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)

Answer: I have completed analysis on categorical columns using the boxplot and barplot and below are a few points we can infer from the visualization.

Inference

season: Most of the bike booking happened in fall season. It was followed by summer and winter. This indicates, season can be a good predictor for the dependent variable.

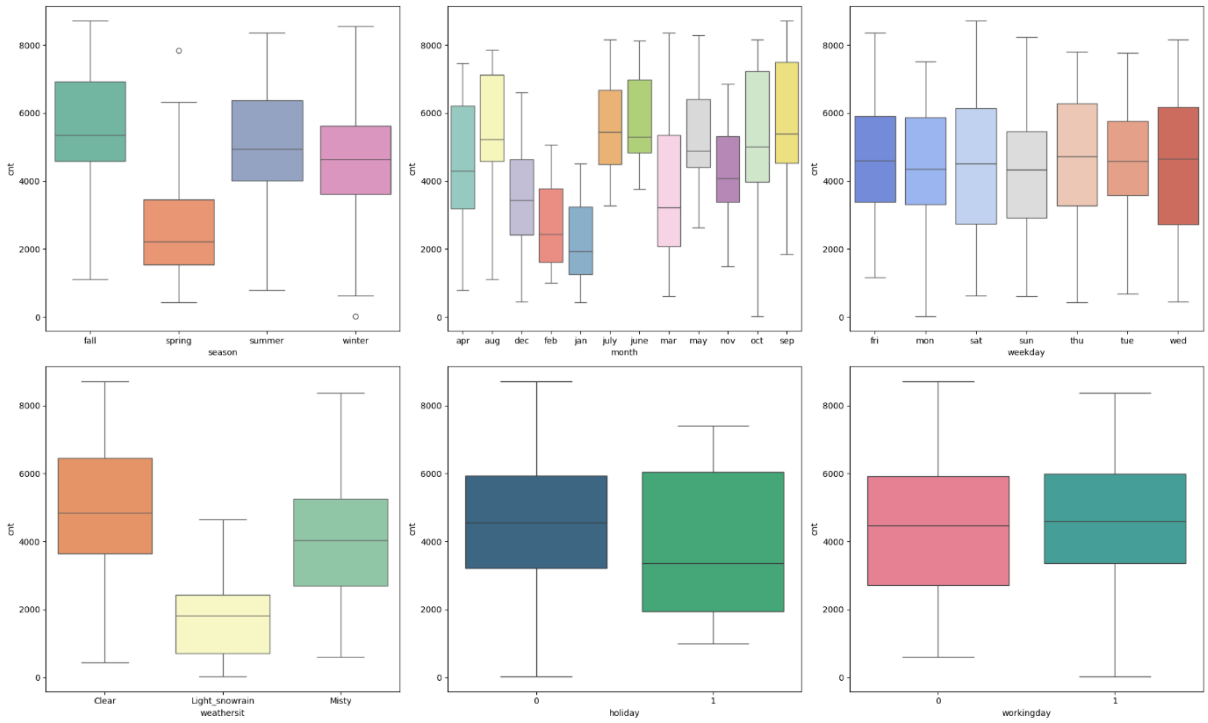
month: Most of the bike booking happened in the months May, June, July, August, September and October. This indicates that Month has some trend for bookings and can be a good predictor for the dependent variable.

weekday: Wed, Thu, Fir, Sat have more number of bookings as compared to the start of the week and can be a good predictor for the dependent variable.

weathersit: Clear weathersit had more number of bookings compared to others. So, it can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking happened on non-holiday day. So, this data can be biased and may not be a good predictor for the dependent variable.

workingday: Most of the bike booking happened on workingday. So, it can be a good predictor for the dependent variable.



