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)

Independent Categorical Variables are 'season', 'mnth', 'weekday', 'weathersit', 'yr', 'holiday', 'workingday'. Dependent Variable is ‘cnt’.

Season: Fall attracts the users count as it has highest user registrations. Fall & Summer has greater than 5000 user count.

Month: May-Oct is good period for business have more than 5000 user count.

Weekdays: Registrations are uniform across the week.

Weather Situation: Clear, Few clouds, Partly cloudy, Partly cloudy weather attracts more customers.

Year: 2019 has more user count than 2018. So, it means business is likely to increase over time.

Holiday: Holiday reduces the bike demands as count decreases on a holiday.

Working Day: Working day has more user count, it is likely working professionals use bike more. It could also me complemented by the fact the count decreases on holidays.

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)

By using **drop_first=True** we are removing one option from the data set which could be evaluated by other dummy variables. For example, if we have colors Red, Blue, Green, we do not need all three dummy variables because knowing it is not Red or Blue it means its Green. Dropping one avoids redundancy and keeps the model simpler and easier to interpret. Thereby , prevents multicollinearity & ensures model interpretability.

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)

‘tempt & ‘atemp’ has correlation with ‘cnt’ which is around ~0.63.

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)

Validation of assumptions of Linear Regression:

1. Relationship between dependent and independent variable should be linear. By plotting residuals against predicted values, we see the residuals are randomly scattered.
2. Residuals should be approximately normally distributed. Using a hist plot we validated this assumption.
3. Homoscedasticity is validated by spread of residuals.

4. Multicollinearity check using VIF. We have VIF below 5 for all feature variables using for model building.

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)

Top 3 features contributing significantly towards explain the shared bike are:

1. 'temp' – Temperature
2. 'weathersit' – Weather situation
3. 'yr' - Year

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)

Liner regression is a supervised learning algorithm used to model relationship between dependent variable (also called target variable) and one or more independent variable (also called predictors or features). The goal for the liner regression is to find a liner relationship that best describes the dependent variable changes with respect to the independent variable.

Algorithm mathematically can be written as below:

Y = beta0 + beta1X1 + beta2X2 + …….. + betanXn + e

Where:

Y is dependent or target variable
X1 to Xn are features
Beta0 to Betan are coefficients for each feature.

Objective of the algorithm is to find betas which would minimize the Mean Squared Error(also called Ordinary Least Square) which is the sum of squares of 'Actual Values' – 'Predicted Values'.

Assumptions:

1. Linearity
2. Normality of Residuals
3. Homoscedasticity
4. Non-multicollinearity or Independence

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 as name says quartet. These data sets have nearly identical simple descriptive statistics but very different distributions and underlying relationship.

This example was created by Francis Anscombe in 1973 to highlight the importance of visualizing data before making conclusions. It is a classical illustration in statistics and data science, emphasizing that statistical summaries (like mean, variance, correlation, etc.) alone can be misleading without deeper analysis and visualization.

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, is a statistical measure that qualifies the liner relationship between two continuous variables. It is denoted as r and provide a value between -1 and +1, which indicates the strength and direction of the relationship.

Pearson's R measure the degree of liner correlation between two variables. It does not capture non-linear relationship.

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 adjusting the values of numerical features in your data to fit within a specific range or to have certain statistical properties. It is crucial when working with machine learning algorithm that rely on the distance between data points, such as liner regression, k-nearest neighbors (KNN), and support vector machines (SVM). It helps make the mode more efficient, improve convergence in gradient descent optimization, and ensures that features contribute equally to the model, especially when features have different units and ranges.

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)

When the value of Variance Inflation Factor becomes infinite (VIF), it means that one of the features is perfectly correlated with another feature in the dataset. This could happen when there is perfect multicollinearity. For example, if we are solving a problem to predict house prices and there are two feature like the size of the house in square feet and in square meters. In this scenario VIF would be infinite as these two features convey the same information.

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)

Q-Q plot also called Quantile-Quantile plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as normal distribution. It is used to visually assess whether the data follows a specific distribution. In a Q-Q plot:

The x-axis represents the quantile of the theoretical distribution.

The y-axis resents the quartile of the observed data.

If the data points in the Q-Q plot lie approximately along a sight line, it suggests the data follows the normal distribution.

We generally use Q-Q plot in linear regression for assessing the normality of residuals.