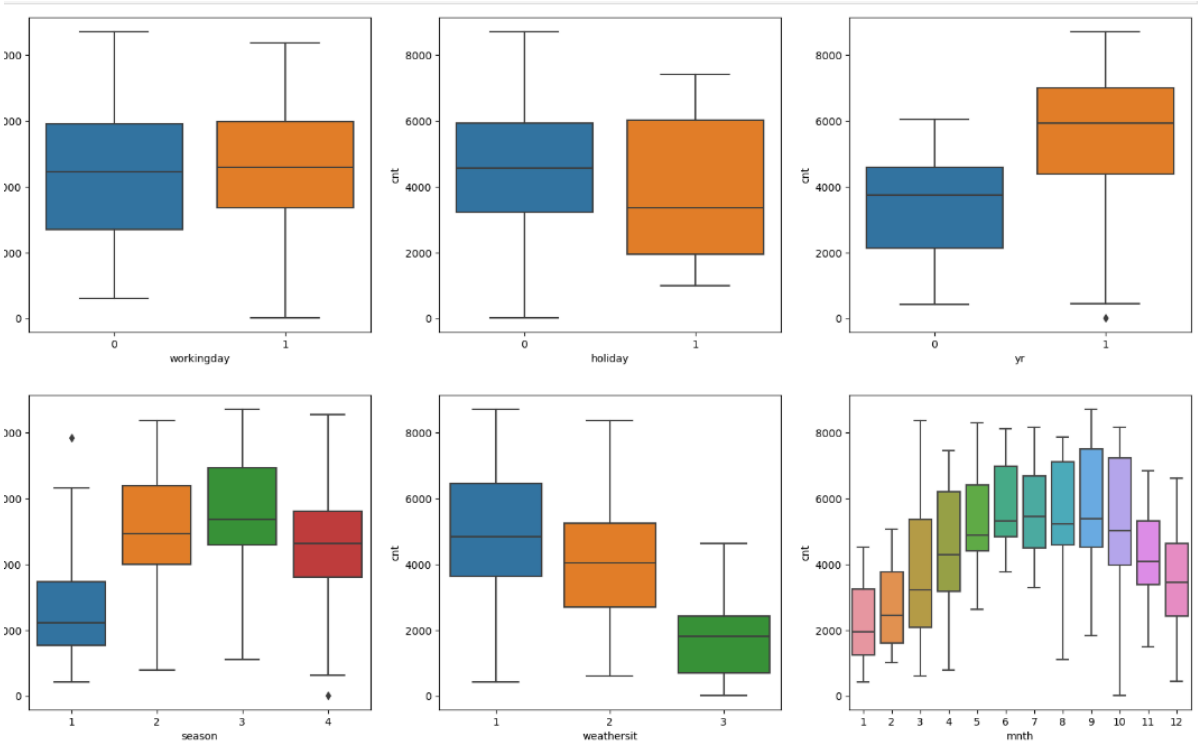
Assignment Based Subjective Questions:

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



We can infer the following from the plots:
1. From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
2. Bikes have been taken out less during the holidays.
3. Compared to 2018, 2019 saw an increase in the users, which might indicate that trends are moving towards bike sharing
4. Fall season is when users find it optimal to use bikes. Spring is when bikes arent that preferred.
5. Weather plays a huge role as can be seen. The days when there is heavy rain, there is no scope for sharing a bike. When the days are clear and when there is mist is when users usually prefer riding bikes
6. The distribution of data across months also supports the inferences from weather and season plots

