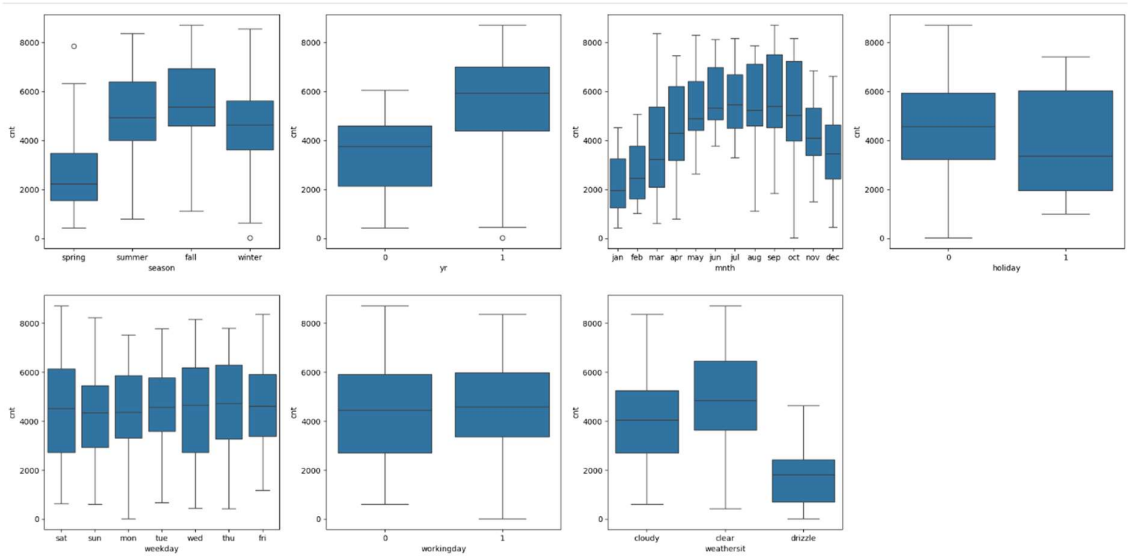
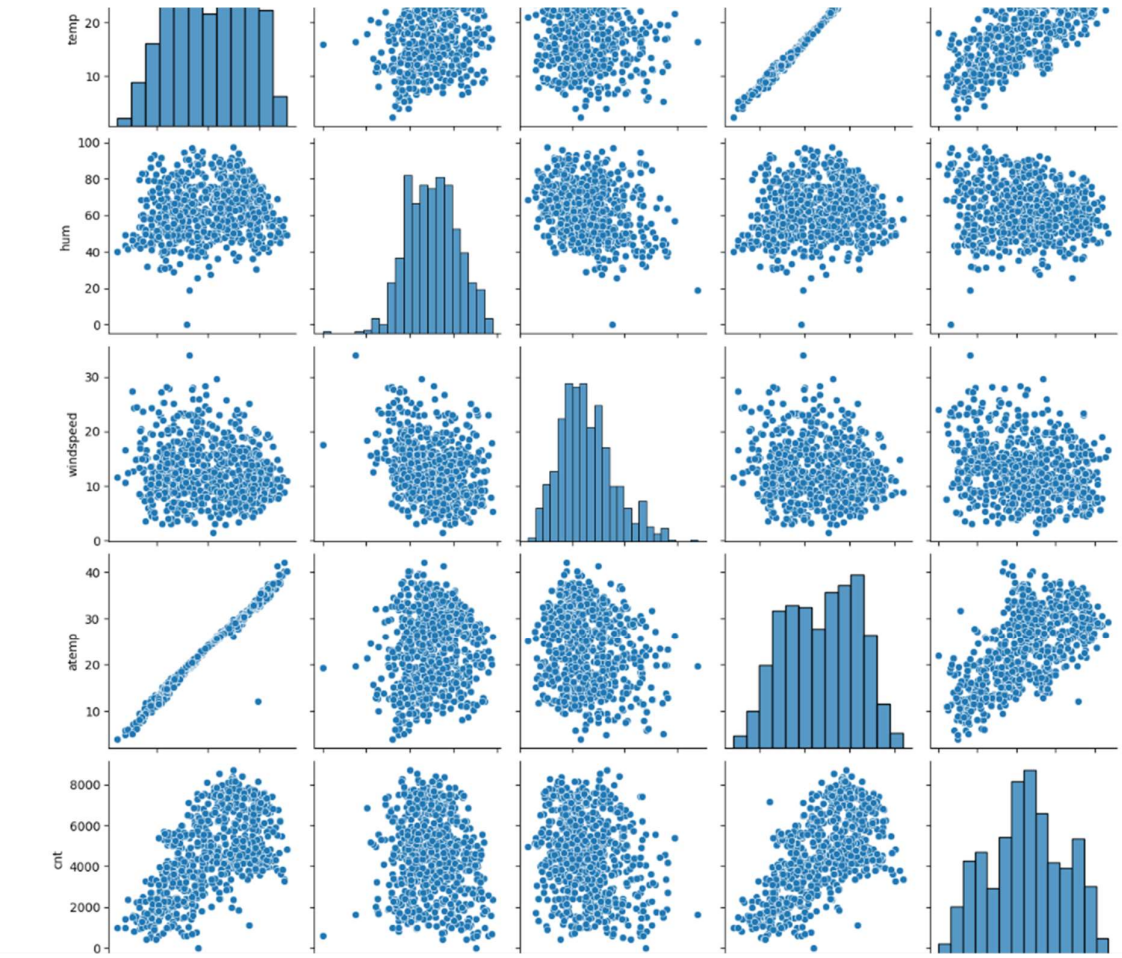
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   ANSWER: Categorical variables play important role in the modelling of the data. Converted these variables to dummy variables during preparation of the data for modelling. Of all the categorical variables present in the dataset, Year, Holiday, Seasons, Months, Weather situation were very significant in the final model that was built. These categorical variables are initially plotted in Box-plot as below:



2. Why is it important to use drop_first=True during dummy variable creation?
   ANSWER: When creating dummy variables, it's crucial to use `drop_first=True` because it ensures that for a categorical variable with 'n' levels, you create only 'n-1' columns. This approach prevents multicollinearity by avoiding the inclusion of redundant information.

