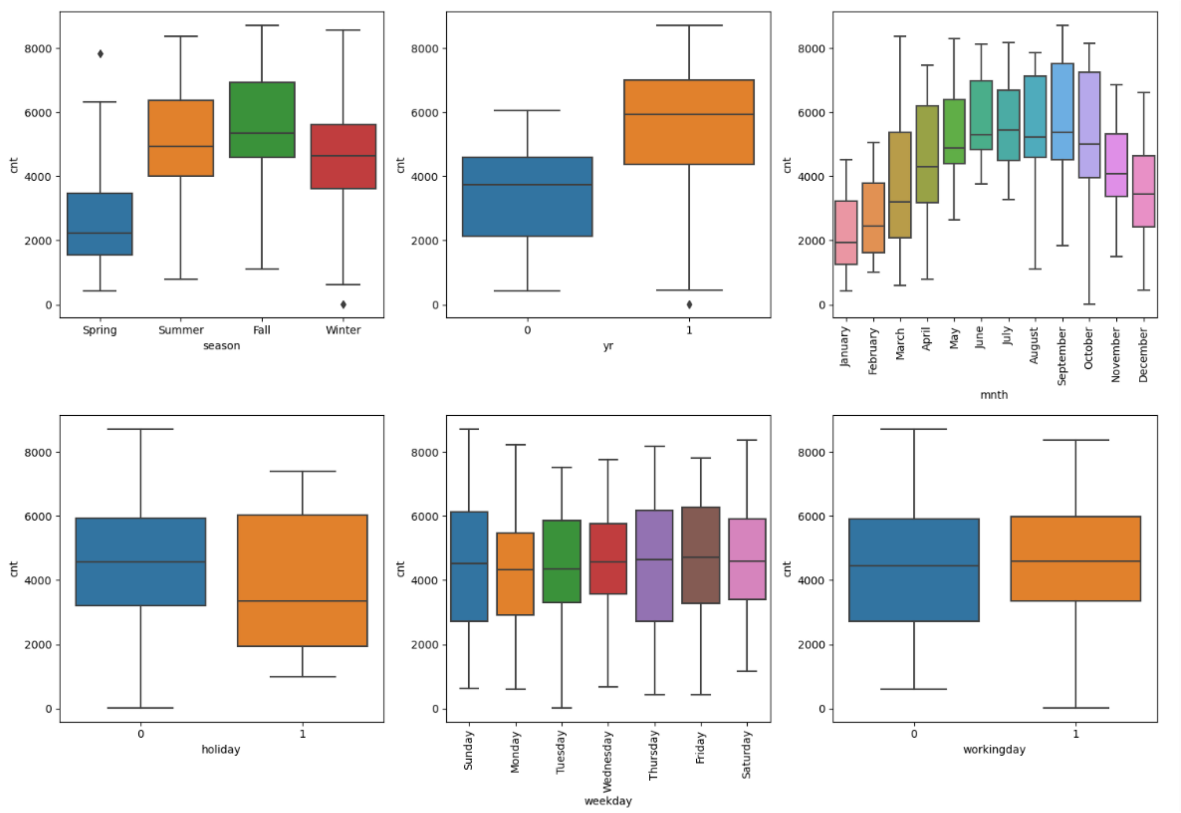
Assignment-based Subjective Questions

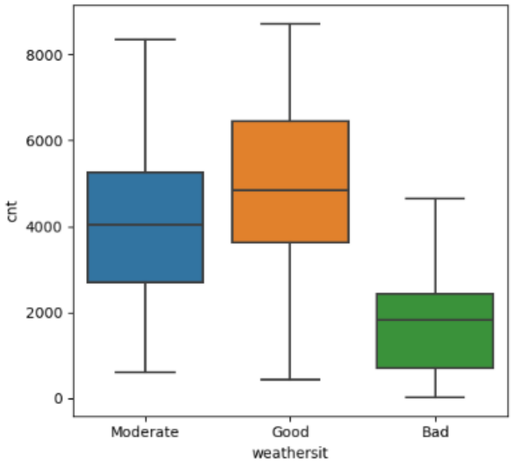
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Effect of categorical variables on the dependent variable (bike demand):