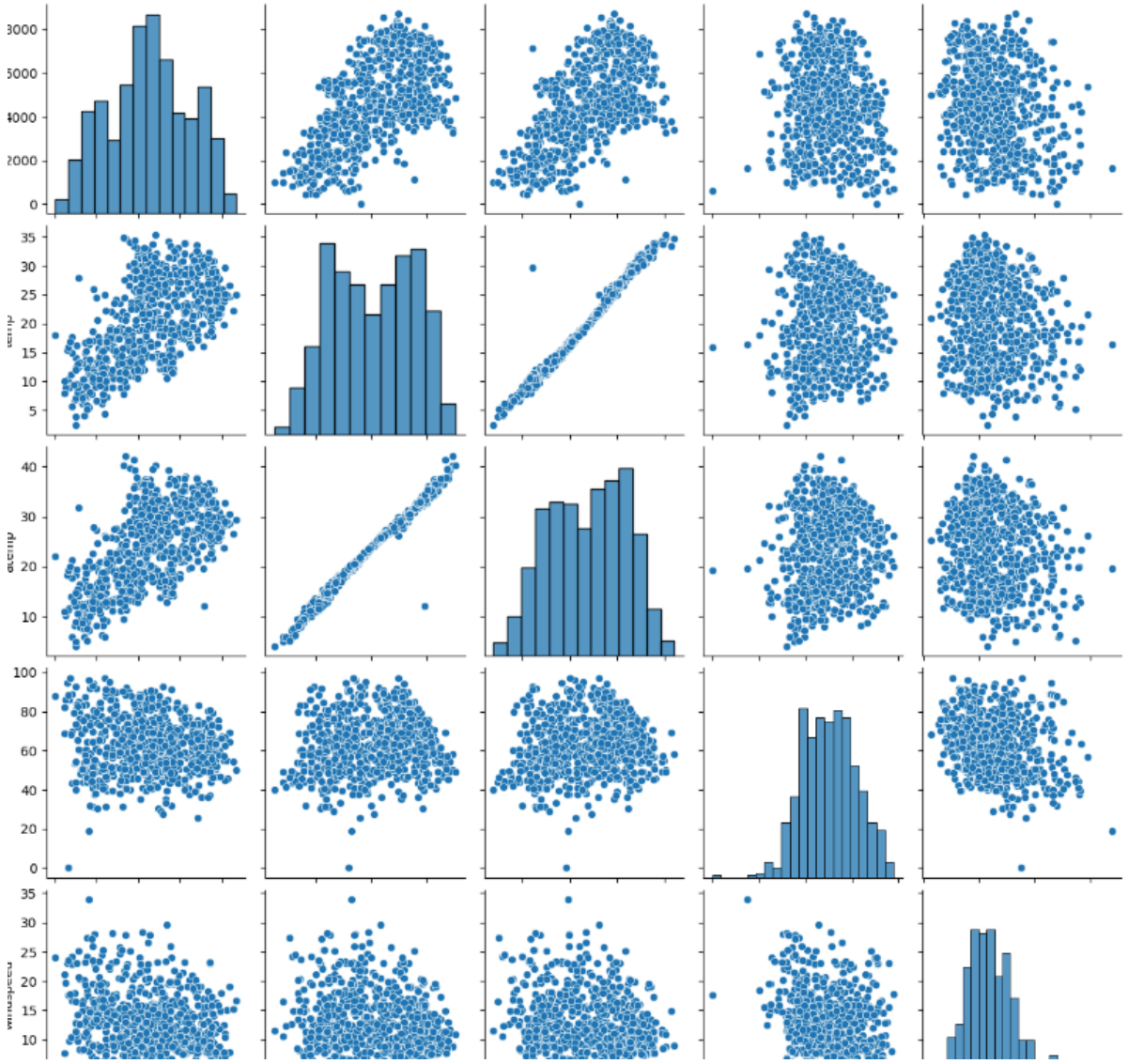
2. **Why is it important to use drop_first=True during dummy variable creation?**
   It is passed to get_dummies() to indicate that the first category should be dropped.
   It is important to avoid multicollinearity which occurs when one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
   It helps in reducing extra column created during dummy variable creation, thus reducing correlations created among dummy variables.
3. **Looking at the pair-plot among the numeric variables, which one has the highest correlation with the target variable?**

Temp and atemp are the two variables which has a higher and linear correlation with The target variable

