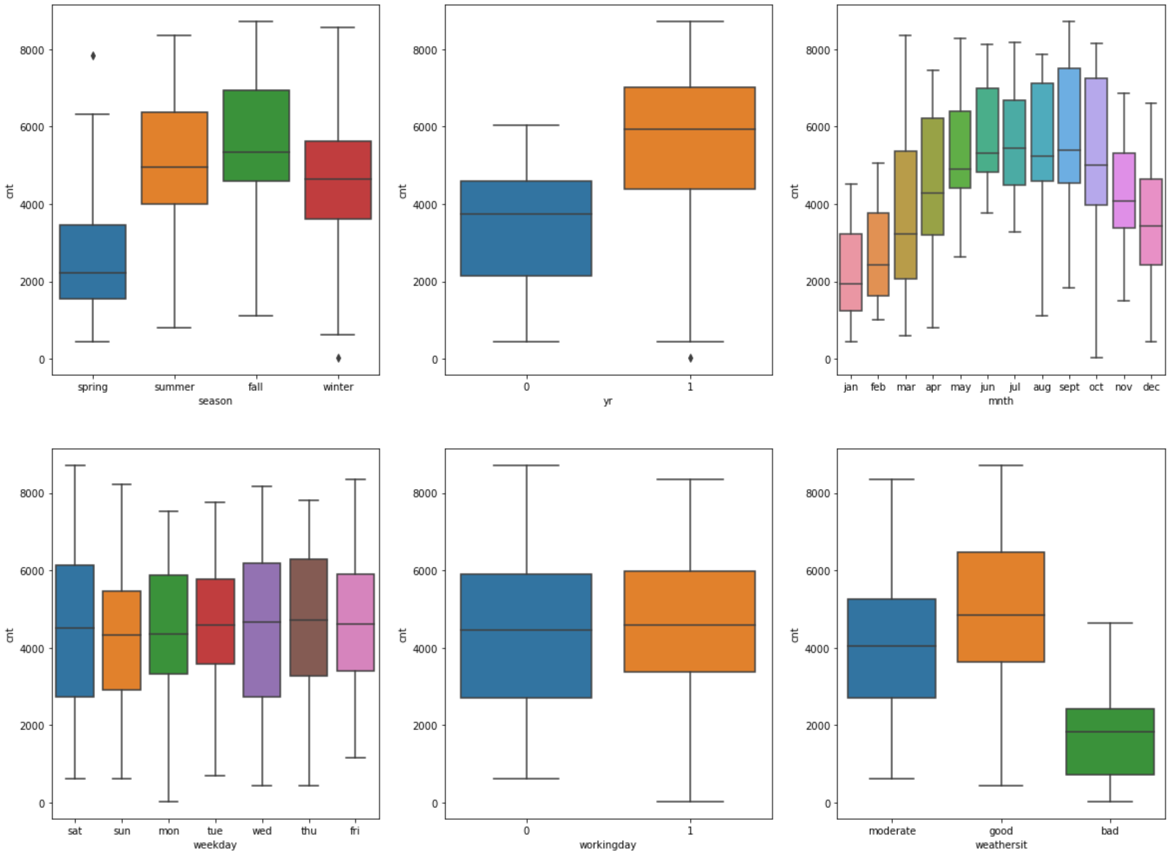
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer) A few categorical variables exist, including season, month, year, weekday, working day, and weather station. The dependent variable 'cnt' is significantly impacted by these category factors. The graph below displays the relationship between the same



