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)

After analysing the effect of each variable on cnt variable using Multiple Linear Regression model, I have below variables with details.

1. Year – Every year, bike sharing is increasing as per analysis.
2. Holiday – Holiday has negative effect on cnt as cnt has lesser counts on holidays.
3. Temperature – Temperature, if it is warm there are more chances of getting more counts of bike sharing.
4. Windspeed – As per negative coefficient for Windspeed, if windspeed is more, there will be lesser counts.
5. Sep – This month has more number of counts compared to other months.
6. Sun – For Sunday, there are less number of bike sharing counts happening each year.
7. Moderate – If climate is moderate, there are more counts for bike sharing.
8. Summer – Summer has positive correlation with cnt.
9. Winter – High Positive coefficient suggests that winter has better bike sharing counts with cnt.

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)

Dummy Variables are created for Categorical data. If Categorical data has n categories which are more than 2 categories we need around n-1 dummy variables.

For creating n-1 variables, we are supposed to drop one variable as the category can easily be understood with n-1 variables. If we don't create and increase number of variables for model building, we will face issue with multicollinearity.

So to prevent this, we have to drop one variable and keep n-1 variables.

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)

Temp and atemp have most correlation with 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)

Validations done for Linear Regression after building the model on training set are:

1. Comparing y_test and y_pred values
2. Finding R2 Value.
3. Residuals plotting to check Homoscedasticity.
4. Multicollinearity using VIF
5. Linearity
6. Independence of Residuals
7. Residuals should be normally distributed and mean to be zero.

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)

Demand of shared bikes is explained by features.

1. Temperature – Positive Coefficient
2. Year – Positively significant
3. Windspeed – Negatively significant

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)

<Your answer for Question 6 goes here>

Linear Regression is a supervised machine learning algorithm used for predicting a continuous target variable (y) based on one or more predictor variables (X). It assumes a linear relationship between the predictors and the target.

It aims to find the best-fitting line that well describes the relation between predictors and target variables.

When we have only 1 independent/predictor Variable – Simple Linear Regression

When we have more than 1 predictor variable – Multiple Linear Regression

Assumptions of Linear Regression are

- Linearity
- Independence
- Homoscedasticity
- Normality
- No Multicollinearity or very little

For creating model, we follow method of OLS – OLS finds the values of the coefficients that minimize the sum of the squared differences between the observed and predicted values.

Once the model is built, it can be evaluated on –

1. R-squared
2. Adjusted R-squared
3. MSE
4. RMSE
5. Probability(F-statistic)

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>
Anscombe's Quartet was created by statistician Francis Anscombe in 1973 to illustrate the importance of plotting data before you analyze it and build your model.

- It is a set of four datasets that, despite having nearly identical simple descriptive statistics (mean, variance, correlation, linear regression line), have very different distributions and scatter plots.
- Dataset 1 – Appears to follow a simple linear relationship
- Dataset 2 – Shows a clear curved relationship
- Dataset 3 – Shows a perfect linear relationship except for one outlier.
- Dataset 4 – Shows that one point is exerting a strong influence, with all other x values being the same.

This experiment highlights that

- Visualization is essential
- Limitations of Correlation and linear Regression
- Impact of Outliers
- Model Selection

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 R also known as Pearson correlation Coefficient (r), is a measure that quantifies the strength and direction of a linear relationship between two continuous variables.
It ranges from -1 to 1.

- R = 1 → Perfect positive correlation
- R = 0 → No correlation
- R = -1 → Perfect negative correlation

It assumes:

- A linear relationship between variables.
- Continuous and normally distributed
- No significant outliers affecting the correlation

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)

<Your answer for Question 9 goes here>

Scaling is process of transforming numerical features in a dataset to make them fall within a specific range or distribution.

Scaling is performed to

- Ensure all features contribute equally.
- Speeds up convergence
- Improve model performance
- Reduces numerical instability

Normalization and Standardization:

- Normalization rescales data to a fixed range. Ex: [-1,1]
  Standardization transforms data to have zero mean and unit variance.
- Normalization is sensitive to Outliers but Standardization is less sensitive as it uses mean and standard deviation.

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)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity in a dataset. Perfect multicollinearity occurs when one predictor variable is an exact linear combination of one or more other predictors.

We can address it by

- Removing perfectly correlated variables.
- Combine collinear variables
- Regularization
- Verify Data

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 assess whether a dataset follows a specific theoretical distribution , most commonly a normal distribution. It plots the quantities of the observed data against the quantities of the theoretical distribution. It is particularly useful for assessing whether a variable or residuals from a model follow a specific distribution.

In linear regression, a Q-Q plot is typically used to check the assumption of normality of residuals. The assumptions underlying linear regression include:

- Residuals should follow normal distribution.
- The normality assumption is crucial for the validity of hypothesis tests (ex: ttt-tests for coefficients) and confidence intervals.