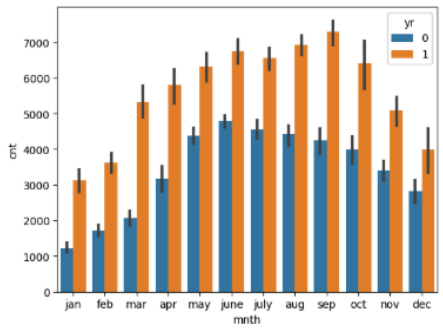
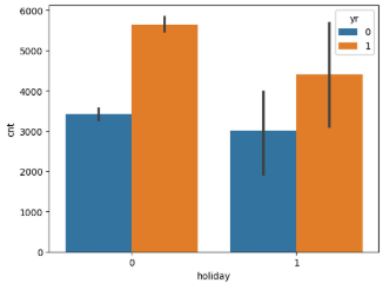
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

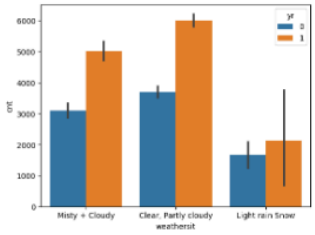
- The number of rented bikes rises during the fall season.
- Bike rentals are higher on non-working days.
- In 2019, the rental bike count exceeded that of 2018.
- More bikes were rented during clear weather conditions.



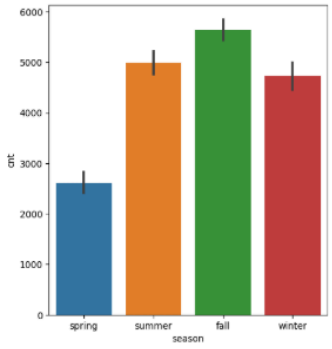


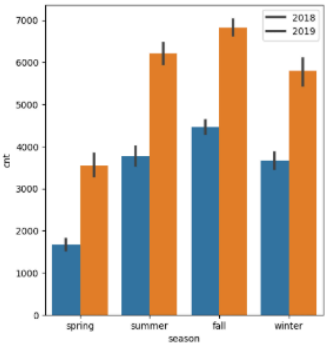
People are more interested is booking for holiday which we guessed intuitively while understanding data

Bookings has increased in year 2019 as compared to 2018 showing a trend of growth and the month june, july, August, September (sep topping the 2019 list) of 2019 has the most of booking against may. june(topping the 2019 list) july august had most booking from 2018



Based on above figure we can conclude that most of the bike rental happen in clear weather and it drops when the weather is rainy or snowy and there has been a significant increase in bookings for each season from 2018 to 2019





From the above figure we can see the fall season attracts more bike rentals

---

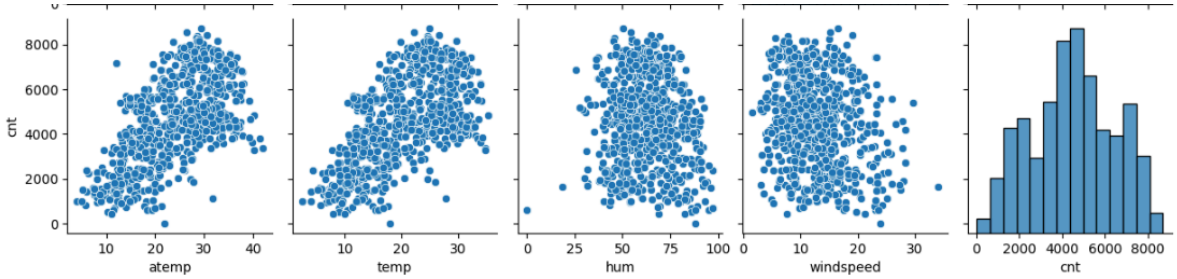
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

- One dummy variable is removed to retain only **k-1** categories.
- This helps reduce collinearity, improving both model performance and interpretability.
- Eliminating the extra column enhances model efficiency while preserving all essential information.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest

correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

- Temp column had the highest correlation with cnt variable making it a good predictor



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

- **Linearity of Relationship:**
  The response variable should have a linear relationship with the predictor variables, meaning any change in the predictors should result in a proportional change in the response.
- **Normality of Error Distribution:**
  The residuals (errors) should follow a normal distribution to ensure the model produces unbiased and reliable predictions.
- **Constant Variance of Errors (Homoscedasticity):**
  The residuals should maintain a consistent variance across all levels of the predictor variables.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- Temp
- Season
- weathersit

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a simple yet powerful machine learning algorithm used for predicting a continuous outcome based on input variables. It finds the best-fitting straight line (also called the **regression line**) that represents the relationship between the dependent variable (what we want to predict) and one or more independent variables (the factors influencing the prediction).

