ASSIGNMENT QUESTIONS

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Effect of Categorical Variables on Bike Rentals (cnt)

Season:

Effect: Bike rentals vary by season. For example:

High in Summer and Fall due to pleasant weather.

Low in Winter because of cold conditions.

Weather Situation (weathersit):

Effect: Rentals are highest in Clear weather and decrease in bad weather like Mist, Light Rain, or Heavy Rain.

Year (yr):

Effect: Rentals might increase in 2019 compared to 2018 due to improved awareness or infrastructure.

Month (mnth):

Effect: Warm months (e.g., May–September) likely see more rentals than colder ones.

Holiday (holiday) and Working Day (workingday):

Effect: Rentals increase on holidays for leisure and on working days for commuting.

Weekday (weekday):

Effect: Higher rentals on weekdays for commuting, with some increase on weekends for leisure.

Categorical variables like season, weathersit, yr, mnth, holiday, and weekday significantly influence bike rentals by reflecting weather, time of year, and people’s behavior.

2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True when creating dummy variables is important to avoid a problem called the dummy variable trap.

drop_first=True helps avoid confusion and makes your regression model work better.

Dummy Variable Trap

The dummy variable trap happens when one dummy variable can be predicted from the others, creating redundancy. This causes problems in regression models because it makes the calculations unstable.

For example, if a variable has 3 categories (A, B, C) and you create dummy variables for all three:

A = [1, 0, 0]

B = [0, 1, 0]

C = [0, 0, 1]

If you know A and B, you can figure out C because $C=1-A-BC = 1 - A - B$. This makes one dummy variable unnecessary.

By dropping one dummy variable (e.g., A), you remove the redundancy. The model can still tell which category each value belongs to without extra variables.

For the above example, with drop_first=True, the model keeps only B and C:

B = [0, 1]

C = [0, 0]

If both B and C are 0, the value belongs to A (the dropped category).

Prevents Redundancy: Keeps the model calculations simple and stable.

Clear Baseline: The dropped category acts as a reference, and the coefficients of the remaining variables show the comparison to this baseline.

- **3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

- The variable with the highest correlation with bike rentals (cnt) is registered users, with a correlation of 0.945411.
- This means the number of registered users has the strongest positive relationship with total bike rentals.

- **4. How did you validate the assumptions of Linear Regression after building the model on the training set?**
- After building the linear regression model on the training set, the assumptions were validated as follows:
- Normality of Residuals:
- The residuals of the model were extracted, and their distribution was visualized using a histogram. This visualization helps check whether the residuals follow a normal distribution, a key assumption of linear regression.
- Multicollinearity:
- To detect multicollinearity among the independent variables, the Variance Inflation Factor (VIF) was calculated for each feature. VIF values greater than 5 indicate significant multicollinearity.
- Correlation Between Features:
- A correlation heatmap was generated to visualize the relationships among the features, helping to identify any highly correlated predictors. This step assists in understanding potential multicollinearity issues and the relationships between predictors.

- **5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
- Top 3 Features contributing to demand:

	Feature	Coefficient
10	casual	1.000000e+00
11	registered	1.000000e+00
12	dteday_year	2.654662e-12

- These are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm:

Linear Regression is a supervised learning algorithm used to predict a continuous target variable based on one or more input features (independent variables). It establishes a linear relationship between the target variable *YY* and the input features *XX*.

The objective is to find the best-fitting straight line (or hyperplane in multiple dimensions) that minimizes the difference between the actual target values and the predicted values. This is achieved using the Ordinary Least Squares (OLS) method.

Assumptions of Linear Regression

To ensure the model's validity, linear regression assumes:

Linearity: The relationship between features and the target variable is linear.

Homoscedasticity refers to the condition where the variance of the residuals (errors) remains constant across all levels of the independent variable(s).

Independence: Residuals are independent of each other.

Normality of Residuals: Residuals follow a normal distribution.

No Multicollinearity: Features are not highly correlated with each other.

Steps of the Linear Regression Algorithm

Feature Selection and Preprocessing

Choose the input features *XX* based on domain knowledge.

Standardize or normalize features if necessary.

Handle missing data and encode categorical variables.

- Splitting Data
  - Divide the dataset into training and test sets.
- Fit the Model
  - Use the training set to estimate the coefficients $\beta0,\beta1,\ldots,\betan$ \beta_0, \beta_1, \dots, \beta_n that minimize the error.
  - The error is minimized using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals: $\text{Residual Sum of Squares (RSS)} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
  - Solution is derived analytically: $\beta = (X^T X)^{-1} X^T Y$
- Model Prediction
  - Use the estimated coefficients $\hat{\beta}$ to make predictions: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$
- Validate Assumptions
  - Check residuals to ensure assumptions (linearity, normality, etc.) hold.
  - Plot residuals against predictions and other diagnostic plots.
- Model Evaluation
  - Evaluate the model on the test set using metrics such as:
    - Mean Absolute Error (MAE): Average absolute error between predicted and actual values.
    - Mean Squared Error (MSE): Average squared error.
    - ~~R-squared ($R2R^2$): Proportion of variance in YY explained by XX.~~

Strengths of Linear Regression

Easy to implement and interpret.

