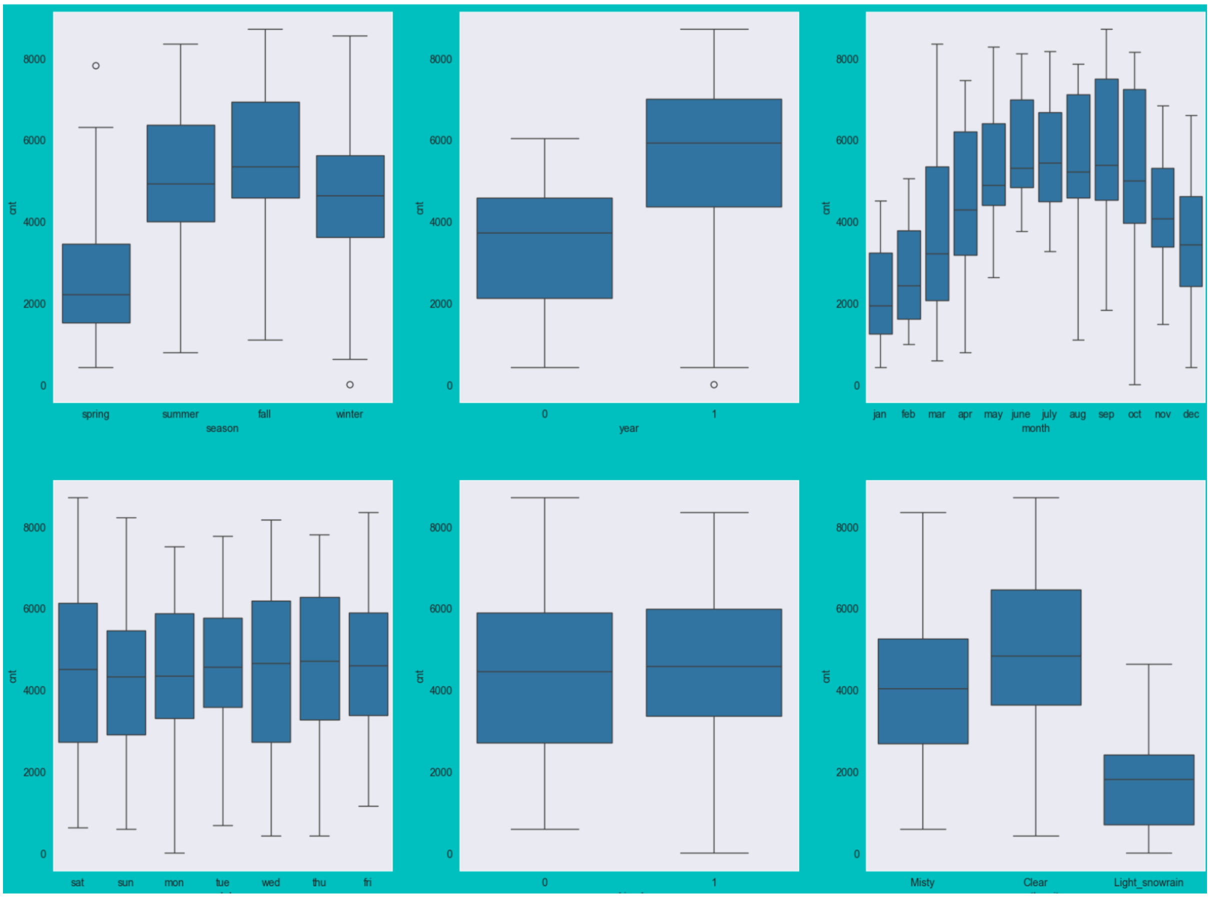
Assignment-based Subjective Questions

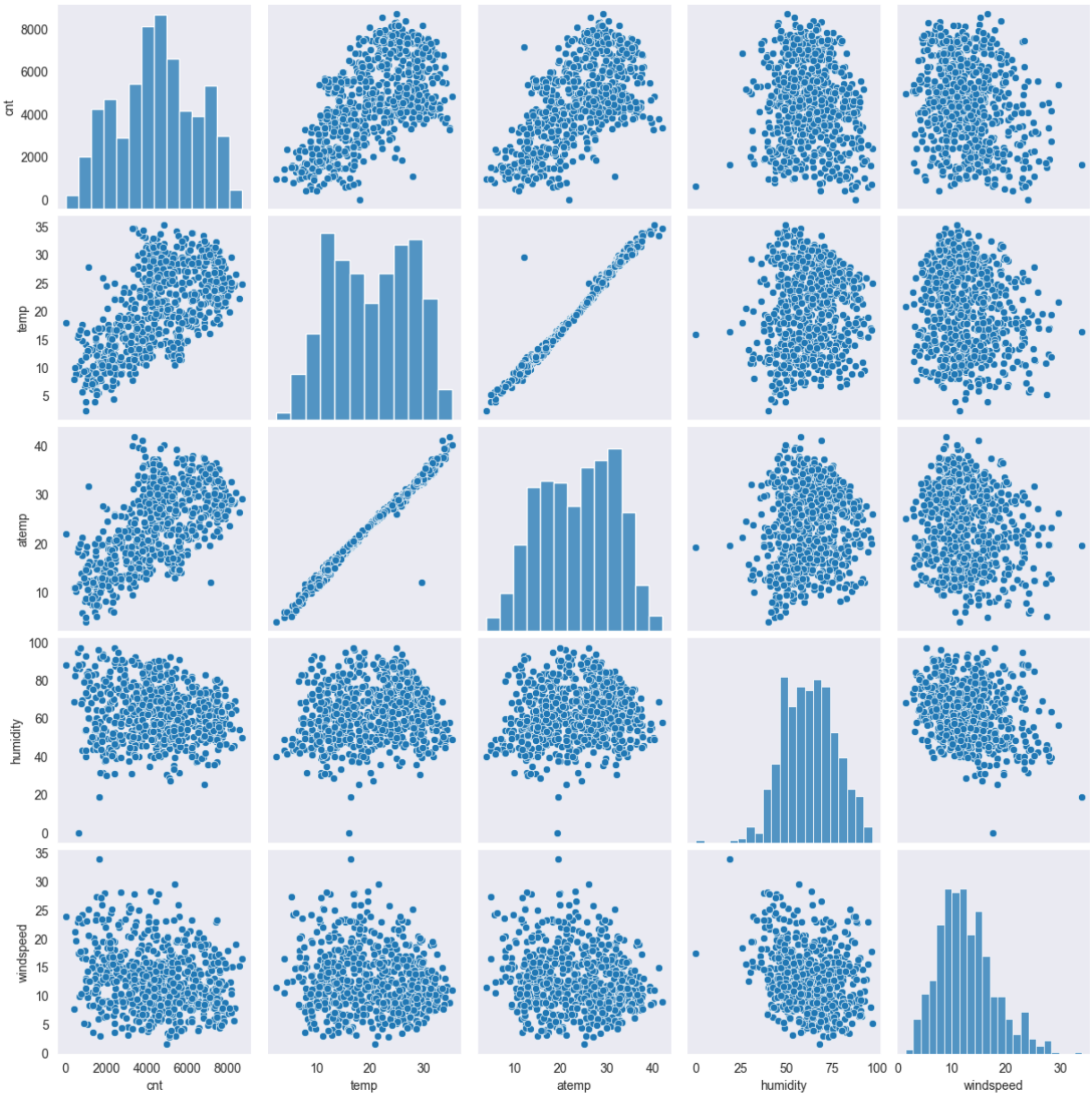
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans- There are a couple of categorical variables namely "season, month, year, weekday, workingday and weathersit". These categorical variables have a major effect on the dependent variable 'cnt'. The below chart showing the correlation among the same.



2. **Why is it important to use drop_first=True during dummy variable creation?**
Ans- drop_first=True helps in reducing the extra column created during dummy variable. dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.
Example-
bike_day_data = pd.get_dummies(data=bike_day_data,columns=["season"],drop_first=True)

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans- 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?**
   Ans- Linear Regression models are validated based on Linearity, No autocorrelation, Normality of error, Homoscedasticity, Multicollinearity.

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

   Ans- op 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 ?**
   Linear regression is a fundamental statistical technique used for modelling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable. A regression line can be a positive linear relationship or a negative linear Relationship.
   Details-
   Assumption of Linearity- Linear regression assumes that there is a linear relationship between the independent variables (features) and the dependent variable (target).
   Simple Linear Regression- In simple linear regression, there is only one independent variable. The relationship between the independent variables.
   - X and the dependent variable.
   - Y can be represented by the equation of a straight line.

   Multiple Linear Regression-In multiple linear regression, there are multiple independent variables. The equation is extended to include multiple predictors.
   Model Evaluation- Once the model is trained, its performance needs to be evaluated.
   Assumptions and Diagnostics- It's important to check whether the assumptions of linear regression are met.
   Predictions- Once the model is validated, it can be used to make predictions on new data by plugging in the values of the independent variables into the regression equation.
2. **Explain the Anscombe's quartet in detail?**
   Anscombe's quartet is created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and not relying solely on summary statistics. Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including means, variances, correlations, and regression lines, but vastly different visual representations. The quartet highlights that datasets with similar summary statistics can exhibit dramatically different patterns when plotted, emphasizing the limitations of relying solely on numerical summaries for understanding data.
3. **What is Pearson's R?**
