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)

From the analysis of the categorical variables in the dataset, we can infer the following about their effects on the dependent variable (cnt):

1. **Season:**
   - Winter and Summer positively influence cnt, indicating higher activity during these seasons.
   - Spring has a negative effect, likely due to unpredictable weather patterns reducing activity.
2. **Month:**
   - September stands out with a strong positive effect, likely due to favorable weather conditions.
   - January and July show negative effects, possibly due to extreme weather (cold or heat) or reduced activity during vacations.
3. **Day of the Week:** Saturday positively affects cnt, indicating higher recreational activity on weekends.
4. **Working Day:** A positive effect suggests that commuting-related activity boosts counts on working days.
5. **Holidays:** A negative effect indicates reduced activity, suggesting the dataset is influenced by commuting patterns, which decline on holiday
6. **Weather:** Mist/Cloudy and Light Snow/Rain have significant negative effects, with adverse weather conditions greatly reducing activity.

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)

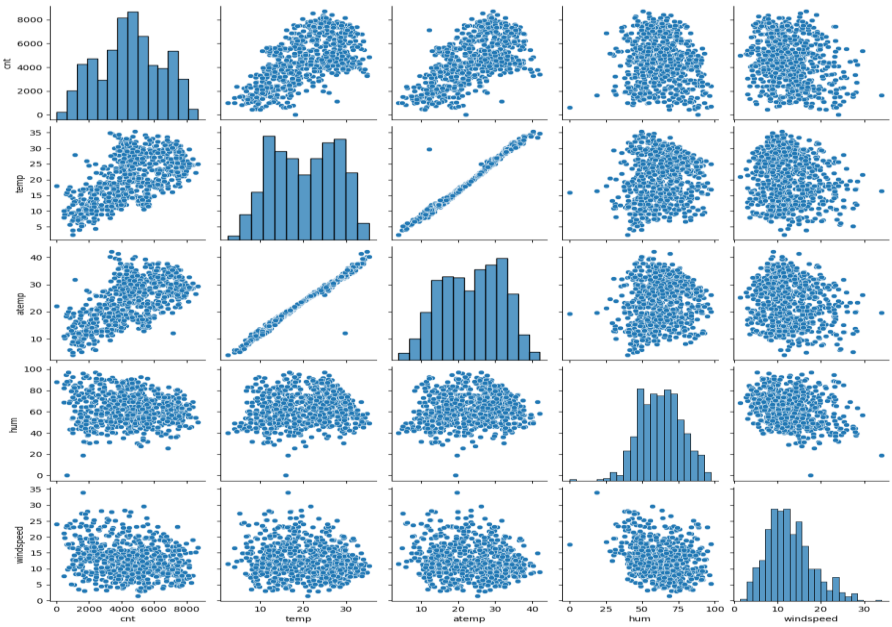
1. **Avoids Dummy Variable Trap:** Prevents multicollinearity by eliminating redundant dummy variables.
2. **Establishes a Reference Category:** Simplifies interpretation by comparing other categories to a baseline.
3. **Ensures Model Stability:** Makes linear models computationally efficient and prevents over-specification.

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)

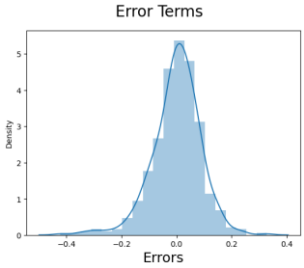
As we can see “temp” and “atemp” numerical variables are highly correlation to the target variable(cnt)



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: <Your answer for Question 4 goes below this line> (Do not edit)

The following tests were done to validate the assumptions of linear regression:

1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualized the numeric variables using a pair plot to see if the variables are linearly related or not. Refer to the notebook for more details.
2. **Normality:** Residuals should follow a normal distribution.



3. **No Multicollinearity:** VIF values should be less than 5. Refer to the notebook for more details.

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)

Based on the coefficients, the top 3 features (both positive and negative) contributing most significantly to explaining the demand for shared bikes (cnt) are:

1. **Temperature (temp)**
   - **Coefficient:** 0.472207 (Positive)
   - **Interpretation:** Temperature has the strongest positive impact on bike demand. As temperature increases, bike demand also increases.
2. **Weather Situation (Light Snow & Rain) (weathersit Light Snow & Rain)**
   - **Coefficient:** -0.290800 (Negative)
   - **Interpretation:** Bad weather, such as light snow or rain, significantly decreases bike demand.
3. **Year (yr)**
   - **Coefficient:** 0.234461 (Positive)
   - **Interpretation:** The demand for bikes increases with the year, likely reflecting growth in bike-sharing usage over time.

