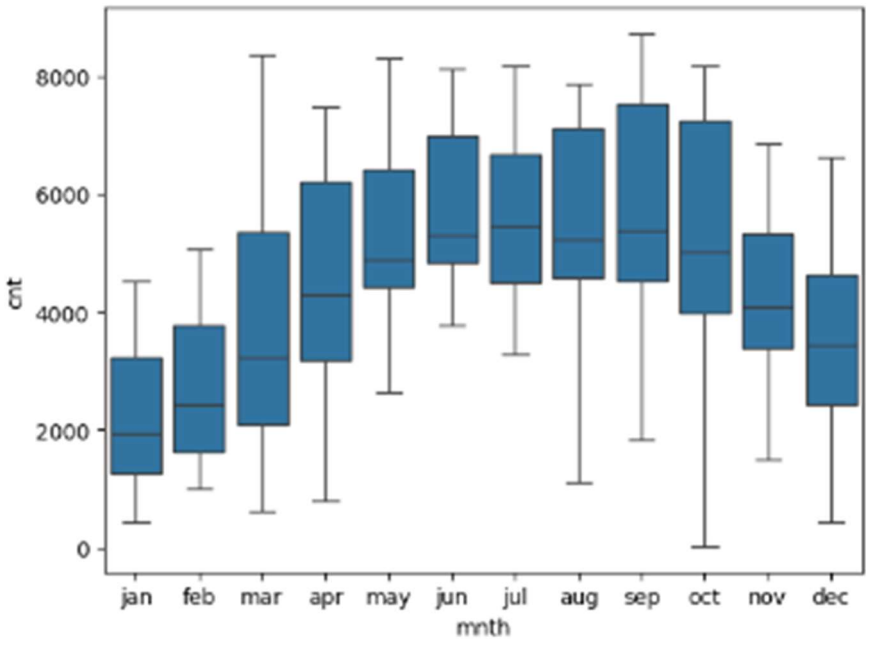
Assignment-based Subjective Questions

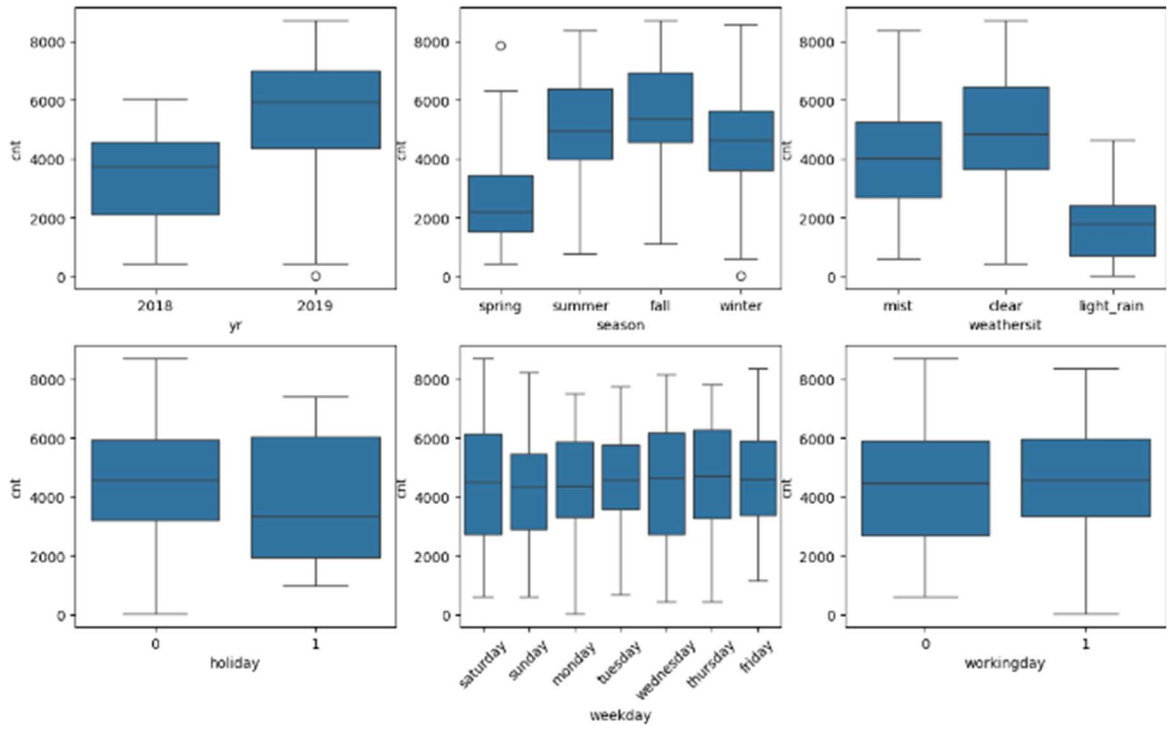
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variable	Impact on dependent variable
dteday	'dteday' variable represents a time trend that is better captured be a separate variable – 'yr'
yr	There is noticeable increase in demand of shared bikes in year 2019 as compared against 2018. It must be owed to the rapidly growing business year over year. Average daily demand in 2019 is roughly 50% higher than in 2018
mnth	Demand tends to increase from Jan to July and then again demand declines until Dec.
weekday	Average demand appears to be mostly similar across all weekdays, having slightly different variance across days, especially on Saturday and Wednesday where variance is highest.
holiday	Bike usage is significantly lower on holidays. It must be due to reduction in usage by office going people on holidays
workingday	Average does not seem to be directly affected much whether it is a working day or not. But variance in daily demand is significantly lower during working days. However working day is determined from two components – whether it is a weekday or holiday
season	Bike usage is significantly higher during fall, followed by summer. It is lowest during spring season
weathersit	Demand for shared bikes is lowest during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds




cnt
yr
2018
2019
season
spring
summer
fall
winter
weathersit
mist
clear
light_rain
holiday
0
1
weekday
saturday
sunday
monday
tuesday
wednesday
thursday
friday
workingday

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

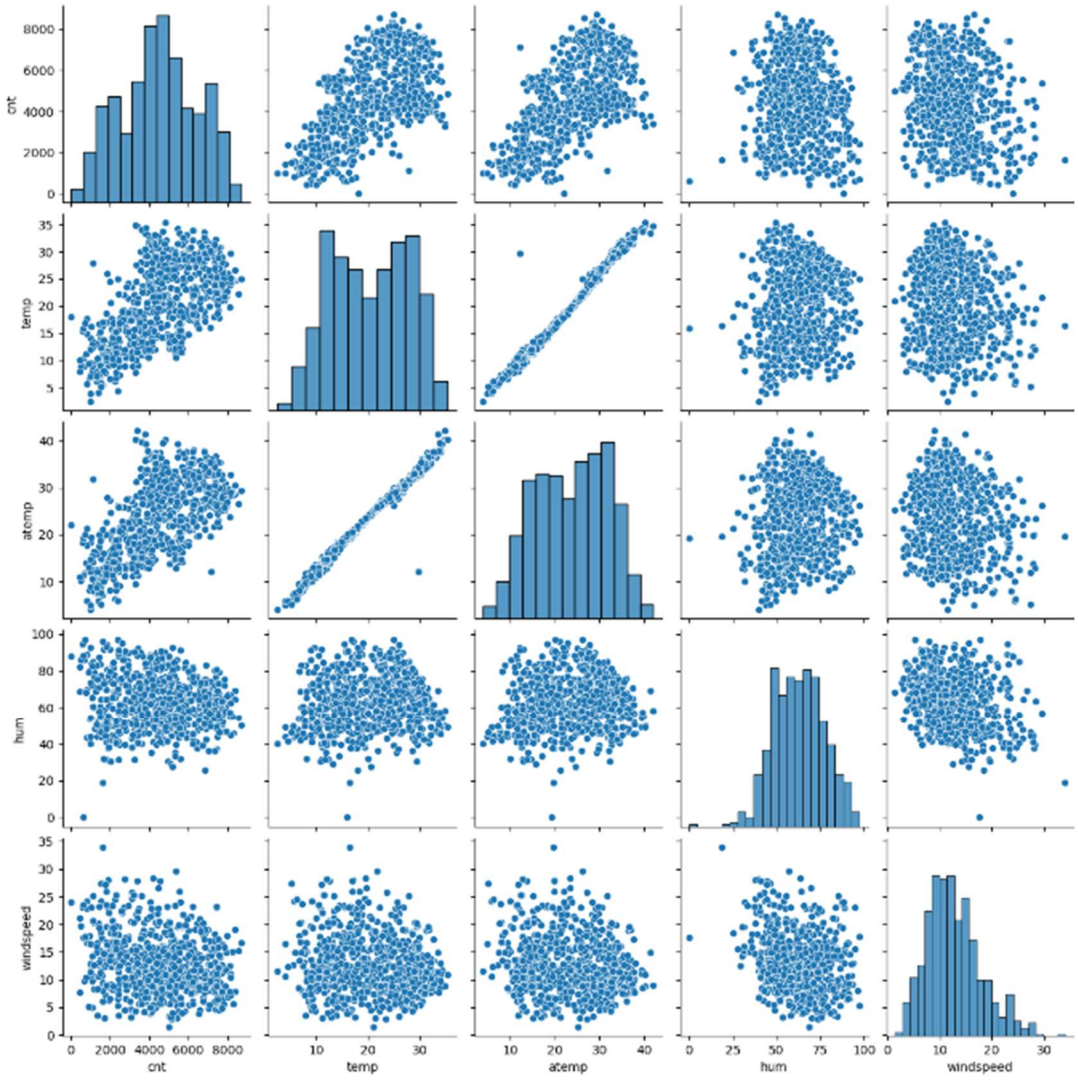
Using **drop_first=True** during dummy variable creation is important because it helps reduce multicollinearity among the dummy variables. When creating dummy variables for a categorical feature with (k) levels, (k) dummy variables are generated. However, one of these dummy variables can be perfectly predicted from the others, leading to multicollinearity, which can distort the results of the regression model.

