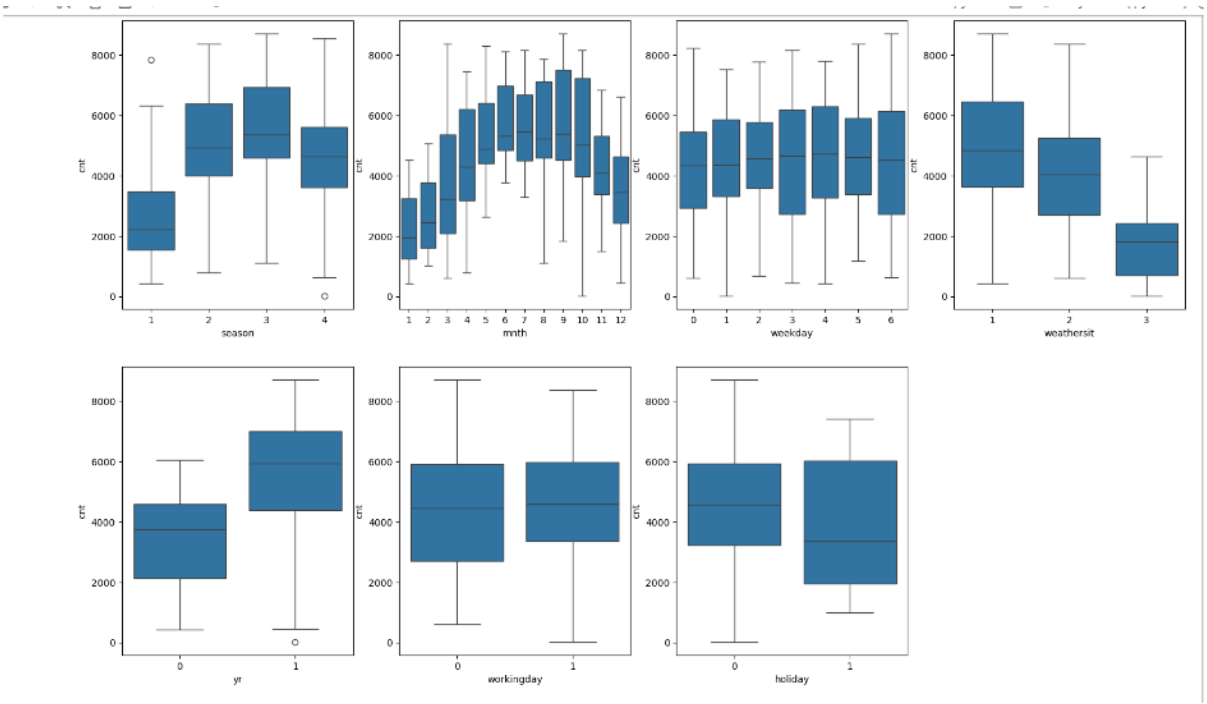
# Assignment-based Subjective Questions

**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:**




1. Season: Season has a significant effect on bike rentals, likely due to weather conditions (e.g., winter being colder).
2. Month: The monthly pattern aligns with seasonal trends, suggesting higher rentals during warmer months.
3. Day of the Week: weekday has little to no effect on the bike rental count, suggesting consistent demand irrespective of the day.
4. Weather Situation: Weather conditions have a strong influence on rentals, with adverse weather significantly reducing demand.
5. Year: There's a noticeable increase in demand over time, which could be due to the growing popularity of bike-sharing services.
6. Working Day: Demand is relatively stable, but there is a slight preference for non-working days, possibly due to leisure activities.
7. Holiday: Holidays may have a modest negative effect, potentially because people avoid commuting or use alternate transport.

**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:**

Using drop_first=True during dummy variable creation is important to avoid the issue of multicollinearity when using regression-based models.

When you set drop_first=True, one of the dummy variables (usually the first category) is dropped, and its absence acts as the **reference category**. The coefficients of the remaining dummy variables are interpreted relative to this reference category.

