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)

Refer to uploaded Categorical_vars_plots.png file

1. **BoomBike demand is lowest in Spring season (median value ~2000).**
2. **BoomBike demand is highest in Fall season (median value ~5500).**
3. **BoomBike demand in 2019-year is higher than that in 2018-year (median value difference of ~2000)**
4. **Between months APR to OCT the BoomBike demand is high.**
5. **BoomBike demand is lowest in JAN month (median value ~2000).**
6. **Throughout the weekdays the demand for BoomBike has approximately similar median values.**
7. **BoomBike demand is very similar in working days and holidays.**
8. **BoomBike demand is lowest (with median value of ~2000) when weather conditions are Light Snow / Light Rain Thunderstorm Scattered Clouds / Light Rain Scattered Clouds.**

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)

It is used to reduce the collinearity between dummy variables and to get to k-1 dummy variables.

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 has the highest correlation with the target variable cnt

---

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)

Linearity relation between y and x, Independence of observations, Homoscedasticity, Residuals are normally distributed, No/Less Multicollinearity (low VIF values).

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. **LightSnow_LightRainThunderstormScatteredClouds_LightRainScatteredClouds**
2. **SEP**
3. **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)

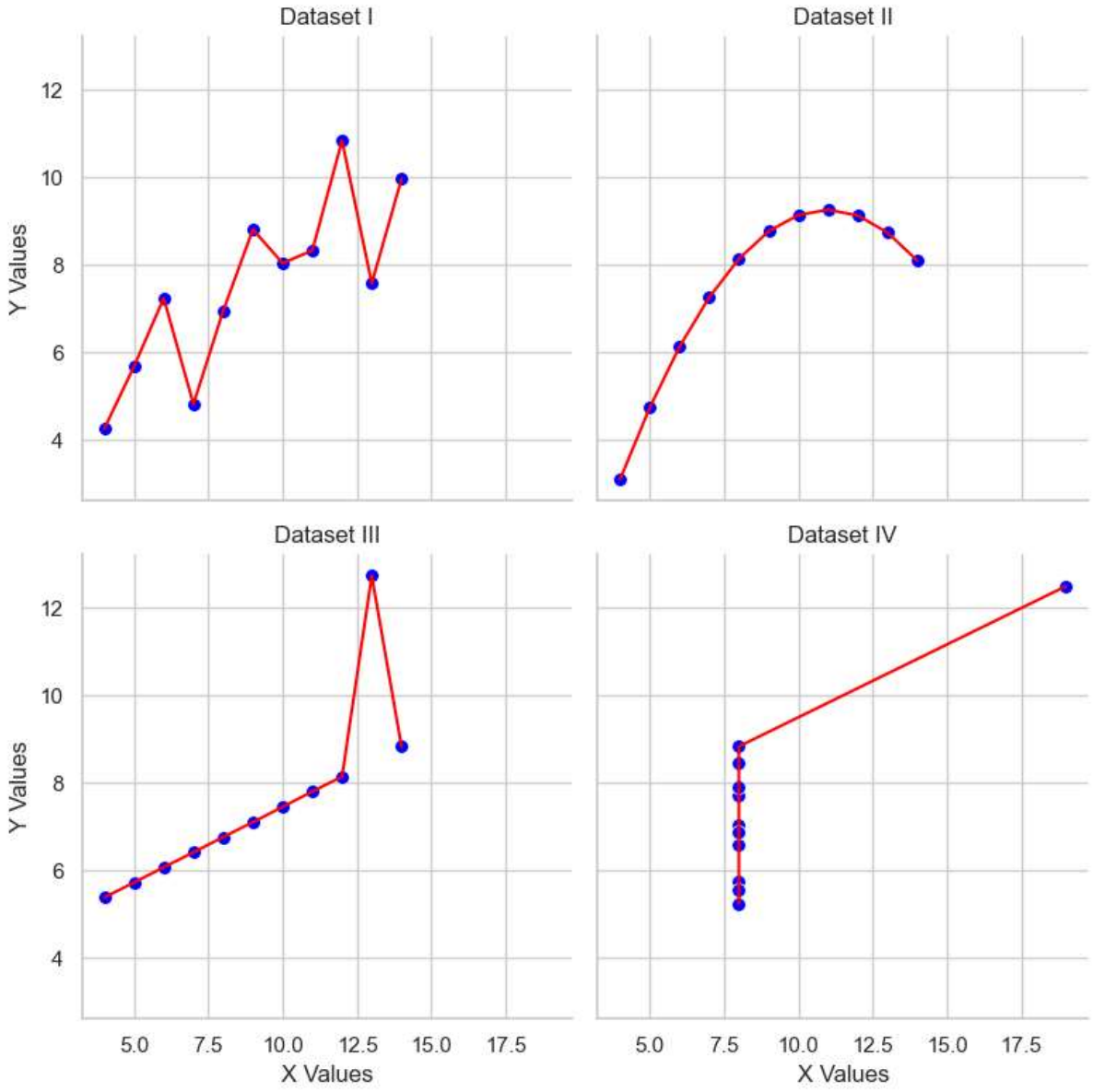
Linear regression aims to find the best-fitting straight line (in the case of simple linear regression) or hyperplane (in the case of multiple linear regression) that minimizes the difference between the predicted values and the actual values.

The key assumptions of Linear Regression are: Linearity relation between y and x, Independence of observations, Homoscedasticity, Residuals are normally distributed, No Multicollinearity.

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 set of four datasets created by the statistician Francis. These datasets are designed to demonstrate the importance of visualizing data rather than relying solely on summary statistics. Despite having nearly identical statistical properties, the datasets display vastly different distributions and relationships when visualized.

Dataset 1: Linear Relationship
Dataset 2: Nonlinear Pattern
Dataset 3: Outlier Effect
Dataset 4: Vertical Line



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 (PCC), is a statistical measure that quantifies the linear relationship between two continuous variables.
$-1 \leq R \leq 1$
R =1 : Perfect positive linear relationship (as X increases, Y increases proportionally).
R = −1 : Perfect negative linear relationship (as X increases, Y decreases proportionally).
R = 0: No linear relationship between X and Y.

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 data preprocessing technique used to adjust the range or distribution of values in a

dataset to bring all features onto a comparable scale.
Scaling is performed to: improve model performance and ensure convergence
Normalization (Min-Max Scaling) rescales data to fit within a specified range (typically [0, 1]).
Standardization (Z-Score Scaling) centers data around 0 with a standard deviation of 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)

The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictors in a regression model. Ri^2 is the coefficient of determination for the regression becomes 1 when the predictor can be perfectly predicted by the others. Hence VIF=1/(1-Ri^2) becomes infinite.

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 plot plots the quantiles of the residuals against the quantiles of a specified theoretical distribution (e.g., normal distribution).**
- ✓ **The x-axis represents the theoretical quantiles.**
- ✓ **The y-axis represents the sample quantiles (the quantiles of the residuals).**

<u>Use of Q-Q Plot in Linear Regression</u>: detection of skewness and kurtosis, normality of residuals, identifying outliers

---