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)

Insights from the Model Design on Categorical Variables

Season: The seasons Spring, Summer, and Winter exhibit negative coefficients, indicating lower bike-sharing demand during these periods.

Month: January shows a negative coefficient, suggesting a decline in bike-sharing demand, while September has a positive coefficient, reflecting the highest demand.

Weather Conditions: Weather scenarios like light snow with rain and mist have negative coefficients, signifying reduced bike-sharing demand under these conditions.

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 Multicollinearity**
   Including all dummy variables creates perfect correlation between the dummy columns and the intercept term in a regression model, known as the "dummy variable trap." This multicollinearity can make it difficult for the model to estimate coefficients accurately. Dropping one column eliminates this problem.
2. **Simplifies the Model**
   By reducing the number of independent variables, the model becomes more efficient and easier to interpret, without losing any critical information about the categories.
3. **Retains Complete Information**
   Dropping one column doesn't result in information loss because the omitted category can be inferred from the remaining columns.

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)

1. A pair plot was generated for the variables Temperature, Humidity, Windspeed, and Count.
2. It is clearly evident that the Temperature variable shows a strong positive correlation with Count (Bike Sharing Demand).

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)

Assumptions of Linear Regression
- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Validation Approach:
Residual Distribution Analysis:
A distribution graph of the residual term (difference between the actual and predicted values of the target variable) was created. This is a strong indicator of whether the model's assumptions are valid. If the center of the residual distribution is around zero, it suggests that the model is well-balanced.

Scatter Plot Analysis:
A scatter plot was created to examine the relationship between the independent (X) and dependent (Y) variables. This helps identify if a clear linear relationship exists, which is essential for validating the model's performance.

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)

Final Model Equation

count= 0.5886+(0.2481*yr)-(0.1889*windspeed)-(0.2599*Spring) - (0.0396*Summer)-(0.0764*Winter)-(0.1034*Jan)+(0.0697*Sep) - (0.0462*Tuesday)-(0.2986*Light_Snow)-(0.0859*Mist)

Top 3 Factors Influencing Bike Sharing Demand:

Season:

The seasons Spring, Summer, and Winter are associated with negative coefficients, indicating lower bike-sharing demand during these periods.

Month:

January shows a negative coefficient, reflecting a drop in demand, whereas September has a positive coefficient, marking it as the month with the highest bike-sharing demand.

Weather Conditions:

Weather scenarios like light snow with rain and mist have negative coefficients, suggesting that bike-sharing demand decreases under these conditions.

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)

Linear Regression Model: Linear regression is a predictive modeling technique that describes the relationship between a dependent (target) variable and one or more independent (predictor) variables.

Simple Linear Regression:
This is the most basic form of linear regression, where the goal is to identify a linear relationship between a single dependent variable and a single independent variable.

Multiple Linear Regression:
A more advanced form of linear regression, where the objective is to find the linear relationship between one dependent variable and multiple independent variables.

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn$

Best Fit Line:
In linear regression, the algorithm aims to determine the coefficients for the independent variables that result in the best-fit line with the minimum residual (error term).

Gradient Descent Process:
To optimize the coefficients for the independent variables, the Gradient Descent method is used. This technique iteratively adjusts the coefficients to minimize the residuals and achieve the best possible fit.

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)

Anscombe's Quartet

Anscombe's Quartet is a classic example used to highlight the importance of data visualization in statistical analysis. Developed by the statistician Francis Anscombe in 1973, it demonstrates that relying solely on statistical properties can be misleading, and emphasizes the need to plot data before making any conclusions.

Example:

Below data set gives an impression that if we do a statistical analysis between x1,y1 or x2,y2 or x3,y3 or x4,y4 then we will get a similar kind of infer from it. But if we follow the Anscombe's quartet guideline and try to visualize these relationships then it will realize that it is not true.

Statistical Analysis of the Four Datasets:

- **Average Value of xxx** = 9
- **Average Value of yyy** = 7.50
- **Variance of xxx** = 11
- **Variance of yyy** = 4.12
- **Correlation Coefficient** = 0.816

- **Linear Regression Equation**: y=0.5x+3y = 0.5x + 3y=0.5x+3

  Credit : Source

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)

Pearson's R

Pearson's R is a widely used correlation coefficient that measures the strength and direction of the relationship between two variables. It is also referred to as Pearson's correlation.

The values of Pearson's R range from -1 to 1:

- A value of 1 indicates a strong positive relationship.
- A value of -1 indicates a strong negative relationship.
- A value of 0 indicates no correlation or relationship between the variables.

The formula to calculate the **Pearson correlation coefficient (r)** is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Where:

- r is the Pearson correlation coefficient.
- n is the number of data points (pairs of values).
- x and y are the individual data points for the two variables being compared.
- $\sum x$ is the sum of all the x-values.
- $\sum y$ is the sum of all the y-values.
- $\sum xy$ is the sum of the product of corresponding x and y values.
- $\sum x^2$ is the sum of the squares of the x-values.
- $\sum y^2$ is the sum of the squares of the y-values.

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)

What is Scaling : Feature scaling is a technique used to transform the values of variables onto a common scale. It is an essential part of the data preprocessing phase in machine learning model building.

Why is scaling : Scaling helps bring the values of all variables within a specified range (e.g., min/max range), enabling the algorithm (such as Gradient Descent) to create a model where each variable contributes equally. Feature scaling becomes particularly important when different variables or independent variables have values with significant differences in their minimum and maximum values. This ensures that the model treats all variables with equal importance, improving its performance and accuracy.

Normalization vs Standardization

Normalization	Standardization
Rescales values to a range between 0 and 1.	Centers data around the mean and scales it to a standard deviation of 1.
Useful when the distribution of the data is unknown or not Gaussian.	Useful when the distribution of the data is Gaussian or unknown.
Sensitive to outliers.	Less sensitive to outliers.
Retains the shape of the original distribution.	Changes the shape of the original distribution.
May not preserve the relationships between the data points.	Preserves the relationships between the data points.
Formula: (x – min)/(max – min)	**Formula**: (x – mean)/standard deviation

---

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)

VIF: Feature scaling (or scaling of variable values) is used to find a correlation (Multicollinearity) between two independent variables.

Value of VIF starts from 1 and has no upper limit.

If **VIF is infinite** that means there is a strong multicollinearity between those two independent

variables, and it is not good for your Model Building

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)

Q-Q Plot

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether two samples of data originate from the same population.

Inferences from a Q-Q Plot:

- Determine if two samples come from the same population.
- Identify whether the two samples have similar tails.
- Check if the two samples share the same distribution shape.
- Evaluate whether the two samples exhibit common location behavior.

Importance of the Q-Q Plot:

- Since a Q-Q plot functions like a probability plot, it allows for comparison between two datasets without requiring equal sample sizes.
- Normalization of the dataset is not necessary, and the dimensions of the values do not affect the plot.