Example:**For Season:**

- Categories: Spring, Summer, Fall, Winter
- Dummy variables with drop_first=True: Summer, Fall, Winter
- Interpretation:
    - Coefficients of summer, Fall, and Winter indicate how these seasons differ from Spring (the dropped category).

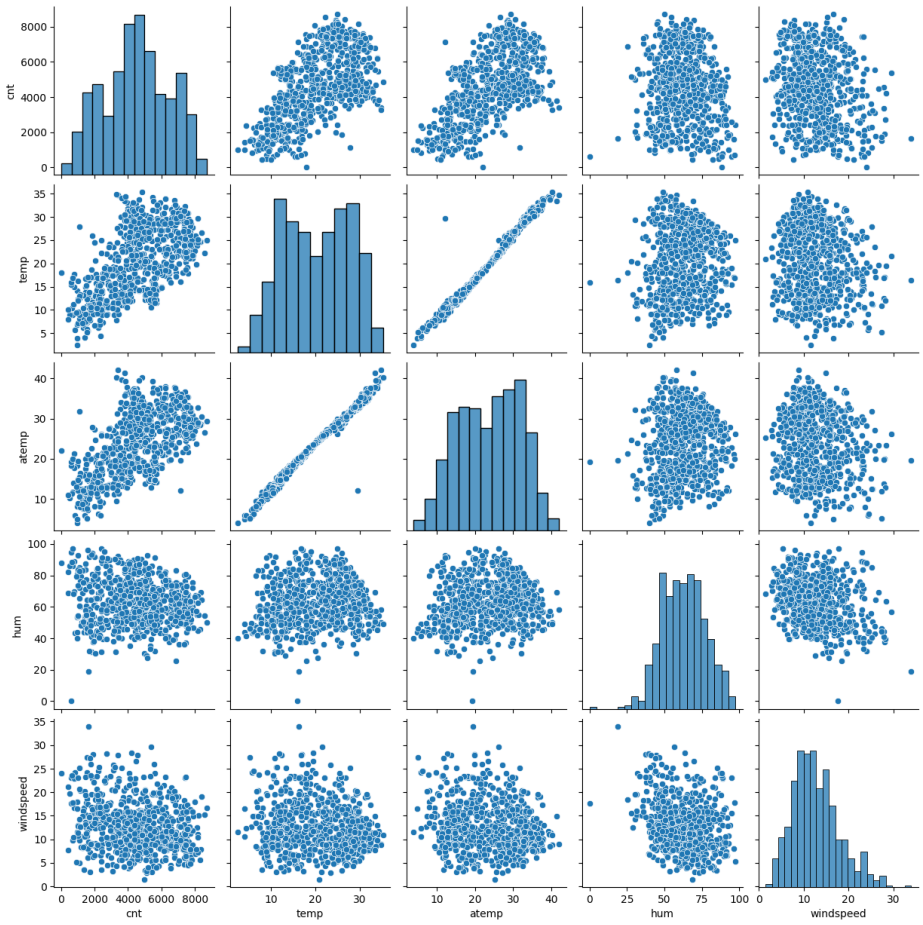
---

**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:

"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).




