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:
As per final Model, the top 3 categorical variables that influences the bike booking are:

- Light snow rain - A coefficient value of `-0.238456` indicate that a unit increase in temp variable decreases the bike hire number by 0.238456
- windspeed- A coefficient value of `-0.192512` indicate that a unit increase in temp variable decreases the bike hire number by 0.192512
- winter - A coefficient value of `0.128776` indicate that a unit increase in temp variable increase the bike hire number by 0.128776

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

Answer:
When creating dummy variables from categorical variables, the drop_first=True parameter is used to prevent multicollinearity in regression models and to enhance model interpretability.

Multicollinearity: Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can cause issues in estimating the individual coefficients' effects on the dependent variable. By dropping one level (category) of each categorical variable, you avoid perfect multicollinearity among the dummy variables, as the dropped category becomes the reference category.

Interpretability: Dropping one level in a categorical variable enhances interpretability by establishing a reference category, simplifying coefficient interpretation. The baseline coefficient represents the omitted category, while coefficients for other dummy variables indicate changes relative to this baseline, streamlining interpretation and minimizing redundancy.

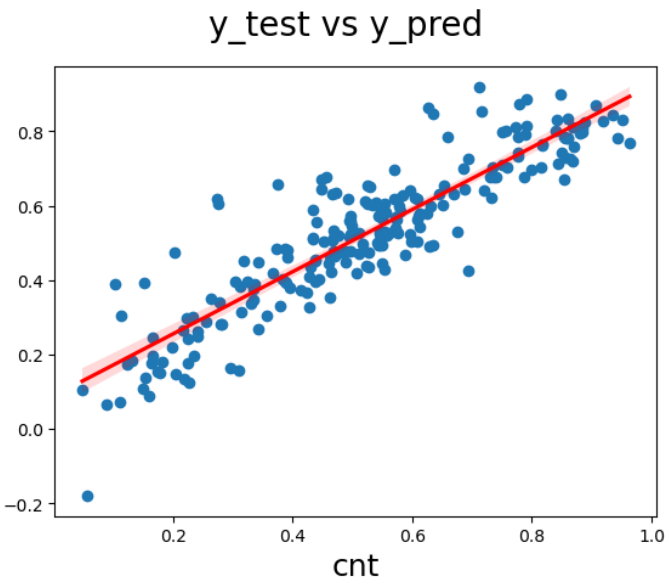
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:
Temperature - A coefficient value of 0.544334 indicate that a unit increase in temp variable increase the bike hire number by 0.544334

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

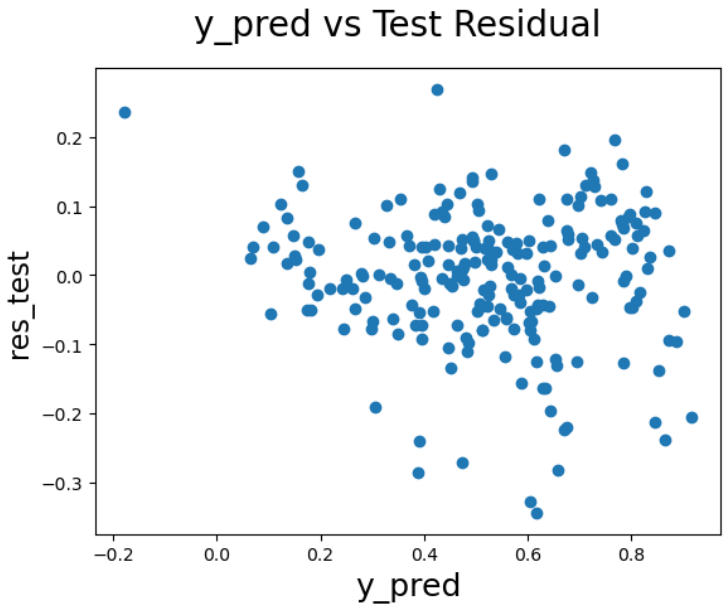
Answer:

1. **Linearity:**
   Check for linearity by plotting the actual versus predicted values. A scatter plot should ideally show a linear relationship. If the plot indicates a pattern, there might be an issue with linearity.



