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 such as **season** and **weather condition** play a crucial role in determining bike demand.

- **Season:** Bike demand fluctuates across different seasons. Demand tends to be highest in fall due to favourable weather conditions, while it drops significantly in winter due to cold temperatures and potential snowfall.
- **Weather Condition:** Favourable weather (clear or few clouds) increases demand, whereas adverse conditions like heavy rain, snow, or thunderstorms reduce it drastically. This insight is essential for planning fleet availability and pricing strategies.

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)

When creating dummy variables for categorical features, there is a risk of **multicollinearity**, which occurs when one predictor variable can be linearly predicted from the others. Using drop_first=True removes one category from each set of dummies, preventing redundant information from being included in the regression model. This ensures better model interpretability and numerical stability, leading to more reliable coefficient estimates.

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)

The variable **temp (temperature)** has the highest correlation with bike demand (cnt).

- Warmer temperatures encourage more people to rent bikes, increasing demand.

- However, extreme temperatures (either very high or very low) might have a negative impact due to discomfort.
- The strong correlation suggests that temperature is a key predictor in the model, influencing how many bikes are rented on a given day.

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)

To ensure the validity of the linear regression model, the following assumptions were checked:

1. **Linearity**: A scatter plot of residuals vs. predicted values was used to confirm that the relationship between the independent variables and the dependent variable is linear.
2. **Homoscedasticity**: A residual plot was analysed to check if the variance of residuals is constant. If the residuals exhibit a funnel shape, heteroscedasticity is present, violating this assumption.
3. **Normality of Residuals**: A Q-Q plot and histogram of residuals were examined to ensure they follow a normal distribution. Any deviation from normality might indicate issues with model assumptions.
4. **No Multicollinearity**: The **Variance Inflation Factor (VIF)** was calculated to detect collinearity among independent variables. A VIF > 5 indicates high multicollinearity, which might impact model accuracy.
5. **No Autocorrelation**: The **Durbin-Watson test** was used to check for autocorrelation in residuals. Values close to 2 suggest no significant autocorrelation, while values approaching 0 or 4 indicate positive or negative autocorrelation, respectively.

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. **temp (Temperature):** Temperature is the most influential factor in determining bike demand. Higher temperatures generally increase rentals as biking becomes more comfortable.
2. **atemp (Feeling Temperature):** Similar to temp, this variable measures perceived temperature, impacting demand.
3. **yr (Year):** The increasing trend in demand over time suggests the growing popularity of bike rentals, indicating an overall upward trajectory in bike-sharing service usage.

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 statistical method used to model the relationship between a dependent variable (**Y**) and one or more independent variables (**X**). The equation for multiple linear regression is:

where:

- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients representing the effect of each independent variable,
- $X_1, X_2, \ldots, X_n$ are the independent variables,
- ε is the error term representing unexplained variance.

The objective is to minimize the **sum of squared residuals (errors)**, which is achieved using the **Ordinary Least Squares (OLS)** method. The assumptions of linear regression include **linearity, homoscedasticity, normality of residuals, and absence of multicollinearity/autocorrelation**.

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 **identical statistical properties** (mean, variance, correlation, regression line), yet they have completely different distributions when plotted. This highlights the importance of **data visualization** rather than relying solely on summary statistics. It demonstrates that:

1. Different data distributions can have the same statistical properties.
2. Outliers and patterns in data can heavily influence results.

3. Always **visualize data** before making model assumptions.

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 continuous variables. It ranges from -1 to +1:

- **+1:** Perfect positive correlation (as one variable increases, the other increases proportionally).
- **0:** No correlation (no linear relationship between the variables).
- **-1:** Perfect negative correlation (as one variable increases, the other decreases proportionally).

It is calculated as:

$r = (\Sigma (X - \bar{X})(Y - \bar{Y})) / \text{sqrt}(\Sigma (X - \bar{X})^2 * \Sigma (Y - \bar{Y})^2)$

Where:

- **X** and **Y** are the two variables being compared.
- $\bar{X}$ and $\bar{Y}$ are their respective means.
- The numerator represents the **covariance** between X and Y.
- The denominator normalizes the covariance by the product of the standard deviations of X and Y.

Pearson's R is widely used in statistical analysis to determine the **strength of relationships** between variables, such as temperature and bike demand. A higher absolute value of **r** indicates a stronger 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)

<Your answer for Question 9 goes here>

Scaling ensures that all features are on a comparable scale to improve model performance. There are two main types:

- **Normalization (Min-Max Scaling):** Scales data to a range of [0,1] using: Used when features have different ranges but no outliers.
- **Standardization (Z-score Scaling):** Transforms data to have **mean = 0** and **standard deviation = 1**: Used when features have different units and contain outliers.

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>

An infinite **Variance Inflation Factor (VIF)** occurs when there is **perfect multicollinearity**, meaning one independent variable is a **perfect linear combination** of others. This makes regression coefficients unreliable. To fix this, we:

- Drop one of the highly correlated variables.
- Use **Principal Component Analysis (PCA)** or **Ridge Regression** to handle multicollinearity.

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 (Quantile-Quantile plot)** compares the **distribution of residuals** against a normal distribution. If the residuals follow a normal distribution, points will align along a **45-degree line**. If not, the regression assumptions may be violated, affecting the reliability of predictions.