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

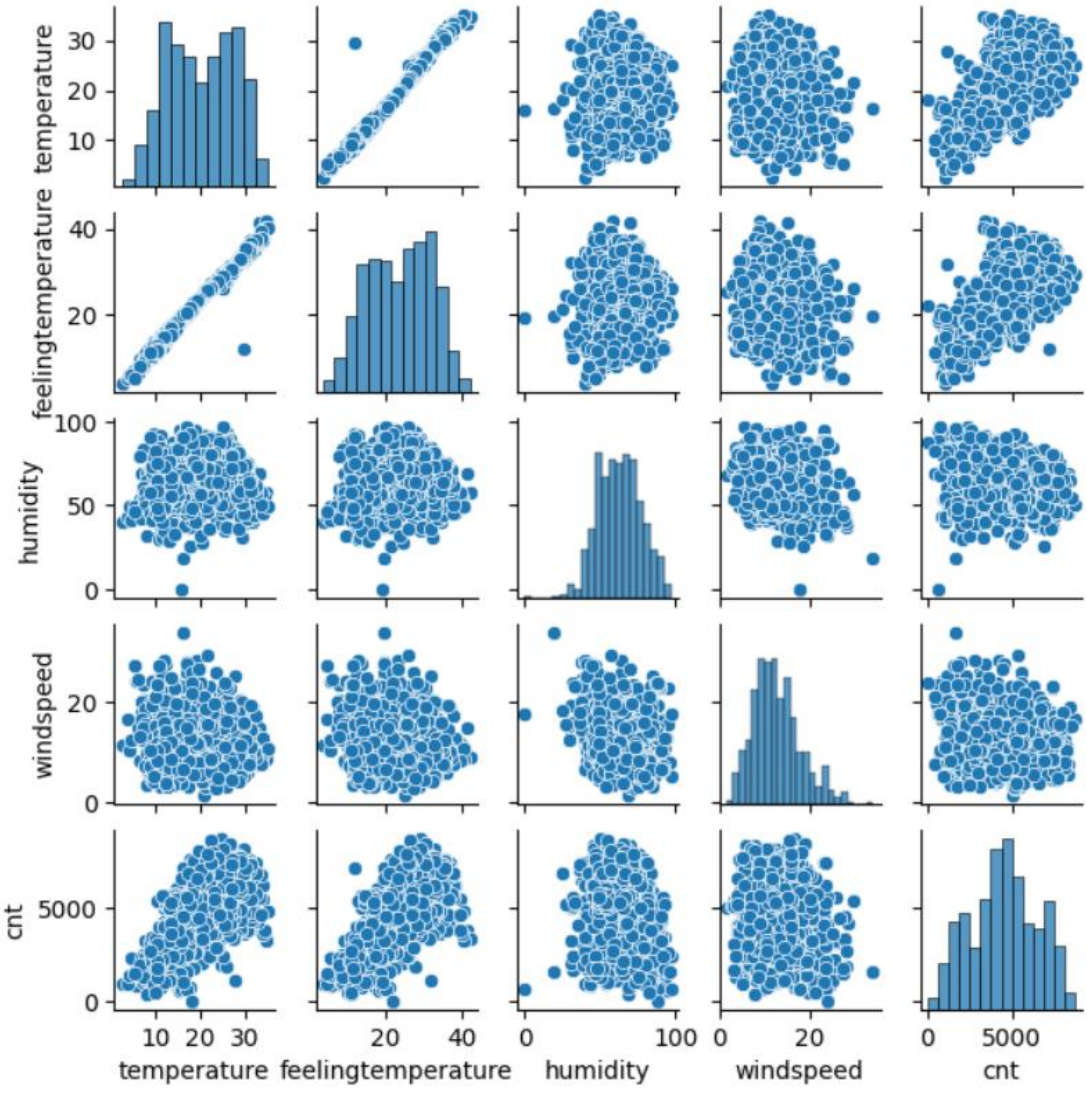
Answer:
When creating dummy variables (one-hot encoding) for categorical data in machine learning models, the drop_first=True parameter is often used to avoid the dummy variable trap and to prevent multicollinearity.

