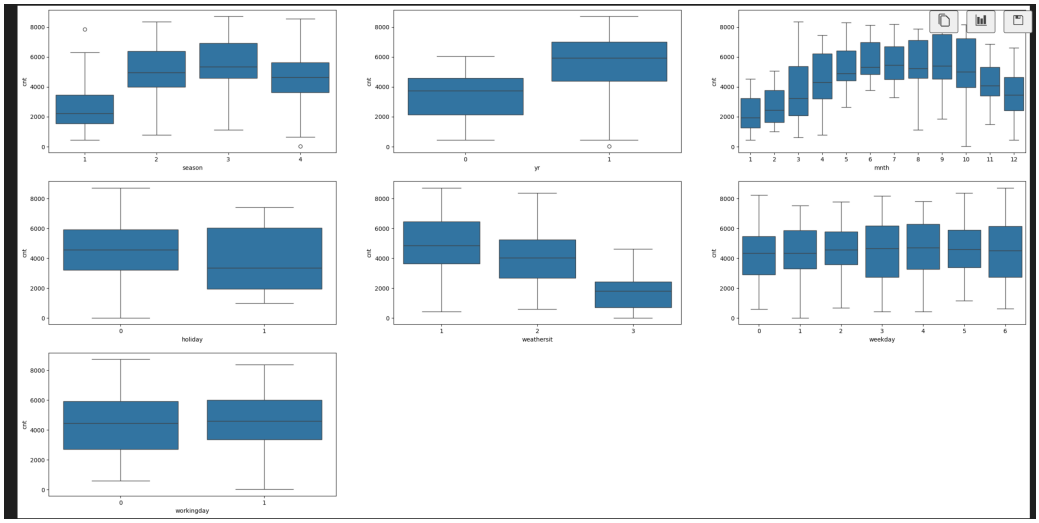
Subjective Question

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



Season: Bike rentals are typically higher during warmer seasons (spring and fall).
Year: Bike rentals are more in 2019 , May be due to fitness / environment conditions
Month: Bike rentals are usually higher in mid months like 3 – 10 . prominently in march , April ,September , October.
Holiday: Bike rentals tend to be lower on holidays.
Weather: Clear weather conditions generally lead to higher bike rentals
Weekday: Weekends might have higher rentals compared to weekdays, but it's essential to consider local culture and work patterns.
Workingday: Working days might have higher rentals, but this can vary based on the city and its demographics.

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

Using drop_first=True in dummy variable creation prevents multicollinearity. It avoids the dummy variable trap by excluding one category from the dummy variables. This is because the information from the excluded category can be inferred from the presence or absence of the other categories.
Example:
Data :

	A	B	C
0	1	0	0
1	0	1	0
2	1	0	0
3	0	0	1

Dummies Table

	B	C
0	0	0
1	1	0
2	0	0
3	0	1

if both B and C are 0, we know that the category must be A.

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

Temperature (or atemp) typically shows the highest correlation with the target variable (bike rentals). Warmer temperatures are often associated with increased bike usage.

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

To validate the assumptions of linear regression, we can use the following methods:
Multicollinearity: Calculate VIF (Variance Inflation Factor) to check for multicollinearity among independent variables.
Residual plots: Check for normality, heteroscedasticity, and autocorrelation.
Normality test: Error Plots should be a normal distribution curve
R2 Score : Check R2 for trained and test data

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

The top 3 features contributing to bike demand typically include:

Temperature (or atemp)
Season
Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables. The best way is fitting in a best-fit line , and minimising the mean square error .
Simple LR deals with one variable , and multiple deals with multiple dependent variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have almost identical statistical properties (mean, variance, correlation, etc.) but look very different when plotted. It demonstrates the importance of visualizing data before applying statistical methods(EDA). It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial.

3. What is Pearson's R?

Pearson's correlation coefficient (R) is a statistical measure that quantifies the linear relationship between two continuous variables. It ranges from -1 to 1, where:

1 indicates a perfect positive correlation
-1 indicates a perfect negative correlation
0 indicates no correlation

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

Scaling is the process of transforming data to a common scale. It is performed to improve the performance of machine learning algorithms, especially those sensitive to feature scales.

Normalized scaling(MinMax scaling): Rescales features to a specific range, typically 0 to 1.
Standardized scaling: Rescales features to have zero mean and unit variance.
Normalization preserves the shape of the original distribution but scales the values to a specific range.

Standardization transforms the data to a standard normal distribution with a mean of 0 and a standard deviation of 1.
It's important to note that both normalization and standardization are preprocessing steps, and the best choice depends on the specific dataset and algorithm used.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
A VIF value of infinity indicates perfect multicollinearity between independent variables. This means one variable is a perfect linear combination of other variables, leading to singularity in the design matrix and making it impossible to estimate the model parameters.

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

A Q-Q plot is a graphical tool used to assess if a dataset follows a particular distribution, often a normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution.

In linear regression, a key assumption is that the residuals (differences between predicted and actual values) are normally distributed. A Q-Q plot of the residuals is crucial to verify this assumption. If the points on the plot closely follow a straight line, it suggests normality. Deviations from the line indicate potential issues with the model, such as outliers or non-normality of residuals, which can affect the reliability of the model's inferences.