Computationally efficient.

Works well for linearly separable data.

Limitations of Linear Regression

Assumes a linear relationship, which may not always hold.

Sensitive to outliers, which can distort predictions.

Performance deteriorates with multicollinearity among features.

Assumes homoscedasticity, which is rare in real-world data.

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a group of four datasets designed by statistician Francis Anscombe in 1973 to show the importance of visualizing data before analyzing it. These four datasets have nearly identical basic statistical properties, such as the mean, variance, correlation, and linear regression line. For example, all four datasets have the same mean ($x=9.0x = 9.0$, $y=7.5y = 7.5$), the same variance, the same correlation coefficient ($r=0.816r = 0.816$), and the same regression equation ($y=3.0+0.5xy = 3.0 + 0.5x$). However, when visualized as scatter plots, they look very different.

The first dataset shows a clear linear relationship.

The second dataset has a curved, non-linear relationship.

The third dataset appears linear but contains an outlier that heavily influences the regression.

The fourth dataset is dominated by a single extreme outlier, and without it, there would be no relationship.

This example demonstrates that relying solely on numerical summaries of data can be misleading, as important patterns, relationships, or anomalies may be hidden. Anscombe's Quartet emphasizes why it is critical to visualize data to better understand its structure and characteristics before performing statistical analyses.

3. What is Pearson's R?

Pearson's R, or the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Its value ranges from -1 to 1, where $R=1R = 1$ indicates a perfect positive linear relationship, $R=-1R = -1$ represents a perfect negative linear relationship, and $R=0R = 0$ suggests no linear relationship. The formula for Pearson's R is the covariance of the two variables divided by the product of their standard deviations. A positive RR means both variables move in the same direction, while a negative RR means they move in opposite directions. For example, a high RR (e.g., 0.8) might show a strong positive relationship, such as between temperature and ice cream sales. Pearson's R assumes the relationship is linear, the data is approximately normal, and outliers are minimal.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a preprocessing technique used in machine learning to adjust the range of features so that they contribute equally to the model. It ensures that features with larger numerical ranges do not dominate those with smaller ranges, especially in distance-based models like k-NN or gradient-based models like logistic regression. There are two common approaches to scaling: normalization and standardization. Normalization (min-max scaling) adjusts the values to a fixed range, usually [0, 1], by rescaling them based on the minimum and maximum values of the feature. It is suitable when the data does not have outliers and is not normally distributed. Standardization scales the data to have a mean of 0 and a standard deviation of 1, making it suitable for datasets with outliers or a Gaussian distribution. Scaling is crucial for improving model performance and ensuring that algorithms converge faster during training.

The primary difference between normalized scaling and standardized scaling is:

Normalized Scaling (Min-Max Scaling): Rescales features to a fixed range, typically between 0 and 1. It is sensitive to outliers and is best suited for data without extreme values.

Standardized Scaling (Z-Score Scaling): Adjusts features to have a mean of 0 and a standard deviation of 1. It is less sensitive to outliers and works well for data with varying ranges or a normal distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of VIF (Variance Inflation Factor) can become infinite when there is perfect multicollinearity among the predictor variables in a regression model. This means that one predictor variable is an exact linear combination of one or more other predictor variables. In such cases, the denominator of the VIF formula, which involves the variance of the independent variable unexplained by the other predictors, becomes zero. Since division by zero is undefined, the VIF value tends to infinity. This situation indicates that the predictors are perfectly correlated, making it impossible for the model to estimate unique regression coefficients for them. To address this issue, one of the highly collinear variables needs to be removed or combined to resolve the multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot (Quantile-Quantile plot) is a visual tool used to compare the distribution of a dataset to a theoretical distribution, commonly the normal distribution. It plots the quantiles of the dataset against the corresponding quantiles of the reference distribution. If the points on the Q-Q plot align closely with a straight diagonal line, it indicates that the data conforms to the theoretical distribution. Any deviations from this line suggest that the data differs from the assumed distribution.

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

Assessing Normality: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps check this assumption by comparing the residuals against a normal distribution.

Identifying Deviations: If the points deviate significantly from the diagonal line in the Q-Q plot, it may indicate non-normal residuals, the presence of outliers, or skewness, which could affect the validity of the regression model.

Model Diagnostics: By verifying the normality of residuals, a Q-Q plot helps ensure that the regression model's p-values and confidence intervals are reliable. Non-normal residuals can lead to inaccurate inferences.

Improving Model Fit: If the Q-Q plot shows non-normality, it signals the need for transformations of the data or the response variable, or for using a different modeling approach.

A Q-Q plot is a vital diagnostic tool in linear regression to validate the assumption of normality for residuals and ensure the reliability of the model's results.