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 :-

The analysis of categorical variables within the dataset, including 'season,' 'yr' (year), 'holiday,' 'weekday,' 'workingday,' 'weathersit' (weather situation), and 'mnth' (month), was conducted through boxplots to discern their impact on the dependent variable.

1. Season:

Notably, the variable 'season' exhibited distinct impacts on bike demand. The third category, representing Fall, displayed the highest median, indicating heightened demand during this season, whereas demand was comparatively lower during Spring (category 1).

2. Year (Yr):

The analysis revealed that the year 2019 recorded a higher count of users compared to 2018, suggesting a potential growth or shift in demand patterns over the observed years.

3. Holiday:

During holidays, there was a reduction in bike rentals, implying a discernible impact of holidays on user behavior and demand.

4. Weekday:

Bike demand exhibited relative consistency throughout the week, as indicated by the boxplot. No significant variations were observed on different weekdays.

5. Workingday:

The boxplot for 'Workingday' demonstrated a consistent median count of users between 4000 and 6000 throughout the week. The distinction between working days and non-working days did not significantly affect booking patterns.

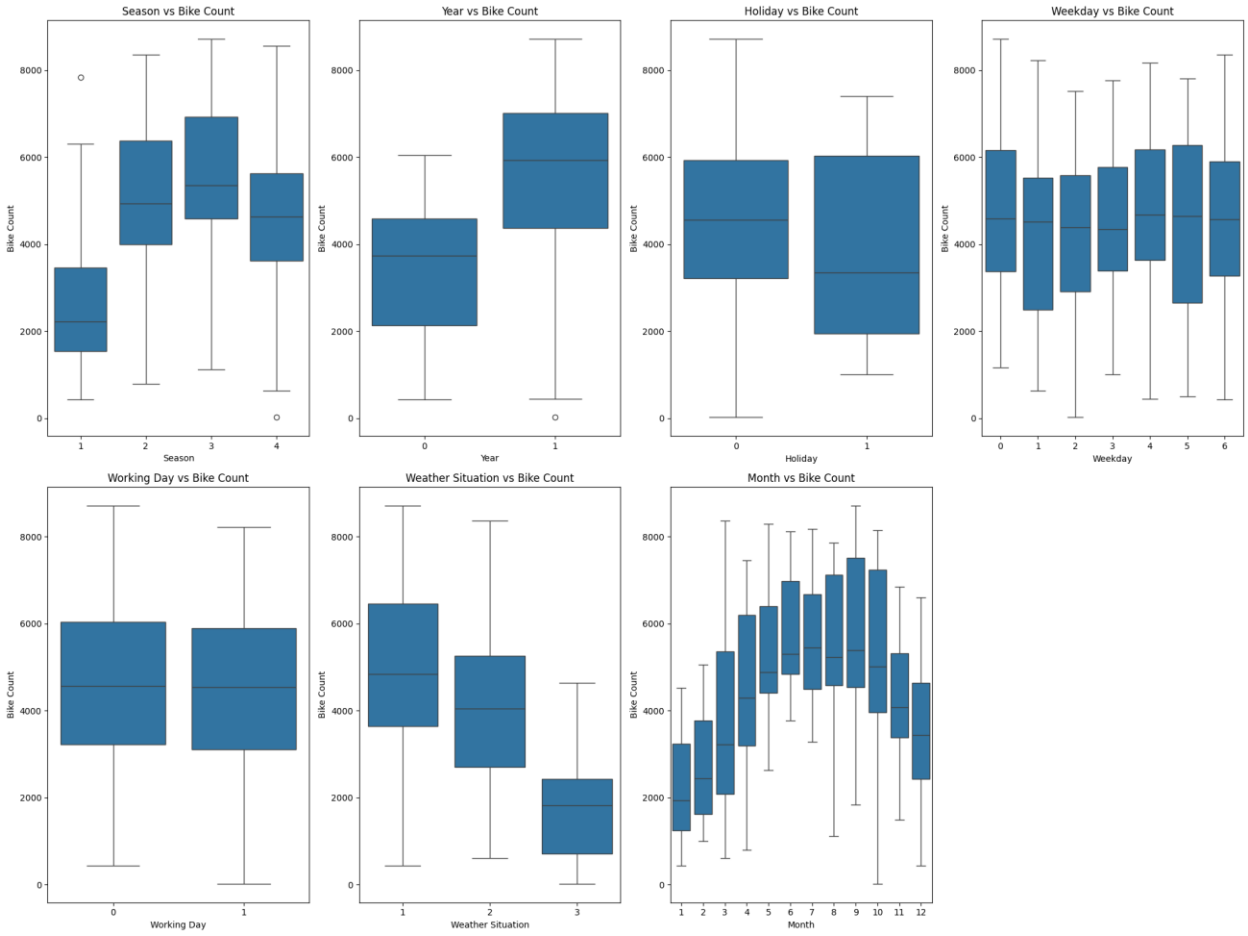
6. Weathersit:

