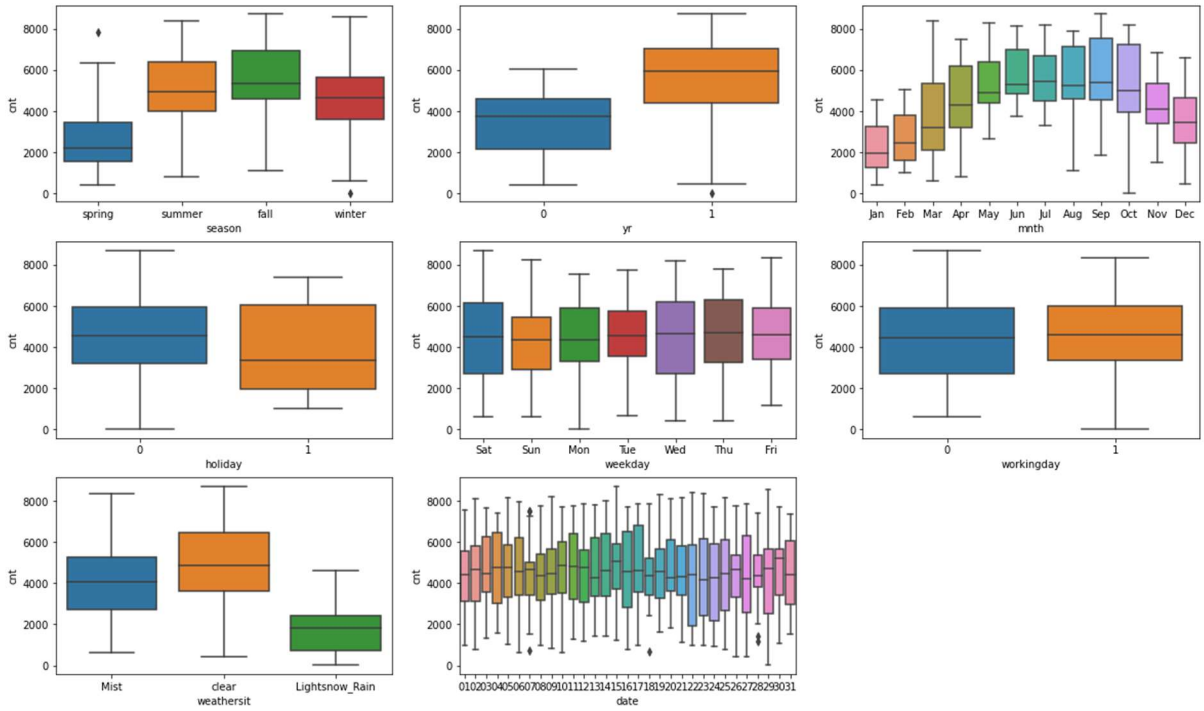
Assignment-based Subjective Questions

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



- The demand for bikes is usually higher during Summer and Fall season than the Spring and Winter seasons
- The demand for bikes is higher in the year 2019
- The demand for bikes is higher when the weather situation is clear
- The demand for bikes is less during winter – from the month of December to February

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

   The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable minus one. If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

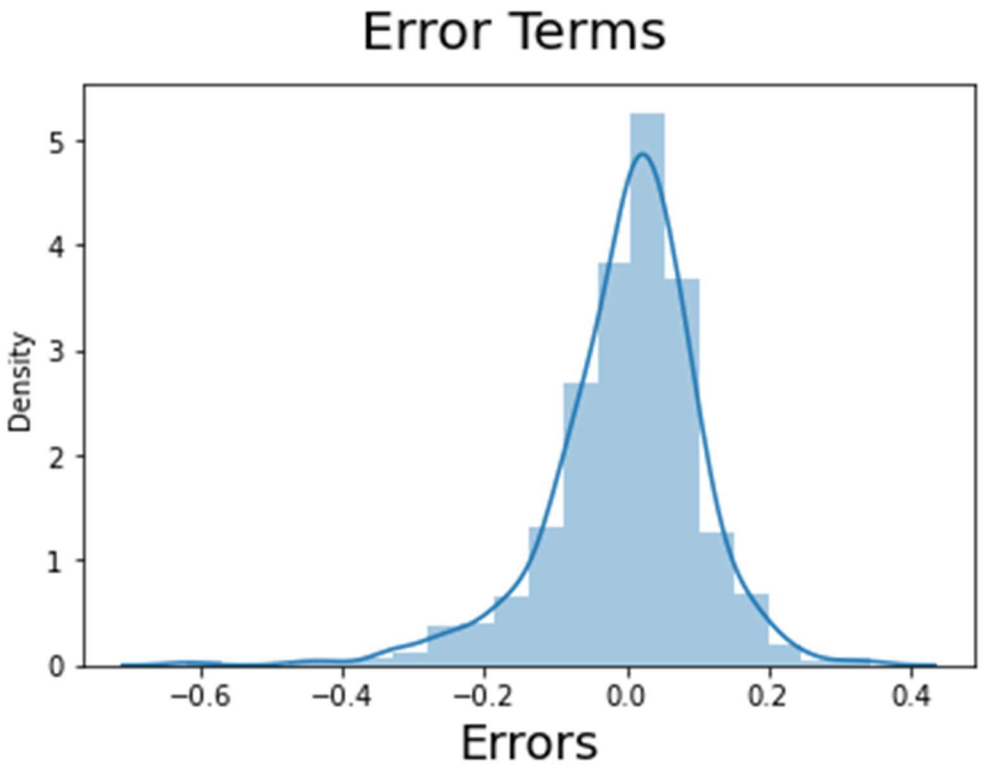
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   The "temp" variable has the highest correlation with the target variable

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

   The assumptions on Error terms are – Zero mean, independent, Normally distributed error terms that have constant variance

   Plotted the error terms that is the y_train - y_train_pred and the plot shows the error term is normally distributed and has Zero Mean



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

   Based on the coefficients of the variables used to build the model, the top three features are

   temp has a coefficient of 0.4793

   year has a coefficient of 0.2395

   holiday has a coefficient of - 0.0753

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   There are two types of linear regression

   - Simple linear regression
   - Multiple linear regression

   The simple linear regression explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

   Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values in independent variables in X

   The strength of the linear regression model can be assessed using 2 metrics: 1. R^2 or Coefficient of Determination 2. Residual Standard Error (RSE)

   Assumptions:

   The Error terms are normally distributed (not X, Y)

   Error terms are independent of each other

   Error terms have constant variance (homoscedasticity)

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that can make the regression model incorrect once you plot each data set.

   Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

3. **What is Pearson’s R?**

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
When there are a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons: 1. Ease of interpretation 2. Faster convergence for gradient descent methods You can scale the features using two very popular method:
1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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 the amount of multicollinearity in a set of multiple regression variables. A large value of VIF indicates that there is a correlation between the variables. If there is perfect correlation, then VIF = infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**

   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data came from same theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.