Assignment-based Subjective Questions

❖ From your analysis of the categorical variables from the data set, what could you infer about their effect on the dependent variable?

- The best weather conditions for bike rentals are during clear skies when the temperature is moderate, humidity is low, and the overall temperature is comfortable.
- The dispersion of the monthly plot reflects the seasonal pattern, with the fall months exhibiting a higher median.
- The median rental activity is nearly identical on both working and non-working days, but the spread is greater for non-working days. This could be attributed to individuals having plans and opting not to rent bikes during those days.

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

To represent a variable with n levels, it is possible to utilize n-1 dummy variables. Hence even by excluding the first column, the data can still be accurately represented. When the values of variables from 2 to n are 0,it signifies that the value of the first variable is 1.

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

atemp and temp both have same correlation with target variable of 0.63 which is the highestamong all numerical variables.

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

Upon analyzing the distribution of residuals through plotting, it was observed that it follows a normal distribution with a mean value of 0.

❖ Based on the final model, which are the top 3 features contributing significantly towards expanding the demand of the shared bikes?

Top 3 features contributing significantly towards explaining the demand of shared bikes:

1. Year (yr): The year has a positive coefficient, indicating that demand tends to increase over time.
2. Temperature (temp): Higher temperatures positively affect the demand for shared bikes.
3. Weather conditions (weather sit_Cloudy): Cloudy weather negatively impacts the demand for shared bikes.

General Subjective Questions

❖ Explain the linear regression algorithm in detail

Linear regression is a powerful supervised machine learning algorithm utilized for predicting continuous numerical values. Its primary objective is to establish a linear connection between an input feature and its corresponding output. By minimizing the sum of squared differences between the actual and predicted values, this algorithm identifies the optimal-fit line. It achieves this by calculating the slope and intercept of the line using the robust least squares method. Subsequently, this model becomes capable of generating predictions by inserting new input values and computing their corresponding outputs. It is important to note that linear regression assumes a linear relationship between the variables and necessitates the independence and normal distribution of the features for accurate results.

❖ Explain the Anscombe’s quartet in detail.

Anscombe's quartet consists of four data sets that share almost identical statistical properties but display distinct patterns when visualized graphically. Each data set comprises 11 (x, y) pairs. Individually analyzed, these data sets showcase similar means, variances, and correlation coefficients. However, their plotted representations unveil significant disparities. Among them, one data set may exhibit a linear relationship, another a quadratic association, while the remaining data sets may display no apparent relationship or demonstrate the presence of outliers. Anscombe's quartet serves as a compelling reminder of the importance of data visualization and sheds light on the limitations of relying solely on summary statistics. It underscores the necessity to explore and comprehend data through graphical analysis.

❖ What is Pearson’s R?

Pearson's R, also referred to as Pearson's correlation coefficient, is a statistical metric that precisely measures the intensity and direction of the linear connection between two continuous

variables. Symbolized by "r," this coefficient ranges between -1 and 1. A value of 1 signifies a perfect positive linear relationship, -1 represents a perfect negative linear relationship, while 0 suggests the absence of a linear relationship. To calculate Pearson's R, the covariance of the two variables is divided by the product of their respective standard deviations. This measure finds extensive application across various disciplines to evaluate the level of association between variables and to leverage observed patterns for predictive purposes.

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

- Scaling refers to the process of transforming variables within a dataset to a consistent scale. This practice ensures that all variables possess comparable ranges and units, facilitating fair comparisons among them. In the realm of machine learning, scaling holds particular significance for algorithms reliant on distance-based calculations or when variables exhibit varying scales.
- One method of scaling, known as normalized scaling or Min-Max scaling, rescales variable values to a range between 0 and 1. This is achieved by subtracting the minimum value and dividing by the range (maximum value minus minimum value). Notably, this transformation preserves the relative order of the data.
- Another method, standardized scaling or Z-score scaling, transforms variable values to have a mean of 0 and a standard deviation of 1. This is accomplished by subtracting the mean and dividing by the standard deviation. Standardized scaling centers the data around the mean while preserving the shape of the distribution.
- The key distinction between normalized scaling and standardized scaling lies in the specific transformation applied to the data. Normalized scaling adjusts the value range, whereas standardized scaling adjusts the mean and standard deviation.

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

- When perfect multicollinearity is present in a data set, it can lead to the occurrence of an infinite value for the Variance Inflation Factor (VIF). Perfect multicollinearity occurs when one or more independent variables in a regression model can be accurately predicted from the other independent variables.

- In such cases, calculating the VIF involves division by zero, resulting in an infinite value. This situation typically arises when there is a linear relationship among the predictors, leading to redundant or duplicated information. When one predictor variable can be

expressed as an exact linear combination of others, it becomes redundant for the model, and the VIF cannot be accurately calculated

- The presence of perfect multicollinearity poses challenges in regression analysis as it hinders the interpretation of coefficients and inflates the standard errors. To address this issue, it is crucial to identify and handle multicollinearity by either removing or transforming correlated variables. Alternatively, alternative regression techniques like ridge regression or principal component analysis can be employed.

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

- A Q-Q plot, which stands for Quartile-Quartile plot, serves as a valuable graphical tool for assessing the adherence of a data set to a specific theoretical distribution, such as the normal distribution. It accomplishes this by comparing the quartiles of the observed data with the expected quartiles derived from the theoretical distribution.

- In the realm of linear regression, a Q-Q plot becomes particularly useful for examining the assumption of normality regarding the residuals. Residuals denote the disparities between the observed and predicted values of the dependent variable. By analyzing the Q-Q plot of the residuals, we can visually evaluate whether they exhibit a normal distribution. If the residuals closely align with the diagonal line in the Q-Q plot, it signifies their conformity to a normal distribution.

- The significance of a Q-Q plot in linear regression lies in its ability to detect deviations from normality. Should the residuals significantly deviate from the anticipated line, it implies a violation of the normality assumption. Such a violation can impact the accuracy of statistical inferences, including hypothesis testing and confidence intervals. Identifying non-normality empowers researchers to explore alternative modeling strategies or consider variable transformations to enhance the model's performance.