1. **Season**:
   - **Fall**: Bike rentals were slightly higher in the fall compared to other seasons.
   - **Winter**: Rentals were lower in winter, likely due to colder weather conditions.
2. **Year (yr)**:
   - **2019**: There were more bike rentals in 2019 compared to 2018, indicating an increasing trend in bike demand over time.
3. **Month (mnth)**:
   - **July to September**: These months saw the highest number of bike rentals, possibly due to favorable weather and holiday seasons.
   - **December**: Rentals were lower, likely due to colder weather and holiday season.
4. **Holiday**:
   - **Non-Holidays**: Bikes were rented more frequently on non-holidays, suggesting that people use bikes more for commuting purposes rather than leisure on holidays.
5. **Weekday**:
   - **Working Days**: Higher rentals were observed on working days, indicating that bikes are primarily used for commuting to work.
   - **Weekends**: Rentals were lower on weekends, suggesting less commuting activity.
6. **Working Day**:
   - **Working Days**: Higher bike rentals were observed on working days compared to non-working days, reinforcing the idea that bikes are mainly used for commuting.
7. **Weather Situation (weathersit)**:
   - **Good Weather**: The highest bike rentals were observed during good weather conditions (clear, few clouds, partly cloudy).
   - **Moderate Weather**: Rentals were lower during moderate weather conditions (mist, mist + cloudy).
   - **Bad Weather**: Rentals were significantly lower during bad weather conditions (light snow, light rain + thunderstorm).

These insights help in understanding how different categorical variables influence bike demand, which can be useful for strategic planning and decision-making for BoomBikes.





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)

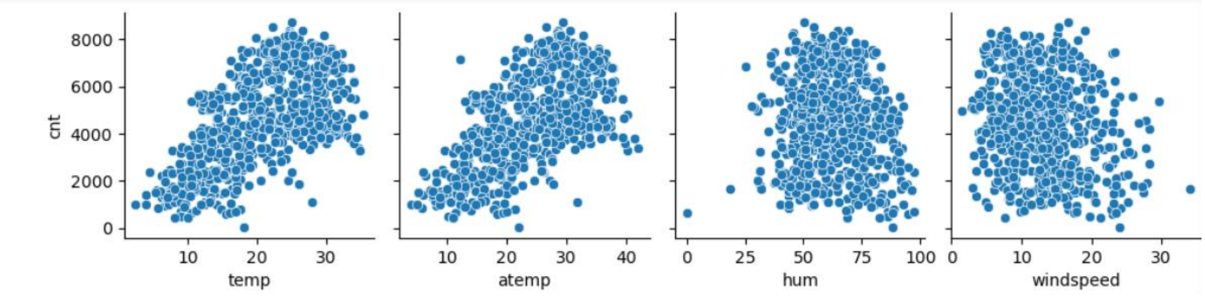
Using drop_first=True during dummy variable creation is important for the following reasons:

1. **Avoiding Multicollinearity**:
   - When creating dummy variables for categorical data, each category is represented by a separate binary (0 or 1) column. If all categories are included,

    it can lead to multicollinearity, where one variable can be perfectly predicted by the others. This can cause issues in regression models.
    - By setting drop_first=True, one category is dropped, which serves as the baseline. This helps in avoiding multicollinearity by ensuring that the dummy variables are independent of each other.
2. **Interpretability**:
    - Dropping the first category makes the model easier to interpret. The coefficients of the remaining dummy variables represent the change in the dependent variable relative to the dropped (baseline) category.
    - For example, if the season variable has categories 'Spring', 'Summer', 'Fall', and 'Winter', and 'Spring' is dropped, the coefficients for 'Summer', 'Fall', and 'Winter' will indicate how much the bike demand changes compared to 'Spring'.
3. **Model Efficiency**:
    - Including all dummy variables can lead to redundancy and unnecessary complexity in the model. Dropping the first category simplifies the model without losing any information.
    - This can also improve the computational efficiency of the model, as there are fewer variables to process.

In summary, using drop_first=True helps in avoiding multicollinearity, improves interpretability, and enhances model efficiency.

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)



From the pair-plot analysis among the numerical variables, the variable atemp **(feels-like temperature)** has the highest correlation with the target variable cnt **(bike demand)**. This indicates that as the feels-like temperature increases, the demand for bikes also tends to increase. This relationship is visually evident in the pair-plot, where atemp shows a strong positive linear trend with cnt.

Both atemp (feels-like temperature) and temp (actual temperature) showed high correlation with the target variable cnt (bike demand). However, due to redundancy, the Recursive Feature Elimination (RFE) approach retained atemp and eliminated temp.

