Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: From the dataset and the correlation heatmap, we can infer that some categorical variables significantly impact bike demand:

- **season**: Winter has the highest positive impact on demand, while spring has the lowest. This makes sense as bike rentals might be lower in spring due to unpredictable weather.
- **yr**: The year variable (yr), which distinguishes between 2018 and 2019 (1), has a strong positive effect, meaning bike demand increased in 2019.
- **weathersit**: Bad weather (light snow/rain) negatively affects demand, which is expected as people avoid biking in bad weather.
- **workingday & holiday**: Holidays negatively impact bike demand, indicating that people use bikes more for commuting than for leisure.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: When creating dummy variables for categorical features, we use drop_first=True to avoid the **dummy variable trap**. This trap occurs when all categories of a variable are included in the regression model, leading to **multicollinearity**. By dropping the first category, we prevent redundancy, ensuring that the model remains stable and interpretable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- From the correlation heatmap, the highest correlation is seen with:
- temp (0.63): Higher temperatures lead to higher bike rentals, as more people prefer biking in good weather.
- atemp (0.63): Feels-like temperature is also highly correlated, confirming that warm weather increases demand.
- yr (0.57): The increase in bike demand in 2019 indicates a growing trend in bike-sharing services.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer:
   After building the model, we checked the following assumptions:

- **Linearity**: Checked using pair plots and scatter plots of predicted values vs. actual values.
- **Normality of Residuals**: Analyzed the residuals' distribution and ensured they followed a normal pattern.
- **Homoscedasticity**: Verified using a residuals vs. fitted values plot to confirm constant variance.

- **Multicollinearity**: Checked using the Variance Inflation Factor (VIF), ensuring no excessive correlation among independent variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   Based on the final model's coefficients, the most significant factors affecting bike demand are:

- weathersit_Light Snow/Rain -> Bad weather strongly reduces demand.
- yr -> Demand significantly increased in 2019 compared to 2018.
- season_spring -> Spring negatively impacts demand, possibly due to variable weather conditions.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear Regression Algorithm

Linear regression is a supervised learning algorithm used for predicting continuous target variables. It models the relationship between the dependent variable (y) and one or more independent variables (X) using a linear equation.

1. Equation:

For simple linear regression (one feature): y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

For multiple linear regression (multiple features): y=β0+β1x1+β2x2+⋯+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilony=β0+β1x1+β2x2+⋯+βnxn+ϵ

β_0 is the intercept, β_1, β_2, ..., β_n are the coefficients, and ε is the error term.

2. Goal:

Minimize the cost function (Mean Squared Error, MSE):

MSE=1n∑(yi−y^i)2MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2MSE=n1∑(yi−y^i)2

To find the best-fit line that minimizes this error, Ordinary Least Squares (OLS) is used.

3. Assumptions:

Linearity: Relationship between X and y is linear.

Homoscedasticity: Constant variance of residuals.

No Multicollinearity: Independent variables should not be highly correlated.

Normality of Residuals: Residuals are normally distributed.

4. Evaluation:

R-squared (R^2): Measures the proportion of variance in y explained by X.

Adjusted R^2: Adjusts for the number of predictors.

Residual Analysis: Examines the difference between actual and predicted values.

2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer:
   Anscombe's Quartet is a powerful example showing that similar summary statistics can mask underlying data patterns. It highlights the need to visualize data, check for outliers, and ensure the appropriateness of the models we use. In practice, it's always wise to pair descriptive statistics with thorough data visualization to avoid making misleading conclusions.
   Structure:
   The quartet consists of four datasets (labeled I, II, III, and IV), each with:
   11 data points (x and y values).
   Identical summary statistics:
   - Mean of x and y.
   - Variance of x and y.
   - Correlation between x and y.
   - The linear regression line.

3. What is Pearson's R? (3 marks)
   Answer:
   Pearson's R is a statistic that measures the strength and direction of the **linear relationship** between two variables. It ranges from **-1 to +1**:

- +1: Perfect positive correlation (both variables move together).
- 0: No linear relationship.
- -1: Perfect negative correlation (variables move in opposite directions).
  It's used to assess how well one variable predicts another and assumes both variables are continuous and normally distributed.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Answer:
   Why Scale?

- Ensures all features contribute equally to distance-based models (e.g., k-NN, SVM) and gradient descent optimizations.
  Types of Scaling:

1. **Normalization (Min-Max Scaling)**:
   - **Range**: [0, 1] or any other range.
   - **Formula**: Xnorm=X−XminXmax−XminX_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}Xnorm=Xmax−XminX−Xmin
   - **Use Case**: When you want all values between 0 and 1.
2. **Standardization (Z-Score Scaling)**:
   - **Mean**: 0, **Standard Deviation**: 1.
   - **Formula**: Xstd=X−μσX_{std} = \frac{X - \mu}{\sigma}Xstd=σX−μ
   - **Use Case**: When features have different units or distributions.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:
   A **Variance Inflation Factor (VIF)** becomes infinite when there is **perfect multicollinearity** between features, meaning one feature is a perfect linear combination of others. This causes the denominator in the VIF formula to be zero, resulting in an infinite VIF value.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer:
   A **Q-Q plot** (quantile-quantile plot) is a graphical tool used to compare the **distribution of residuals** to a normal distribution. In linear regression, it's essential because it helps check the assumption that **residuals are normally distributed**, which is critical for valid confidence intervals and hypothesis tests.
   If the points in a Q-Q plot lie on or close to the straight line, it indicates that residuals are normally distributed. Significant deviations suggest potential issues, such as non-normality, that could affect model performance.