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:
Based on the analysis of the categorical variables from the dataset, here are some key insights into their effect on the dependent variable:

I. **Seasonal Patterns**:

Fall and **summer months** (May through October) see the highest number of bookings or rentals, with a notable increase in mid-year.

II. **Weather Impact**:

Clear weather drives higher bookings, while **adverse weather conditions** (e.g., light snow or rain) reduce them.

III. **Weekly and Holiday Trends**:

Weekends and **holidays** are associated with more bookings, as people are more available for recreational activities.

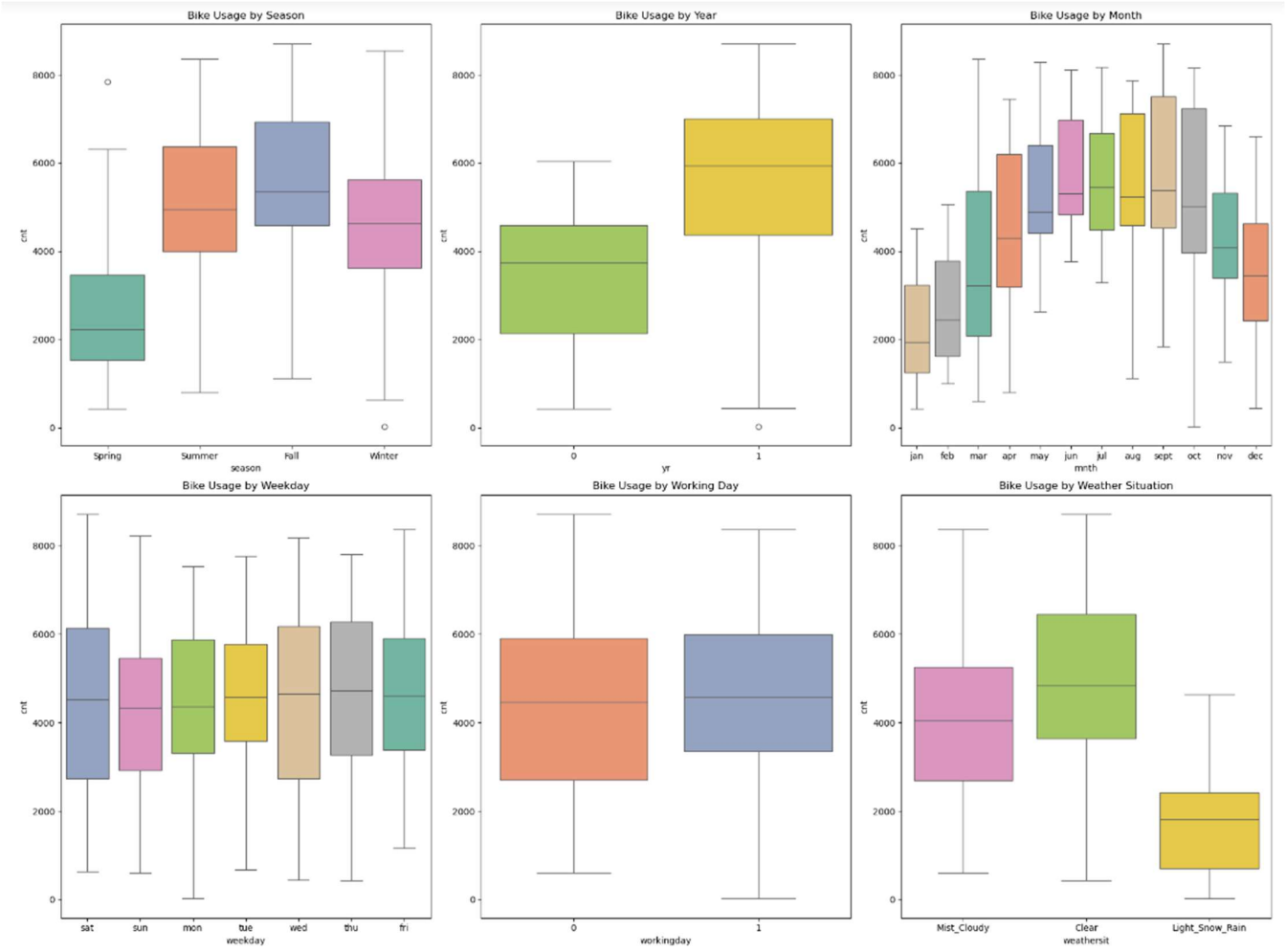
IV. **Yearly Growth**:

Bookings increased significantly in **2019** compared to the previous year, indicating growth in the service.

