- Linear regression line: \(y = 3.00 + 0.500x \)

Despite these similarities, visual inspection reveals substantial differences in the data's distribution. Below are the datasets:

Dataset 1

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Dataset 2

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

Dataset 3

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

Dataset 4

x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

Visualizing Anscombe's Quartet

To fully appreciate the differences between these datasets, it is essential to visualize them:

Dataset 1

Dataset 1 forms a roughly linear pattern with some variance around the line. This dataset fits well with a linear regression model, which is why its statistical properties align closely with the regression line.

Dataset 2

Dataset 2 presents a non-linear relationship. The data points form a curve, and although the linear regression line and statistical properties are the same, the fit is not appropriate for a linear model.

Dataset 3

Dataset 3 contains an outlier which significantly influences the regression line. Without this single influential point, the other data points would suggest a different relationship.

Dataset 4

Dataset 4 is unique in that it has most of its x-values equal, except for one outlier. This outlier completely dictates the slope of the regression line, even though the majority of the data does not support this trend.

The Lessons of Anscombe's Quartet

Importance of Data Visualization

Anscombe's quartet highlights the critical importance of visualizing data before drawing conclusions based on statistical measures. It demonstrates that datasets with identical statistical properties can have very different distributions and relationships when plotted. Therefore, visual tools such as scatter plots should always accompany statistical analysis to capture the true nature of the data.

Understanding Outliers

Each dataset in the quartet also shows how outliers can significantly impact statistical properties and regression models. Outliers can distort the true relationship between variables and lead to misleading conclusions if not identified and handled appropriately.

Limits of Summary Statistics

The quartet underscores the limitations of relying solely on summary statistics like mean, variance, and correlation. While these measures provide valuable information, they do not capture the full story. Detailed analysis and visualization are necessary to understand the underlying patterns and relationships in the data.

Conclusion

Anscombe's quartet serves as a powerful reminder of the importance of graphical analysis in statistics. It emphasizes that while summary statistics are useful, they are not sufficient to fully understand data. Visualization plays a crucial role in identifying patterns, relationships, and anomalies that may not be apparent from numerical measures alone. By combining both statistical analysis and visualization, one can achieve a more comprehensive and accurate interpretation of data.

Anscombe's work remains highly relevant in the field of data science and analytics, reinforcing the need for a balanced approach that leverages both quantitative and visual tools to uncover the true insights hidden within datasets.

Q3: What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that evaluates the linear relationship between two continuous variables. This coefficient quantifies the degree to which a straight line can describe the relationship between the variables. Pearson's R is widely used in fields such as statistics, data science, economics, and social sciences due to its simplicity and interpretability.

Calculation

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula is given by:

$$r = \Sigma[(X - \mu X)(Y - \mu Y)] / \sqrt{[\Sigma(X - \mu X)^2 \Sigma(Y - \mu Y)^2]}$$

Where:

- X and Y are the two variables being compared.
- μX and μY are the means of the variables X and Y, respectively.
- Σ denotes the sum over all data points.

Pearson's R ranges from -1 to +1, where:

- +1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

The closer the coefficient is to either extreme, the stronger the linear relationship between the variables. A positive value signifies that as one variable increases, the other tends to increase as well, while a negative value indicates that as one variable increases, the other tends to decrease.

Assumptions

For Pearson's R to be a valid measure, certain assumptions must be met:

- Linearity: The relationship between the variables should be linear.
- Homoscedasticity: The variance of the variables should be roughly constant across levels of the independent variable.
- Normality: The variables should be approximately normally distributed.

Applications

Pearson's R is used in various applications, including:

- Research: To identify the strength and direction of relationships between variables.
- Finance: To assess the correlation between asset returns.
- Social Sciences: To evaluate relationships between behavioral variables.

Limitations

While Pearson's R is a powerful tool, it has limitations:

- Outliers: The presence of outliers can significantly affect the value of the coefficient.

- Non-linearity: It only measures linear relationships; non-linear relationships require different methods.
- Confounding Variables: The correlation does not imply causation and can be influenced by external variables.

Conclusion

Pearson's R is a fundamental statistical tool that provides valuable insights into the linear relationships between variables. Despite its limitations, it remains a cornerstone in the analytical toolbox, offering a straightforward method to quantify correlations. To fully understand data, Pearson's R should be used in conjunction with other statistical and visualization techniques, ensuring a comprehensive analysis.

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

What is Scaling?

Scaling is a crucial preprocessing step in data analysis that involves adjusting the range of data values. The primary goal is to ensure that features contribute equally to analysis and model training, especially when dealing with algorithms sensitive to the magnitude of data values. Scaling helps to normalize the data, making it easier to compare and analyze.