Syntax
drop_first: bool, default value = False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

E.g.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:
The temp(temperature) variable has the highest correlation with target variable "cnt".



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:
I have validated the assumption of Linear Regression Model based on below 4 assumptions:

1. Linearity
Assumption: The relationship between the independent variables and the dependent variable is linear.
Validation:
Residual vs. Fitted Plot: I plotted the residuals (errors) against the fitted values. If the relationship is linear, the plot should show no clear pattern; residuals should be randomly scattered around zero.

2. Independence
Assumption: The residuals are independent of each other.
Validation:
Plotting Residuals: Plotted residuals over time to check for patterns indicating autocorrelation.

3. Homoscedasticity
Assumption: The residuals have constant variance.
Validation:
Residuals vs. Fitted Values Plot: A residual plot should show no pattern. If there is a funnel shape or other patterns, it may indicate heteroscedasticity.

4. No Multicollinearity
Assumption: Predictor variables are not highly correlated with each other.
Validation:
Variance Inflation Factor (VIF): Calculated the VIF for each predictor. High VIF values (typically > 5) indicate multicollinearity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:
Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a statistical technique used to model the relationship between a dependent variable (target variable) and one or more independent variables (predictors or features). The goal of linear regression is to find the best-fitting linear line (regression line) that predicts the dependent variable from the independent variables. If there is a single input variable (x), such linear regression is called simple linear regression and if there is more than one input variable, such linear regression is called multiple linear regression.

Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

The linear regression model produces a sloped straight line that represents the relationship between the variables. This regression line can indicate either a positive or negative linear relationship. The aim of the linear regression algorithm is to determine the optimal values for a0 (the intercept) and a1 (the slope) to find the line of best fit, which minimizes errors. In linear regression, methods like Recursive Feature Elimination (RFE) or the Mean Squared Error (MSE) cost function are used to identify the best values for a0 and a1, ensuring the line of best fit accurately represents the data points.

Assumptions of Linear Regression

Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

Homoscedasticity: The variance of the error terms should be constant i.e. the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise, they will be constant.

Independence: The residuals (errors) are independent.

Normality: The residuals of the model should be normally distributed.

