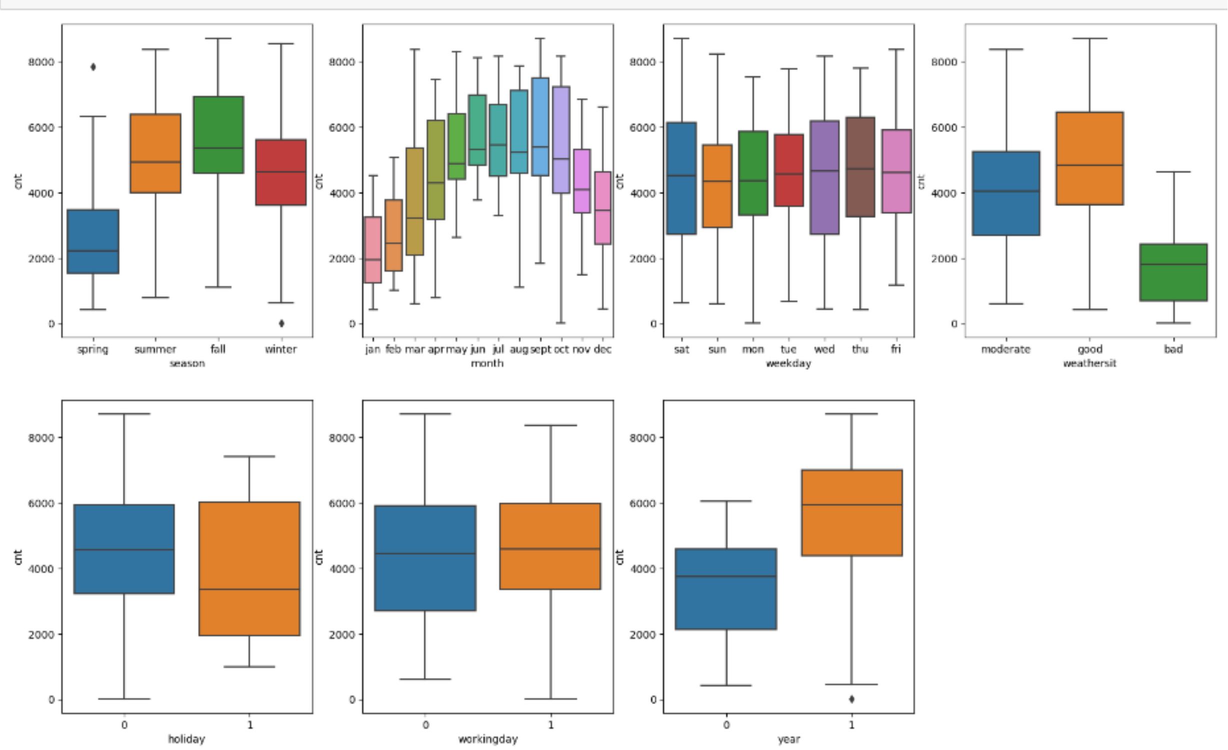
Assignment-based Subjective Questions

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

The given dataset contains several categorical variables such as season, month, year, weekday, working day, and weather situation that have a significant impact on the dependent variable 'cnt'. According to the correlation shown in the figure:

- The fall season appears to attract more bookings than other seasons.
- Most of the bookings have been made during the months of May, June, July, August, September, and October.
- Clear weather conditions seem to attract more bookings than other weather conditions.
- Thursdays, Fridays, Saturdays, and Sundays have more bookings compared to the beginning of the week.
- Bookings are less frequent when it's not a holiday, which is reasonable as people may want to spend time at home with family on holidays.
- The number of bookings appears to be almost equal on working and non-working days.
- 2019 had more bookings than the previous year, indicating progress in business.



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

It is important to use the **drop_first = True** parameter when creating dummy variables, as it helps to eliminate the extra column created during the process. This, in turn, helps to reduce the correlations that may be created among the dummy variables. The syntax for drop_first is a boolean value, with the default set to False. Setting it to True will generate k-1 dummies out of k categorical levels by removing the first level.

For example, if we have three types of values in a categorical column and we want to create a dummy variable for that column, we don't need the third variable to identify the value as it can be easily inferred from the absence of the first two values.

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

The variable “temp” has the highest correlation.

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

I have assessed the validity of the Linear Regression Model based on the following five assumptions:

- Normality of error terms:
  - The error terms should be normally distributed.
- Multicollinearity check:
  - There should be no significant multicollinearity among the variables.
- Linear relationship validation:
  - There should be a visible linear relationship among the variables.
- Homoscedasticity:
  - There should be no discernible pattern in the residual values.
- Independence of residuals:
  - There should be no autocorrelation present among the residuals.

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

The top three features are Temperature, Year, and Windspeed.

General Subjective Questions

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

Linear Regression is a widely used statistical method that is used to predict a continuous outcome variable based on one or more predictor variables. The basic idea behind linear regression is to establish a relationship between the outcome variable and one or more predictor variables in such a way that we can use this relationship to predict the outcome variable for new observations.