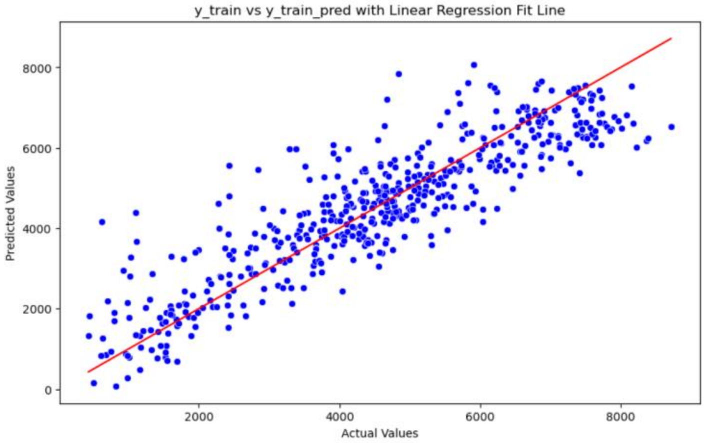
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

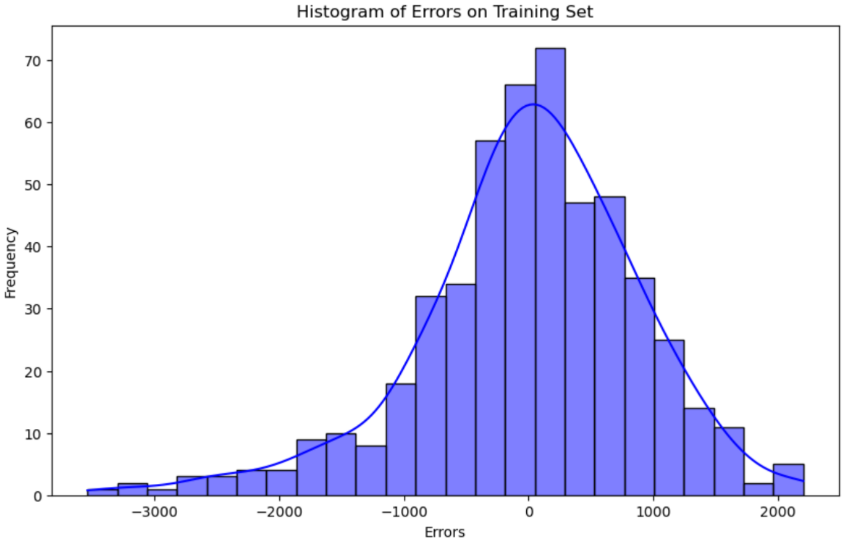
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Steps taken to validate the assumptions of Linear Regression after building the model on the training set:

1. **Linearity**:
   - **Residual Plot**: A plot of residuals versus fitted values was created to check for any patterns. A random scatter of points indicates that the linearity assumption is met.



2. **Normality of Residuals**:
   - **Histogram**: A histogram of residuals was also plotted to visually inspect the distribution.



3. **Multicollinearity**:

  - **Variance Inflation Factor (VIF)**: VIF values were calculated for each predictor variable. VIF values greater than 5 indicate potential multicollinearity issues, and steps were taken to address them by removing or combining variables.

4. **Homoscedasticity**:
   - **Residual Plot**: The same residual plot was used to check for homoscedasticity. If the spread of residuals is constant across all levels of the independent variables, the assumption is satisfied.

These steps ensured that the assumptions of Linear Regression were validated, leading to a reliable and robust model.

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)

Equation of the best fit line:

cnt = 4552.95*const + 981.40*yr + 143.60*workingday + 1010.94*atemp + -382.43*hum + -252.30*windspeed + -450.87*season_Spring + 238.21*season_Winter + -115.35*mnth_December + -203.70*mnth_July + 122.47*weekday_Sunday

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. **Feels-Like Temperature (**atemp**)**:
   - **Coefficient**: 1010.94
   - **Interpretation**: Higher feels-like temperatures are associated with increased bike demand.
2. **Year (**yr**)**:
   - **Coefficient**: 981.40
   - **Interpretation**: The demand for shared bikes increased significantly in 2019 compared to 2018.
3. **Spring (**season_Spring**)**:
   - **Coefficient**: -383.27
   - **Interpretation**: Bike demand decreases during the spring season.

These features help explain the variations in bike demand and can guide strategic decisions for the bike-sharing service.

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 statistical method used to model the relationship between a dependent variable and one or more independent variables. Here’s a detailed explanation:

1. **Definition**:
   - Linear regression aims to find the best-fitting straight line (regression line) through the data points that minimizes the sum of squared differences between observed and predicted values.
2. **Types**:
   - **Simple Linear Regression**: Involves one independent variable.
   - **Multiple Linear Regression**: Involves two or more independent variables.
3. **Equation**:
   - The linear regression equation is: $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$
     - Y: Dependent variable
     - $\beta0$: Intercept
     - $\beta1,\beta2,...,\beta n$: Coefficients of independent variables
     - $X1,X2,...,Xn$: Independent variables
     - ϵ: Error term
4. **Assumptions**:
   - **Linearity**: The relationship between the dependent and independent variables is linear.
   - **Independence**: Observations are independent of each other.
   - **Homoscedasticity**: Constant variance of errors.
   - **Normality**: Errors are normally distributed.
5. **Steps**:
   - **Data Collection**: Gather data for the dependent and independent variables.
   - **Data Preprocessing**: Handle missing values, outliers, and scale the data if necessary.
   - **Model Building**: Fit the linear regression model to the training data.
   - **Model Evaluation**: Assess the model using metrics like R-squared, Adjusted R-squared, Mean Squared Error (MSE), and Variance Inflation Factor (VIF).
   - **Prediction**: Use the model to make predictions on new data.
6. **Interpretation**:
   - **Coefficients**: Indicate the change in the dependent variable for a one-unit change in the independent variable.
   - **R-squared**: Measures the proportion of variance in the dependent variable explained by the independent variables.
   - **P-values**: Assess the significance of each coefficient.
7. **Applications**:
   - Linear regression is widely used in various fields such as economics, finance,