2. Explain the Anscombe's quartet in detail. (3 marks)

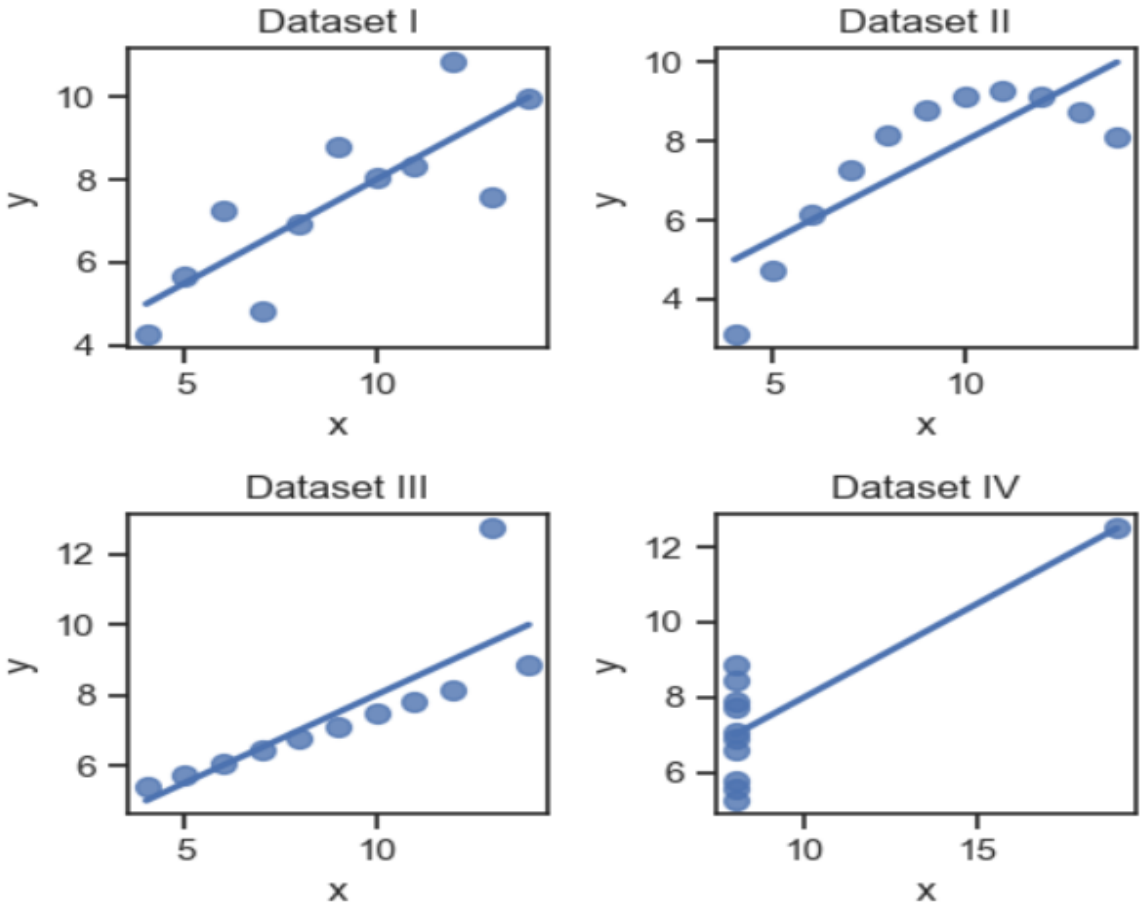
Answer:

Anscombe's Quartet is a collection of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, and correlation, but appear very different when graphed. Created by statistician Francis Anscombe in 1973, the quartet demonstrates the importance of graphing data before analyzing it and illustrates how relying solely on summary statistics can be misleading.

<u>The Four Datasets</u>

Let's look at each dataset in the quartet:

1. Dataset 1: This dataset appears to follow a roughly linear trend with some random noise. It is a typical example of data that would be well-represented by linear regression.
2. Dataset 2: This dataset also appears to have a linear relationship, but with an obvious outlier. The outlier skews the data, affecting the linear regression and correlation values.
3. Dataset 3: This dataset has a clear non-linear relationship. It resembles a quadratic curve, and the linear regression line is not an appropriate model, even though the correlation coefficient is the same as the others.
4. Dataset 4: This dataset is vertical clustering around a single value of xxx, with an extreme outlier. The outlier determines the linear regression line, although it doesn't represent the overall pattern of the majority of the data.



3. What is Pearson's R? (3 marks)

Answer:

Pearson's R, also known as the Pearson correlation coefficient (or simply Pearson's correlation), is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most widely used correlation coefficients in statistics.

Formula for Pearson's R

Pearson's R is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual sample points for variables X and Y, respectively.
- $\bar{x}$ is the mean of the variable X.
- $\bar{y}$ is the mean of the variable Y.
- $\sum$ denotes the summation.

Interpretation of Pearson's R

- Perfect Positive Correlation (r = +1): Both variables move in the same direction together perfectly.
- Perfect Negative Correlation (r = -1): One variable increases while the other decreases perfectly.
- Strong Positive Correlation (0.7 < r < 1): The variables are strongly positively related.
- Strong Negative Correlation (-1 < r < -0.7): The variables are strongly negatively related.
- Moderate Correlation (0.3 < |r| < 0.7): The variables have a moderate linear relationship.
- Weak Correlation (0 < |r| < 0.3): The variables have a weak linear relationship.
- No Correlation (r = 0): There is no linear relationship between the variables.