These three features—Temperature (positive), Light Snow & Rain weather (negative), and Year (positive)—are the most significant in explaining bike demand.

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 statistical method used for predicting a continuous dependent variable based on one or more independent variables. The primary goal of linear regression is to establish a linear relationship between the dependent variable and the independent variables.

Formula:

- For **simple linear regression** (one independent variable), the formula is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- For **multiple linear regression** (multiple independent variables), the formula is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- y = predicted value
- $X_1, X_2, \ldots, X_n$ = independent variables
- β_0 = intercept term
- $\beta_1, \beta_2, \ldots, \beta_n$ = coefficients for the independent variables
- ϵ = error term

Steps in Linear Regression:

- Data Preparation: Clean and preprocess the data (e.g., handle missing values, convert categorical variables).
- Model Training: Train the model to estimate the coefficients that minimize the error, typically using the Least Squares method.
- Optimization: Use techniques like Gradient Descent to minimize the error between predicted and actual values.
- Model Evaluation: Evaluate the model using metrics like R-squared (R^2) and Mean Squared Error (MSE).
- Prediction: Use the trained model to make predictions on new data.

Assumptions of Linear Regression:

- Linearity: The relationship between dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of errors remains constant.
- Normality of Errors: The errors are normally distributed.
- No Multicollinearity: The independent variables should not be highly correlated.

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 simple descriptive statistics, yet they are visually very different when plotted. The purpose of the quartet is to show the importance of graphical analysis and the potential dangers of relying solely on statistical summaries (like mean, variance, and correlation) without visualizing the data.

It was created by the statistician Francis Anscombe in 1973 to demonstrate the need for data visualization in statistical analysis. Each of the four datasets consists of 11 data points, and they are structured in such a way that they share the same key statistics but exhibit different relationships when plotted.

1. **Dataset 1 (Standard Linear Relationship)**:

    - This dataset follows a simple linear trend with a positive correlation between X and Y.
    - **Plot:** The data points are scattered along a straight line with some noise.
    - **Characteristics:** This is a typical scenario where linear regression works well.

2. **Dataset 2 (Non-linear Relationship with Outlier)**:

    - This dataset also shows a positive correlation between X and Y, but with a clear **curved (non-linear) pattern**, and one **outlier** that significantly affects the fit of the regression line.
    - **Plot:** The data points are arranged in a curve, and the outlier skews the regression line.

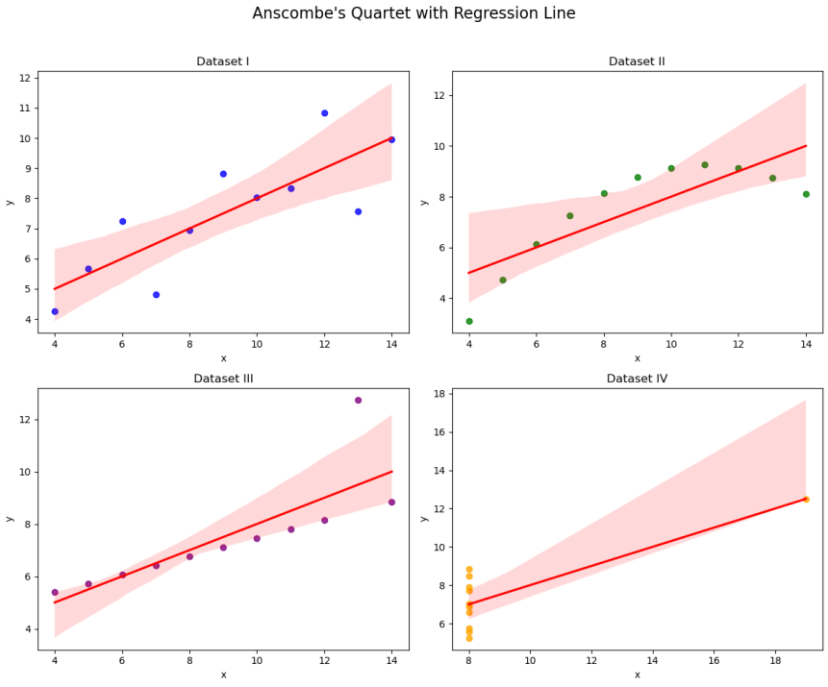
- **Characteristics:** A linear regression model may not be the best fit for this type of data, and the outlier can lead to misleading results.