V. **Modeling Insights**:

Including categorical variables like year and season improves model performance, highlighting their importance in explaining variance.

In summary, the dataset shows that season, weather, and timing significantly affect bookings or rentals, with a positive trend observed in recent years. The below fig shows the correlation among the same.



These variables are visualized using bar plot and Box plot both.

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

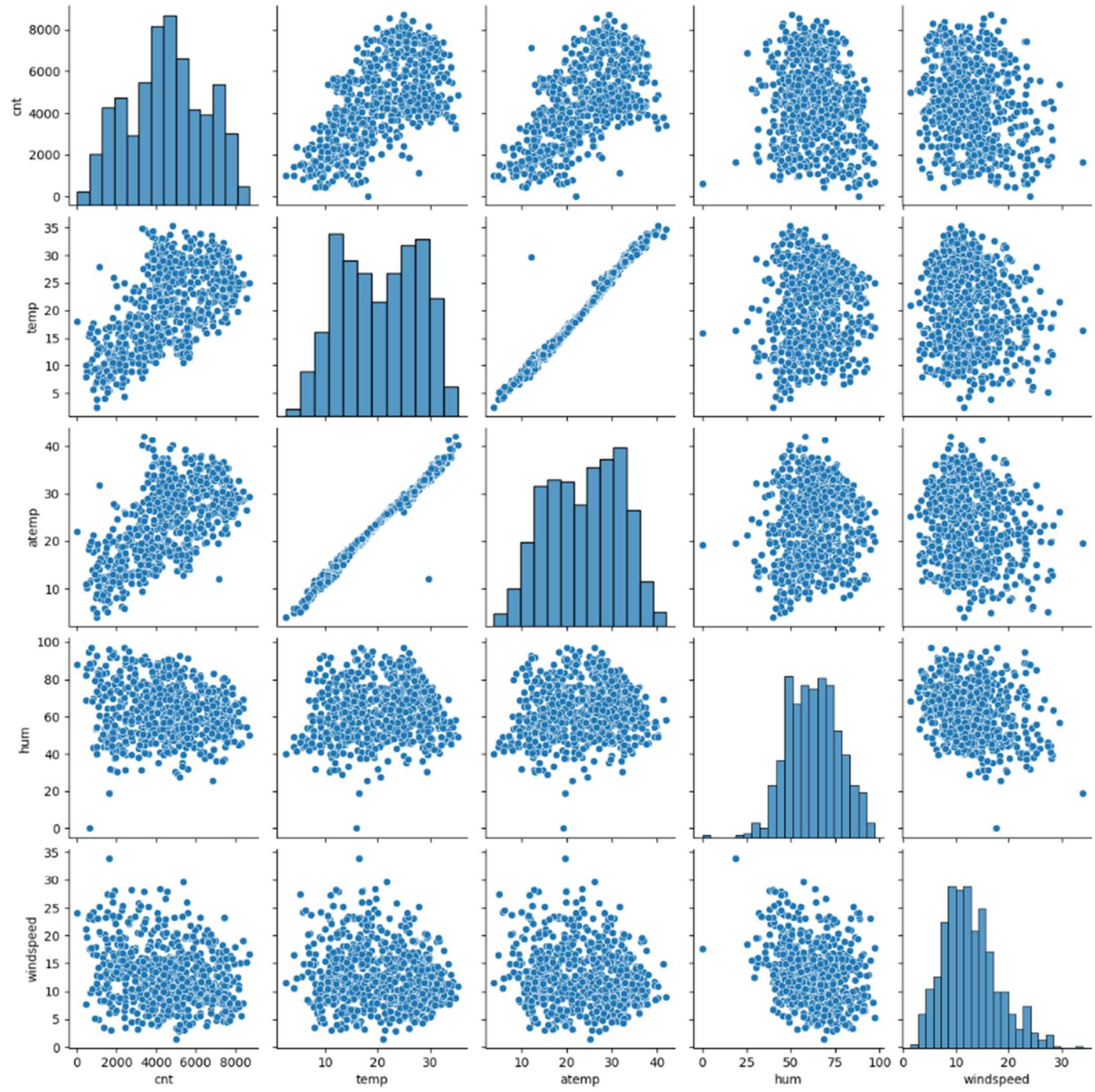
Answer:

