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)

Categorical variables have a significant effect on the target variable as many categorical variables like season (spring, winter), month (July, September), weekday (Monday) and weather situation (Misty) have ended up in the final model with very low p-value indicating their high significance in predicting the dependent variable. Their low p-value indicates each variables' significance towards the prediction of the target variable with very less multicollinearity effect as indicated by low VIF.

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)

This parameter is used to reduce the number of columns needed for any categorical variables, so to measure the effect of categorical variables, we convert them into numerical variables mostly by a combination of 0s and 1s, let's say there are n unique values for a categorical variable, all the possible values for the categorical variable can be represented in n – 1 columns, the above parameter drops the first column from the columns created as part of the dummy variables creation for the categorical variable.

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)

As per the pair-plot and correlation matrix among the numerical variables, 'atemp' and 'temp' 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)

We validated the assumption by doing residual analysis which is the difference between the predicted and the actual value of the target variable, so after plotting the distribution of the residuals we found out that the distribution follows a normal distribution with mean centered at 0, which are the two major assumptions made while doing the linear regression. Also, the Durbin-Watson statistic came out to be 2.000 for our final model which signifies no autocorrelation in the

residuals of the regression model which is the other assumption made while building the regression model.

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)

The top three features which are significantly explaining the demand of the shared bikes are: temp, hum and yr, while 'temp' and 'yr' have a positive impact on the demand of the bikes but 'hum' have a negative impact on the demand of bikes

---

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 algorithm is used to predict a dependent variable based on a set of independent variables, when the target variable is available, continuous and follows a linear relationship with the independent variables, linear regression can be employed for predicting the target variable. Following is few of the assumptions made while building a linear regression model:

1. The residuals (difference between the predicted and actual values of the target variable) follow a normal distribution.
2. The distribution of the residuals of the predicted values has a mean centered at 0.
3. The residuals are not correlated and doesn't follow any pattern.
4. The variance in the values of the target variable is constant across the distribution.

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)

<Your answer for Question 7 goes here>
The Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but appear very different when visualized. It was created to highlight the importance of data visualization in statistical analysis. While all four datasets share similar summary statistics, their scatter plot very different relationships. This shows that relying only on numerical statistic can be

misleading and visualizing data is crucial for accurate interpretation. Visualization helps detect patterns, outliers and relationships. Linear regression mayn't always be the best fit for data, outliers and non-linearity can distort conclusions.

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)

<Your answer for Question 8 goes here>
Pearson's correlation coefficient (r) measures the strength and direction of the linear relationship between two numerical variables. It ranges from -1(perfect negative correlation) to 1(perfect positive correlation). It works best for the linear relationship and is sensitive to outliers. We should use it when we need to quantify the strength of linear relationship, the data is normally distributed, and both the variables are numerical.

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 the process of bringing all the predictor variables on a common scale, as the predictor variable need to be on a common scale for model accuracy, while numerical variables can be scaled directly, categorical variables are first converted to numerical values using dummies and then considered for scaling based on the requirement. Normalized scaling brings the variable values between 0 and 1 whereas standardized scaling scales the values such that they have 0 mean and 1 as the standard deviation, also standardized scaling is less sensitive to outliers as compared to normalized scaling.

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)

Such high value of VIF indicates the issue of multicollinearity in the regression model, when a model is built there could be parameters, like in our linear regression model as well, that are highly correlated with each other causing the VIF values for the respective parameters to be very high. Typically, we remove the variable with very VIF value as they are in a way redundant to the regression model and not adding any additional value to the model prediction process. But we always remove one variable at a time, as removing one variable with high VIF can trigger VIF reduction in the VIF values of the other correlated variable.

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)

<Your answer for Question 11 goes here>

A Q-Q plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically the normal distribution. If the data follows a normal distribution, the points in a Q-Q plot will form a straight diagonal line. A Q-Q plot visually checks if data follows a normal distribution. In linear regression it ensures that the residuals are normally distributed, a key assumption. Deviations from the diagonal indicates skewness, heavy tails, or outliers.