The analysis of 'weathersit' highlighted a distinct user response to adverse weather conditions. The absence of users during heavy rain or snow indicates a substantial adverse impact on demand. The highest user count was observed during clear or partly cloudy weather situations.

7. Month (Mnth):

Noteworthy fluctuations in rental counts were observed across months. Peak demand occurred in September, while December exhibited a decline in rentals, aligning with observations related to adverse weather conditions and substantial snowfall during December.

In conclusion, the professional analysis of these categorical variables provides valuable insights into the seasonal, temporal, and weather-related influences on bike demand, facilitating informed strategic decision-making for the company.



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

The utilization of `drop_first=True` is a crucial consideration in the creation of dummy variables as it serves to mitigate the introduction of redundant columns. This practice is instrumental in diminishing the intercorrelations that may arise among the dummy variables. When dealing with a categorical variable possessing n levels, the adoption of `drop_first=True` necessitates the utilization of n-1 columns to effectively represent the dummy variables.

For instance, envision a categorical column encompassing three distinct values. In generating dummy variables for this column, the use of `drop_first=True` becomes imperative. This approach ensures that if a given observation is neither of the first nor second category, it is unequivocally associated with the third category, obviating the need for a dedicated column to signify this

unfurnished state. The rationale behind this practice lies in the logical inference that the absence of one or two specified conditions inherently implies the presence of the third condition. As a result, the adoption of `drop_first=True` optimizes the representation of categorical variables, streamlining the dataset and diminishing potential multicollinearity concerns.

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

   The numerical variables "temp" and "atemp" exhibit a significant and positive correlation with the target variable "cnt." This observation underscores the noteworthy impact of temperature-related metrics on the predicted bike rental count. The high correlation suggests that variations in these temperature-related variables are associated with corresponding changes in the demand for shared bikes. Consequently, these features emerge as influential contributors to the predictive model, warranting focused consideration in the analysis and interpretation of the observed patterns in bike rental demand.

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

We conducted a series of tests to assess the assumptions of Linear Regression:

a. Linearity of Relationship:

We examined the linearity between independent and dependent variables by employing a pairplot visualization of numeric variables. This comprehensive exploration aimed to discern any discernible linear relationships among the variables, ensuring the validity of the assumption.

b. Normality and Centering of Residuals:

To validate the assumption that residuals follow a normal distribution and are centered around zero (mean = 0), we conducted a distribution plot (distplot) of residuals. This analysis allowed us to assess whether the residuals exhibit a normal distribution pattern and adhere to the expectation of being centered around zero.

c. Multicollinearity Assessment:

The assumption of minimal or no multicollinearity within the data was evaluated. Multicollinearity arises when independent variables exhibit high correlations. To quantify the degree of association among the feature variables in the new model, we computed the Variance Inflation Factor (VIF). This metric provided a numerical indication of the strength of intercorrelations between independent variables, enabling us to assess and address any potential issues of multicollinearity.