Using drop_first=True during dummy variable creation is important to avoid multicollinearity and redundancy. It reduces the number of dummy variables by one, thus preventing the dummy variables from being perfectly multicollinear (i.e., their sum equals one, which makes them linearly dependent). This practice also simplifies model interpretation by setting a reference category against which other categories are compared.

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

Answer:

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.



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

Answer:

To validate the assumptions of Linear Regression after building the model on the training set, I did the following:

a) **Check Linearity**: Ensure there is a linear relationship between independent variables and the target variable.
b) **Assess Normality of Errors**: Verify that the residuals (errors) are normally distributed.
c) **Evaluate Homoscedasticity**: Confirm that residuals have constant variance and show no patterns.
d) **Inspect Multicollinearity**: Use metrics like Variance Inflation Factor (VIF) to ensure that there is no significant multicollinearity among the predictors.
e) **Check Independence of Residuals**: Ensure that residuals are independent and there is no autocorrelation.

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

Answer:

The top 3 features are:

- Temp
- weathersit_light_snow_rain
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a supervised machine learning technique used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It aims to predict the value of the dependent variable based on the given independent variables.

a) **Types of Linear Regression:**
   - **Simple Linear Regression**: Involves a single independent variable. The model is represented by the equation $Y = \beta_0 + \beta_1 X$, where β_0 is the intercept and β_1 is the slope of the line.
   - **Multiple Linear Regression**: Involves two or more independent variables. The model is represented by $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ where β_0 is the intercept and β_1 , β_2 , … , β_n are the coefficients of the predictors.

b) **Objective**: The goal is to find the best-fit line or plane that minimizes the error between the predicted values and the actual values. This is usually achieved by minimizing the Mean Squared Error (MSE) or a similar cost function.

c) **Assumptions**:
   - **Linearity**: There is a linear relationship between the dependent and independent variables.
   - **Normality**: The residuals (errors) of the model should be normally distributed.
   - **Homoscedasticity**: The residuals should have constant variance (no pattern in residuals).
   - **Independence**: Residuals should be independent of each other.
   - **No Multicollinearity**: Independent variables should not be too highly correlated with each other.

d) **Process**:
   - **Model Fitting**: Determine the coefficients (slope and intercept) that best fit the data.
   - **Prediction**: Use the fitted model to make predictions on new data.
   - **Evaluation**: Assess model performance using metrics like R-squared, MSE, or adjusted R-squared.

Thus, we can conclude that Linear Regression helps in understanding the relationship between variables and making predictions based on historical data.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet consists of four datasets with identical statistical properties but different patterns, created by Francis Anscombe to highlight the importance of data visualization:

1. **Dataset I**: Shows a clear linear relationship.
2. **Dataset II**: Linear, but with an influential outlier affecting the regression line.
3. **Dataset III**: Displays a non-linear (quadratic) relationship.
4. **Dataset IV**: Linear, with an outlier impacting the fit.

Statistical Properties: Each dataset has the same mean, variance, and correlation coefficient (~0.82), despite differing visual patterns.

Lessons: Anscombe's Quartet illustrates that relying solely on statistical summaries can be misleading. Visualization through scatterplots reveals underlying patterns and anomalies that summary statistics alone might miss, emphasizing the need for both statistical analysis and graphical exploration to fully understand data.

