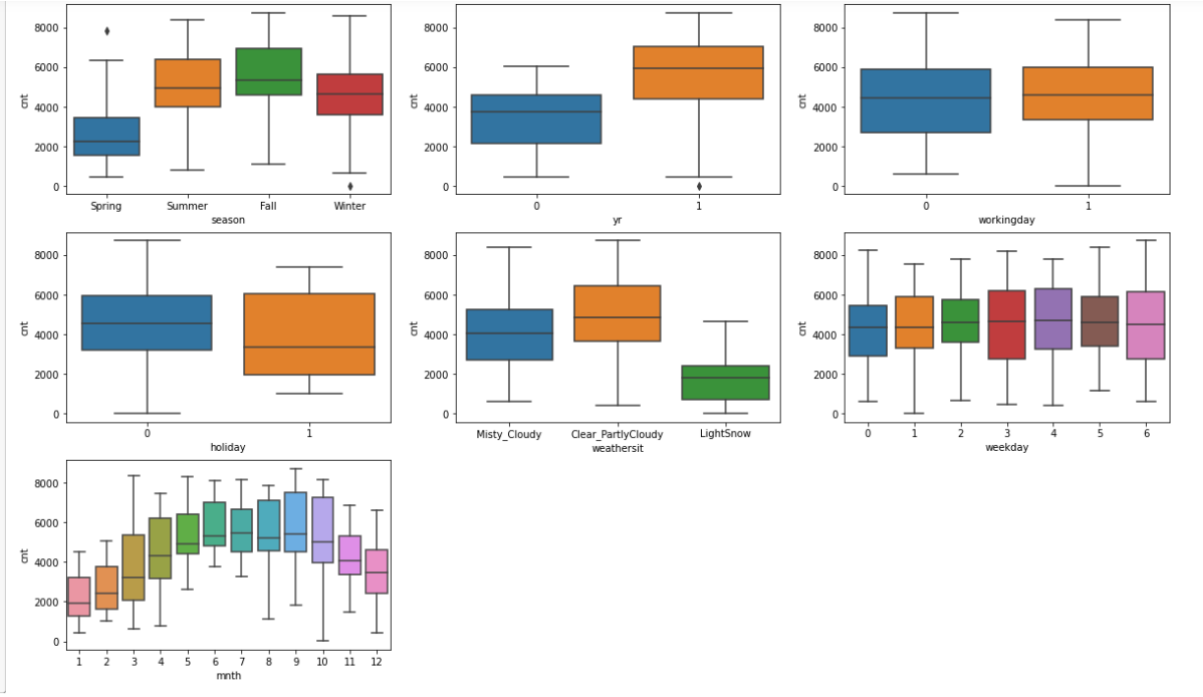
Assignment-based Subjective Questions

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 categorical variables in the dataset are-

- Season-Spring season had least value of bike demands 'cnt' and fall season had the maximum, summer and winter had intermediate values.
- WeatherSit-No users when there is heavy rain/snow indicating this weather is extremely unfavourabble. Highest count is seen in weather which is Clear and Partly Cloudy
- Mnth-September had highest no of rentals, and least in december and Jan months due to extreme cold and snow.
- Yr- Year 2019 has more no of rentals compared to 2018
- Weekday- cnt is almost same throught the week
- Holiday- rentals reduced during holiday
- WorkingDay – almost same throughout the week

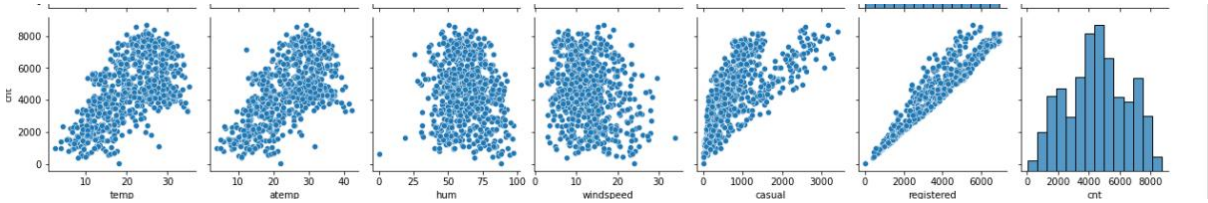
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

If we don't drop the first column then there would be the problem of multi collinearity in regression models which can cause unstable and unreliable model estimates. Dropping one category using drop_first=True, ensures the remaining dummy variables represent the effect relative to the dropped category.

