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?

   - yr column-> Both year have almost same level of demand.
   - holiday column-> On holiday demand was low.
   - temp-> When temperature>30, low demand.
   - hum-> When humidity is high, demand was low for bikes.
   - windspeed-> When windspeed<20, bike demand was high.
   - summer-> In summer, bike demand was 2nd highest.
   - winter-> In winter, bike demand is lowest.
   - jul-> In July month, bike demand was high.
   - sept-> In sept, bike demand was lower then July.
   - sun-> On Sunday, bike demand was low.
   - moderate-> In moderate weather, bike demand was at medium level.

2. Why is it important to use drop_first=True during dummy variable creation?
   - When we created dummy columns of non-binary categorical columns, then it is important to drop one column among each columns dummy created columns to avoid the issue of multicollinearity. And to avoid this issue we use syntax 'drop_first=True' to drop one column from created dummy columns for each category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - temp, atemp has the highest correlation with the target column(cnt).

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - After building the model on the training set we perform following steps to validate the assumptions of linear regression->
     - Check linearity: To check if the relationship between the independent variables and the dependent variable is linear, you can create scatter plots of each independent variable against the dependent variable. If the relationship is not linear, you may need to transform the variables or use a different type of model.

- Check normality: To check if the residuals (the differences between the predicted values and the actual values) are normally distributed, you can create a histogram or a Q-Q plot of the residuals. If the residuals are not normally distributed, you may need to transform the dependent variable or use a different type of model.

- Check for homoscedasticity: To check if the variance of the residuals is constant across all values of the independent variables, you can create a scatter plot of the residuals against the predicted values. If the variance of the residuals increases or decreases as the predicted values increase, then the model may not meet the assumption of homoscedasticity. You may need to transform the dependent variable or use a different type of model.

- Check for multicollinearity: To check if there is high correlation among the independent variables, you can create a correlation matrix of the independent variables. If there is high correlation, it may be necessary to remove some of the variables or use a different type of model.

- Evaluate the model performance: To evaluate the performance of the model, you can use the testing set to generate predictions and calculate metrics such as R-squared, mean squared error, and root mean squared error. These metrics can help you understand how well the model is performing on new data.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   - Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables.
   - The equation for simple linear regression is:

     $Y = \beta0 + \beta1 * X + \varepsilon$

     where Y is the dependent variable, X is the independent variable, β0 is the intercept or constant term, β1 is the slope or coefficient of X, and ε is the error term.

   - Multiple linear regression is an extension of simple linear regression, where there are multiple independent variables and the equation is:

     $Y = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \betan * Xn + \varepsilon$

     where Y is the dependent variable, X1 to Xn are the independent variables, β0 is the intercept or constant term, β1 to βn are the coefficients of X1 to Xn respectively, and ε is the error term.

   - The goal of linear regression is to estimate the values of the coefficients β0 to βn that best fit the data. This is done by minimizing the sum of the squared errors between the predicted values and the actual values in the training dataset.

   - The most commonly used method for estimating the coefficients is Ordinary Least Squares (OLS), which involves finding the values of the coefficients that minimize the sum of the squared errors. This is done by taking the partial derivatives of the error function with respect to each coefficient and setting them to zero.

   - Once the coefficients have been estimated, we can use the linear regression model to make predictions on new data by plugging in the values of the independent variables into the equation and calculating the corresponding value of the dependent variable.

   - There are some assumptions that need to be met ->

     1. Linearity: The relationship between the dependent variable and the independent variables is linear.
     2. Independence: The errors are independent of each other.

3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
4. Normality: The errors follow a normal distribution.
If these assumptions are not met, the accuracy of the model may be compromised and additional steps may be required to improve its performance.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of four datasets that have identical statistical properties, but appear very different when plotted graphically. This set of datasets was first introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on numerical summaries.

- Each dataset in Anscombe's quartet contains 11 pairs of x and y values, and they all have the same mean and variance for both x and y, the same correlation coefficient, and the same linear regression line. However, they have very different distributions and patterns when plotted graphically.

- Dataset I: This dataset has a linear relationship between x and y, with no outliers or extreme values. The linear regression line fits the data well, and the correlation coefficient is high.

- Dataset II: This dataset has a non-linear relationship between x and y, with a single outlier that has a significant effect on the regression line. The linear regression line is not a good fit for the data, and the correlation coefficient is still high despite the outlier.

- Dataset III: This dataset has a strong relationship between x and y, but it is not linear. The relationship can be modeled using a quadratic equation. The linear regression line is a poor fit for the data, and the correlation coefficient is low.

- Dataset IV: This dataset has four distinct clusters of data, each with a linear relationship between x and y. When all the data is considered together, the linear regression line is a poor fit for the data, and the correlation coefficient is close to zero.