By setting **drop_first=True**, the first level of the categorical variable is dropped, creating (k-1) dummy variables instead of (k). This approach ensures that the dummy variables are independent of each other, which helps in providing more reliable and interpretable regression coefficients.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Among all the numerical variables, 'temp' variable has the highest correlation with the target variable.

Please note that 'temp' variable denotes temperature in Celsius, where as 'atemp' variable denotes feeling temperature in Celsius. By definition itself, both these variables denote a very similar attribute, and have strong correlation (0.63 approx.) with the target variable.



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of linear regression is a critical step to ensure that the model's results are reliable and interpretable. After building the model on the training set, the assumptions of linear regression have been validated using diagnostic checks and plots. Here's a breakdown of the steps for validation:

1. **Linearity**: The relationship between the independent variables (predictors) and the dependent variable (target) should be linear. This can be checked using a residual plot, where the residuals (difference between observed and predicted values) are plotted against the predicted values. If the residuals are randomly scattered around zero with no clear pattern, the linearity assumption holds. Curved or non-random patterns suggest non-linearity.

   Validation – Residuals vs predicted values plot indicate that residuals are randomly scattered around zero, and don't show any discernible pattern

2. **Independence of Errors**: The residuals (errors) should be independent of each other, meaning there is no correlation between them. This can be checked using the Durbin-Watson test, which provides a test statistic to detect autocorrelation in residuals. Values close to 2 indicate no autocorrelation, while values closer to 0 or 4 indicate positive or negative autocorrelation, respectively.

   Validation – Durbin-Watson test value is 2.04, which indicates no autocorrelation

3. **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables. This can be checked using a residuals vs. fitted plot. If the spread of residuals is uniform, the assumption holds. A funnel shape (increasing or decreasing spread) indicates heteroscedasticity. The Breusch-Pagan test can also be used to statistically test for heteroscedasticity.

   Validation – Residuals vs predicted values plot indicate that variance of residuals remain reasonably constant across all levels of the predicted values

4. **Normality of Errors**: The residuals should be approximately normally distributed. This can be checked using a histogram of residuals or a Q-Q plot (quantile-quantile plot), which compares the residuals' quantiles with those of a normal distribution. Residuals should closely follow the 45-degree line. Statistical tests like the Shapiro-Wilk test or Kolmogorov-Smirnov test can also be used to check normality.

   Validation – Residuals clearly follow normal distribution, centered around zero

5. **No Multicollinearity**: Independent variables should not be highly correlated with each other. This can be checked using the Variance Inflation Factor (VIF). VIF values greater than 10 indicate potential multicollinearity.

   Validation – All the variables have VIF fairly below 10, indicating no severe multicollinearity

Summary of Validation Steps:

- **Linearity**: Residual Plot - Random scatter around zero.
- **Independence**: Durbin-Watson Test - DW ≈ 2 indicates no autocorrelation.
- **Homoscedasticity**: Residual Plot / Breusch-Pagan Test - Uniform spread of residuals; p-value > 0.05.
- **Normality of Errors**: Histogram / Q-Q Plot / Tests - Residuals normally distributed.
- **No Multicollinearity**: VIF - VIF < 5 (or <10) is acceptable.