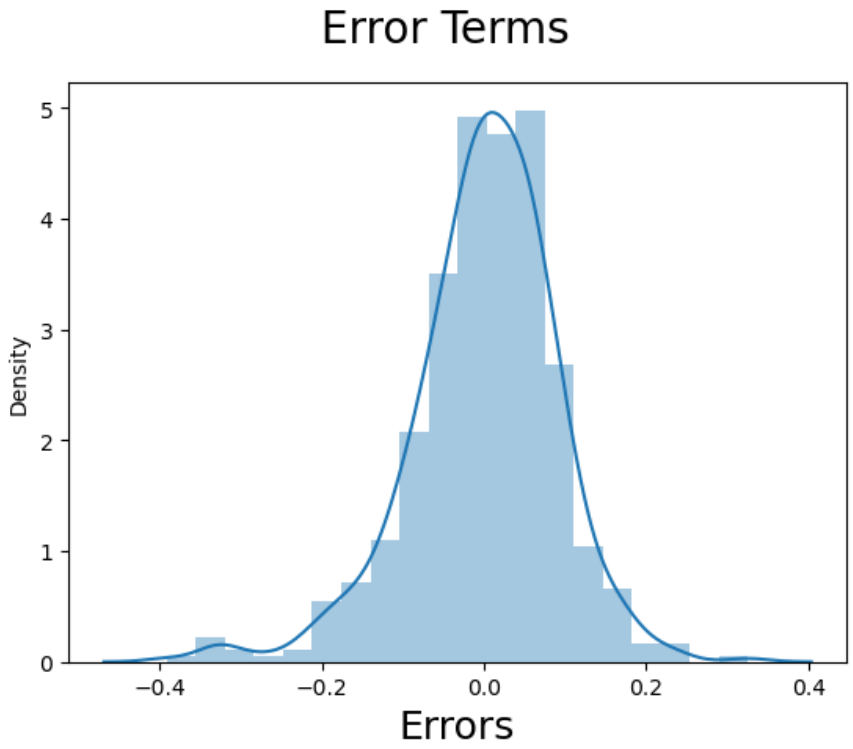
In our model, shows linear relation between actual and predicted values

2. **Residuals Analysis**
   Examine the residuals (the differences between actual and predicted values). Residual plots should be random and not exhibit any clear patterns



Above chart is the residuals of predicted and actual values, from our model it is clear that plots show random and not exhibit any clear patterns

3. **Normality of Residuals:**
   Assess the normality of residuals using a histogram or a Q-Q plot. Normally distributed residuals are important for the validity of statistical inferences. If residuals are not normally distributed, transformations or other techniques may be needed.



The above chart is the histogram of the error terms. It clear that the model which I created exhibit normal distribution of residual

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

Answer:

- Temperature - A coefficient value of `0.544334` indicate that a unit increase in temp variable increase the bike hire number by 0.544334
- Light snow rain - A coefficient value of `-0.238456` indicate that a unit increase in temp variable decreases the bike hire number by 0.238456
- Year - A coefficient value of `0.229886` indicate that a unit increase in temp variable increase the bike hire number by 0.229886

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data. The fundamental equation for a simple linear regression with one independent variable is:

$$Y = \beta 0 + \beta 1 \cdot X + \epsilon$$

Y - dependent variable (target),
X - independent variable (feature),
B0 - y-intercept (constant term),
β1 - slope of the line, representing the coefficient for the independent variable,
ϵ - error term, accounting for unobserved factors affecting Y.

The goal of linear regression is to estimate the values of β0 and β1 that minimize the sum of squared differences between the observed and predicted values. This is often done using the method of least squares.

$$Y=\beta 0 + \beta 1 \cdot X1 + \beta 2 \cdot X2 + \ldots + \beta n \cdot Xn + \epsilon$$

Linear regression assumes a linear relationship between the variables, independence of errors, homoscedasticity (constant variance of errors), and normally distributed errors.

The model is trained on a dataset to find the optimal coefficients, and once trained, it can be used to make predictions on new data by substituting the values of independent variables into the equation. The model's performance is often assessed using metrics like Mean Squared Error (MSE) or R-squared. Linear regression is widely used for predictive modelling and understanding the relationships between variables in various fields such as economics, finance, and natural sciences.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a set of four datasets created by statistician Francis Anscombe in 1973. These datasets are fascinating because they share identical basic summary statistics (mean, variance, correlation coefficient) and regression line, yet visually appear very different when plotted. This serves as a powerful

reminder of the limitations of relying solely on summary statistics and emphasizes the importance of data visualization in exploratory analysis.

- Each dataset consists of 11 data points (x, y).
- All four datasets share the same:
  - Mean of x values
  - Mean of y values
  - Variance of x values
  - Variance of y values
  - Correlation coefficient between x and y
  - Regression line slope and intercept
- Despite these identical statistics, the datasets show vastly different characteristics when visualized:
  - Dataset 1: Linear relationship, well-fitted by the regression line.
  - Dataset 2: Curved, non-linear relationship, poorly fitted by the line.
  - Dataset 3: Outlier with a significant impact on the regression line.
  - Dataset 4: Two influential points heavily affecting the line, creating an illusion of linear association.

Importance:

- Highlights the dangers of relying solely on summary statistics: They can be misleading and fail to capture important features of the data.
- Emphasizes the importance of data visualization: Plotting reveals crucial information about the distribution, trends, outliers, and potential issues that wouldn't be apparent from statistics alone.
- Serves as a cautionary tale in regression analysis: Just because a model fits well statistically doesn't necessarily mean it represents the true relationship in the data.

