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)

Yr - 2019 attracted more number of booking from the previous year.
Season - fall season has the most count of people using boom bikes. Spring season has the least
Mnth - Most of the bookings have done in the month of June but there is not much difference to the other months
Weekday - Sunday has the least users and the users increases from monday to friday. This may be because people might be using it for work to their offices.
Holiday - Bikes are more used when there is no holiday.

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)
During dummy variable creation is important to avoid the dummy variable trap, which occurs when one or more dummy variables are highly correlated (multicollinear) with each other.

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)
Temp

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)
Multicollinearity check - There should be insignificant multicollinearity among variables.
Normality of error terms - Error terms should be normally distributed.
Linear relationship - Linearity should be visible among variables.
Homoscedasticity - There should be no visible pattern in residual values.
Independence of residuals - No auto-correlation

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)

temp
yr
Winter

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 is a supervised learning algorithm used to model the relationship between one or more independent variables (features) and a dependent variable (target). In simple linear regression, the model takes the form:

$Y=b0+b1X+\epsilon Y$

- b0 (intercept): The expected value of Y when X is 0.
- b1 (slope): The change in Y for every unit increase in X.
- e (error term): Accounts for randomness or unmodeled factors that affect Y.

In multiple linear regression, multiple independent variables are used, represented as:

$Y=b0+b1X1+b2X2+\cdots+bnXn+eY$

The algorithm tries to minimize the sum of squared residuals (SSE), which is the difference between the observed and predicted values. This optimization is achieved using Ordinary Least Squares (OLS), which provides the best-fit line by minimizing the errors.

Linear regression assumes:

1. Linearity: The relationship between predictors and target is linear.
2. Independence: Observations are independent.
3. Homoscedasticity: Constant variance of errors.
4. Normality: Errors follow a normal distribution.

It performs well with linear relationships but is sensitive to outliers and multicollinearity, which can affect model performance.

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 is a group of four datasets, each with nearly identical statistical properties (such as mean, variance, correlation, and regression line) but very different distributions when visualized. It was introduced by Francis Anscombe to highlight the importance of data visualization in statistical analysis.

- Mean of X and Y are almost the same across all four datasets.
- The variance of X and Y is similar.
- The correlation coefficient between X and Y is around 0.816 in all cases.
- The linear regression line is nearly identical: Y=3+0.5XY = 3 + 0.5XY=3+0.5X.

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 measures the strength and direction of the linear relationship between two variables. Its value ranges from -1 ie perfect negative correlation to 1 ie perfect positive correlation, with 0 indicating no linear relationship.

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)

Scaling is the process of transforming features to a specific range or distribution to ensure they contribute equally to a model. It's performed to improve the performance of machine learning algorithms, especially those sensitive to feature magnitudes (e.g., regression, k-NN).

- Normalized scaling scales values to a fixed range, typically [0, 1], using Min-Max Scaling.
- Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1, ensuring features follow a standard normal distribution.

Both methods address different needs: normalization for bounded scales and standardization for normal-like distributions.

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)

A VIF becomes infinite when there is perfect multicollinearity, meaning one predictor variable is an

exact linear combination of others. This makes the regression model unstable and the variance of coefficients unmeasurable.

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)

A Q-Q (Quantile-Quantile) plot compares the quantiles of a dataset with a theoretical distribution (often normal). In linear regression, it helps assess if the residuals are normally distributed, a key assumption. If the points lie close to the 45-degree line, the residuals are normally distributed; deviations indicate potential issues with normality.
