Assignment-based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A. The categorical variables in the dataset, like season, year, month, and weather conditions, show significant effects on bike demand. For instance, bike rentals tend to be higher in favorable weather conditions and during certain seasons ex- summer. The year variable indicates growth trends in bike usage, and working days typically see different usage patterns compared to weekends.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

A. Using `drop_first=True` during dummy variable creation avoids the dummy variable trap, which occurs due to multicollinearity. By dropping the first category, we prevent redundant information that would otherwise inflate the model's coefficients and make it hard to interpret.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A. The `registered` variable has the highest correlation with the target variable `cnt`. This is because the total count (`cnt`) includes both casual and registered users, and registered users usually make up a significant portion of the total rentals.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A. To validate the assumptions of linear regression, I did:

- **Linearity**: By plotting residuals against predicted values.
- **Independence of errors**: By examining residual plots for patterns.
- **Homoscedasticity**: By ensuring the spread of residuals is consistent across levels of the predicted values.
- **Normality of residuals**: Using a Q-Q plot to check if residuals are normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A. The top three features contributing to bike demand are:

- **Temperature** (`temp`): Warmer weather increases bike usage.

- **Year** (`yr`): Indicates growth in bike-sharing over time.
- **Working Day** (`workingday`): Affects usage patterns as commuting needs change between weekdays and weekends.

General Subjective

1. Explain the linear regression algorithm in detail. (4 marks)

A. Linear regression predicts the value of a dependent variable based on one or more independent variables. It finds the best-fit line by minimizing the differences (residuals) between observed and predicted values, using a straight-line formula
y=β0+β1x+ϵy = \beta0 + \beta1 + \epsilony=β0+β1x+ϵ.

2. Explain the Anscombe's quartet in detail. (3 marks)
A. Anscombe's quartet consists of four datasets with nearly identical statistical properties, but with very different distributions when graphed. It shows the importance of visualizing data, as relying solely on statistics can be misleading.

3. What is Pearson's R? (3 marks)
A. Pearson's R is a measure of the linear correlation between two variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
A. Scaling adjusts the range of features in the data to ensure they contribute equally to the model. Normalization scales data between 0 and 1, while standardization scales data to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
A. A VIF (Variance Inflation Factor) can be infinite when two or more predictor variables are perfectly correlated. This means the model cannot distinguish between them, causing problems with multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A. A Q-Q plot (quantile-quantile plot) compares the distribution of residuals to a normal distribution. It helps check the normality assumption in linear regression, where points should align along a straight line if residuals are normally distributed.