Why is Scaling Performed?

Scaling is performed for several reasons:

- Improving Model Performance: Many machine learning algorithms, such as gradient descent-based methods, perform better and converge faster when features are on a similar scale.
- Enhanced Interpretability: When data is scaled, it becomes easier to visualize and interpret relationships between variables.
- Equal Contribution: Features with larger ranges can dominate the analysis, overshadowing features with smaller ranges. Scaling ensures that all features contribute equally.
- Dimensional Consistency: In multivariate analysis, having features on the same scale ensures consistency and comparability across dimensions.

Normalized Scaling vs. Standardized Scaling

There are different methods to scale data, and two commonly used techniques are normalized scaling and standardized scaling.

Normalized Scaling

Normalized scaling, also known as Min-Max Scaling, involves rescaling the data to fit within a specific range, typically [0, 1]. The formula for Min-Max Scaling is:

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

Where:
- X’ is the normalized value.
- X is the original value.
- Xmin and Xmax are the minimum and maximum values of the feature, respectively.

Advantages:

- Preserves the relationships between values.
- Suitable for algorithms that require data within a specific range, such as neural networks and certain distance-based methods.

Disadvantages:

- Sensitive to outliers, as they can skew the range.

Standardized Scaling

Standardized scaling, or z-score normalization, involves transforming the data to have a mean of zero and a standard deviation of one. The formula for standardized scaling is:

$$z = \frac{x - \mu}{\sigma}$$

Where:
- z is the standardized value.
- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Advantages:

- Reduces the impact of outliers since it centers the data around the mean.
- Commonly used in algorithms that assume normally distributed data, such as linear regression and principal component analysis (PCA).

Disadvantages:

- May not work well if the data is not normally distributed.

Conclusion

Scaling is a fundamental process in data analysis that ensures features contribute equally to model training and analysis. Both normalized scaling and standardized scaling have their unique advantages and are suitable for different types of data and algorithms. Understanding the differences between these methods allows analysts to choose the most appropriate scaling technique for their specific needs, ultimately leading to more accurate and reliable results.

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

One important aspect to consider in data analysis is the Variance Inflation Factor (VIF), which measures the extent of multicollinearity in regression analysis.

Sometimes, analysts encounter scenarios where the value of VIF becomes infinite.

This typically occurs when there is perfect multicollinearity among the predictors in the model. In other words, one predictor variable can be expressed as an exact linear combination of other predictors. This perfect linear relationship causes the denominator in the VIF calculation, which involves the determination coefficient r^2 to be zero, thereby making VIF approach infinity.

To address this issue, analysts often need to reassess the model and ensure that predictors are not excessively correlated. Techniques such as removing redundant predictors, using principal component analysis (PCA), or applying regularization methods can help mitigate the impact of multicollinearity and avoid infinite VIF values.

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a set of data follows a particular distribution. It is particularly useful in linear regression analysis for checking the assumption of normality, which is crucial for valid hypothesis testing and accurate confidence intervals.

What is a Q-Q Plot?

A Q-Q plot compares the quantiles of the data sample to the quantiles of a theoretical distribution, such as the normal distribution. If the data follows the theoretical distribution, the points on the plot will lie approximately along a straight line. Deviations from this line indicate departures from the specified distribution.

Constructing a Q-Q Plot

To create a Q-Q plot:

1. **Sort** the data in ascending order.
2. **Calculate** the theoretical quantiles from the specified distribution.
3. **Plot** the sample quantiles against the theoretical quantiles.

The resulting plot provides a visual assessment of how well the data conforms to the theoretical distribution.

Use of Q-Q Plots in Linear Regression

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. Violations of this assumption can lead to unreliable estimates and hypothesis tests.

Checking Normality of Residuals

A Q-Q plot helps verify the normality of residuals:

- Normal Residuals: Points lie on or close to the straight line.
- Non-Normal Residuals: Points deviate significantly from the line, indicating skewness or kurtosis.

Identifying Outliers

Q-Q plots can also identify outliers and heavy-tailed distributions. Outliers appear as points that deviate markedly from the line. Recognizing these can help refine the model and improve its accuracy.

Importance of Q-Q Plots

Q-Q plots are essential for:

- Diagnostic Checking: Ensuring the residuals meet the normality assumption.
- Model Validation: Confirming the appropriateness of the linear regression model.
- Insights into Data: Providing insights into data distribution and potential issues.

By visually assessing the distribution of residuals, analysts can make informed decisions about the suitability of their model and take corrective actions if necessary.

In summary, Q-Q plots are a powerful diagnostic tool in linear regression analysis, helping to validate assumptions and identify outliers.