Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: From what I observed, the seasons impact bike demand—summer and fall seem to have higher demand, which makes sense since people are probably more willing to rent bikes when the weather is nice. The weather condition (weathersit) also plays a role; clear weather boosts bike rentals, while bad weather (like heavy rain or snow) understandably deters people. The year (yr) is interesting too—demand is higher in 2019 than in 2018, which might be because bike-sharing was becoming more popular or accessible.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: When you're converting categorical variables to dummy variables, you don't want to keep all the dummy columns because it can cause issues with multicollinearity—basically, it can confuse the model because one category can be perfectly predicted from the others. By dropping the first category (using drop_first=True), we avoid this problem and keep the model simpler.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Temperature (temp) has the strongest relationship with the number of bike rentals.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model, I checked a few things to make sure everything was in order:

First, I looked at the plot of actual vs. predicted values to confirm that the relationship is linear.

I also checked that the residuals (errors) don't show any weird patterns, which would suggest that the errors are independent and evenly spread out.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:

Temperature (temp): Warmer weather leads to more bike rentals.

Year (yr): There was more demand in 2019, probably due to increased popularity.

Season: Especially summer and winter, which are peak times for biking.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear regression is one of the simplest ways to model the relationship between a dependent variable (like the number of bike rentals) and one or more independent variables (like temperature, season, etc.). The idea is to fit a straight line through the data points that best represents the relationship. The algorithm does this by minimizing the difference between the observed data points and the points on the line—this difference is called the "residual." In a nutshell, it's about finding the best-fit line that predicts the dependent variable based on the independent variables.

Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is an example of why just looking at summary statistics like the mean and correlation isn't enough. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset illustrates the importance of visualizing data before analysing it and the effect of outliers and other anomalies on statistical properties.

Q3. What is Pearson's R?

Ans: Pearson's R measures the strength and direction of a linear relationship between two continuous variables. The coefficient ranges from -1 to 1, where 1 means a perfect positive linear correlation, -1 means a perfect negative linear correlation, and 0 means no linear correlation.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is about bringing all the features in your dataset onto the same level so that no single feature dominates the others just because it has a larger range. For example, if one feature is in the range of 0 to 1 and another is in the range of 0 to 1000, the model might give more importance to the second one just because of its larger values. There

are two main ways to scale: normalization (scaling values to a range of 0 to 1) and standardization (scaling so that the feature has a mean of 0 and a standard deviation of 1). Scaling is important for models that are sensitive to the magnitude of input features.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: An infinite VIF (Variance Inflation Factor) occurs when features are perfectly multicollinear, i.e., one independent variable is a perfect linear combination of another. This typically indicates redundant data and requires feature selection or dimensionality reduction.

Q6. What is a Q-Q Plot?

Ans. A Q-Q plot is a way to check if your data follows a certain distribution—in linear regression, we usually check if the residuals follow a normal distribution. In the plot, if the points line up along a straight line, it means our data is normal. It's important because the validity of the linear regression model assumptions (like the normality of errors) depends on it.