## 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:**

Effect of categorical variables on the dependent variable (cnt):

- Season has a strong effect on bike rentals, with highest demand in summer/fall and lowest in winter
- Year shows an increasing trend from 2018 to 2019, indicating growing popularity
- Weather situation significantly impacts rentals with clear weather having highest demand
- Working days show higher rental counts compared to holidays/weekends
- The month variable indicates seasonal patterns with peak rentals in summer months

---

**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)

- It helps avoid the dummy variable trap by preventing multicollinearity.
- When creating dummy variables, if we include all categories, it leads to perfect multicollinearity.
- By dropping one category (first), we avoid this while still retaining all the information, as the dropped category becomes the reference level which can be inferred from the other dummy variables.

---

**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)

Highest correlation with target variable:
Looking at the correlation matrix, temperature ('temp') has the highest positive correlation with the target variable 'cnt', followed closely by 'atemp' (feels-like temperature).

---

**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)

Validation of Linear Regression assumptions:

- Linearity: Checked through scatter plots between predictors and target
- Independence: Verified through residual plots showing no patterns
- Normality: Validated using Q-Q plot showing residuals follow normal distribution
- Homoscedasticity: Residual plot shows relatively constant variance
- No multicollinearity: Checked through correlation matrix and VIF values

---

**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 contributing significantly: Based on the feature importance analysis:

- Weather situation (clear weather)
- Temperature
- Season (fall)

---

## 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 to model the relationship between a dependent variable (also known as the target or outcome) and one or more independent variables (features). It's widely used in predictive modeling and machine learning for forecasting and understanding variable relationships. Here's a detailed breakdown of the algorithm:

## 1. Objective of Linear Regression

- Linear Regression aims to find the best-fit line, or "regression line," that minimizes the difference (error) between predicted values and actual values. This line represents the linear relationship between input features and the target output, typically represented by the equation:

$$y=\beta 0 +\beta 1\ x1 +\beta 2\ x2 +\cdots+\beta n\ xn +\epsilon$$

where:

y is the predicted output,
β0 is the y-intercept,
β1 ,β2 ,…,βn are coefficients for each feature x
ϵ is the error term.

## 2. Cost Function and Error Minimization

- To determine the best-fit line, Linear Regression uses a **cost function**—typically **Mean Squared Error (MSE)**, which is the average of the squared differences between predicted and actual values:
-

$$MSE=N1\ i=1\sum N\ (yi\ -y\hat{}\ i\ )2$$

where

yi is the actual value,

y^ i is the predicted value, and

N is the number of observations.

The model adjusts the coefficients to minimize this cost function using optimization techniques like **Gradient Descent**.

## 3. Gradient Descent for Optimization

- **Gradient Descent** is an iterative optimization algorithm that adjusts the model's parameters (coefficients) to minimize the cost function. It calculates the gradient (slope) of the cost function with respect to each parameter and updates the parameters in the opposite direction of the gradient:

$$\beta j =\beta j -\alpha\partial\beta j\ \partial\ MSE$$

where

α is the learning rate.

This process continues until convergence, where further updates make negligible changes to the cost.

## 4. Assumptions of Linear Regression

- For Linear Regression to perform optimally, certain assumptions should ideally be met:
- **Linearity**: The relationship between input features and output is linear.
- **Independence**: Observations are independent.
- **Homoscedasticity**: The variance of errors remains constant across all levels of independent variables.
- **Normality of Errors**: Errors should be normally distributed for reliable predictions and confidence intervals.

This straightforward, interpretable model is foundational in machine learning and statistics due to its simplicity, making it a useful starting point in data analysis.

**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 appear very different when graphed. It was created by statistician Francis Anscombe to demonstrate the importance of visualizing data before performing statistical analysis. Here's a breakdown of each dataset in the quartet:

1. **Identical Summary Statistics**: All four datasets have nearly identical values for the following:
    - Mean of the x and y values
    - Variance of x and y values
    - Correlation between x and y
    - Linear regression line ($y = mx + c$)
2. Despite these similarities, the datasets differ in structure and relationship when graphed, which underscores how summary statistics can be misleading without visual context.
3. **Individual Differences**:
    - **Dataset I**: A simple linear relationship, which is well-suited for linear regression. The data points are close to the line.
    - **Dataset II**: A nonlinear relationship. The points form a curve, showing that the linear model doesn't fit well.
    - **Dataset III**: A linear relationship with an outlier, which skews the regression line despite the rest of the data aligning closely.
    - **Dataset IV**: A nearly vertical line where most points have the same x value, making correlation misleading as a measure of association.
4. **Key Takeaway**: Anscombe's quartet illustrates that relying solely on numerical summaries without data visualization can lead to incorrect conclusions. It emphasizes that visualizing data is crucial for understanding the true nature of the relationships within it.

