Assignment-based Subjective Questions

1. Weather plays a major role in people renting a bike. When the weather is clear, people more of than not rent a bike. Time of the year, season, as defined in the data is also a significant factor for people renting out a bike. Fall and summer have the highest number of bike rentals. The same can be inferred from the monthly data too, where bike rentals are at the highest in months from May to Sep. People tend to rent a bike on holidays.

2. drop_first=True it is implemented during creation of dummy variables to avoid dummy variables trap. Dummy variable trap is a situation where the independent variables after dummy variable creation are highly correlated. This typically results into multicollinearity. By setting drop_first=True, we try to eliminate one of the dummy variables for each categorical feature, ensuring that they are not perfectly correlated.
   In our case while generating dummy variables for columns like 'season', 'weathersit', 'mnth' and 'weekday' we have used drop_first=True

3. Variable atemp is highest correlated according to the pair-plot

4. Checked the linearity of the model by creating a scatterplot for each of the independent variable against our dependent variable. Linear pattern would mean linearity. Check for VIF and remove features having a high VIF (>5). Check and remove outliers.

5. ‘yr’,’atemp’,’season_Spring’ are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

1. Linear Regression is a statistical technique used to understand a relationship between dependent and independent variable by fitting a linear equation. The aim of the model is to find a best fitted line by reducing or minimizing errors between the actual and predicted values. We typically use mean squared error for this. The model is evaluated using metrics like R2 and adjusted r2. Adjusted r2 reduces with adding of redundant features to the model. Linear regression with more than 1 variable is called multiple linear regression. The Outcome of the model is usually a numerical variable. For example, to predict **how** much it would rain in Delhi on a given day.

2. Anscombe’s quartet consists of 4 similar datasets with similar statistics however, are very different when visualized. It helps understand the power of visualization. Anscombe’s quartet helps understand why looking at only statistical aspects may be sufficient It helps the data analysts to understand the reason for looking at a visual representation of the data for a comprehensive data understanding.

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.

   The Pearson correlation coefficient ranges from -1 to 1, where:

   I. indicates a perfect positive linear relationship: as one variable increases, the other variable also increases proportionally.

II. -1 indicates a perfect negative linear relationship: as one variable increases, the other variable decreases proportionally.

III. 0 indicates no linear relationship between the variables.

4. Scaling is a pre-processing step that involves adjusting the numerical values of variables to a specific range . The primary reasons for scaling are:
    a. Bringing all variables in a comparable state and size
    b. Improves interpretability

    Normalized scaling is used when we want the data to be between 0 and 1, for data we need to have in a specified range and on the other hand standardized scaling transforms the data to a standard normal distribution. It is often preferred when the distribution of the variable is approximately normal.

    Both, normalization and standardization are techniques used to scale variables. Normalization is useful when the specific range of the data is important, while standardization is more appropriate when the focus is on achieving a standard normal distribution or when dealing with algorithms that assume normally distributed data.

5. The Variance Inflation Factor (VIF) is a measure used in regression analysis to gauge the multicollinearity within independent variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to understand or calculate the individual effects of each variable on the dependent variable. The VIF helps understand the variance of an estimated regression coefficient increases when predictors are correlated.

    If the VIF for a variable is infinite, it indicates perfect multicollinearity, meaning, the correlation between certain variables is so high that one variable can be expressed as an exact linear function of the others. This implies that the regression coefficients cannot be uniquely determined, which affects the model.

6. A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to understand whether a dataset follows a particular distribution. It is mostly useful in linear regression for verifying the assumption of normality of errors. If the data perfectly follows the theoretical distribution, the points on the Q-Q plot should fall along a straight line. In linear regression, the Q-Q plot is often used to assess the normality of residuals. Residuals are the differences between the observed values and the values predicted by the regression model. The normality of residuals is an important assumption for making valid statistical inferences and constructing reliable prediction intervals. In linear regression, we make predictions about one thing based on another (like predicting a person's weight based on their height). The Q-Q plot, in this case, helps us check if the "errors" we make in our predictions (the differences between our predictions and the actual data) look like they come from a normal pattern.