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)

1) Impact of Season categorical variable – Bike sharing increases during summer & fall season and lowest in spring – possibly due to avoid rains.
2) Impact of Month categorical variable – Specific months like July, August, September, October has observed higher number of bookings possibly as we see summer and fall observe high bookings
3) Impact of weather situation categorical variable – Clear weather has been the preferred choice of people to choose bikes as transport option.
4) Year 2019 has observed higher number of bookings then previous year i.e 2018.
5) All days of week has seen similar trend and booking count.

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, all category values are converted to columns. Dropping first one, just avoid redundancy as it can calculated as 1 if all other variables are 0.

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/Atemp has highest correlation (0.63)

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)

After building the model, Residual analysis is done to see how the error terms (difference between predicted values and actual values from the training dataset) are distributed. If it follows

1) A normal distribution with mean as 0.
2) No pattern in the error terms distribution
3) The residuals (errors) are independent of each other.
4) The variance of the residuals is constant across all levels of the independent variables.
5) No Multicollinearity among the independent variables.

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)

1) Temp (positive impact)
2) LightSnow (Weather Situation) – negative impact
3) Year 2019 – positive impact
4) Windspeed – negative impact

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 predictive modeling. It models the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation to the observed data.

The general form of a linear regression model is Y = B0 + B1*X1 + B2*X2……. + Bn*Xn + E

To ensure reliable results, linear regression relies on the following assumptions:

Linearity: The relationship between dependent and independent variables is linear.
Independence: Observations are independent of each other.
Homoscedasticity: Constant variance of residuals across all levels of predictors.
Normality of Residuals: Residuals should be normally distributed.
No Multicollinearity: Independent variables should not be highly correlated.

The goal of linear regression is to estimate the coefficients (B0, B1, B2 … Bn) such that the model minimizes the error between the predicted values (Y^) and the actual values (Y).

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, such as mean, variance, correlation coefficient, and linear regression line, yet appear very different when visualized.

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 strength and direction of the linear relationship between two continuous variables. It is represented by the symbol r and ranges from −1 to +1.

r = +1 (Perfect positive correlation)
r = -1 (perfect negative correlation)
r = 0 (No linear correlation)
0 < r < 1 (Positive linear correlation)
-1 < r < 0 (Negative linear correlation)

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>

Scaling involves modifying the range and distribution of feature values in a dataset to ensure all features equally to a model's learning process by aligning them to a common scale.

Scaling is crucial for the following reasons:

1) Enhancing Model Performance: Features with larger ranges can dominate the learning process, potentially biasing the results. Scaling prevents this imbalance.
2) Accelerating Training: Algorithms like Gradient Descent benefit from scaling, as it reduces distortions caused by varying magnitudes of features, leading to faster convergence.
3) Avoiding Numerical Instability: Handling large or very small feature values can cause numerical issues such as overflow or underflow in models, which scaling helps to mitigate.

Difference between Normalized and Standardized Scaling methods:

1) Normalized scaling uses minimum and maximum values of features to scale whereas as Standardized scaling uses mean and standard deviation of feature values.
2) Normalized scaling adjust the data in range of -1 to 1 whereas in standardized scaling, there is no range bound.
3) Normalized scaling is highly sensitive to outliers whereas standardized scaling is not.

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>

We can get inf values for VIF due to the perfect multicollinearity. This happens when two or more independent variables in a model are perfectly linearly dependent. That is, one independent variable in the model can be entirely predicted by another independent variable. It can also be understood by the formula of VIF which is

VIF for Xi = 1/1-R-Squarei

R2 is 1 when they is perfect collinearity and hence the denominator value goes 0 and we get infinite VIF.

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 plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution (commonly the normal distribution). It plots the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will lie approximately on a straight 45-degree line.