**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 known as the Pearson correlation coefficient, is a statistic that measures the strength and direction of a linear relationship between two continuous variables. Here are the key points:

1. **Range and Interpretation**:
- Pearson's r

  r ranges from -1 to +1.

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

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

  r=0: No linear relationship between the variables.

2. **Formula and Calculation**:
- The formula for Pearson's r is

  $$r = n \sum X Y - \sum X \sum Y \, ( n \sum X 2 - ( \sum X ) 2 ) \cdot ( n \sum Y 2 - ( \sum Y ) 2 ) .$$

  It essentially calculates the covariance of the two variables normalized by their standard deviations, giving a unitless measure of association.

**Application**:
- Pearson's r is widely used in statistics and data analysis to assess linear relationships.
- However,
- it only captures linear associations, so it may be misleading if the relationship between variables is nonlinear.

**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 a data preprocessing technique used to adjust the range and distribution of numerical features. It's done to ensure all features contribute equally to the model, especially when they have varying units or scales, which can improve the accuracy and efficiency of machine learning models.

1. **Purpose of Scaling**:
   - Scaling improves the performance and convergence of many algorithms (like gradient descent-based methods).
   - It ensures that features with large magnitudes do not dominate features with smaller ones.
   - It helps in maintaining a consistent range, which is especially important for distance-based models (e.g., K-Nearest Neighbors).
2. **Types of Scaling**:
   - **Normalized Scaling**: This process, often called *Min-Max Scaling*, rescales features to a specific range, typically [0, 1]. The formula is:
   - Xnormalized =Xmax –Xmin X–Xmin I
   - t's useful when you need to bound values to a specific range, like in image processing or neural networks.

   - **Standardized Scaling**: Also called *Z-score normalization*, it transforms the data to have a mean of 0 and a standard deviation of 1:

     Xstandardized =σX–μ This approach is often better for algorithms assuming normally distributed data (e.g., linear regression).

In summary, normalized scaling constrains values within a fixed range, while standardized scaling transforms data to a distribution with mean 0 and standard deviation 1.

**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 assess multicollinearity in a regression model. VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. Sometimes, the value of VIF becomes infinite, and here's why that happens:

1. **Perfect Multicollinearity**:
   - An infinite VIF occurs when there is *perfect multicollinearity* in the dataset, meaning one predictor variable can be expressed as an exact linear combination of other predictor(s).
   - This results in a denominator of zero in the VIF formula, which causes the VIF value to approach infinity.
2. **Formula and Cause**:
   - VIF for a predictor Xj is calculated as:

     VIF(Xj )=1−Rj2 1 where Rj2 R2 value obtained by regressing Xj on all other predictors

     Rj2 =1 (perfect linear relationship), the denominator becomes zero, making VIF infinite.
3. **Implication**:
   - An infinite VIF signals severe multicollinearity, which means that one or more predictors are redundant.
   - To resolve this, you typically need to remove or combine collinear variables to stabilize the model.

**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 plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, often the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution.

1. **Use in Linear Regression**:
    - In linear regression, one key assumption is that the residuals (errors) are normally distributed. The Q-Q plot helps verify this by comparing the distribution of residuals to a normal distribution.
    - If the residuals follow a normal distribution, the points in the Q-Q plot should lie approximately along a straight diagonal line.
2. **Interpreting a Q-Q Plot**:
    - **Straight Line**: Points following a straight line indicate that the residuals are approximately normally distributed.
    - **Curvature**: Deviations from the line (e.g., S-shaped or curved patterns) suggest skewness or other non-normal characteristics.
    - **Outliers**: Points that stray far from the line at the ends indicate outliers in the data.
3. **Importance in Linear Regression**:
    - A Q-Q plot is essential for diagnosing the normality of residuals in regression analysis. Non-normally distributed residuals can violate the assumptions of linear regression, impacting the reliability of confidence intervals, hypothesis tests, and the validity of the model.

---