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. Year 2019 has high bike rentals than 2018
2. Demand is high in middle of the years(May to October)
3. Fall season has the highest rentals followed by summer
4. Rentals on non-holidays are higher
5. Rentals are slightly more on Thursday's and Friday's
6. Rentals are high on clear weather

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)

To avoid multicollinearity and to prevent dummy variable trap

---

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)

atemp

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)

1. By using Residual analysis
2. Using histplot for error terms to observe normal distribution and zero mean
3. Using Scatter plot to observe error terms are independent or not
4. Using Scatter plot to observe Homoscedaticity

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
2. weathersit_light_precipitation
3. yr

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)

1. Linear regression is a supervised learning algorithm
2. Two types of regressions
    1. Simple linear regression ($y=\beta0 +\beta1\ x+\epsilon$) where $\beta0$ represents intercept and $\beta1$ is coefficient(slope)
    2. Multiple linear regression($y=\beta0 +\beta1\ x1 +\beta2\ x2 +...+\beta n\ xn +\epsilon$) where $\beta0$ represents intercept and $\beta1$, $\beta2$ etc., represents coefficients of features
3. Target variable is a continuous value
4. Linear regression is used to find the best fit line so that the sum of error between the target and predicted is minimum

---

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 used to illustrate the importance of EDA and what will happen if we depend only on summary statistics.It also explains the importance of using data visualisation to observe trends, outliers etc., that might not be get from summary statistics

---

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 also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is denoted as r and ranges from -1 to 1

1. r=1 perfect positive correlation(as one variable increases, the other increases)
2. r=-1 perfect negative correlation(as one variable increases, the other decreases)
3. r=0 no correlation(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 a preprocessing step in machine learning where numerical features are transformed to ensure they are on a comparable scale.

1. Improves model performance
2. Prevents bias towards large scale features
3. Accelerates convergence in gradient descent
4. Ensures better interpretation

Normalization rescales the data to fixed range between 0 and 1 where as the standardization centres the data to mean 0 and scales it to unit variance

---

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)

When there is a perfect multicollinearity then VIF become infinite as R-squared is 1 for perfect multicollinearity

---

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)

The Q-Q plot is used to visually compare the quantiles of your data to the quantiles of a theoretical distribution, which can reveal deviations from the expected distribution. A Q-Q plot is a valuable tool for assessing the distribution of data, especially in the context of linear regression. It helps evaluate normality assumption, detect skewness and outliers and guide model improvements if deviations are observed

---