Validating these assumptions ensures that the linear regression model provides reliable predictions and interpretations.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Top features contributing significantly towards explaining the demand of the shared bikes (in decreasing order of impact):

Original Variable	Feature in final model	Coefficient	Interpretation
temp	temp	0.4396	Demand for shared bikes increases with increasing temperature
weathersit	light_rain	-0.2953	Demand for shared bikes decrease significantly in case of Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
yr	2019	0.2344	Overall demand in 2019 is found to be significantly higher than demand in 2018

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a fundamental algorithm in machine learning and statistical modeling. It is used to model the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors or features). The goal is to find the best-fitting straight line through the data points that minimizes the differences between the predicted and actual values.

Simple Linear Regression

In simple linear regression, we have one independent variable and one dependent variable. The relationship between these variables is modeled using a straight line, which can be represented by the equation:

[y = b_0 + b_1x

Here:

- y is the dependent variable.
- x is the independent variable.
- b_0 is the y-intercept of the line.
- b_1 is the slope of the line.

The slope (b_1) represents the change in the dependent variable for each unit change in the independent variable, while the intercept (b_0) represents the predicted value of the dependent variable when the independent variable is zero.

Multiple Linear Regression

Multiple linear regression extends simple linear regression by using two or more independent variables to predict the dependent variable. The equation for multiple linear regression is:

y = b_0 + b_1x_1 + b_2x_2 + … + b_nx_n

Here:

- y is the dependent variable.
- x_1, x_2, … x_n are the independent variables.
- b_0 is the y-intercept.
- b_1, b_2, … b_n are the coefficients for the independent variables.

Cost Function

The cost function, also known as the loss function, measures how well the linear regression model fits the data. The most commonly used cost function for linear regression is the Mean Squared Error (MSE), which is defined as:

MSE = (1/n) * Σ (actual value - predicted value)^2

Here:

- n is the number of data points.
- actual value is the true value of the dependent variable.
- predicted value is the value predicted by the model.

The goal of linear regression is to minimize the MSE by finding the optimal values for the coefficients (b_0, b_1, … b_n).

Optimize the Coefficients

Gradient descent is an optimization algorithm used to minimize the cost function. It iteratively adjusts the coefficients to reduce the MSE.

Train the Model

Use the training data to compute the optimal values for the coefficients (b_0, b_1, … b_n).

Make Predictions

Use the trained model to predict target variable (y) for new data:

y = b_0 + b_1x_1 + b_2x_2 + … + b_nx_n

Model Evaluation

To assess the performance of a linear regression model, we use evaluation metrics such as:

- **R-squared (R^2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared (R^2)**: Modified version of R-squared that adjusts for the number of predictors in a regression model.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, providing a measure of the average error in the same units as the dependent variable.

Assumptions of Linear Regression

Linear regression makes several key assumptions:

1. **Linearity**: The relationship between the independent and dependent variables is linear.
2. **Independence**: The residuals (errors) are independent.
3. **Homoscedasticity**: The residuals have constant variance.
4. **Normality**: The residuals are normally distributed.
5. **No Multicollinearity**: Predictors are not highly correlated

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets that were constructed by the statistician Francis Anscombe in 1973. The purpose of these datasets is to demonstrate the importance of graphing data before analyzing it and to show how different datasets can have identical statistical properties but look very different when graphed.

Key Points of Anscombe's Quartet

1. **Identical Statistical Properties**: All four datasets in Anscombe's quartet have nearly identical simple descriptive statistics. This includes the mean of the x and y values, the variance of the x and y values, the correlation between x and y, and the linear regression line. Despite these similarities, the datasets are very different when visualized.
2. **Different Distributions**: When plotted on a scatter plot, each dataset reveals a different pattern. This highlights the limitations of relying solely on statistical summaries without visualizing the data.
3. **Importance of Visualization**: Anscombe's quartet emphasizes the need to visualize data to understand its underlying structure and to identify any anomalies or patterns that might not be apparent from statistical summaries alone.