  The purpose of Anscombe's quartet is to illustrate the importance of visualizing data and the limitations of relying solely on numerical summaries. While all four datasets have the same statistical properties, they have very different patterns when plotted graphically, which can have a significant impact on how we interpret the data and draw conclusions from it. Therefore, it is important to

always visualize data and not rely solely on numerical summaries to understand the patterns and relationships in the data.

3. What is Pearson’s R?
   - Pearson's r, also known as Pearson correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to 1.

   - A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear fashion. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear fashion. A value of 0 indicates no correlation between the two variables.

   - Pearson's r is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is sensitive to outliers and assumes that the relationship between the variables is linear.

   - Pearson's r is widely used in various fields such as psychology, economics, and biology to measure the strength of the relationship between two continuous variables. It can be used to assess the association between variables, to determine the predictive power of one variable on the other, or to identify potential confounding factors in a study.

   - Despite its popularity, Pearson's r has some limitations. It only measures the strength of linear relationships and does not capture non-linear relationships. Additionally, it assumes that the variables are normally distributed and have equal variances. When these assumptions are not met, alternative correlation coefficients such as Spearman's rank correlation or Kendall's tau may be used.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - Scaling is a data preprocessing technique used to transform the features of a dataset so that they have a similar scale and distribution. The purpose of scaling is to make sure that all features contribute equally to the analysis and prevent certain features from dominating or skewing the results of a machine learning model.

   - Scaling is performed to ensure that features with larger values and/or larger ranges of values do not dominate over features with smaller values or ranges.

This is especially important when using machine learning algorithms that use distance-based measures, such as k-nearest neighbors, or when using gradient descent-based algorithms to optimize the model parameters.

- There are two commonly used methods of scaling, normalized scaling and standardized scaling.

- Normalized scaling, also known as min-max scaling, scales the features so that they have a range of values between 0 and 1. This is done by subtracting the minimum value of the feature and dividing by the range of the feature.

- Standardized scaling, also known as z-score scaling, scales the features to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the feature and dividing by the standard deviation of the feature.

- The main difference between normalized scaling and standardized scaling is that normalized scaling uses the range of the feature to scale the data, while standardized scaling uses the standard deviation of the feature. Normalized scaling is useful when the range of the features is known and fixed, and when the minimum and maximum values are meaningful. Standardized scaling is useful when the mean and standard deviation of the feature are important, and when the distribution of the feature is close to a normal distribution.

  In summary, scaling is performed to ensure that all features contribute equally to the analysis and to prevent certain features from dominating or skewing the results of a machine learning model. Normalized scaling and standardized scaling are two commonly used scaling methods, with differences in their approach to scaling the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - Variance Inflation Factor (VIF) is a measure of multicollinearity between predictor variables in a linear regression model. It quantifies how much the variance of the estimated regression coefficient is increased due to collinearity among the predictor variables. A high value of VIF indicates a high degree of multicollinearity.
   - In some cases, the VIF value can be infinite. This happens when one of the predictor variables in the model is a perfect linear combination

of other predictor variables. When this happens, the regression model is no longer identifiable, which means that the regression coefficients cannot be uniquely estimated.

- In other words, if there is perfect collinearity among the predictor variables, the regression model cannot distinguish the individual effects of the predictor variables on the outcome variable. In such cases, one of the predictor variables must be dropped from the model to avoid perfect multicollinearity.

- It is important to note that infinite VIF values can also occur due to computational errors. This can happen if the algorithm used to calculate VIF has an issue with precision or rounding errors. Therefore, it is important to check the data carefully and ensure that there is no perfect collinearity among the predictor variables before concluding that an infinite VIF value is due to multicollinearity.

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 method used to compare the distribution of a sample of data to a theoretical distribution. The Q-Q plot compares the quantiles of the two distributions and plots them against each other to visually assess whether the data fits the theoretical distribution.

   - In linear regression, Q-Q plots are often used to assess the assumption of normality of the residuals. The residuals are the differences between the actual values of the dependent variable and the predicted values from the linear regression model. If the residuals follow a normal distribution, this indicates that the linear regression model is a good fit for the data.

   - To create a Q-Q plot, the residuals are sorted from smallest to largest and plotted against the theoretical quantiles of a normal distribution. If the points on the Q-Q plot form a straight line, this indicates that the residuals follow a normal distribution. If the points deviate from a straight line, this indicates that the residuals are not normally distributed.

- The use of Q-Q plots in linear regression is important because normality of the residuals is one of the assumptions of linear regression. If the assumption of normality is violated, the linear regression model may not provide accurate or reliable predictions. Therefore, the use of Q-Q plots can help identify potential issues with the assumption of normality and allow for corrective measures to be taken, such as transforming the data or using a different statistical model.

  In summary, a Q-Q plot is a graphical method used to compare the distribution of a sample of data to a theoretical distribution. In linear regression, Q-Q plots are used to assess the assumption of normality of the residuals, which is important for ensuring the accuracy and reliability of the linear regression model.