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' and 'atemp' are 2 variables which are highly correlated with target variable 'cnt' among all the variables. (0.65)

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)
The following tests are done to validate the assumptions of linear regression-

1. Residuals distribution – residuals should follow normal distribution and should be centered around 0. we do this by plotting a distplot of residuals/Q-Q plot and see if residuals follow that or not.
2. Multi-collinearity Check- Independent variables should not be correlated and should have Variation Inflation Factor<=5.
3. Homoscedasticity- Plot Residuals vs. Fitted values to check constant variance of residuals.
4. Linearity (The relationship between independent and dependent variables is linear)- Plot Residuals vs. Fitted values using sns.residplot(), If the residuals form a pattern (curve or trend), linearity is violated.

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)
Top 3 features in my model are-

1. Temp- 0.415477
2. Yr- 0.236191
3. Weathersit_LightSnow(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)- -0.288847

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 fundamental supervised learning algorithm used for predicting a continuous dependent variable (Y) based on one or more independent variables (X). It establishes a relationship between the dependent and independent variables by fitting a straight line to the data.

Types of Linear Regression

a) Simple Linear Regression

- Involves only **one** independent variable (X).
- The relationship is modeled using a straight-line equation: $Y=mX+cY = mX + c$Y=mX+c where:
  - YYY = Predicted value (dependent variable)
  - XXX = Independent variable
  - mmm = Slope of the line (coefficient)
  - ccc = Intercept (constant)

b) Multiple Linear Regression

- Involves **multiple** independent variables.
- The equation extends to: $Y=m1X1+m2X2+...+mnXn+cY = m_1X_1 + m_2X_2 + ... + m_nX_n + c$Y=m1 X1 +m2 X2 +...+mn Xn +c where:
  - $X1,X2,...,XnX_1, X_2, ..., X_n$X1 ,X2 ,...,Xn are independent variables.
  - $m1,m2,...,mnm_1, m_2, ..., m_n$m1 ,m2 ,...,mn are their corresponding coefficients.
  - ccc is the intercept.

2. How Linear Regression Works

The algorithm finds the **best-fitting line** by minimizing the difference between the **actual values (Y)** and **predicted values (**$Y^|hat{Y}$**Y^)**.

Step 1: Compute the Predicted Value

For each input XiX_iXi , the model predicts:

$Y^i=mXi+c|hat{Y}_i = mX_i + c$Y^i =mXi +c

Step 2: Calculate the Error (Residuals)

The error (residual) is the difference between actual and predicted values:

$Error=Yi−Y^i|text{Error} = Y\_i$ - $|hat{Y}_i$Error=Yi −Y^i

Step 3: Find the Best-Fitting Line Using Cost Function

The most common cost function in linear regression is the **Mean Squared Error (MSE)**:

$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$MSE=n1 i=1∑n (Yi −Y^i)2

where nn is the number of observations.

Step 4: Optimize Parameters Using Gradient Descent

To minimize MSE, we update m and c iteratively using **gradient descent**:

m=m− α ∂MSE/∂ m ,c=c−α ∂MSE / ∂ c

where **α** is the learning rate.

3. Assumptions of Linear Regression

1. **Linearity** – The relationship between X and Y should be linear.
2. **Independence** – Observations should be independent.
3. **Homoscedasticity** – Constant variance of residuals.
4. **No Multicollinearity** – Independent variables should not be highly correlated.
5. **Normally Distributed Residuals** – Errors should be normally distributed.