3. **Dataset 3 (Vertical Line with Outlier)**:

   - In this case, the dataset consists of mostly vertical data points with a few outliers.
   - **Plot:** A majority of the points fall along a vertical line, while a single outlier is far from the others.
   - **Characteristics:** The high correlation value of 0.82 is driven mainly by the outliers, and the data doesn't follow any meaningful linear relationship. In such a case, correlation may not be a good measure of the relationship.

4. **Dataset 4 (Horizontal Line with Outlier)**:

   - This dataset has a majority of points that lie on a horizontal line.
   - **Plot:** The data points lie along a horizontal line, and an outlier far away from the other points.
   - **Characteristics:** Even though the correlation appears to be high, the linear regression model does not fit well. Like Dataset 3, correlation here is not representative of the data's true nature.



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>

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Formula for Pearson's R:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Where:

- r = Pearson's correlation coefficient
- X and Y = the two variables being compared
- n = the number of data points
- $\sum X$ = sum of the values of variable X
- $\sum Y$ = sum of the values of variable Y
- $\sum XY$ = sum of the product of X and Y for each data point
- $\sum X^2$ = sum of the squared values of X
- $\sum Y^2$ = sum of the squared values of Y

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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 refers to the process of adjusting the range of features (variables) in a dataset to a standard or common scale. This is crucial when applying machine learning algorithms that are sensitive to the magnitudes of features

1. **Improves Convergence in Algorithms**: Many machine learning algorithms, like **gradient descent**, converge faster when the features are scaled to similar ranges.
2. **Prevents Dominance of Features**: Features with larger numeric ranges can dominate the model training process, leading to biased results. Scaling ensures that all features contribute equally to the model.
3. **Equal Contribution**: Scaling ensures that each feature contributes equally to the prediction, particularly when features have different units (e.g., height in cm vs. weight in kg).

Types of Scaling

Normalized Scaling (Min-Max Scaling)
Standardized Scaling (Z-Score Scaling)

Key Differences:

Aspect	Normalized Scaling (Min-Max)	Standardized Scaling (Z-Score)
Range	[0, 1] (or another fixed range)	Any range, with mean = 0 and std = 1
Effect of Outliers	Sensitive to outliers (can distort results)	Less sensitive but can still be affected
Application	Suitable when data needs to be within a fixed range	Best for algorithms assuming normal distribution
Interpretation	Easier to interpret in context (e.g., data in a 0 to 1 scale)	More abstract, but often preferred for many models

---

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) is a measure used to detect multicollinearity in regression models. It quantifies how much the variance of the estimated regression coefficients is inflated due to collinearity with other independent variables. Specifically, VIF tells us how much a given predictor variable's variance is inflated because of the linear relationships with the other predictor variables.

A VIF value can become infinite when multicollinearity is perfect. This means that one predictor variable is perfectly correlated with one or more of the other predictor variables.

How VIF is Calculated:

VIF for a particular variable is calculated using the following formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where:

- VIF_i is the Variance Inflation Factor for the i^{th} variable.
- R_i^2 is the **coefficient of determination** (R-squared value) when the i^{th} variable is regressed on all other variables in the model.

When R_i^2 approaches **1** (i.e., perfect correlation), the denominator becomes very small, causing **VIF to approach infinity**.

1. **Perfect Multicollinearity:**
   - If one independent variable is a perfect linear combination of other variables in the model (i.e., there is exact correlation), the VIF for that variable will be infinite.
   - For example, if Variable A = 2 * Variable B, there is perfect collinearity, and the VIF for Variable A (or Variable B) will be infinite.
2. **Identical Variables:**
   - If two variables are exactly the same (e.g., you mistakenly add a column that is a duplicate of another), this will cause perfect correlation, leading to infinite VIF for those variables.
3. **Rank Deficiency:**
   - A dataset that is rank deficient means there is a linear dependence among the variables, such that the design matrix is not of full rank. This results in perfect collinearity and causes infinite VIF.

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 assess if a dataset follows a particular theoretical distribution, most commonly the **normal distribution**. It compares the **quantiles** of the sample data against the quantiles of the chosen distribution (e.g., normal distribution).

- **X-axis**: Quantiles from the theoretical distribution (e.g., standard normal distribution).
- **Y-axis**: Quantiles from the sample data.

If the data follows the specified distribution, the points in the Q-Q plot will fall approximately along a **straight line**.

Use and Importance of a Q-Q Plot in Linear Regression

In the context of **linear regression**, the Q-Q plot is mainly used to **check the residuals** (the difference between observed and predicted values) to ensure the assumptions of the regression model are met.

