Assignment-based Subjective Questions

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

Answer: - After analysing the categorical variables, I found that some of them had a strong effect on the bike demand. For example, the variable **season** showed that spring tends to have lower bike usage, while winter has higher usage compared to the reference category. Similarly, **weekday** revealed that Saturdays were associated with a significant rise in bike demand whereas weather like Light Snow/Light Rain or Mist/Cloudy cause a significant drop in bike demand . On the other hand, some categories like **month** values did not show significant influence and were later removed during model refinement. This shows that not all categorical variables have equal importance, but some clearly affect the demand pattern.

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

Answer: - Using **drop_first=True** helps prevent a problem called the dummy variable trap. This trap happens when one dummy variable can be predicted from the others, causing multicollinearity in the regression model. For example, for gender with two type of category male or female, if one is not female then he would be male. By dropping the first category, we keep the model clean and avoid confusion in coefficient estimation. It also makes the model more stable and easier to interpret.

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

Answer: - From the pair-plot, I observed that the variable atemp (feeling temperature) had the highest positive correlation with the target variable cnt (total bike demand). As feeling temperature increased, the demand for bikes also increased.

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

Answer: - After building the model, I validated the key assumptions of linear regression using residual analysis. I plotted residuals versus predicted values to check for linearity and constant variance (homoscedasticity). The residuals appeared randomly scattered, which supports these assumptions. I also created a histogram and a Q-Q plot of the residuals to check for normality. The histogram showed a roughly bell-shaped curve, and the Q-Q plot followed a roughly straight line, confirming that the residuals were approximately normal. Lastly, the Durbin-Watson value was around 2, indicating no autocorrelation in residuals.

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

Answer: - The top 3 features in the final model that contributed significantly to explaining the demand were yr (year), season_Spring (spring session), and weathersit_Light Snow/Light Rain. The year showed a strong increasing trend in demand whereas spring season and bad weather conditions like light snow or rain had a strong negative effect on bike usage.

General Subjective Questions

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

Answer: - Linear regression is a supervised learning algorithm. It is used to predict a continuous value based on one or miltiple input features. The goal is to find the best-fit straight line that shows the relationship between the dependent variable (what we want to predict) and the independent variables (features). This line is defined using an equation of the form:

$$Y= b_0+ b_1x_1+ b_2x_2 + \ldots + b_nx_n$$

where b_0 is the intercept, and b_1, b_2, ..., b_n are the coefficients that represent the effect of each feature on the target. The algorithm calculates the best values of these coefficients by minimizing the **sum of squared differences** between the actual and predicted values, this is called the **least squares method**. After building the model, we can use it to predict values and also analyze how each feature influences the outcome.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: - Anscombe's quartet is a set of four different dataset with similar mean, variance correlation and regression line, however they are very different when plotted graphically. Though they are producing the same regression output, the scatterplot shows very different pattern – one with nonlinear curve, with only one outlier. It shows importance of graphically plotting the data during analysis, only statistical data is not sufficient. It is created by statistician Fancic Anscombe, to convey that we should not rely only on statistical metrics, we must look the data visually to better understand the relations.

3. What is Pearson's R? (3 marks)

Answer: - Pearson's R, also called the Pearson correlation coefficient, is a measure that shows how strongly two variables are linearly related. Its value ranges from -1 to +1. A value of +1 means a perfect positive linear relationship, -1 means a perfect negative linear relationship, and 0 means no linear relationship. For example, if temperature and ice cream sales have a Pearson's R of 0.9, it means as temperature increases, ice cream sales also increase strongly. It helps in understanding how closely variables move together in a straight-line pattern.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: - Scaling is the process of bringing all numerical features to a similar range so that no feature dominates others due to its large values. It is important when features have different units or ranges, especially in algorithms like linear regression.

Normalized scaling (also called min-max scaling) brings values to a fixed range, usually between 0 and 1. **Standardized scaling** (also called Z-score scaling) transforms data to have a mean of 0 and a standard deviation of 1.

The key difference is that normalization compresses values within a fixed range, while standardization centers and scales data based on its distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: - VIF, or Variance Inflation Factor, becomes infinite when a feature in the dataset is perfectly correlated with one or more other features. This means that the feature can be exactly predicted using a combination of the other features. In this case, the denominator in the VIF formula becomes zero, leading to an infinite value. It indicates perfect multicollinearity, which causes problems in linear regression as the model cannot separate the effect of one variable from the other. To fix this, we need to remove or combine the highly correlated variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: - A Q-Q plot, or Quantile-Quantile plot, is a graph that compares the distribution of your data (usually residuals) with a normal distribution. In linear regression, we use it to check if the residuals (errors) are normally distributed — this is one of the important assumptions of regression. If the residuals follow a straight 45° line in the Q-Q plot, it means they are roughly normal. If the points curve away from the line, it indicates skewness or outliers. Using a Q-Q plot helps us know whether linear regression results can be trusted or if another method is needed.