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)

- **Season** : Considerable increase in bookings in **summer, fall & winter** as compared to **spring.** High correlation with 'cnt'

- **yr** : 2019 saw greater number of booking as compared to 2018. Expecting high correlation with 'cnt'

- **holiday** : Greater bookings when there is **NO HOLIDAY** as compared to on Holidays. So holiday has good correlation with 'cnt'

- **weekday** : **No significant** variation among weekdays. A very weak correlation or dependence with 'cnt'

- **workingday** : **No significant** variation or trend observed. A very weak correlation with 'cnt'

- **weathersit** : High bookings observed for **Clear(1) and Cloudy/Misty(2)** conditions and Lower bookings on **Snowy/Rainy (3)** conditions. NO BOOKINGS made on Heavy Rains/Foggy/High Snow conditions. Expecting good correlation with 'cnt'

- **mnth** : Relatively very High bookings made in **summer to early winter months (4th to 10th month)**and then bookings drop significantly in **winter months (11th to 3rd)**. Expecting a high correlation between 'mnth' and '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)

By dropping the first column of each categorical variable, we prevent the creation of dummy variables that are linearly dependent on each other. This helps in maintaining the independence of predictors, which is crucial for accurate model estimation. This also reduces the redundancy by reducing number of dummy variables. In short, we prevent multicollinearity and redundancy of data.

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)

For the numerical variables from the pairplot, highest correlation of the target variable is with the variable **'atemp'**

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 basically do a number of checks on the residuals (y_train - y_train_pred) & VIF-

1) Linearity of Residuals : Residuals should be randomly scattered around zero without a pattern. We check this through drawing a scatterplot of residuals.
2) Homoscedasticity : Residuals spread/variance should not show any trend when plotted
3) Mean of Residuals should be zero
4) Normal Behavior : Residuals should exhibit Normality. This is seen by plotting a histogram of residuals or a QQ plot
5) VIF of Variables should not be more than 5

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)

1) **atemp (0.645413)**: As the apparent temperature (atemp) increases, the count of bike rentals increases significantly.
2) **yr (0.239980)**: The count of bike rentals increases significantly over time, possibly due to the growing popularity of bike-sharing programs or improvements in service availability.
3) **weathersit_3 (-0.218072)**: Adverse weather conditions (weathersit_3) significantly decrease the count of bike rentals, showing that people are less likely to rent bikes in poor weather.

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 Process:

1. Collect Data:
   - Gather dataset with dependent and independent variables.
2. Preprocess Data:
   - Identify and remove null values.
   - Look for outliers and handle them.
   - Handle missing data.
3. Plot (boxplots, pairplots, correlations, scatterplot etc.) data (numerical & categorical variables) to find any obvious trends
4. Encode Categorical Variables:
   - Convert categorical variables to numerical using OneHotEncoder or similar techniques.
   - Use `drop_first=True` to avoid multicollinearity.

5. Split Data:

- Divide the data into training and testing sets.

6. Scale Data:

- Normalize or standardize features to remove effect of scaling of variables
- Do it for both training and test data

7. Run RFE : To get initial or some assessment/ranking of relevant variables. Remove low(>1) ranked variables.

8. Build & Train Model on training data:

- Initialize the linear regression model.
- Run OLS data to get model parameters (R-square, F-stat, p-Values etc)
- Run VIF to find multicollinear variables
- Remove variables which are high VIF (>5) or high p-Value (>5%)
- Remove variables until all variables have low VIF and low p-Value along with high enough r-square

9. Check Model:

- Do checks on Residuals to check for linearity, zero mean, normality and homoscedasticity

10. Test Model:

- Test your model on Test Data
- Check residuals
- Calculate performance metrics (R-squared, RMSE).
- Validate assumptions (linearity, homoscedasticity, normality, independence, multicollinearity).

11. Interpret Results:

- Analyze model coefficients and significance of predictor variables.

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 consists of four datasets with nearly identical summary statistics but very different distributions when graphed. It demonstrates the importance of visualizing data to understand its true patterns and relationships, as summary statistics alone can be misleading.

Key Datasets:
Dataset I: Linear relationship.
Dataset II: Curvilinear relationship.
Dataset III: Linear relationship with an outlier.
Dataset IV: Vertical line with an outlier.

Importance of Visualization: Identical summary statistics can be misleading; visualizing the data helps to uncover underlying patterns and relationships.

Outliers and Patterns: Outliers and non-linear patterns can significantly affect statistical measures and interpretations.

Contextual Understanding: Graphical representation provides a deeper understanding of the data's context and behavior.

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's correlation coefficient, measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.

Pearson's R helps understand how strongly two variables are related and whether they move together in a linear fashion.

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: It is the process of transforming the features of a dataset to a specific range or distribution. It improves the performance and accuracy of machine learning algorithms by ensuring features are on a comparable scale.

Why Scaling:
Prevents Dominance: Ensures that no single feature dominates others due to its scale.
Improves Convergence: Helps gradient-based algorithms converge faster by providing a consistent scale for all features.
Enhances Interpretability: Makes the model's coefficients more interpretable by having features on similar scales.
There are two main methods for scaling : Normalized Scaling & MinMax Scaing
Normalized Scaling: Rescales features to a range, typically [0, 1].
Standardized Scaling: Transforms features to have a mean of 0 and a standard deviation of 1.

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>
When the value of Variance Inflation Factor (VIF) is infinite, it indicates perfect multicollinearity among the predictor variables.
Reasons:
Perfect Multicollinearity: Predictors are perfectly linearly dependent.
Duplicate Columns: Dataset contains duplicate or linear combination features.

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 (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps assess whether the data follows the expected distribution.

Use in Linear Regression:

Assumption Validation: Q-Q plots are used to check the normality of residuals, which is an important assumption in linear regression.

Visualization: It visually compares the quantiles of the residuals to the quantiles of a normal distribution. Points should lie approximately along a straight line if the residuals are normally distributed.

Importance:

Model Reliability: Ensures that the residuals are normally distributed, which is crucial for making valid inferences about the model parameters.

Identifies Deviations: Detects deviations from normality, such as skewness or kurtosis, that could impact the model's performance.