In essence, Anscombe's Quartet underscores the need to combine statistical analysis with visual exploration to accurately interpret data and uncover meaningful insights.

3. What is Pearson's R?

Answer:

Pearson's R (Pearson's correlation coefficient) is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. Here's a concise explanation suitable for a 3-mark answer:

a) Definition: Pearson's R is a correlation coefficient that ranges from -1 to +1. It measures the degree to which two variables move in relation to each other. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

b) Formula: It is calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

c) where x_i and y_i are individual data points, and $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

d) Interpretation:
   - +1: Perfect positive correlation (as one variable increases, the other also increases).
   - -1: Perfect negative correlation (as one variable increases, the other decreases).
   - 0: No linear correlation (no discernible linear relationship between the variables).

Pearson's R provides insight into the linear relationship between variables, aiding in the assessment of how changes in one variable are associated with changes in another.

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

Answer:

Scaling adjusts the range and distribution of feature values in a dataset to a common scale. This process improves the performance and convergence of machine learning algorithms by ensuring that no single feature disproportionately influences the model due to its magnitude. Scaling also enhances numerical stability, making the optimization process more efficient.

Normalized Scaling rescales feature values to a range between 0 and 1. This method is useful when features have different units or when it is important to bound values within a specific range. It ensures that all features contribute equally to the model, particularly in algorithms sensitive to the scale of input data.

Standardized Scaling adjusts feature values to have a mean of 0 and a standard deviation of 1. This method centres and scales the data, which is beneficial for algorithms assuming normally distributed data. Unlike normalization, standardization is less sensitive to outliers and is preferred when data needs to be transformed to fit a standard normal distribution.

Difference: Normalized scaling adjusts values to a fixed range [0, 1], making it suitable for algorithms that require bounded input. Standardized scaling adjusts values to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms that assume normally distributed data.

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

Answer:

When the Value of Inflation Factor (VIF) is infinite, it indicates perfect multicollinearity among predictor variables in a regression model. This situation arises when one predictor variable is an exact linear combination of one or more other predictors. In mathematical terms, this results in a correlation coefficient of 1 or -1, which makes the R-squared value for the regression of that predictor on the other predictors equal to 1. Consequently, the VIF formula 11−R2\frac{1}{1 - R^2}1−R21 yields an infinite value.

a) **Perfect Correlation**: An infinite VIF indicates perfect correlation among variables, meaning one or more predictors are redundant. This occurs when the R-squared value for a predictor regressed on other predictors is 1, leading to an infinite VIF.
b) **Implications**: A high VIF, such as 4, suggests that multicollinearity is inflating the variance of the coefficient estimates. Infinite VIF highlights perfect multicollinearity and implies that the variable in question can be expressed exactly as a linear combination of other variables.
c) **Resolution**: To address perfect multicollinearity, you should drop one or more of the offending variables to eliminate redundancy and ensure that each predictor contributes unique information to the model.

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

Answer:

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the sample data against the quantiles of a theoretical distribution, creating a scatterplot where each point represents a pair of quantiles.

Use of a Q-Q Plot in Linear Regression:

a) **Checking Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (errors) of the model are normally distributed. A Q-Q plot helps visualize whether this assumption holds. If the residuals are normally distributed, the points in the Q-Q plot will approximately lie on a straight line. Deviations from this line suggest departures from normality.

b) **Model Diagnostics**: The Q-Q plot is essential for diagnosing potential issues with the regression model. If the residuals are not normally distributed, this might indicate problems such as model misspecification, outliers, or heteroscedasticity. Addressing these issues can lead to more reliable model results and better predictions.

Importance of a Q-Q Plot:

a) **Validation of Assumptions**: It provides a visual method to check the normality assumption of residuals, which is crucial for valid hypothesis testing and confidence intervals in linear regression.

b) **Model Reliability**: By ensuring residuals are normally distributed, a Q-Q plot helps validate the appropriateness of the regression model and the reliability of its inferences.