2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

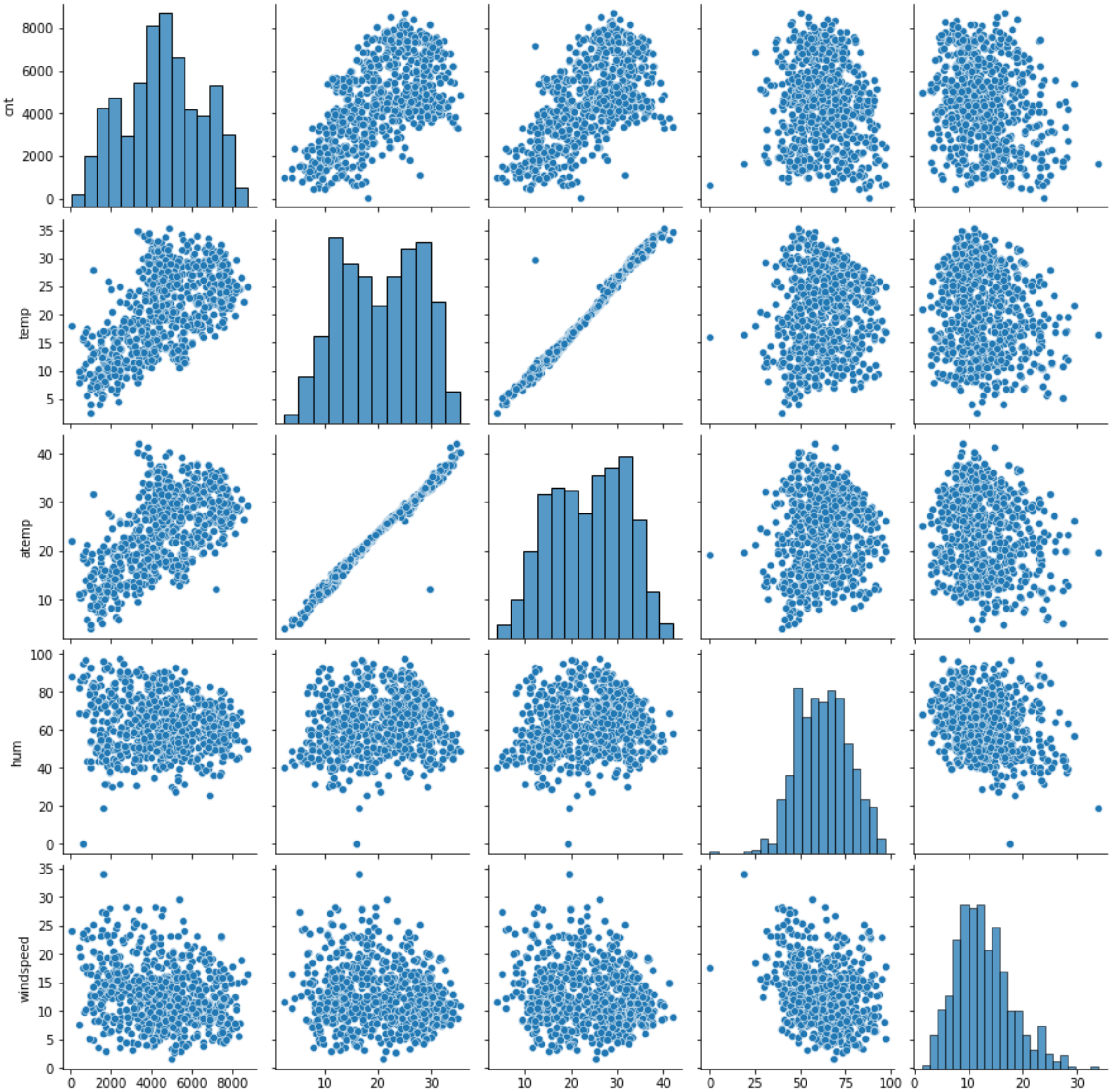
Answer) The idea behind the dummy variable is that given a category variable with 'n' levels, you construct 'n-1' additional columns, each of which indicates whether that level exists or not by using a zero or one.

Drop_first=True is thus utilised in order for the outcome to line up with levels n-1. As a result, it lessens the connection between the dummy variables.

For instance, if there are three levels, drop_first will remove the top column.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   (1 mark)

Answer)



4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer) verified the following five assumptions used in the linear regression model:

- Normality of error terms
    - Error terms should be normally distributed
- Multicollinearity check
    - There should be insignificant multicollinearity among variables.
- Linear relationship validation
    - Linearity should be visible among variables
- Homoscedasticity
    - There should be no visible pattern in residual values.
- Independence of residuals

- No auto-correlation

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

Answer) Top 3 features that has significant impact towards explaining the demand of the shared
bikes are

- Temperature
- year
- season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.** **(4 marks)**

Answer) The value of a dependent variable can be predicted using the statistical procedure known as linear regression based on one or more independent variables. The dependent variable is a linear function of the independent variable(s), and it is assumed that there is a linear relationship between the independent and dependent variables.

In other words, linear regression seeks to identify the line that fits the independent and dependent variables the best. By reducing the sum of squared errors between the dependent variable's actual and predicted values, the best-fit line is found.
Simple linear regression and multiple linear regression are the two different forms of linear regression. While multiple linear regression involves two or more independent variables, simple linear regression simply requires one.

The equation for simple linear regression is:
$Y = β0 + β1X + ε$
where Y is the dependent variable, X is the independent variable, β0 is the intercept, β1 is the slope, and ε is the error term.