**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

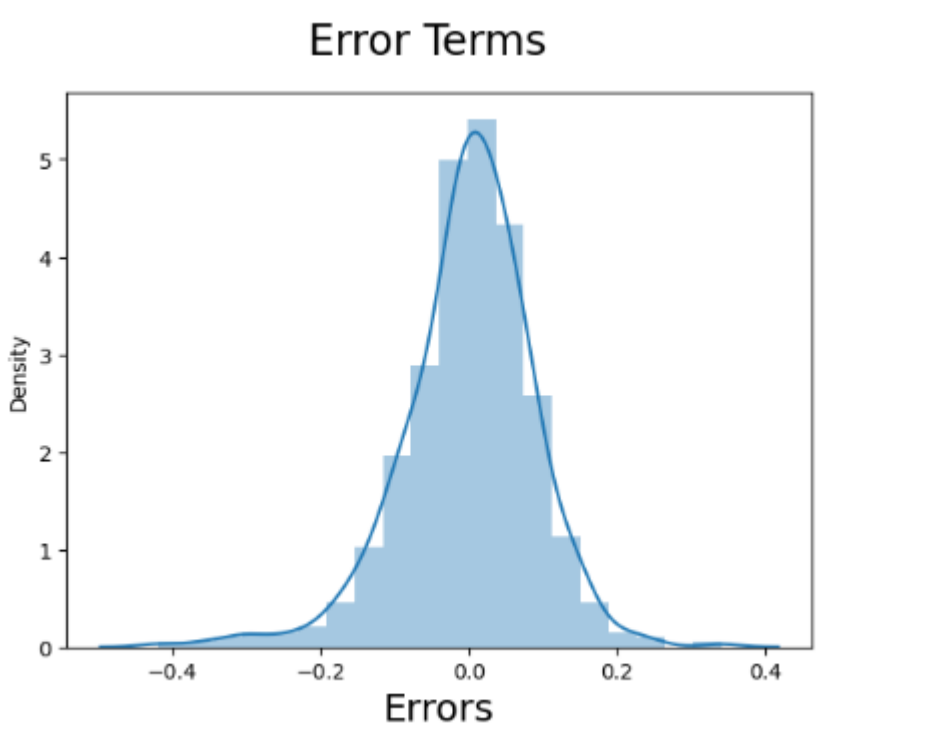
**Total Marks:** 3 marks (Do not edit)

**Answer:**

We have done following tests to validate assumptions of Linear Regression:

a. There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. (ref. see above question's pairplot)

b. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.




c. Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

---

**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: Based on the above methodology, the top 3 features are typically:

1.Temperature (temp): Positively correlates with demand 0.527708

2.Year(yr) : positive correlates with demand 0.229763

3. Weather (weathersit_Light Snow): Negatively correlates with demand -0.245009

---

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:**
Linear regression is a supervised machine learning algorithm which helps in predicting continues variables. It tries to find the best fit line which minimizes the error.

### Types of Linear Regression

- **Simple Linear Regression**: Involves one independent variable.

  y=mx+c

  - y: Dependent variable (target).
  - x: Independent variable (predictor).
  - m:slope
  - c:const
- **Multiple Linear Regression**: Involves multiple independent variables.

  Y=m1 x 2 + m2 x 2 + ……..+Mn xn +c

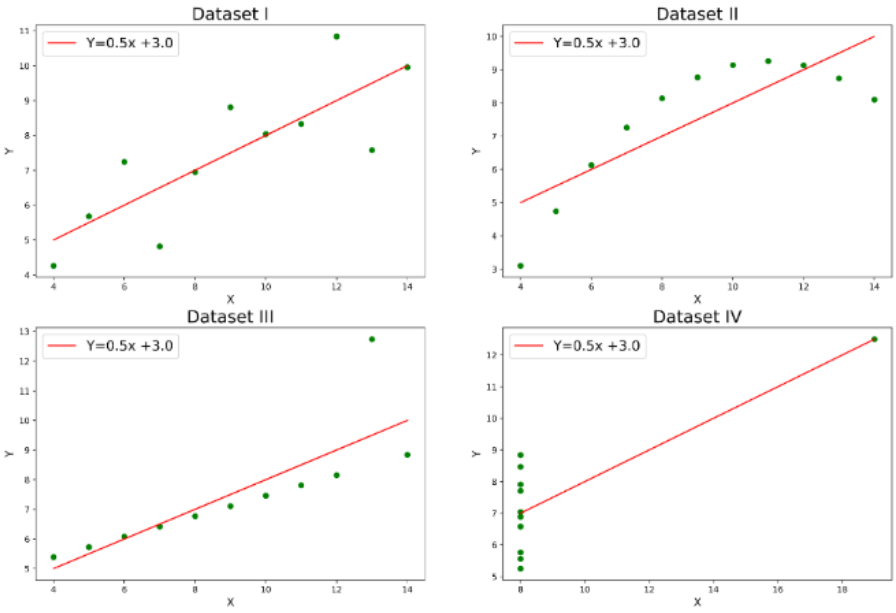
### Assumptions of Linear Regression

- **Linearity**: The relationship between predictors and the target is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The variance of the residuals (errors) is constant across all levels of the independent variables.
- **Normality of Errors**: Residuals are normally distributed.
- **No Multicollinearity**: In multiple regression, predictors should not be highly correlated with each other.