3. What is Pearson's R?

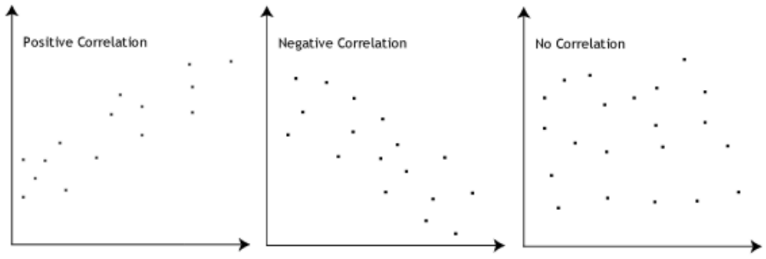
Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure used to quantify the **linear relationship** between two continuous variables. It ranges from **-1 to 1**, with:

- -1 indicating a perfect negative correlation: As one variable increases, the other decreases proportionally.
- 0 indicating no correlation: There's no linear relationship between the variables.
- +1 indicating a perfect positive correlation: As one variable increases, the other also increases proportionally.

Key points about Pearson's R:

- Linear association: It only measures linear relationships, not other types like curves or complex patterns.
- Continuous variables: It's best suited for data measured on a continuous scale (e.g., height, weight, temperature).
- Strength and direction: The value indicates both the strength (absolute value) and direction (sign) of the relationship.
- Interpretation: A closer value to 1 (positive or negative) implies a stronger association. Closer to 0 indicates a weaker or no association.



Examples:

- Positive correlation: Study hours and exam scores might have a positive correlation, indicating higher study hours tend to lead to higher scores.
- Negative correlation: Age and video game playing time might have a negative correlation, indicating older people tend to play less.
- No correlation: Shoe size and intelligence might have no correlation, meaning there's no linear relationship between them.

Limitations:

- Linearity: It doesn't capture non-linear relationships.
- Outliers: Outliers can significantly influence the value.
- Assumptions: Normality of data is often assumed, but deviations can affect results.

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

Answer:

Scaling is the process of transforming the numerical features of a dataset to a standard range or distribution. In the context of linear regression, it involves adjusting the scale of independent variables. Common scaling techniques include normalization and standardization.

Why is Scaling Performed?

Scaling is performed in linear regression for several reasons:

- Magnitude Consistency: Scaling ensures that all variables have similar magnitudes, preventing variables with larger scales from dominating the learning process.
- Convergence: It can help optimization algorithms converge faster during the model training process.
- Equal Contribution: Each variable contributes equally to the model fitting, avoiding bias towards variables with larger scales.
- Interpretability: Coefficients in a linear regression model become comparable when variables are on the same scale, aiding interpretation.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling:

- Range**:** Scales the values to a specific range, typically [0, 1].
- Formula: X(normalized) = X-min(X)/ max(X)-min(X)
- Advantage: Useful when the data distribution is unknown or skewed.

Standardized Scaling (Z-score normalization):

- Mean and Standard Deviation: Scales the values to have a mean of 0 and a standard deviation of 1.
- Formula: X standardized = X - mean(X)/std(X)
- Advantage: Preserves information about the shape of the distribution, suitable when data follows a normal distribution.

Both normalized and standardized scaling are methods to bring features to a common scale, but they differ in the specific transformations applied. The choice between them depends on the characteristics of the data and the requirements of the modelling process. Standardized scaling is often preferred in linear regression as it maintains information about the distribution and is less sensitive to outliers compared to normalized scaling.

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

Answer:
The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple regression model. VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A VIF of 1 indicates no multicollinearity, and as the VIF increases, the correlation among predictors becomes more problematic.

Here's why VIF can become infinite:

- **Perfect Linear Relationship**: If one predictor variable can be expressed as a perfect linear combination of one or more other predictors, the correlation between them is perfect (1 or -1).

- **Matrix Inversion Issue**: When calculating VIF, it involves the inversion of the correlation matrix of the predictors. If the correlation matrix is not invertible due to perfect multicollinearity, the inversion process fails, leading to an undefined or infinite value.

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

Answer:
A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess the normality of a distribution by comparing the quantiles of the observed data to the quantiles of a theoretical normal distribution. In the context of linear regression, Q-Q plots are often employed to examine the normality of the residuals (the differences between observed and predicted values) because normality of residuals is a crucial assumption in linear regression modelling.

How to Interpret a Q-Q Plot:

- If the points on the Q-Q plot closely follow a straight line, it suggests that the residuals are approximately normally distributed.
- Deviations from a straight line indicate departures from normality. If the points deviate significantly from the line, it may indicate non-normality in the residuals.

Use and Importance in Linear Regression:

- Normality of Residuals: The assumption of normality is important in linear regression as it influences the validity of statistical tests, confidence intervals, and hypothesis testing associated with the model parameters.

- Model Assumption Check: Q-Q plots provide a visual check of the normality assumption in the residuals. If the residuals are not normally distributed, it may impact the reliability of the statistical inferences drawn from the regression analysis.
- Identifying Outliers: Q-Q plots can help identify outliers and influential data points that may affect the normality assumption. Deviations from the straight line in specific regions of the plot may indicate the presence of outliers.
- Model Diagnostics: Q-Q plots are part of a suite of diagnostic tools used to assess the overall fit and assumptions of the linear regression model. They complement other diagnostic techniques like residual plots and leverage plots.

Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals, a key assumption for valid statistical inferences. Detecting deviations from normality through Q-Q plots prompts further investigation and may lead to adjustments or transformations to improve the model's validity and reliability.