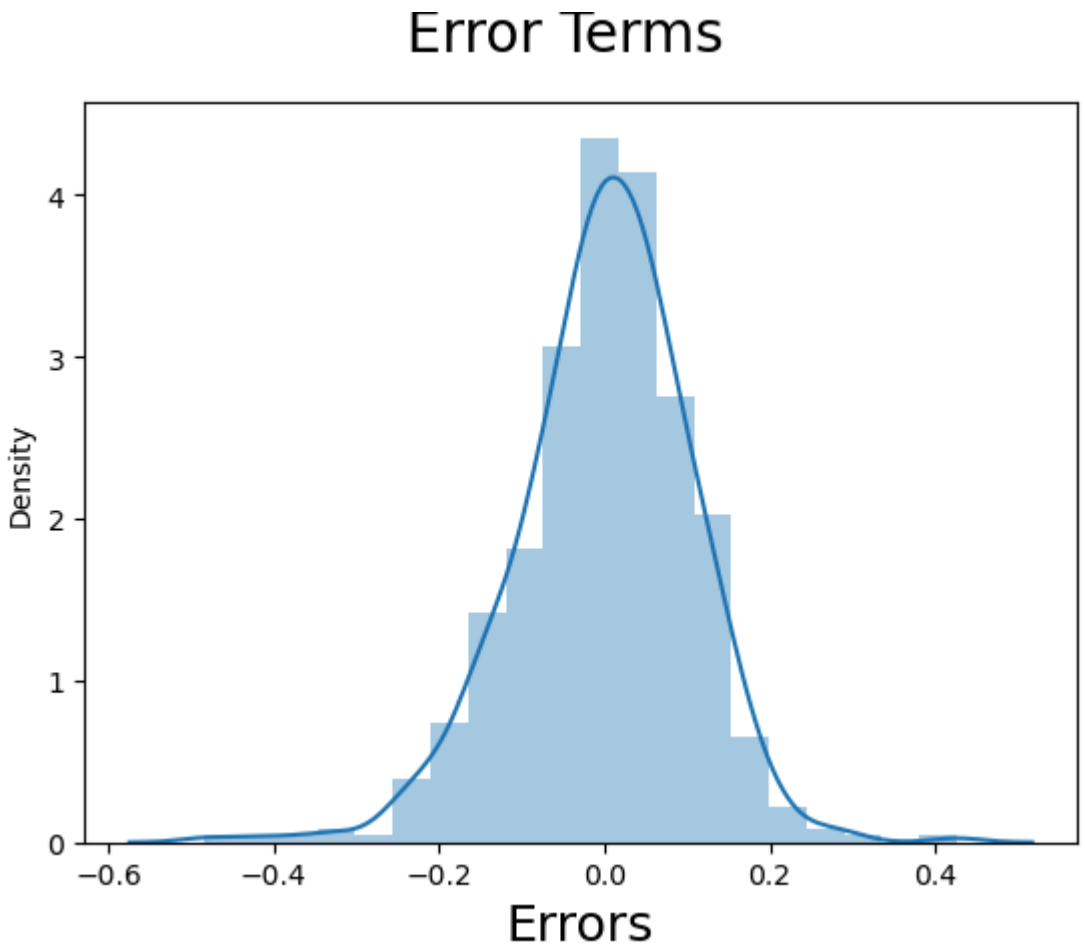
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   We validated the assumptions of Linear Regression as follows:

   1. There should be linear relationship between dependent and independent variables. We visualised the numeric variables using pairplot to check the linearity.



   2. Residuals should follow a normal distribution with mean centered around 0. We checked this by plotting a distplot of residuals



3. Linear Regression assumes that there is no multicollinearity in the data. We calculated the VIF to validate how strongly the feature variables are associated with one another.

	Features	VIF
2	windspeed	4.04
1	workingday	3.29
3	season_spring	2.65
4	season_summer	2.00
0	yr	1.88
5	season_winter	1.73
6	mnth_Jan	1.60
10	weathersit_Mist	1.57
8	weekday_Saturday	1.56
7	mnth_Sep	1.18
9	weathersit_Light Rain	1.08

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

   The top three features contributing towards the demand of shared bikes are:

   1. The year- Compared to 2018, 2019 saw an increase in bike rentals
   2. Workingday - Workingday had an effect on the bike rentals as compared to holidays.
   3. Windspeed - The more the windspeed the lesser the count of bikes rented.

General Subjective Questions:

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

   Sure, I'd be happy to explain linear regression in detail!

   Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship that can be used to predict the dependent variable.

   Types of Linear Regression:

1. Simple Linear Regression: Involves one independent variable and one dependent variable.
2. Multiple Linear Regression: Involves two or more independent variables.

The objective of linear regression is to determine the coefficients that minimize the difference between the predicted values and the actual values. This difference is measured using a cost function, commonly the Mean Squared Error (MSE)

To find the optimal coefficients, the model uses optimization techniques such as:
1. Ordinary Least Squares (OLS): A method for finding the coefficients that minimize the MSE. This involves solving a set of linear equations derived from the cost function.
2. Gradient Descent: An iterative optimization algorithm that updates the coefficients in the direction of the negative gradient of the cost function. It's used when the number of features is large or when an analytical solution is difficult to compute.

Linear regression relies on several key assumptions:
1. The relationship between the independent and dependent variables is linear.
2. Observations are independent of each other.
3. The variance of the residuals (errors) is constant across all levels of the independent variables.
4. The residuals of the model are normally distributed.