The Four Datasets

1. **Dataset I**: This dataset fits a linear regression model quite well. The data points are scattered around a straight line, and the relationship between x and y is linear.
2. **Dataset II**: This dataset shows a clear non-linear relationship between x and y. A linear regression model would not be appropriate here, as it would not capture the true relationship between the variables.
3. **Dataset III**: This dataset includes an outlier that significantly affects the linear regression line. The presence of this outlier demonstrates how a single data point can influence the results of a statistical analysis.
4. **Dataset IV**: This dataset has a high-leverage point that creates a high correlation coefficient, even though the other data points do not indicate any relationship between x and y. This shows how influential points can distort the interpretation of data.

Conclusion

Anscombe's quartet serves as a powerful reminder of the importance of data visualization in statistical analysis. By plotting data, we can gain insights that are not apparent from statistical summaries alone, and we can avoid being misled by outliers or unusual data points. This principle is crucial for accurate data analysis and effective decision-making

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies both the strength and direction of this relationship. The value of 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.

Pearson's R assumes that the relationship between variables is symmetrical. For example, the correlation between X and Y is the same as Y and X.

Calculation

The formula for Pearson's R is:

$$r = (\sum (xi - \bar{x})(yi - \bar{y})) / (\sqrt{\sum (xi - \bar{x})^2} * \sqrt{\sum (yi - \bar{y})^2})$$

Here:

- xi and yi are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.
- $\sum$ denotes the summation over all data points.

Interpretation

- **Positive Correlation**: If (r) is positive, it means that as one variable increases, the other variable also increases. The closer the value is to 1, the stronger the positive linear relationship.
- **Negative Correlation**: If (r) is negative, it means that as one variable increases, the other variable decreases. The closer the value is to -1, the stronger the negative linear relationship.
- **No Correlation**: If (r) is close to 0, it means there is no linear relationship between the variables.

Assumptions

Pearson's R makes several key assumptions:

1. Linearity
- The relationship between the two variables must be linear.
- If the relationship is non-linear, Pearson's rrr may underestimate or fail to detect the association.
- Validation: Use scatter plots to visualize the relationship.

2. Continuous Variables
- Both variables should be continuous (e.g., interval or ratio scale).
- Categorical data is not appropriate for Pearson's correlation.

3. Normality of Variables
- Both variables should be approximately normally distributed.
- This assumption is particularly important for significance testing (e.g., p-values).
- Validation: Use histograms, Q-Q plots, or normality tests (e.g., Shapiro-Wilk).

4. Homoscedasticity
- The variance of one variable should remain constant across the levels of the other variable.
- If heteroscedasticity exists, Pearson's rrr might not accurately reflect the strength of the relationship.
- Validation: Examine a scatter plot of residuals.

5. Independence of Observations
- Each pair of observations should be independent.
- Correlated or dependent data (e.g., repeated measures) violate this assumption.

6. No Significant Outliers
- Outliers can disproportionately influence the value of Pearson's rrr, leading to misleading results.
- Validation: Identify outliers using scatter plots or statistical tests (e.g., Z-scores).

Example

Let's consider an example to illustrate Pearson's R. Suppose we have the following data points for two variables, X and Y:

X	Y
1	2
2	3
3	4
4	5
5	6

To calculate Pearson's R, we first compute the means of X and Y, then use the formula to find the correlation coefficient. In this case, the correlation coefficient would be 1, indicating a perfect positive linear relationship.

Practical Use

Pearson's R is widely used in various fields such as finance, economics, and social sciences to determine the strength and direction of the relationship between two variables. It helps in understanding how changes in one variable are associated with changes in another variable.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a preprocessing technique in machine learning that transforms the numerical features in a dataset to a common scale or range. This is crucial because different features in a dataset may have different ranges or units, which can affect the performance of machine learning algorithms.

Purpose of Scaling

Scaling is performed for several reasons:

1. **Improves Model Performance**: Some machine learning algorithms, such as those that use gradient descent (e.g., linear regression, logistic regression, neural networks), perform better when the data is scaled. This is because scaling ensures that the gradient descent converges more quickly and smoothly.
2. **Prevents Bias**: Features with larger ranges can dominate the learning process, leading to biased models. Scaling ensures that all features contribute equally to the model.
3. **Enhances Interpretability**: Scaling makes it easier to interpret the coefficients of the model, especially in linear models.
4. **Reduces Impact of Outliers**: Scaling can reduce the impact of outliers by bringing all features to a similar scale.

