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)

Ans: In the bike-sharing dataset, categorical variables like weathersit and season significantly influenced the target variable cnt. For example, during weathersit_3 (Light Snow), bike hire numbers decreased, while season trends like season_Spring showed different patterns. Including categorical features such as yr and season in the model increased both R-squared and adjusted R-squared values, indicating these variables explained more variance in bike bookings. This highlights the importance of including categorical variables to capture patterns that numeric features alone may miss, improving the model's accuracy.

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)

Ans: When creating dummy variables, each category of a categorical variable is converted into a binary (0 or 1) variable. If all categories are included, one of the can be perfectly predicted from the others, leading to **multicollinearity** (a situation where independent variables are highly correlated). This distorts the model's ability to estimate the true relationship between variables. By using drop_first=True, we drop one category, serving as the baseline, and avoid this issue, ensuring the model is not biased and produces reliable coefficient estimates.

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)

Ans: Both temp and atemp have the highest correlation with cnt, as they have the highest correlation values (around 0.627 and 0.628).

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)

Ans: After building the model on the training set, I conducted the following analysis:

1. Assumed a linear relationship between the independent variables (X) and the dependent variable (Y).
2. Checked that the error terms are normally distributed with a mean of zero (not X or Y).
3. Performed residual analysis on the training data, confirming that the residuals are normally distributed.
4. Validated the assumptions for linear regression based on these checks.
5. Used Variance Inflation Factor (VIF) and p-values to selectively include or exclude independent variables, ensuring that multicollinearity is avoided

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)

Ans: As per the final model, the top three predictor variables influencing bike bookings are:

1. **Temperature (temp)**: The coefficient of 0.4411 indicates that a unit increase in temperature leads to an increase of 0.4411 bike hires, suggesting that warmer weather encourages more bookings.

2. **Weather Situation 3 (weathersit_3)** (e.g., Light Snow, Light Rain + Thunderstorm, Scattered Clouds, Light Rain + Scattered): The coefficient of -0.3282 implies that a unit increase in weathersit_3 results in a decrease in bike hires by 0.3282 units, indicating that certain weather conditions negatively impact bookings.

3. **Year (yr)**: The coefficient of 0.2310 suggests that as the year progresses (e.g., from one year to the next), bike bookings increase by 0.2310 units. Thus, these variables should be given utmost importance in planning to maximize bike bookings.

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>

Ans: Linear regression is a method used to identify the best linear relationship between independent variables and a dependent variable. In your analysis,

Temperature was found to be the most significant feature affecting the business positively, while other environmental factors such as **Rain**, **Humidity**, **Windspeed**, and **Cloudiness** were seen to negatively affect the business.**Residual Sum of Squares (RSS)**:
RSS is the sum of squared differences between the actual values (yi) and the predicted values (ŷi). It is used to measure the error of the model and is given by:

$$RSS=i=1\sum n\ (yi - y\hat{}\ i\ )2$$

where: $y\hat{}\ i$ is the predicted value for each data point.

1. **Cost Function (J)**:
   The cost function for linear regression is:
   $J(ß1\ ,ß0\ )=\sum(yi - ß1\ xi - ß0\ )2$
   This function is minimized to find the optimal values of the coefficients

2. **OLS (Ordinary Least Squares)**:
   OLS is used to minimize the residual sum of squares and estimate the coefficients. The goal of OLS is to find the line that minimizes the total error (squared differences between actual and predicted values).

3. **Error Terms (Residuals)**:
   The error term for each data point is: $ei = yi - y\hat{}\ i$
   These residuals represent the difference between the actual and predicted values. The OLS method minimizes the sum of these squared errors to fit the best line.

Methods for Solving:

- **Closed Form**: The solution to the minimization problem can be obtained algebraically using matrix operations.
- **Gradient Descent**: An iterative approach to find the minimum by updating the coefficients based on the gradient of the cost function.

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>

Ans: Anscombe's quartet consists of four datasets that appear quite similar when you look at their summary statistics, but the plots reveal very different patterns. Each dataset contains 11 pairs of (x, y) values, and the summary statistics are as follows:

1. The average of x is 9.
2. The average of y is 7.5.
3. The variance of x is 11 and y is 4.12.
4. The correlation between x and y is 0.816.

5. The line of best fit is y
   y = 0.5x + 3
   y=0.5x+3.

However, when you graph each dataset, you see unique behaviors:

1. **First plot**: This shows a simple linear relationship where y is normally distributed, and the Pearson correlation is meaningful. It can be modeled with linear regression.

2. **Second plot**: Here, although a relationship between the variables is apparent, it's not linear. The Pearson correlation isn't helpful, and a more generalized regression model would be better.

3. **Third plot**: This shows a linear distribution, but there's an outlier affecting the regression line. A robust regression method would be needed to avoid the influence of the outlier.

4. **Fourth plot**: This dataset demonstrates a high correlation due to a single high-leverage point, even though the rest of the data doesn't show a clear relationship.

Anscombe's quartet emphasizes the importance of not just relying on summary statistics but also carefully analyzing the data visually to uncover any underlying patterns.

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>
Ans: Pearson's R is a measure of the strength and direction of the linear relationship between two variables, ranging from -1 to +1.

- **r = +1**: Perfect positive linear relationship (both variables increase together).
- **r = -1**: Perfect negative linear relationship (one variable increases while the other decreases).
- **r = 0**: No linear relationship.
- **r between 0 and 0.5**: Weak positive correlation.
- **r between 0.5 and 0.8**: Moderate positive correlation.
- **r between 0.8 and 1**: Strong positive correlation.

Pearson's R evaluates how closely the data points fit a line of best fit, showing the strength of the association. Positive values indicate variables move in the same direction, while negative values show they move in opposite directions.

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>

Ans: Scaling is a pre-processing step that ensures comparability across features with different magnitudes or units. It improves model accuracy and performance by standardizing the data.

- **Normalization (Min-Max Scaling)**: Scales data between 0 and 1. Useful when features have different ranges.
- **Standardization (Z-Score Scaling)**: Centers data with a mean of 0 and standard deviation of 1. Useful for normally distributed data.

Scaling helps avoid bias in models, especially distance-based algorithms, and ensures all features contribute equally. It affects model coefficients but not statistical measures like p-values or R-squared.

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>

Ans: The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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>

Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used to check if a dataset follows a specific distribution (like Normal) or if two datasets come from the same distribution. In linear regression, it helps confirm if the training and test datasets have similar distributions.

Advantages:

- Works with small sample sizes.
- Detects shifts in location, scale, symmetry, and outliers.

The plot compares the quantiles of two datasets. If they come from the same distribution, the points will form a straight line.