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 designed by **Francis Anscombe (1973)** to demonstrate the importance of **data visualization** in statistical analysis. Each dataset has **identical summary statistics** but exhibits very different distributions when graphed.

1. The Importance of Anscombe's Quartet

Often, analysts rely only on **summary statistics** (mean, variance, correlation, regression line) to interpret data. However, Anscombe's Quartet shows that **datasets with the same statistics can have very different patterns**, emphasizing the need for **visual inspection**.

2. The Four Datasets in Anscombe's Quartet

Each dataset consists of **11 (X, Y) pairs** and shares the following statistical properties:

- **Mean of X:** ≈ 9.0
- **Mean of Y:** ≈ 7.5
- **Variance of X:** ≈ 11
- **Variance of Y:** ≈ 4.1
- **Correlation coefficient (r):** ≈ 0.816
- **Linear regression equation:** $Y=3+0.5XY = 3 + 0.5X$Y=3+0.5X

However, **when plotted**, they reveal vastly different distributions.

3. Explanation of Each Dataset

Dataset 1: A Typical Linear Relationship

- Points follow a **linear pattern** with some random noise.
- The least-squares regression line is a **good fit**.
- The dataset behaves as expected in regression analysis.

Dataset 2: A Non-Linear Relationship

- Appears **quadratic** rather than linear.
- The regression line does **not** represent the data well.
- The correlation is misleading because a simple linear model is inappropriate.

Dataset 3: An Outlier-Dominated Dataset

- Most data points fit the linear trend **except for one extreme outlier**.
- The outlier significantly **influences the regression line**.
- Without visualization, the model may be misinterpreted.

Dataset 4: A Vertical Outlier

- All X-values are **identical** except for one point.
- The **correlation is meaningless** because correlation relies on variability in X.
- A single point dictates the regression line.

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 (r)**, is a statistical measure that quantifies the **strength and direction** of a **linear** relationship between two variables. It ranges from **-1 to +1**.

1. Formula for Pearson's R

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:

- X_i, Y_i = Data points
- $\bar{X}, \bar{Y}$ = Mean of X and Y
- r = Correlation coefficient

Assumptions of Pearson's R

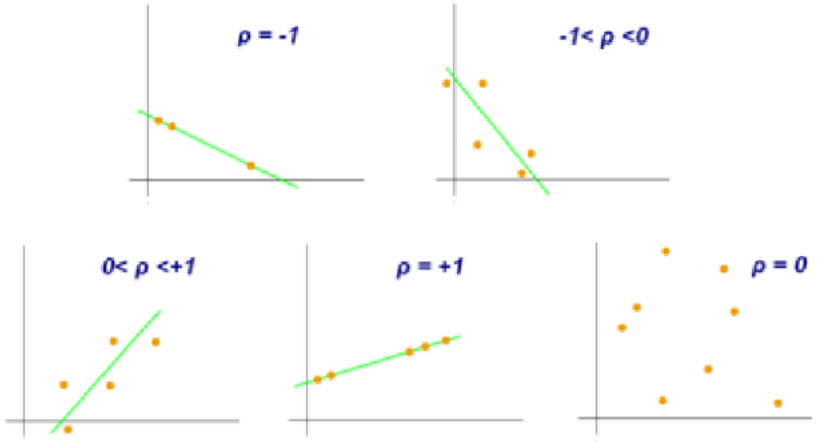
1. **Linear relationship** – Works only for linear associations.
2. **Continuous variables** – Both X and Y should be numerical.
3. **No extreme outliers** – Outliers can distort results.
4. **Normally distributed variables** – Works best if X and Y follow normal distributions.

As can be seen from the graph below,

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association



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 features so they fall within a specific range. This ensures that all features contribute equally to a machine learning model, preventing dominance by variables with larger magnitudes.

Why is Scaling Performed?