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:**
**Anscombe's quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.




*Anscombe's quartet Plot*

**1. Dataset 1**

- A typical linear relationship between x and y
- Data points are evenly distributed around the regression line.
- Suitable for linear regression.

**2. Dataset 2**

- A clear non-linear relationship.
- y remains constant for most x values but increases sharply at one point.
- Indicates a curve-like trend rather than a straight line.

**3. Dataset 3**

- Contains a strong linear trend, but one outlier skews the regression line.
- This shows how outliers can affect statistical summaries.

**4. Dataset 4**

- Most points are identical, except for a single influential outlier.
- The outlier drives the regression line, making it unrepresentative of the overall data.

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)

**Pearson's R**, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is widely used in statistics, data analysis, and machine learning to assess how strongly two variables are related.

**Formula for Pearson's R:**

The formula for calculating Pearson's R is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$r$ = correlation coefficient
$x_i$ = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
$y_i$ = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

**Properties of Pearson's R:**

**1. Range**: The value of r lies between −1 and 1.

r = 1: Perfect positive linear relationship (as one variable increases, the other increases proportionally).

r = −1: "Perfect negative linear relationship (as one variable increases, the other decreases proportionally).

r = 0: No linear relationship between the variables.

**2. Symmetry**: The correlation of X with Y is the same as Y with X.

**3. Unit less**: Pearson's R is a dimensionless measure and does not depend on the scale of the variables.

**4. Linear Relationships**: It only captures the strength of linear relationships, not non-linear ones.

Use Cases:

- Identifying relationships between variables (e.g., income and expenditure).

- Evaluating features for predictive models in machine learning.
- Hypothesis testing to assess if a correlation is statistically significant.

## Limitations:

- It assumes a linear relationship between variables; non-linear relationships may result in misleading values.
- Sensitive to outliers, which can distort the correlation value.
- It does not imply causation (a high r value doesn't mean one variable causes the other to change).

---

**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)

**Scaling** refers to the process of adjusting the range or distribution of data features so they fit within a specific range or distribution. This is especially important in machine learning and statistical analysis to ensure features are on comparable scales, which helps algorithms converge faster and perform better.

Scaling is performed for:

1. **Improved Model Performance**: Many algorithms (like gradient descent-based models and distance-based models) are sensitive to feature magnitudes. Scaling ensures no feature dominates others due to its larger range.
2. **Faster Convergence**: In optimization problems, scaled features help in faster convergence during training.
3. **Prevention of Bias**: Without scaling, models may assign higher importance to variables with larger scales, leading to biased results.
4. **Comparability**: Scaling ensures that different features are comparable and reduces the risk of numerical instabilities.
5. **Algorithm-Specific Requirements**: Algorithms like SVMs, K-means, PCA, and neural networks require scaled data for proper functionality.

**Diffrence between Normalized Scaling and Standardized Scaling:**

| Aspect | Normalized Scaling | Standardized Scaling |
|---|---|---|
| **Definition** | Rescales data to a fixed range, typically [0, 1]. | Transforms data to have a mean of 0 and a standard deviation of 1. |
| **Formula** | $X' = \frac{X - Xmin}{Xmax - Xmin}$ | $z = \frac{x_i - \mu}{\sigma}$ |
| **Range** | Typically [0, 1] (or [-1, 1] depending on the scaling function). | No fixed range; data is centered around 0 and scaled by standard deviation. |
| **Purpose** | Ensures all data falls within the same range, useful for feature comparison. | Ensures data is standardized to have the same scale and center. |
| **Sensitivity to Outliers** | Highly sensitive to outliers (outliers can distort the range). | Less sensitive to outliers since it uses mean and standard deviation for scaling. |
| **Use Cases** | Used in algorithms where a fixed range is needed (e.g., image processing, neural nets). | Used in algorithms assuming Gaussian distributions (e.g., PCA, logistic regression). |
| **Example Algorithms** | K-Nearest Neighbors (KNN), Neural Networks. | Logistic Regression, PCA, Linear Regression. |

