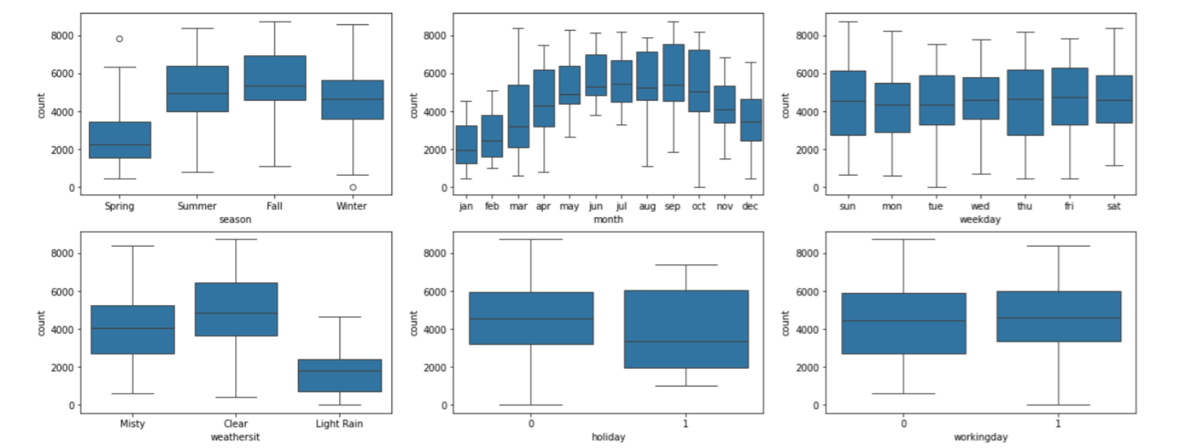
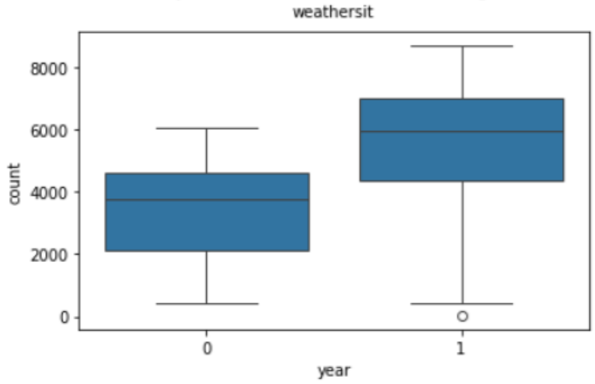
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Visualising Categorical columns to see the coorelation of predictor variable with the target variable using Box plots were done





From the visualizations, we can derive several insights:

- Seasonal Trends: The demand for rental bikes peaks in the fall season.

- Monthly Trends: The demand for rental bikes rises steadily each month until June, with September experiencing the highest demand before it starts to decline.

- Weekdays and Working Days: There isn't much variation in demand during weekdays and working days.

- Weather Conditions: Clear weather conditions see the highest demand for bike rentals.

- Holidays: The demand for bike rentals decreases on holidays.

- Yearly Growth: There is noticeable growth in demand for the following year.

Question 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. By dropping the first category, we prevent the dummy variables from being perfectly collinear, which can cause issues in the regression model. This ensures that the model remains stable and interpretable.

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

Answer: From the pair-plot analysis, it is observed that the variable temp (temperature) has the highest correlation with the target variable cnt (count of bike rentals). This indicates that temperature is a strong predictor of bike rental demand.

Question 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, the following steps were taken:

1. **Linearity**: Checked by plotting the residuals versus the fitted values to ensure there is no pattern.
2. **Normality**: Verified using a Q-Q plot to ensure the residuals are normally distributed.
3. **Homoscedasticity**: Ensured by plotting the residuals versus the fitted values to check for constant variance.
4. **Multicollinearity**: Assessed using Variance Inflation Factor (VIF) to ensure no high correlation among the independent variables.

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

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

1. **Temperature (temp)**
2. **Year**
3. **Light Rain**

These features have the highest coefficients and are statistically significant in predicting bike rental demand.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The algorithm aims to find the best-fitting line through the data points by minimizing the sum of the squared differences between the observed and predicted values. The equation of the line is given by (y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n), where (\beta_0) is the intercept and (\beta_1, \beta_2, ..., \beta_n) are the coefficients of the independent variables. The coefficients are estimated using the least squares method.

Question 7. Explain the Anscombe's quartet in detail.

Answer: 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 Francis Anscombe to demonstrate the importance of graphing data before analyzing it and the effect of outliers and the distribution of data on statistical properties. Each dataset in the quartet has the same mean, variance, correlation, and linear regression line, but their scatter plots reveal different patterns, highlighting the importance of visual data analysis.

Question 8. What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations.

Question 9. 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 features of a dataset to a common scale. It is performed to ensure that all features contribute equally to the model and to improve the convergence of gradient-based optimization algorithms. Normalized scaling transforms the data to a range of [0, 1], while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: The value of VIF can be infinite when there is perfect multicollinearity among the independent variables, meaning that one variable is a perfect linear combination of others. This causes the denominator in the VIF calculation to be zero, resulting in an infinite value. Perfect multicollinearity indicates that the model cannot uniquely estimate the coefficients of the independent variables.

Question 11. 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 whether a dataset follows a particular distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. In linear regression, a Q-Q plot is used to check the normality of the residuals. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed, which is an important assumption for the validity of the regression model.