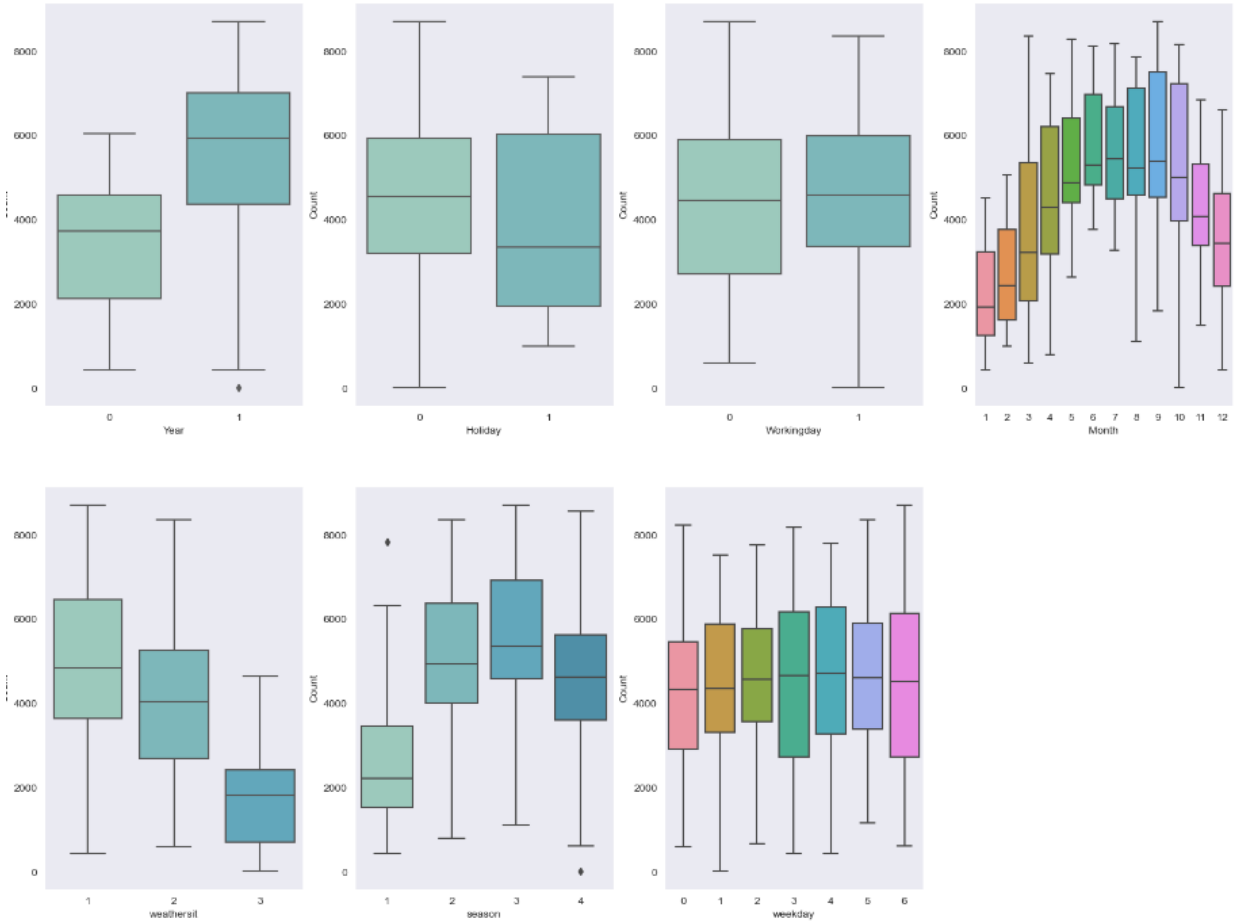
Bike Demand – Boombike Assignment Questions

Assignment-based Subjective Questions & Answers

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

Answer 1 - Categorical variables in the dataset are Year, Holiday, Workingday, Month, weathersit,

Season & weekday



- Bike Demand are more in the year 2019 compared to 2018

- Bike Demand are more Aug, Sep and Oct month in comparison to other month

- Bike Demand are more in clear weather then mist

- Bike Demand are more during the Fall season and then in summer

- Bike Rentals are more on wednesday, thursday and Saturday

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

Answer 2 - Setting drop_first=True when creating dummy variables for categorical features helps prevent multicollinearity by omitting one reference category, improves interpretability by providing a clear baseline for comparisons, and reduces model complexity, potentially enhancing generalization by avoiding overfitting. This practice ensures more effective and understandable models in machine learning.

