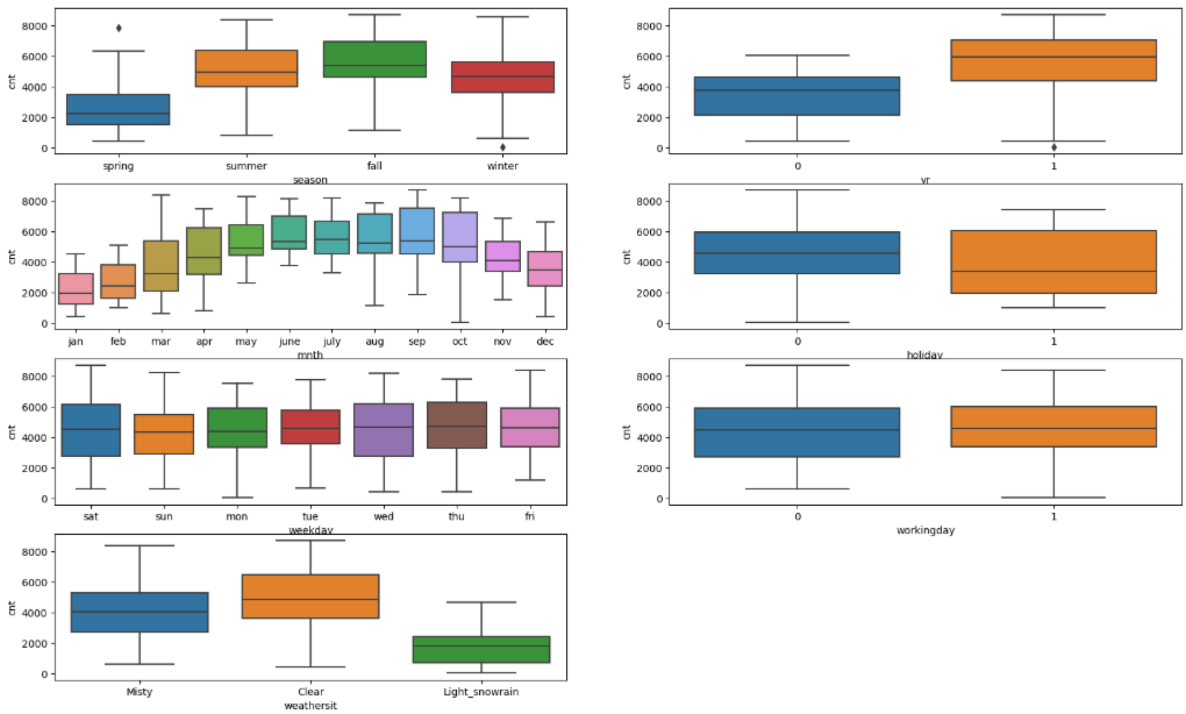
Assignment-based Subjective Questions

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

Ans: The categorical variables in the dataset are:

"season", "yr", "mnth", "holiday", "weekday", "workingday", "weathersit"

When performed an analysis of these variables against the dependent variable “cnt” below are the observations.



Bike rental rates are likely to be higher in summer and the fall season, when the weather is clear are more prominent in the months of September and October, more so in the days of Sat, Wed and Thu and in the year of 2019. Additionally, we could observe that bike rentals are higher on holidays.

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 the dummy variable creation and hence it reduces the correlations created among dummy variables.

Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then it is obvious C. So we do not need 3rd variable to identify the C.

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

Ans: ‘temp’ variable has the highest correlation with the target variable ‘cnt’.

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

Ans: The assumptions of Linear Regression that are validated are

- There should be a linear relationship between dependent and independent variables. (X and Y)
- Error terms are normally distributed with mean zero
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)
- Multicollinearity Check

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

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- Temp
- Aug
- Spring

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression belongs to Supervised Machine learning category, meaning it learns from labeled data (data with known inputs and outputs) to make predictions about future data. The goal is to model the relationship between one or more independent variables (predictors, often denoted as X) and a dependent variable (response, denoted as Y), with the aim of predicting the value of Y given new values of X. It assumes a linear relationship between the variables, meaning they can be plotted as a straight line.

Algorithm Steps:

1. Data Collection: Gather a dataset containing pairs of values for X and Y.
2. Data Visualization: Plot the data to visualize the relationship between X and Y. This helps assess linearity and identify potential outliers.
3. Model Training:
    - Define a linear equation: $Y = \beta_0 + \beta_1 X + \varepsilon$ (where β_0 is the intercept, β_1 is the slope, and ε is the error term).
    - Use a technique called Ordinary Least Squares (OLS) to find the best-fitting line for the data. This involves minimizing the sum of squared residuals (the vertical distances between data points and the line).
4. Model Evaluation:
    - Calculate statistical measures like R-squared (how well the model explains the data) and p-value (significance of the relationship).
    - Visually inspect the residuals (differences between actual and predicted values) to check for patterns or biases.
5. Prediction: Use the trained model to predict Y values for new X values.

Types of Linear Regression:

- Simple Linear Regression: Involves a single independent variable (X).
- Multiple Linear Regression: Involves multiple independent variables (X1, X2, X3, etc.).

Additional Considerations:

- Feature Scaling: Transforming variables to have similar scales can sometimes improve model performance.
- Regularization: Techniques like L1 or L2 regularization can prevent overfitting (when the model fits too closely to the training data and doesn't generalize well to new data).
- Assumptions: Linear regression assumes linearity, independence of errors, normality of errors, and homoscedasticity (constant variance of errors). It's important to check these assumptions before using the model.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an

outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Eg: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is alsouseful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.