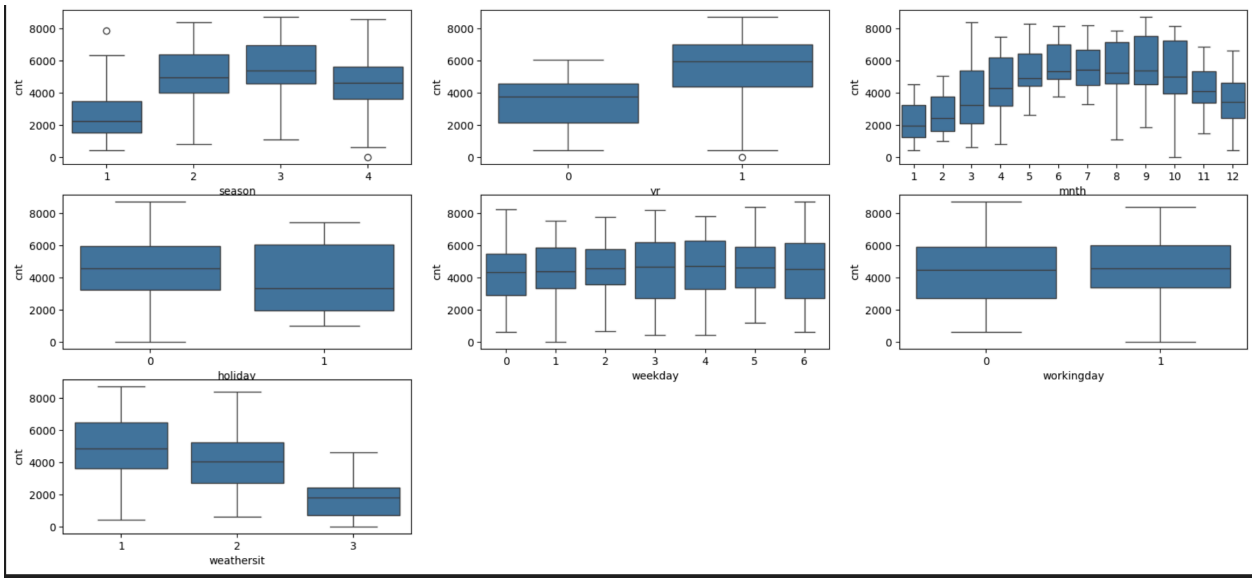
Assignment-based Subjective Questions

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

Ans:

Analysis of the unconditional variables from shows that bike rent rates are inclined be bigger in

· Demand is bigger in the summertime and fall seasons, exceptionally when the weather is clear and acceptable.· Rental rates are taller all along the months of March, May, June, July, September, and October.· Saturdays and Fridays have the chief demand for bike rentals.· Bike rent counts have raised from 2018 to 2019.



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

Ans: Using drop_first = ture during dummy variable creation is crucial because it prevents multicollinearity by dropping one category and treating it as the reference. This avoids the dummy variable trap, ensuring that the model can be estimated correctly. Additionally, it simplifies the model and makes the interpretation of coefficients clearer, as they represent the change relative to the reference category.

Q3- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable

Temp/atemp

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

Ans: - Error terms must be normalized

The histogram of residuals determines a normal distribution.

There should be insignificant multicollinearity among variables
No auto-correlation

Q5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. Temp/atemp
2. Winter
3. Clear weather

General Subjective Questions

Q1- Explain the linear regression algorithm in detail.

Linear Regression

Linear regression is a supervised learning algorithm that analyzes the relationship between input (X) and output (Y) variables using labeled data. It aims to identify the relationship between these variables and make predictions about future outcomes based on historical data.

Simple Linear Regression

Simple linear regression involves plotting a line on a graph to represent the relationship between an independent variable (X) and a dependent variable (Y). This method predicts the value of the dependent variable using the independent variable.

Linear Regression Formula

$Y=mx+cY=mx+c$

In this formula, mm represents the slope, cc is the intercept, and YY is the predicted value.

Correlation Coefficient (R-squared Value)

The R-squared value, also known as the correlation coefficient, evaluates the goodness of fit for the model. It ranges from 0 to 1, where 0 indicates no correlation and 1 indicates perfect correlation.

Q2- Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a well-known example in statistics and data visualization, created by Francis Anscombe in 1973. It was designed to illustrate that summary statistics alone may not provide a complete understanding of the data.
The quartet comprises four datasets, each containing 11 data points with two numerical variables (X and Y). What makes Anscombe's Quartet remarkable is that these four datasets have identical simple descriptive statistics (such as means, variances, and correlation coefficients) for X and Y, yet they exhibit vastly different graphical representations and relationships between the variables.
It highlights the importance of data visualization and graphical exploration, revealing hidden patterns, outliers, and relationships that might not be apparent from summary statistics alone.
Anscombe's Quartet is frequently used in statistics and data science education to emphasize the necessity of visualizing data before drawing conclusions or making decisions based on statistical summaries.

Q3 What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the **Pearson's R**, also known as Pearson's correlation coefficient, is a statistical metric that quantifies the strength and direction of the linear relationship between two continuous variables. Represented by the letter "r," its values range from -1 to 1, where:

- **1** signifies a perfect positive linear relationship,
- **-1** signifies a perfect negative linear relationship, and
- **0** signifies no linear relationship.

The formula for calculating Pearson's R is:

r=cov(X,Y)σXσYr=σXσYcov(X,Y)

where:

- cov(X,Y)cov(X,Y) is the covariance between variables X and Y,
- σXσX is the standard deviation of X,
- σYσY is the standard deviation of Y.

Pearson's R is widely used across various fields to assess the strength and direction of the linear relationship between two variables, providing valuable insights into their correlation.

Q4- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a pre-processing technique used in building a machine learning model to standardize the independent feature variables within a fixed range.
In a dataset, features can vary greatly in magnitude and units. Without scaling, these discrepancies can lead to incorrect modeling due to mismatched units among the features involved in the model.
The difference between **normalization** and **standardization** is that normalization adjusts all data points to fall within a range between 0 and 1, while standardization replaces the values with their Z scores.

Q5- You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Perfect multicollinearity: This occurs when two or more independent variables in a regression model are perfectly correlated. In such cases, one variable can be exactly predicted by a linear combination of the others, resulting in an R2R2value of 1 and an infinite Variance Inflation Factor (VIF).
High multicollinearity: Even if multicollinearity is not perfect but is very high, the R2R2 value can still approach 1, leading to a very large VIF.
Small sample size: When the sample size is small relative to the number of independent variables in the model, it can lead to unstable estimates and high VIF values, potentially reaching infinity.

Q6 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 evaluate whether a dataset follows a specified distribution, typically a normal distribution.

Axes:

- **X-axis:** Theoretical quantiles from the specified distribution.
- **Y-axis:** Sample quantiles from the dataset.

In linear regression, it is important to validate several assumptions to ensure the reliability of the model. One key assumption is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot is instrumental in assessing this assumption.

Importance of Q-Q Plots in Linear Regression:

- Ensuring that the residuals are normally distributed is crucial for the validity of hypothesis tests and confidence intervals in linear regression.
- Q-Q plots serve as a diagnostic tool to identify potential issues with the model, such as outliers or non-normality.