biology, and social sciences for predictive modeling and understanding relationships between variables.

Linear regression is a powerful and interpretable method for modeling relationships between variables, making it a cornerstone of statistical analysis and machine learning.

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 have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading. Here are the key points:

1. **Datasets**:
   - Each dataset consists of 11 (x, y) points.
   - They have the same mean, variance, correlation, and linear regression line.
2. **Visual Differences**:
   - **Dataset 1**: Appears as a simple linear relationship with some random noise.
   - **Dataset 2**: Forms a perfect curve, indicating a non-linear relationship.
   - **Dataset 3**: Contains a linear relationship but with an outlier that affects the regression line.
   - **Dataset 4**: Shows a vertical line with one outlier, indicating a different type of relationship.
3. **Importance**:
   - **Graphing Data**: Highlights the necessity of visualizing data to understand its true nature.
   - **Misleading Statistics**: Demonstrates that identical statistical properties can represent very different datasets.
   - **Data Analysis**: Emphasizes the importance of using multiple methods to analyze data, including visual inspection.

Anscombe's quartet is a powerful reminder that relying solely on summary statistics can be misleading, and visualizing data is crucial for accurate analysis.

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 variables. Here are the key points:

1. **Definition**:
   - Pearson's R quantifies the strength and direction of the linear relationship between two continuous variables.

   - It 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.
2. **Calculation**:
   - Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations.
3. **Interpretation**:
   - **Positive Correlation**: As one variable increases, the other also increases.
   - **Negative Correlation**: As one variable increases, the other decreases.
   - **No Correlation**: No linear relationship between the variables.
4. **Use in Linear Regression**:
   - Pearson's R helps in identifying the strength of the relationship between the independent and dependent variables.
   - It is useful for feature selection and understanding the impact of variables on the target.

Pearson's R is a fundamental tool in statistics for understanding relationships between variables.

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)

1. **Scaling**:
   - Adjusting the range of features to a standard scale.
2. **Why Scaling is Performed**:
   - **Consistent Feature Magnitudes**: Ensures that all features contribute equally to the model.
   - **Improved Model Performance**: Helps algorithms converge faster and perform better.
   - **Reduced Bias**: Prevents features with larger scales from dominating the model.
   - **Enhanced Interpretability**: Makes it easier to compare the importance of different features.
3. **Normalized Scaling**:
   - **Definition**: Rescales features to a range of [0, 1].
   - **Formula**: $X_{\text{normalized}} = X - \min(X) / \max(X) - \min(X)$

4. **Standardized Scaling**:
    - **Definition**: Centers features around the mean with unit variance.
    - **Formula**: $\mathrm{X}_{\mathrm{standardized}} = X - \text{mean}\ (X)\ /\ \text{std}(X)$

These techniques ensure that features 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)

- **Perfect Multicollinearity**:
    - Occurs when one or more independent variables are perfectly correlated with others.
    - This means one variable can be expressed as a perfect linear combination of others.
- **Matrix Inversion Issues**:
    - In the computation of VIF, there's an attempt to invert a matrix.
    - Perfect multicollinearity leads to the matrix being singular (non-invertible), causing the VIF to be infinite.
- **Redundant Information**:
    - Redundant variables provide no additional information.
    - This redundancy results in computational problems, leading to an infinite VIF value.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a theoretical distribution, like the normal distribution.

1. **Definition**:
    - Compares quantiles of observed data against a theoretical distribution.
    - Points along a straight line suggest the data fits the distribution.

2. **Use in Linear Regression**:
    - **Normality Assessment**: Checks if residuals are normally distributed.
    - **Identifying Outliers**: Detects outliers by deviations from the line.
    - **Model Fit Assessment**: Visualizes how well residuals conform to normality.
    - **Validity of Statistical Tests**: Ensures accurate p-values and confidence intervals.
3. **Interpretation**:
    - Points along the line indicate normality.
    - Deviations suggest departures from normality.

Q-Q plots are essential for diagnosing residual normality, identifying outliers, and ensuring valid statistical inferences in linear regression.