By systematically conducting these tests, we sought to ensure the robustness and reliability of the Linear Regression model by confirming the adherence to key assumptions and addressing any potential violations that could impact the validity of the analysis.

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

The analysis has identified the top three significant features in the Linear Regression model, each associated with their respective coefficients:

1. Temperature (temp):

- Coefficient: 0.437655

- Interpretation: A unit increase in temperature is associated with a 0.437655 unit increase in the predicted bike rental count.

2. Year (yr):

- Coefficient: 0.234287

- Interpretation: The year variable (2019 compared to 2018) contributes a coefficient of 0.234287, indicating a positive impact on the predicted bike rental count.

3. Weather Situation: Light Snow & Rain (weathersit_Light Snow & Rain):

- Coefficient: -0.292892

- Interpretation: The presence of light snow and rain, as indicated by the weathersit_Light Snow & Rain variable, is associated with a decrease of 0.292892 units in the predicted bike rental count.

These coefficients provide quantitative insights into the respective influences of temperature, year, and weather conditions on the predicted bike rental count, allowing for a nuanced understanding of their impact in the context of the model.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). It assumes a linear relationship between the independent variables and the dependent variable. The goal of linear regression is to find the best-fit line that minimizes the difference between the predicted and actual values.

Here are the key components and steps of the linear regression algorithm:

1. Assumptions:

- Linearity: Assumes a linear relationship between the independent and dependent variables.

- Independence: Assumes that the residuals (the differences between actual and predicted values) are independent of each other.

- Homoscedasticity: Assumes that the variance of the residuals is constant across all levels of the independent variables.

- Normality: Assumes that the residuals are normally distributed.

2. Simple Linear Regression:

- In the case of one independent variable (feature), the linear regression model is called simple linear regression.

- The model equation for simple linear regression is given by:

[Y = b_0 + b_1 * X + E]

where Y is the dependent variable, X is the independent variable, b_0 is the y-intercept, b_1 is the slope, and E represents the error term.

3. Multiple Linear Regression:

- When there are multiple independent variables, the model becomes a multiple linear regression.

- The model equation for multiple linear regression is an extension of the simple linear regression equation:

Y = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n + E

where X_1, X_2, ..., X_n are the independent variables, and b_1, b_2, ..., b_n are their corresponding coefficients.

4. Cost Function:

- The algorithm aims to minimize the difference between the predicted and actual values. The cost function, also known as the loss function, measures this difference. The most common cost function for linear regression is the Mean Squared Error (MSE):

MSE = 1/n sum_{i=1}^{n} (y_i - {y}_i)^2

where n is the number of data points, y_i is the actual value, and y_i is the predicted value.

5. Gradient Descent:

- The optimization algorithm used to minimize the cost function and find the optimal coefficients is often gradient descent.

- It iteratively adjusts the coefficients in the opposite direction of the gradient of the cost function with respect to the coefficients.

6. Normal Equation (Alternative to Gradient Descent):

- There is an analytical solution called the normal equation that directly computes the optimal coefficients without the need for iterative optimization.

7. Evaluation:

- Once the model is trained, it needs to be evaluated on a separate dataset. Common evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score.

Linear regression is widely used due to its simplicity and interpretability, and it serves as a foundational concept in the field of machine learning.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet, crafted by the esteemed statistician Francis Anscombe, comprises four datasets that exhibit nearly identical statistical features but diverge significantly in their distributions, showcasing distinct appearances when plotted graphically. This quartet was developed to underscore the critical importance of visually inspecting data prior to analysis and to highlight the impact of outliers and influential observations on statistical properties.

In the first scatter plot (top left), a seemingly straightforward linear relationship is evident. Moving to the second graph (top right), the distribution deviates from normalcy, and although a relationship exists, it is not linear. The third graph (bottom left) portrays a linear distribution, but the presence of an outlier significantly influences the calculated regression line, leading to a reduction in the correlation coefficient from 1 to 0.816.