---

**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:**
The **Variance Inflation Factor (VIF)** measures how much the variance of a regression coefficient is inflated due to multicollinearity among the predictors in a regression model. A VIF value of **infinity** indicates perfect multicollinearity. Here's why this happens:

$$VIF = \frac{1}{1 - R^2}$$

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in “infinity” The numerical value for VIF tells you what percentage the variance is inflated for each coefficient.

For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

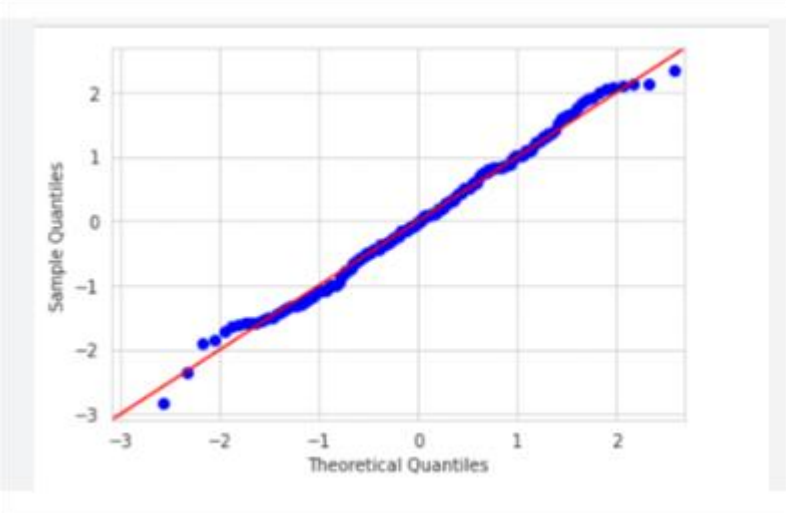
- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

---

**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:**

A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the **quantiles** of the observed data against the **quantiles** of the theoretical distribution.

- If the data follows the theoretical distribution (e.g., normal), the points in the Q-Q plot will align approximately along a **straight 45-degree line**.
- Deviations from the line indicate deviations from the assumed distribution.




**Use of Q-Q Plots in Linear Regression**

In linear regression, one of the assumptions is that the residuals (errors) follow a normal distribution. A Q-Q plot is used to **visually assess this assumption**.

Steps in Linear Regression Analysis:

1. **Obtain Residuals**: After fitting a regression model, calculate the residuals (y actual−y predicted).
2. **Create a Q-Q Plot**: Plot the quantiles of the residuals against the theoretical normal quantiles.
3. **Interpretation**:
    - Points align along the line: Residuals are normally distributed.
    - Systematic deviations:
        - **S-shaped curve**: Indicates heavy tails (data has outliers).
        - **Upward or downward curves**: Indicates skewness (data is not symmetric).

**Importance of Q-Q Plot in Linear Regression**

1. **Assumption Validation**:
    - Normality of residuals is a key assumption in linear regression for reliable hypothesis testing and accurate confidence intervals.
2. **Model Diagnostics**:
    - Identifies deviations from normality, which can suggest problems with the model, such as:
        - Presence of outliers.
        - Incorrect functional form.
        - Heteroscedasticity (non-constant variance of residuals).
3. **Model Improvement**:
    - If residuals are not normal:
        - Apply transformations (e.g., log, square root).
        - Use robust regression methods that do not assume normality.
4. **Visual Simplicity**:
    - Provides a quick and intuitive way to assess normality without relying solely on numerical tests (e.g., Shapiro-Wilk test).