1. **Improves Model Performance** – Many ML algorithms (e.g., gradient descent, k-means, SVM) work better when features are on a similar scale.
2. **Faster Convergence** – Scaling speeds up optimization in models like logistic regression and neural networks.
3. **Prevents Feature Bias** – Features with larger values shouldn't disproportionately influence results.
4. **Enhances Distance-Based Algorithms** – Algorithms like k-NN and K-Means use distance calculations, which are affected by unscaled data.
5.

Types of Scaling: Normalization vs. Standardization

Method	Formula	Range	When to Use?
Normalization (Min-Max Scaling)	$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	[0,1] or [-1,1]	When data is **not normally distributed** and has **bounded values**. Common in deep learning.
Standardization (Z-score Scaling)	$X' = \frac{X - \mu}{\sigma}$	Mean = 0, SD = 1	When data is **normally distributed** or has **outliers**. Common in linear models and SVM.

Key Differences Between Normalization & Standardization

1. **Normalization (Min-Max Scaling)**
    a. Rescales data to **[0,1]** or **[-1,1]** range.
    b. Sensitive to **outliers**.
    c. Works well for **bounded datasets** (e.g., image pixels).
2. **Standardization (Z-score Scaling)**
    a. Centers data with **mean = 0** and **variance = 1**.
    b. Less sensitive to **outliers** than min-max scaling.
    c. Works well for **normally distributed** features.

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>

Variance Inflation Factor (VIF) quantifies **multicollinearity** in a regression model. It measures how

much the variance of a regression coefficient is **inflated** due to correlation with other independent variables.

$VIF=1/1-R2$
Where R2 is the coefficient of determination when X is regressed on all other independent variables.

Why Does VIF Become Infinite?

VIF becomes **infinite** when $R2=1,$ for any independent variable, meaning that the variable is **perfectly correlated** with one or more other variables.

$VIF=1/1-1=1/0=infinty$ This happens in cases like:

1. **Perfect Multicollinearity**
    a. When one independent variable is an **exact linear combination** of other variables.
    b. Example: If $X3=2X1+5X2X\_3 = 2X\_1 + 5X\_2$**X3 =2X1 +5X2** , then VIF for $X3X_3$X3 will be infinite.

2. **Duplicated Features**
    a. When the same feature is included twice in the dataset (e.g., **two identical columns**).
    b. Example: Having both "Height (cm)" and "Height (meters)" as separate columns.

3. **Highly Correlated Features**
    a. If two or more features have an almost **perfect correlation** (e.g., **correlation coefficient near ±1**).
    b. Example: "Total Price" and "Quantity × Unit Price" in a sales dataset.

4. **Small Sample Size with Many Variables**
    a. If the number of observations is **too small** compared to the number of predictors, some variables might appear **highly correlated** by chance.

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 graphical tool used to compare the distribution of a dataset with a theoretical distribution (usually a normal distribution). It helps determine whether a variable follows a specific distribution by plotting quantiles of the dataset against quantiles of the theoretical distribution.

Interpreting a Q-Q Plot?

Points on the straight line → Data follows the expected distribution (good!).
Points curve upward → Data has **more extreme values** than expected (long tails).
Points curve downward → Data has **fewer extreme values** than expected (short tails).
Points bend left or right → Data is **skewed** (not symmetrical).

Use & Importance of Q-Q Plot in Linear Regression

In linear regression, a key assumption is that the **residuals (errors) should be normally distributed**. A **Q-Q plot of residuals** helps check this assumption.

Why is it Important?

Validates Normality Assumption → Ensures residuals are approximately normal, which is required for accurate confidence intervals and hypothesis tests.
Detects Skewness & Heavy Tails → Helps identify deviations that may affect model performance.
Spots Outliers → Outliers appear as extreme points far from the reference line.
Improves Model Selection → If residuals are **not normal**, transformations or alternative models (e.g., robust regression) may be needed

---