The linear regression algorithm can be summarized in the following steps:

- **Data Collection**: The first step is to collect the data that you will use to build the linear regression model. The data should contain the outcome variable and one or more predictor variables.
- **Data Preprocessing**: This step involves cleaning and preprocessing the data to prepare it for analysis. This may include removing missing values, identifying and handling outliers, and scaling or normalizing the data.
- **Splitting the Data:** The next step is to split the data into training and testing datasets. The training dataset is used to build the model, while the testing dataset is used to evaluate the model's performance.
- **Building the Model**: In this step, we use the training data to build the linear regression model. The model tries to establish a linear relationship between the outcome variable and the predictor variables. The most common approach is to use the least squares method to find the coefficients of the linear equation that best fits the data.
- **Evaluating the Model**: After building the model, we evaluate its performance on the testing dataset. The most commonly used metric for evaluating a linear regression model is the R-squared value, which measures the proportion of variance in the outcome variable that can be explained by the predictor variables.
- **Using the Model**: Once we have built and evaluated the model, we can use it to predict the outcome variable for new observations. We simply plug in the values of the predictor variables for the new observation into the linear equation and get the predicted value of the outcome variable.

In summary, the linear regression algorithm involves collecting and preprocessing data, splitting the data into training and testing datasets, building the model using the training dataset, evaluating the model's performance using the testing dataset, and using the model to make predictions for new observations.

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

Anscombe's quartet is a collection of four datasets, each consisting of 11 (x, y) pairs, created by the statistician Francis Anscombe in 1973. The four datasets appear to have nearly identical simple statistical properties, such as the mean and variance of x and y, the correlation coefficient between x and y, and the line of best fit when plotted. However, when plotted graphically, the datasets are distinctly different, and illustrate the importance of graphically exploring and visualizing data before drawing conclusions based solely on summary statistics.

The four datasets of Anscombe's quartet are:

- Dataset I: This dataset has a linear relationship between x and y, and all of the statistical properties of the dataset are nearly identical to those of the other three datasets. When plotted, the dataset appears to be a simple linear relationship between x and y, with a strong correlation coefficient.
- Dataset II: This dataset has a non-linear relationship between x and y, but the statistical properties are still nearly identical to those of the other three datasets. When plotted, the

dataset appears to have a curvilinear relationship between x and y, with a lower correlation coefficient than Dataset I.

- Dataset III: This dataset has a linear relationship between x and y, but one outlier point in the dataset significantly affects the line of best fit and the correlation coefficient. When plotted, the dataset appears to have a clear linear relationship between x and y, but with a clear outlier point.
- Dataset IV: This dataset has a curvilinear relationship between x and y, and two distinct groups of points that appear to follow different relationships. When plotted, the dataset appears to have a clear curvilinear relationship between x and y, with two distinct groups of points.

Anscombe's quartet highlights the importance of graphically exploring and visualizing data, and shows that summary statistics can be misleading when used alone. It is often used as a demonstration in statistics courses to emphasize the importance of visualizing data before making conclusions.

3. What is Pearson’s R? (3 marks)

Pearson’s R is a statistical measure that represents the strength and direction of the linear relationship between two continuous variables. It is also known as the Pearson correlation coefficient or simply the correlation coefficient. Pearson’s R ranges from -1 to +1, where -1 indicates a perfectly negative linear relationship, 0 indicates no linear relationship, and +1 indicates a perfectly positive linear relationship. The formula for Pearson’s R involves dividing the covariance between the two variables by the product of their standard deviations. It is widely used in fields such as psychology, social sciences, and economics to analyze the relationship between variables and to make predictions based on that relationship.

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

Scaling is a preprocessing step in data analysis where the numerical values of features in a dataset are transformed to fit within a specific scale. This transformation ensures that the features have a similar range and do not dominate each other in the analysis process. Scaling is performed to ensure that the features are on a common scale and to avoid biased analysis due to differences in feature magnitudes.

Normalized scaling transforms the features so that they fall within the range of 0 and 1. This is done by subtracting the minimum value of the feature from each value and then dividing by the range of the feature.

Standardized scaling, on the other hand, transforms the features to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the feature from each value and then dividing by the standard deviation of the feature. The main difference between these two methods is that normalized scaling preserves the shape of the distribution of the feature values, while standardized scaling transforms the feature values to have a mean of 0 and a standard deviation of 1.

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

VIF measures the degree to which the variance of the estimated regression coefficient is increased because of collinearity in the model. When the VIF value is infinite, it indicates that:

- The independent variables are perfectly collinear, meaning that they can be perfectly predicted from each other.
- This often happens when there is a linear relationship between two or more independent variables, leading to an unstable and poorly defined regression model.
- Infinite VIF values can also be caused by having too many independent variables relative to the sample size, leading to a numerical problem in the calculation of the VIF.

It is important to identify and address the issue of multicollinearity to ensure the accuracy and reliability of the regression analysis.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a sample to a theoretical distribution, such as a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will fall along a straight line. If the residuals are not normally distributed, the points on the Q-Q plot will deviate from a straight line. The use and importance of a Q-Q plot in **linear regression are:**

- Checking Normality Assumption: A Q-Q plot is a simple and effective way to check the normality assumption of the errors in linear regression. It allows us to see if the residuals are normally distributed, which is important because the normality assumption is necessary for accurate hypothesis testing and confidence interval construction.
- Outlier Detection: A Q-Q plot can also be used to detect outliers. If there are outliers in the data, the Q-Q plot will show deviations from the straight line in the tails.
- Model Comparison: A Q-Q plot can be used to compare the residuals of two or more linear regression models. If the residuals of two models have similar Q-Q plots, it suggests that they have similar distributions, and therefore, similar goodness of fit.