   Pearson correlation coefficient (PCC)[a] is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling means transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes,

units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.
Difference between Normalizing Scaling and Standardize Scaling:

- In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
- Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
- Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
- Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
- Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
- Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   The Variance Inflation Factor (VIF) measures the degree of multicollinearity among predictor variables in a regression analysis. When the VIF is infinite for a particular predictor variable, it indicates perfect multicollinearity with other predictor variables in the model.

   Perfect multicollinearity occurs when one or more independent variables in the regression model can be perfectly predicted from the others. In other words, there is a linear relationship among the independent variables. This leads to problems in estimating the regression coefficients because the regression model cannot distinguish the individual effects of the perfectly collinear variables.

   The most common scenario leading to infinite VIF values is when one variable in the model can be expressed exactly as a linear combination of the other variables. For example, if one predictor variable is a constant multiple of another variable, or if one variable is the sum of two or more other variables, perfect multicollinearity occurs.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**.
   A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, typically the normal distribution.
   Steps-
   - Arrange the data in ascending order.
   - Compute the quantiles of the dataset.
   - Calculate the expected quantiles of the specified distribution (e.g., normal distribution) based on the sample size.
   -
   - Plot the observed quantiles against the expected quantiles. If the data perfectly follows the specified distribution, the points will fall along a straight line.

   Importance-

- Assumption Checkin- To determine if the residuals from a linear regression model—that is, the differences between actual and predicted values—follow a normal distribution, Q-Q plots are utilised. This is important because the normality of the residuals is assumed by many statistical techniques, such as linear regression.
- Finding Deviations from Normality- The residuals may not have a normal distribution if the points in the Q-Q plot substantially differ from the straight line. Deviations from the norm may point to misspecification, heteroscedasticity, or outliers as possible problems with the model.
- Model Validation- One diagnostic technique for validating linear regression models is the Q-Q plot. Analysts can evaluate the dependability and validity of the model's assumptions by evaluating the residuals' normalcy and making any required modifications.