Linear regression assumes that:

1. The residuals (errors) are **normally distributed**.
2. There is **homoscedasticity** (constant variance of residuals).
3. There is **no autocorrelation** in residuals.

A Q-Q plot helps validate the **normality assumption** of residuals, which is important for:

- **Statistical Inference**: Validating normality allows for accurate confidence intervals and hypothesis tests on the regression coefficients (e.g., t-tests and F-tests).
- **Predictive Accuracy**: If residuals deviate significantly from normality, the model may not be properly fitted, leading to less accurate predictions.

Types of Q-Q plots

There are several types of Q-Q plots commonly used in statistics and data analysis, each suited to different scenarios or purposes:

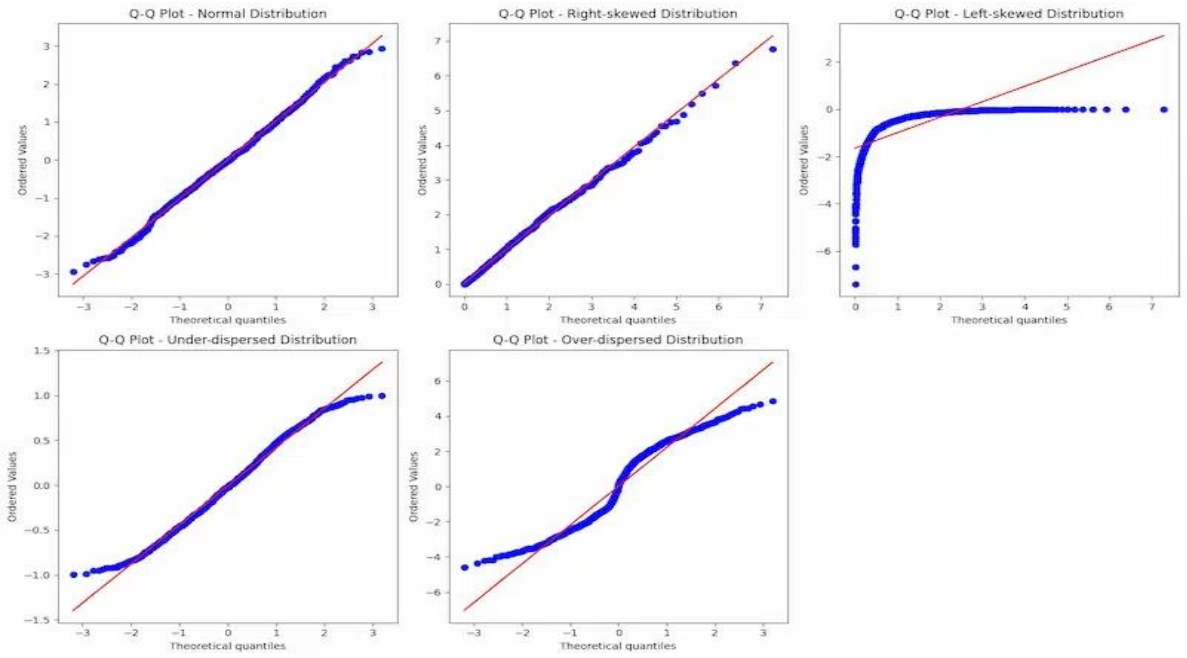
1. **Normal Distribution**: A symmetric distribution where the Q-Q plot would show points approximately along a diagonal line if the data adheres to a normal distribution.

2. **Right-skewed Distribution**: A distribution where the Q-Q plot would display a pattern where the observed quantiles deviate from the straight line towards the upper end, indicating a longer tail on the right side.

3. **Left-skewed Distribution**: A distribution where the Q-Q plot would exhibit a pattern where the observed quantiles deviate from the straight line towards the lower end, indicating a longer tail on the left side.

4. **Under-dispersed Distribution**: A distribution where the Q-Q plot would show observed quantiles clustered more tightly around the diagonal line compared to the theoretical quantiles, suggesting lower variance.

5. **Over-dispersed Distribution**: A distribution where the Q-Q plot would display observed quantiles more spread out or deviating from the diagonal line, indicating higher variance or dispersion compared to the theoretical distribution.


Q-Q Plot - Normal Distribution
Q-Q Plot - Right-skewed Distribution
Q-Q Plot - Left-skewed Distribution
Q-Q Plot - Under-dispersed Distribution
Q-Q Plot - Over-dispersed Distribution
Ordered Values
Theoretical quantiles