The equation for multiple linear regression is:
$Y = β0 + β1X1 + β2X2 + … + βnXn + ε$
where Y is the dependent variable, X1, X2, …, Xn are the independent variables, β0 is the intercept, β1, β2, …, βn are the coefficients, and ε is the error term.

The ordinary least squares (OLS) regression technique is used to estimate the coefficients. Between the dependent variable's anticipated and actual values, OLS seeks to minimise the sum of squared errors. The best-fit line is the regression line that minimises this sum of squared errors.

The dependent variable's value can be predicted using the estimated coefficients for fresh values of the independent variable(s) once they have been calculated. The new numbers are entered into the regression equation to accomplish this.

The assumptions made by linear regression are that the independent and dependent variables have a linear connection, that the independent variables are not multicollinear, that the errors are normally distributed, and that the errors are homoscedastic (have a constant variance). The accuracy and dependability of the predictions provided by the model can be impacted by violations of these underlying premises.

2. **Explain the Anscombe's quartet in detail.** **(3 marks)**

Answer) The Anscombe quartet is a collection of four datasets with almost similar means, variances, correlations, and regression lines. Francis Anscombe, a statistician, developed the quartet in 1973 to highlight the value of charting data before statistical analysis.

The four datasets in Anscombe's quartet are intended to demonstrate the drawbacks of understanding data connections exclusively through summary statistics. For instance, the first dataset has 11 points that seem to be related linearly. However, the dataset clearly shows an outlier when it is displayed, and this outlier has a significant impact on both the regression line and correlation coefficient.

The second dataset was created to demonstrate the correlation coefficient's limitations as a measure of connection. It is made up of a group of data points that have the same correlation coefficient as the first dataset but follow a non-linear connection.

The third dataset emphasises how crucial it is to look at the distribution of data around the regression line. It comprises of a group of data that exhibit a linear connection, but there is one outlier that causes the variability around the regression line to increase significantly.

The fourth dataset, which is the last one, aims to demonstrate the impact of a single point on summary statistics. When one point is added, the regression line changes from appearing to follow a perfect quadratic connection to one that is a perfect linear relationship.

Anscombe's quartet emphasises the value of data visualisation before using statistical tests or models in their entirety. It emphasises the need of data exploration through graphs and visualisations and draws attention to the shortcomings of summary statistics in reflecting the complexity of relationships in data.

3. **What is Pearson's R?** **(3 marks)**

Answer) The Pearson correlation coefficient, sometimes referred to as Pearson's R, is a statistical indicator that assesses the linear connection between two continuous variables. It runs from -1 to +1 and assesses the strength of the correlation between two variables. Perfect positive correlation is represented by a value of 1, perfect negative correlation by a value of 1, and no correlation by a value of 0. By dividing the covariance of the two variables by the sum of their standard deviations, the Pearson's R value is determined. It is frequently used in research, especially in the social and behavioural sciences, to examine the link between variables and evaluate causality hypotheses. It should be emphasised, nevertheless, that Pearson's R does not account for nonlinear interactions and only evaluates the strength of linear relationships.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Answer) A technique used in data pre-processing to convert data to a common range of values is scaling. To avoid giving any feature or variable undue weight, scaling aims to put all characteristics or variables on the same scale.

The data is scaled in order to normalise it and provide machine learning algorithms additional insight into it. By ensuring that all characteristics have a consistent scale, it helps to improve the models' accuracy and effectiveness.

Scaling may be divided into two categories: standardised scaling and normalised scaling.

The data is scaled between 0 and 1 using normalised scaling, sometimes referred to as min-max scaling. It works by dividing by the difference between the maximum and minimum values of the feature after deleting the minimum value of the feature from each observation.

Data are scaled using standardised scaling to have a mean and standard deviation of 0 and 1, respectively. It operates by dividing each observation by the feature's standard deviation after deducting the feature's mean value from each observation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer) VIF = infinite if there is perfect correlation. A high VIF score denotes a strong connection between the variables. The existence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove one of the variables from the dataset that is the source of this ideal multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer) A graphical method for assessing if two data sets originate from populations with a similar distribution is the quantile-quantile (q-q) plot.

Use of Q-Q plot: The quantiles of the first data set are shown against the quantiles of the second dataset in a q-q figure. A quantile is the percentage of points that fall below the specified number.

In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is drawn. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The further the two data sets deviate from this reference line, the more evidence there is that they came from populations with distinct distributions.

Importance of the Q-Q plot: It is frequently desirable to determine whether the presumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference may be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.