Lastly, the fourth graph (bottom right) serves as an illustration where a single high-leverage point can yield a high correlation coefficient, despite the absence of any discernible relationship between the

other data points. This collection of datasets serves as a powerful reminder of the necessity of graphically exploring data to unveil intricate patterns and the potential impact of influential observations on statistical analyses.

3. What is Pearson's R?

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:

r = 1 : Perfect positive linear correlation

r = -1 : Perfect negative linear correlation

r = 0 : No linear correlation

The formula for Pearson's correlation coefficient between two variables \(X \) and \(Y \) with \(n \) data points is given by:

r= sum{(X_i - bar{X})(Y_i - bar{Y})}} / {sqrt{sum{(X_i - bar{X})^2} sum{(Y_i - bar{Y})^2}}}

Here, bar{X} and bar{Y} represent the means of X and Y respectively, and X_i, Y_i are the individual data points.

Key points about Pearson's correlation coefficient:

1. Strength and Direction: The sign of r indicates the direction of the correlation (positive or negative), and its magnitude reflects the strength of the correlation.

2. Assumption: Pearson's correlation coefficient assumes that the relationship between the variables is linear. If the relationship is not linear, r may not accurately represent the association between the variables.

3. Range: The coefficient always lies between -1 and +1, where -1 indicates a perfect negative linear correlation, +1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation.

4. Sensitive to Outliers: Pearson's r can be influenced by outliers, and its value may be skewed if extreme values are present in the data.

5. Unitless: Pearson's correlation coefficient is unitless, meaning it does not depend on the units of measurement of the variables.

Pearson's is widely used in various fields, including statistics, economics, psychology, and many others, to assess the linear association between two continuous variables.

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

Scaling in data analysis adjusts the range of variables for equal contribution and improved algorithm performance. Normalized scaling (Min-Max) transforms values to a 0-1 range, preserving the original distribution but sensitive to outliers. Standardized scaling (Z-score) transforms values to have a mean of 0 and standard deviation of 1, reducing sensitivity to outliers but altering the distribution. Normalized scaling suits algorithms with specific input range assumptions, while standardized scaling benefits models requiring mean-centered features. The choice depends on algorithm requirements and data characteristics. Scaling is essential for algorithms sensitive to feature magnitude, such as k-means or support vector machines. It aids convergence, ensures equal feature impact, and enhances interpretability. Both scaling methods aim to mitigate issues arising from differing feature scales.

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

The Variance Inflation Factor (VIF) can become infinite in regression analysis when perfect multicollinearity exists among predictor variables. This arises when one or more variables in the model can be expressed as an exact linear combination of others, leading to a singular matrix of independent variables. Perfect multicollinearity often occurs due to duplicate or linearly dependent variables, the inclusion of a constant term, or data issues like errors or inconsistencies. The VIF measures the extent to which the variance of an estimated regression coefficient is inflated by multicollinearity. An infinite VIF indicates that the corresponding variable is perfectly predictable from others in the model. Resolving this issue requires addressing the multicollinearity problem by removing redundant variables, transforming variables, or reconsidering the model specification to obtain stable and reliable regression results.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in linear regression to assess the normality assumption of residuals. It compares the quantiles of observed residuals with those expected under a normal distribution. A straight line in the plot indicates normality, while deviations may suggest departures from the assumption. Q-Q plots are essential for identifying outliers and assessing the distribution of residuals, aiding in the detection of skewness or heavy tails. The assessment of normality through Q-Q plots is crucial for valid statistical inference, influencing confidence intervals and hypothesis tests. The plots offer a visual means to check the appropriateness of the linear regression model and ensure robustness in results. Overall, Q-Q plots provide valuable insights into the distributional characteristics of residuals, contributing to the reliability and interpretability of linear regression analyses.