The equation for linear regression is:
Y = mX + b (for simple regression)
or
Y = b0 + b1X1 + b2X2 + ... + bnXn (for multiple regression),
where:

- **Y** is the predicted value,
- **X** represents the input variables,
- **m** (or **b1, b2, ...**) are the coefficients that determine the impact of each variable,
- **b** (or **b0**) is the intercept (the value of Y when all Xs are zero).

Linear regression works by minimizing the difference between actual and predicted values using a method called **least squares**, which reduces the sum of squared errors. It assumes that the relationship between variables is **linear**, errors are **normally distributed**, and there is **constant variance** in the residuals. This algorithm is widely used in finance, healthcare, and various industries to make predictions and analyze trends.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties, such as mean, variance, correlation, and linear regression line, but look very different when graphed. It was created by statistician **Francis Anscombe** to show the importance of visualizing data instead of relying only on summary statistics.

Each dataset in the quartet has the same average **x** and **y** values, the same regression equation, and the same correlation, yet their scatter plots reveal very different patterns: one follows a linear trend, another is curved, one has an outlier that affects the regression, and one shows a vertical cluster of points. This highlights why data visualization is crucial in analysis, as numbers alone can be misleading without graphical representation.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also called the Pearson correlation coefficient, is a measure of the strength and direction of a linear relationship between two numerical variables. It ranges from -1 to 1:

- 1 means a perfect positive correlation (as one variable increases, the other also increases).
- -1 means a perfect negative correlation (as one increases, the other decreases).
- 0 means no correlation (no relationship between the variables).

Pearson's R helps in understanding how closely two variables are related, but it only works for linear relationships and doesn't imply causation. It's widely used in statistics, finance, and science to analyze trends and relationships in data.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is the process of transforming numerical data so that it falls within a specific range, making it easier for machine learning models to process. It ensures that no variable dominates due to differences in units or magnitude.

Scaling is performed to **improve model performance**, **speed up training**, and **prevent bias** toward larger values. It is especially important in algorithms that rely on distance measurements, such as k-NN, SVM, and gradient descent-based models.

There are two main types of scaling:

1. **Normalization (Min-Max Scaling):** Rescales data to a fixed range, usually **0 to 1**, using the formula:
   $$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$
   It preserves the distribution but can be sensitive to outliers.
2. **Standardization (Z-Score Scaling):** Transforms data to have a **mean of 0** and a **standard deviation of 1** using:

$$X' = \frac{X - \mu}{\sigma}$$

It handles outliers better and is useful when data follows a normal distribution.

Normalization is ideal when the data has fixed boundaries, while standardization is preferred for datasets with varying scales.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The **Variance Inflation Factor (VIF)** measures how much a predictor variable is correlated with other predictors in a regression model. A **VIF value of infinity** occurs when there is **perfect multicollinearity**, meaning one variable is an exact linear combination of others.

This happens when:

- Two or more variables are **highly correlated** (e.g., one is a duplicate or derived from another).
- There is **redundant information** in the dataset (e.g., including both temperature in Celsius and Fahrenheit).
- Dummy variables are incorrectly encoded (e.g., not dropping one category in one-hot encoding).

When VIF is infinite, the model cannot estimate coefficients properly, leading to unstable predictions. To fix this, remove or combine highly correlated variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) plot is a graph used to check if a dataset follows a particular

distribution, usually a normal distribution. It compares the quantiles (percentiles) of the dataset with the expected quantiles of a theoretical distribution. If the points lie along a straight 45-degree line, the data is normally distributed.

In linear regression, Q-Q plots are important for checking if the residuals (errors) follow a normal distribution, which is a key assumption for valid statistical inferences. If the plot shows significant deviations (curves or outliers), it suggests issues like skewness, heavy tails, or non-normality, which can impact model accuracy and confidence in predictions.