Difference Between Normalized Scaling and Standardized Scaling

Normalization and **Standardization** are two common techniques for scaling data:

1. **Normalization**:
    - **Definition**: Normalization rescales the feature values to a specific range, often between 0 and 1.
    - **Formula**: x_normalized = (x - x_min) / (x_max - x_min)
      Here:
        - x" is the original value,
        - x_min is the minimum value in the data set, and
        - x_max is the maximum value in the data set
    - **Use Case**: Normalization is useful when you want to ensure that all features have the same scale, especially when the features have different units or ranges.
    - **Example**: If you have a dataset with features like age (0-100) and income (in thousands), normalization will bring both features to a common scale.
2. **Standardization**:
    - **Definition**: Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
    - **Formula**: x_standardized = (x - mean(x)) / std(x)
      Here:
        - x is the original value,
        - mean(x) is the mean of x variable, and

    - std(x) is standard deviation of x variable
- **Use Case**: Standardization is useful when the data follows a normal distribution or when you want to ensure that the features have the same scale but retain their original distribution.
- **Example**: If you have a dataset with features like height and weight, standardization will ensure that both features have the same scale but retain their original distribution.

Practical Example

Let's consider a dataset with two features: height (in cm) and weight (in kg). The height ranges from 150 to 200 cm, and the weight ranges from 50 to 100 kg. Without scaling, the model might give more importance to the weight feature because it has a larger range. By applying normalization, both features will be scaled to a range of 0 to 1, ensuring that they contribute equally to the model

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) can sometimes take on an infinite value, and this typically occurs due to perfect multicollinearity in the dataset. Perfect multicollinearity happens when one independent variable in a regression model can be perfectly predicted by one or more other independent variables. In other words, there is an exact linear relationship between the variables.

Reason for Infinite VIF

The VIF for a given independent variable is calculated using the formula:

VIF = 1 / (1 - R^2)

Here R^2 is the coefficient of determination from the regression of that independent variable on all other independent variables in the model. When (R^2) is equal to 1, it means that the independent variable is perfectly predicted by the other variables, leading to the denominator becoming zero. As a result, the VIF value becomes infinite.

Implications of Infinite VIF

An infinite VIF indicates that there is perfect multicollinearity, which means that the regression model cannot uniquely estimate the coefficients of the independent variables. This situation makes the model unstable and unreliable, as the presence of perfect multicollinearity violates the assumptions of the regression analysis.

Example

Consider a regression model where with two independent variables, (X_1) and (X_2), and (X_2) is a perfect linear combination of (X_1) (e.g., (X_2 = 2X_1)). In this case, the VIF for both (X_1) and (X_2) would be infinite because each variable can be perfectly predicted by the other.

Addressing Infinite VIF

To address the issue of infinite VIF, we can:

1. **Remove one of the perfectly collinear variables**: This will eliminate the perfect multicollinearity.
2. **Combine the collinear variables**: If the variables are measuring the same underlying construct, we can combine them into a single variable.

Understanding and addressing multicollinearity is crucial for building reliable and interpretable regression models

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.

Use and Importance in Linear Regression

In the context of linear regression, a Q-Q plot is primarily used to assess whether the residuals (errors) of the model are normally distributed. This is important because one of the key assumptions of linear regression is that the residuals are normally distributed. If this assumption is violated, the results of the regression analysis, including hypothesis tests and confidence intervals, may not be valid.

How to Interpret a Q-Q Plot

1. **Straight Line**: If the points on the Q-Q plot lie along a straight line, it indicates that the residuals are normally distributed.
2. **Deviations from Line**: If the points deviate from the straight line, it suggests that the residuals are not normally distributed. The nature of the deviations can provide insights into the type of non-normality, such as skewness or kurtosis.

Practical Example

After fitting a linear regression model to data, a Q-Q plot of the residuals can be created. If the points on the Q-Q plot lie along a straight line, it can be concluded that the residuals are normally distributed, and the assumption of normality is satisfied. If the points deviate significantly from the line, it may be necessary to consider transforming the data or using a different model.

Conclusion

The Q-Q plot is a valuable diagnostic tool in linear regression. It helps validate the assumption of normality of residuals, which is crucial for the validity of the regression results. By visually assessing the distribution of residuals, informed decisions can be made about the appropriateness of the model and the need for any adjustments.