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

Answer 3 - temp is the variable which has the highest correlation with target variable count.

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

Answer 4 - By checking below two points in residual analysis

- We can see Error Terms are normally distributed with mean Zero. Hence Model is actually following the assumption of Normality.
- We can see there is no specific pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other.

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

Answer 5 – It can seen temp, workingday & weather (windspeed, mist & light)

General Subjective Questions

Question 1 - Explain the linear regression algorithm in detail.

Answer 1 - Linear regression is a supervised machine learning algorithm used for predicting continuous output based on input features. It assumes a linear relationship between inputs and the output. The model's equation is $y = b0 + b1*x1 + b2*x2 + ... + bn*xn + \varepsilon$, where y is the predicted output, b0 is the intercept, b1 to bn are coefficients for input features x1 to xn, and ε represents error. The goal is to minimize the sum of squared differences between predicted and actual values.

During training, the algorithm adjusts coefficients using methods like gradient descent. It iteratively updates coefficients to minimize the error between predictions and actual values. Common evaluation metrics include MSE or RMSE.

Assumptions include linearity, independence, homoscedasticity, normality, and minimal multicollinearity. Extensions include multiple linear regression and polynomial regression for nonlinear relationships.

Regularization techniques like Ridge and Lasso address multicollinearity and overfitting by adding penalty terms to the cost function. Once trained, the model can make predictions by inputting new values into the equation.

Linear regression's simplicity and interpretability make it a foundational tool, but it may not handle complex relationships well. Careful consideration of assumptions and potential model improvements is crucial. Other regression techniques like decision trees, random forests, or neural networks might be more suitable for intricate data relationships.

Question 2 - Explain the Anscombe's quartet in detail.

Answer 2 - Anscombe's quartet is a set of four distinct datasets with nearly identical statistical properties, showcasing the importance of visualization in data analysis. Created by statistician Francis Anscombe in 1973, each dataset consists of 11 pairs of x and y values. While having similar means, variances, and correlations, the quartet reveals diverse patterns when graphed. This highlights that relying solely on summary statistics can lead to erroneous conclusions, emphasizing the necessity of visual exploration to apprehend the underlying relationships within data. Anscombe's quartet underscores the significance of data visualization in uncovering hidden complexities and avoiding unwarranted assumptions in statistical analysis.

Question 3 - What is Pearson's R?

Answer 3 - Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation. Pearson's r is sensitive to outliers and assumes that the relationship between variables is linear. It's widely used in various fields, including statistics, social sciences, economics, and natural sciences, to assess the degree of association between two variables and to provide insights into their potential connections.

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

Answer 4 - Scaling is a process used in data analysis and machine learning to adjust the values of numerical features so that they're within a consistent and manageable range. It's crucial because many algorithms are sensitive to the scale of features. Scaling ensures that no single feature dominates others due to its magnitude, which can skew results.

There are two main scaling techniques: normalized scaling and standardized scaling. Normalization scales features to a range between 0 and 1, maintaining the original relationships between data points. Standardization transforms features to have a mean of 0 and a standard deviation of 1, accommodating features with diverse scales and distributions. Normalization is ideal for maintaining relative relationships, while standardization centers data for better performance in algorithms sensitive to variations in scale. The choice depends on data characteristics and analysis requirements.

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

Answer 5 - An infinite Variance Inflation Factor (VIF) usually occurs due to a phenomenon called "perfect multicollinearity." This happens when two or more predictor variables in a regression analysis are so highly correlated that they essentially provide the same information to the model. When this correlation is perfect, it means one variable can be exactly predicted from the others using a linear relationship.

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

Answer 6 - A Q-Q plot, or Quantile-Quantile plot, is a graph that helps us determine if a dataset follows a certain expected pattern, like a normal distribution. In linear regression, it's used to check if the differences between actual and predicted values (residuals) are normally distributed. The Q-Q plot compares these residuals' quantiles with those of a normal distribution. If the points fall roughly along a straight line, it suggests the residuals are normally distributed, which is important because many statistical methods rely on this assumption. If the points deviate, it indicates a departure from normality. Thus, Q-Q plots help us decide if our linear regression model is valid or if adjustments are needed to ensure accurate results and reliable conclusions.