Limitations of Pearson's R

1. Only Measures Linear Relationships: Pearson's R only assesses the strength of a linear relationship. It does not measure non-linear relationships.
2. Sensitive to Outliers: Outliers can significantly affect the value of Pearson's R, leading to misleading interpretations.
3. Assumes Homoscedasticity: Assumes that the variance of the two variables is constant across values.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a data preprocessing technique used to transform features to a common scale. This is important because many machine learning algorithms perform better or converge faster when features are on a similar scale. Scaling ensures that each feature contributes equally to the analysis and prevents features with larger ranges from dominating the results.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling)
   Normalized scaling, or Min-Max scaling, rescales the features to a fixed range, usually [0, 1]. The formula for normalization is:

$$x_{\text{norm}} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Where:

- x is the original feature value.
- $\min(X)$ and $\max(X)$ are the minimum and maximum values of the feature in the dataset.

2. Standardized Scaling (Z-Score Normalization)
   Standardized scaling, or Z-score normalization, rescales the data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

Where:

- x is the original feature value.
- μ is the mean of the feature values.
- σ is the standard deviation of the feature values.

Example

Suppose you have a feature with the following values: [10, 20, 30, 40, 50]

- Normalized Scaling (Min-Max Scaling):
  - Min = 10, Max = 50
  - Normalized values: [0, 0.25, 0.5, 0.75, 1]
- Standardized Scaling (Z-score Normalization):
  - Mean (μ) = 30, Standard Deviation (σ) ≈ 15.81
  - Standardized values: [-1.27, -0.68, 0, 0.68, 1.27]

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

The Variance Inflation Factor (VIF) is a measure used to quantify the extent of multicollinearity in a regression model. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other. High multicollinearity can make it difficult to determine the individual effect of each predictor on the dependent variable and can lead to unstable estimates of regression coefficients.

Formula

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Where:

- Ri^2 is the R-squared value obtained by regressing the i-th predictor against all other predictors in the model.

VIF might be infinite

A VIF value becomes infinite when the denominator of the VIF formula equals zero. This happens when:

$$1 - R_i^2 = 0$$

or equivalently,

$$R_i^2 = 1$$

Causes of Infinite VIF

1. Duplicate Variables: Including identical or very similar variables in the model. For example, if two variables are perfectly correlated, one can be predicted perfectly from the other.
2. Linear Combinations: Including variables that are linear combinations of other variables in the model. For example, if you include both X and 2X as predictors, this creates perfect multicollinearity.
3. Data Issues: Problems in the dataset, such as errors or inconsistencies, that create exact linear dependencies among predictors.

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

(3 marks)

Answer:

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, commonly the normal distribution. It compares the quantiles of the dataset's distribution against the quantiles of the theoretical distribution.

Use of Q-Q Plots in Linear Regression

In linear regression, Q-Q plots are primarily used to assess whether the residuals (errors) of the model are normally distributed. This assumption is crucial for:

1. Validating Model Assumptions:
    - Normality of Residuals: Linear regression assumes that residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of residuals helps verify this assumption.
    - Model Diagnostics: If residuals are not normally distributed, it may indicate model specification issues or the need for data transformations.
2. Evaluating Model Fit:
    - Goodness-of-Fit: A Q-Q plot helps evaluate whether the model is a good fit for the data. If the residuals are normally distributed, it suggests that the model fits the data well. If not, it might imply that the model is missing important variables or interactions.
3. Assessing Homoscedasticity:
    - Error Variance: While the Q-Q plot is primarily used to check normality, deviations from the straight line can also signal issues with homoscedasticity (constant variance of residuals).

Importance in Linear Regression

1. Model Validity: Checking residuals' normality helps validate the assumptions of linear regression, ensuring that the model's results (such as hypothesis tests and confidence intervals) are reliable.
2. Diagnostics: Identifying non-normality in residuals can prompt further investigation and adjustments, such as transforming variables or adding interaction terms.