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 categorical variables, the following inferences can be made about their effect on the dependent variable (cnt - total bike demand):

Season (season) – The demand for bikes varies significantly across seasons. Winter (season_4) shows the highest demand, while spring (season_2) has comparatively lower demand.

Year (yr) – Bike demand has increased from 2018 (yr_0) to 2019 (yr_1), indicating a growing trend in bike usage over time.

Weather Situation (weathersit) – Poor weather conditions (weathersit_3) negatively impact bike demand, as seen from the negative coefficient in the regression model.

These categorical variables significantly influence bike rentals and should be considered in business decision-making.

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)

When dealing with a binary categorical variable like "holiday" (where values can be "yes" or "no"), using 0 and 1 to represent these categories can sometimes lead to multicollinearity issues in models. In such cases, using drop_first=True while creating dummy variables ensures that only one variable is used to represent the presence or absence of the category, thereby avoiding redundancy. That is "holiday_0" is deleted.

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)

Before removing the unwanted columns, the column with the most correlation with the target variable (cnt) was the 'registered' column, as it directly reflects the total number of rentals by registered users. However, after removing the irrelevant and redundant features, the column with the highest correlation to cnt became the 'temp' (temperature) column. This suggests that temperature has a significant impact on bike rental demand, highlighting its importance in predicting future bike-sharing usage.

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)
Answer:
After building the Linear Regression model on the training set, I validated its assumptions using the following methods:

Linearity:
Plotted the predicted values vs. actual values to check if they follow a linear pattern.
Normality of Residuals:
Used a histogram and Q-Q plot to verify if the residuals are normally distributed.
No Multicollinearity:
Calculated the Variance Inflation Factor (VIF) for independent variables.
Ensured that all VIF values were below 5 to avoid high multicollinearity.
Independence of Errors (No Autocorrelation):
Ensured no clear patterns in the residual plot.

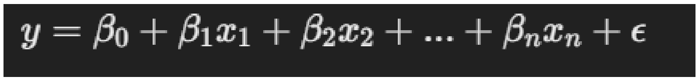
Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
Based on the final model, the top 3 features that significantly contribute to explaining the demand for shared bikes are:

Year (yr_1) – The coefficient of 0.2302 with the lowest p-value indicates that bike demand increased significantly in the second year (2019).
Temperature (temp) – The coefficient of 0.5350 shows that temperature has a strong positive impact on bike demand, meaning higher temperatures lead to more rentals.
Season (season_4) – The coefficient of 0.1347 suggests that demand is most affected in winter (season 4), making it one of the top influencing factors.

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 supervised learning algorithm used for predicting a continuous dependent variable (Y) based on one or more independent variables (X). It establishes a linear relationship between the variables using the equation:



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

The goal is to find the values of the coefficients (β_0, β_1, ..., β□) that minimize the difference between the predicted output and the actual target values.

This difference is typically measured using the Mean Squared Error (MSE) or Residual Sum of Squares (RSS).

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 (mean, variance, correlation, and regression line) but look completely different when plotted. It was created by Francis Anscombe in 1973 to highlight the importance of data visualization in statistical analysis.

Key Insights from Anscombe's Quartet:
Although each dataset has:

The same mean (for x and y)
The same variance
The same correlation coefficient
The same linear regression equation
Their scatter plots reveal drastically different relationships.

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 statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by r and ranges from -1 to +1.

r = +1 → Perfect positive correlation (as one variable increases, the other also increases).
r = 0 → No correlation (no linear relationship between the variables).
r = -1 → Perfect negative correlation (as one variable increases, the other decreases).

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>

What is Scaling?

Scaling is the process of transforming numerical features into a specific range or distribution to ensure that all variables contribute equally to a model. It is essential when features have different units or magnitudes.

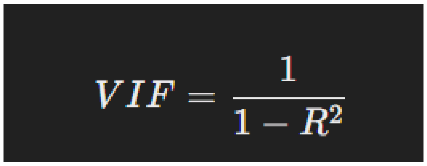
Example:
Original data: [20, 40, 60, 80, 100]
Min = 20, Max = 100
Scaled: [0, 0.25, 0.5, 0.75, 1]

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>

Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between independent variables. This happens when one feature is an exact linear combination of another, causing the denominator in the VIF formula to become zero.

This happens due to :
Duplicate Columns: If two or more columns contain identical values, one can be perfectly predicted from the other.
Dummy Variable Trap: When creating dummy variables for categorical features, keeping all categories can lead to redundancy. This is why drop_first=True is used in pd.get_dummies().
Linear Dependence: If a feature is a perfect linear combination of others (e.g., X3 = X1 + X2), the VIF becomes infinite.



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

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 compare the distribution of a dataset to a theoretical distribution, usually the normal distribution. It helps assess whether the residuals of a model follow a normal distribution, which is a key assumption in linear regression.

Use of Q-Q Plot in Linear Regression:
In linear regression, the assumption of normally distributed residuals ensures that:

The model’s predictions are unbiased and reliable.
Confidence intervals and hypothesis tests (such as p-values) are valid.
The model's errors are homoscedastic (constant variance).