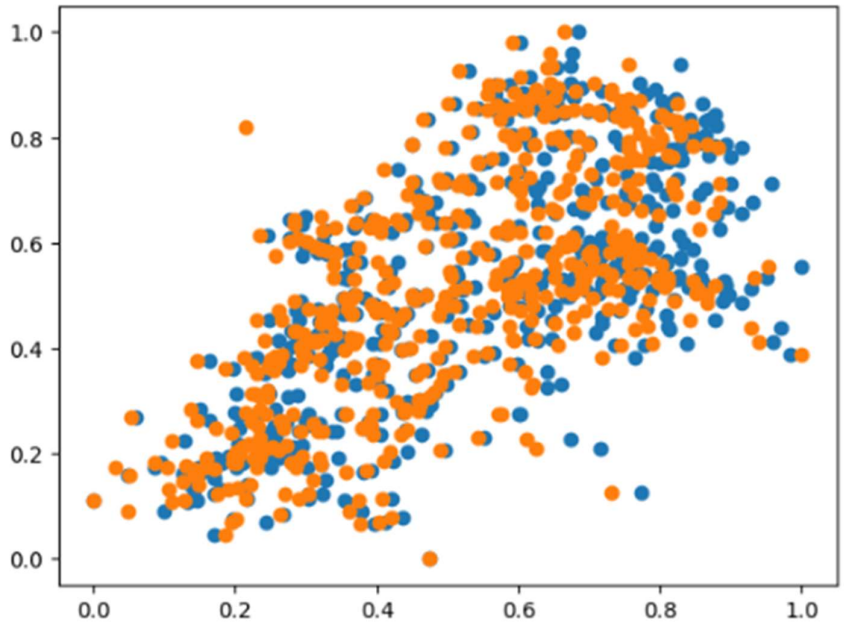
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   ANSWER: The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.



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

   ANSWER: Linear Regression models are validated based on below aspects:

   Linearity between cnt and temp, atemp



We can observer there is a linear relation between Bikes cnt and temp, atemp features

Low VIF values of the variables in the final model

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   ANSWER: Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   ANSWER: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to fit a straight line (in the case of simple linear regression) or a hyperplane (in multiple linear regression) that minimizes the sum of squared residuals—the differences between observed and predicted values. The coefficients of the model are determined using the least squares method. The primary assumptions include linearity, independence, homoscedasticity, and normality of residuals. The model is widely used for predicting continuous outcomes and understanding relationships between variables.

2. Explain the Anscombe's quartet in detail.
   ANSWER: Anscombe's quartet comprises four datasets that, despite having nearly identical summary statistics—such as mean, variance, and correlation— demonstrate vastly different distributions when graphed. The quartet was designed to emphasize the importance of visualizing data before interpreting it. Each dataset reveals different

patterns, including linear, non-linear, and outlier-influenced relationships, highlighting that relying solely on statistical measures without visualization can lead to misleading conclusions. The quartet underscores the need for graphical analysis in understanding data structures and relationships, reinforcing that context and distribution matter in statistical analysis.

3. What is Pearson's R?
   ANSWER: Pearson's R, or Pearson correlation coefficient, quantifies the linear relationship between two continuous variables. It ranges from -1 to 1, where +1 indicates a perfect positive linear correlation, -1 signifies a perfect negative correlation, and 0 denotes no linear relationship. The coefficient is calculated as the covariance of the variables divided by the product of their standard deviations. Pearson's R is widely used in statistical analysis to measure the strength and direction of a linear relationship, but it assumes the data is normally distributed and that the relationship is linear, which limits its applicability in some cases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   ANSWER: Scaling adjusts the range and distribution of features in data to improve model performance, especially in algorithms sensitive to feature magnitude (e.g., SVMs, k-NN). Normalized scaling rescales data to a fixed range, typically [0, 1], making all features comparable. Standardized scaling adjusts data to have a mean of 0 and a standard deviation of 1, useful for data following a Gaussian distribution. Scaling ensures that all features contribute equally to the model, preventing features with larger ranges from dominating. It also accelerates convergence in gradient-based optimizers and improves the model's interpretability.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   ANSWER: The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables. When the VIF is infinite, it indicates perfect multicollinearity—meaning one independent variable is an exact linear combination of others. In such cases, the regression model cannot distinguish between the correlated variables, leading to unreliable and unstable coefficient estimates. Perfect multicollinearity prevents the model from accurately estimating the relationship between the predictors and the outcome, often necessitating the removal of one or more correlated variables to resolve the issue.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   ANSWER: A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. By plotting the quantiles of the data against the quantiles of the theoretical distribution, the Q-Q plot reveals deviations from normality. In linear regression, it's used to assess whether the residuals are normally distributed—a key assumption for hypothesis testing and confidence intervals. If the points on the plot fall along a straight line, the residuals

are normally distributed. Deviations from this line indicate departures from normality, suggesting potential issues with the model's assumptions