To evaluate the performance of a linear regression model, you can use various metrics:
1. R-squared : Measures the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
2. Adjusted R-squared: Adjusted for the number of predictors in the model, useful for comparing models with different numbers of predictors.
3. Root Mean Squared Error (RMSE): The square root of MSE, providing the standard deviation of the residuals

Linear regression is used in various fields including economics (predicting income), biology (predicting growth rates), and engineering (predicting product performance), among others.

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

Anscombe's Quartet is a set of four datasets that were constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization in statistical analysis. Despite having nearly identical statistical properties, each dataset reveals a different underlying relationship when visualized. Anscombe's Quartet is often used to illustrate why it's crucial to look at data visually rather than relying solely on statistical summaries.

The quartet consists of four datasets, each with 11 data points. They all have the same statistical properties:

a. Mean of x values

b. Mean of y values
c. Variance of x values
d. Variance of y values
e. Correlation between x and y
f. Regression line slope and intercept

Anscombe's Quartet serves as a valuable educational tool, showing that while statistical measures are important, understanding the underlying data through visualization is crucial for accurate analysis and interpretation.

3. **What is Pearson's R?**

Pearson's correlation coefficient, commonly denoted as rrr, is a measure of the strength and direction of the linear relationship between two continuous variables. It quantifies how well the data points fit a straight line and is one of the most widely used statistics in correlation analysis.

The Pearson correlation coefficient r is calculated using the formula:

$$r = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}$$

Where:

- n is the number of data points.
- x_i and y_i are the individual data points of the two variables.

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

Scaling is a preprocessing step in data analysis and machine learning where numerical features are adjusted to a common scale. This is done to ensure that features contribute equally to the analysis or modeling process, especially when the features have different units or ranges. Scaling is crucial for many algorithms, especially those that rely on distance metrics or gradient-based optimization.

There are two primary types of scaling methods: normalization and standardization.

1. Normalization, or Min-Max scaling, rescales the features to a fixed range, usually [0, 1], but it can be any range.
2. Standardization, or Z-score normalization, transforms features to have a mean of 0 and a standard deviation of 1.

Normalization is preferred when you need to transform the features into a specific range and when the features have different units. It's commonly used in algorithms like neural networks where input data is often expected to be between 0 and 1.

Standardization is preferred when you need features to have a normal distribution with zero mean and unit variance. It's often used in algorithms that assume normally distributed data, such as linear regression and logistic regression.

5. **You might have observed that sometime the value of VIF is infinite, why does this happen?**
   A **Variance Inflation Factor (VIF)** is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when independent variables are highly correlated with each other, which can make the estimates of regression coefficients unstable and unreliable.

   The VIF can become infinite when Ri2=1R_i^2 = 1Ri2=1. This situation indicates perfect multicollinearity. Here's why it happens:

   a. **Perfect Multicollinearity**: If a predictor variable is a perfect linear combination of other predictor variables, the Ri2R_i^2Ri2 value will be 1. For example, if one predictor variable can be expressed exactly as a combination of other predictor variables in the model, this creates a perfect collinearity situation.
   b. **Redundant Predictors**: Sometimes, this situation arises if two or more predictor variables are identical or if there is an exact linear dependency among the predictors. For instance, if two predictors are perfectly correlated (i.e., one is a scaled version of the other), then Ri2R_i^2Ri2 will be 1 for one of them when regressed on the other, leading to an infinite VIF

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a theoretical distribution, such as the normal distribution. It compares the quantiles of the data distribution with the quantiles of a theoretical distribution.

In the context of linear regression, a Q-Q plot is primarily used to assess the **normality of residuals**. Residuals are the differences between the observed and predicted values of the dependent variable. Checking the normality of residuals is important for several reasons:

1. **Model Assumptions**:
   - **Normality**: One of the key assumptions of linear regression is that the residuals are normally distributed. This assumption is important for the validity of hypothesis tests and confidence intervals for the regression coefficients.
   - If the residuals are not normally distributed, it might indicate that the model is not appropriately specified or that some key predictors are missing.
2. **Evaluating Model Fit**:
   - A Q-Q plot can help evaluate how well the linear regression model fits the data by checking if the residuals exhibit the assumed normal distribution. Deviations from the straight line in the Q-Q plot suggest departures from normality.

3. **Detecting Outliers and Influential Points**:
    - The Q-Q plot can also help identify outliers or influential points. Outliers will appear as points far from the line, indicating that they do not fit well with the assumed normal distribution of residuals