Subjective Questions - Assignment

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

- **Year:** Bike rentals saw a noticeable increase in 2019 compared to 2018, suggesting significant year-over-year growth in demand.
- **Season:** The fall season recorded the highest bike rentals, followed closely by summer. Spring had the lowest number of rentals, reflecting a strong seasonal impact on demand.
- **Weather:** Rentals peaked during sunny weather, while rain and snow had a strong negative effect, significantly decreasing bike rentals.
- **Holiday:** There were fewer bike rentals on holidays compared to non-holidays, indicating that holidays reduce demand.
- **Weekday/Working Day:** There was no significant difference in bike rentals between weekdays and weekends or between working and non-working days, showing a fairly even distribution throughout the week.
- **Month:** September saw the highest rentals, with consistent demand from April to October, and the lowest demand occurred during the winter months, particularly January & February.

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

- It helps avoid the dummy variable trap, which can cause multicollinearity in the model.
- For a categorical variable with n categories, we only need n-1 dummy variables, as the nth category can be inferred from the others.
- This reduces redundancy in the model and improves its stability and interpretability.
- It prevents perfect multicollinearity, which can cause issues in estimating coefficients in regression models

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

- Temperature (temp) has the highest correlation with the target variable (bike rentals count).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- **Linearity:** Scatter plots of residuals versus predicted values were examined to confirm that a linear relationship exists between predictors and the target variable.
- **Normality of residuals:** A Q-Q plot and histogram of the residuals were used to verify that the residuals are normally distributed.
- **Homoscedasticity:** The residuals vs. fitted values plot was analyzed to check for constant variance, confirming no significant patterns in the spread of residuals.
- **Multicollinearity:** The Variance Inflation Factor (VIF) was calculated to ensure no high correlations among independent variables.

1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

I. Temp (coef: 0.4784):
II. Year (coef: 0.2442)
III. Rainy weather (coef: -0.1909)

General subjective questions:

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

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The algorithm aims to find the best-fitting line or hyperplane that represents this relationship.

Key components:

- Equation: $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn + \varepsilon$ Where $\beta0$ is the intercept, $\beta1...\betan$ are coefficients, and ε is the error term.
- Goal: Minimize the sum of squared residuals (Ordinary Least Squares method).
- Cost function: $J(\beta0, \beta1) = \Sigma(yi - \hat{y}i)^2 / n$ Where yi is the actual value and $\hat{y}i$ is the predicted value.
- Optimization: Use calculus to find the minimum of the cost function.

Assumptions:

- Linearity
- Independence of observations
- Homoscedasticity
- Normality of residuals
- No multicollinearity

Types:

- Simple linear regression (one independent variable)
- Multiple linear regression (two or more independent variables)

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets created by Francis Anscombe in 1973 to demonstrate the importance of data visualization in statistical analysis. Key points:

- All four datasets have nearly identical simple descriptive statistics (mean, variance, correlation, regression line).
- When graphed, they appear very different, revealing unique patterns:
    1. Dataset 1: Simple linear relationship with some noise
    2. Dataset 2: Clear non-linear relationship
    3. Dataset 3: Linear relationship with a single outlier
    4. Dataset 4: Vertical line with one outlier drastically affecting correlation

The quartet illustrates that relying solely on summary statistics can be misleading, and visual inspection of data is crucial for proper analysis and interpretation.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as Pearson's correlation coefficient, is a measure of linear correlation between two variables X and Y. Key points:

- Range: -1 to 1
  - 1: Perfect positive linear correlation
  - 0: No linear correlation
  - -1: Perfect negative linear correlation
- Formula: $r = \Sigma(Xi - \bar{X})(Yi - \bar{Y}) / \sqrt{[\Sigma(Xi - \bar{X})^2 * \Sigma(Yi - \bar{Y})^2]}$
- Interpretation: Indicates strength and direction of linear relationship between variables.
- Use: Commonly used in statistics to measure the degree of linear dependence between two variables.

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

Scaling refers to the process of adjusting the range of values in features within your dataset to ensure they contribute equally to the model. This process is particularly important because many machine learning algorithms (e.g., gradient descent-based models) are sensitive to the scale of input data. Without scaling, certain features might dominate others purely due to the difference in their ranges.

There are two common methods of scaling:

- **Normalization (Min-Max Scaling):** This method rescales the data to a fixed range, typically between 0 and 1. This is useful when we know the boundaries of the data or need to ensure that all features lie within the same range.
  The formula for Min-Max Scaling is:
  - $X_{scaled} = (X - X_{min}) / (X_{max} - X_{min})$

- **Standardization (Z-score Scaling):** This method rescales the data so that it has a mean of 0 and a standard deviation of 1. This is often preferred when the data follows a normal distribution and is especially useful in algorithms that assume normality, like Principal Component Analysis (PCA) or Logistic Regression.
  The formula for Z-score Scaling is:
  - $X_{scaled} = (X - \mu)/\sigma$

Scaling is performed to ensure that the learning algorithm treats all features equally, rather than giving preference to features with higher magnitudes. In addition, algorithms that rely on distance metrics (e.g., K-Nearest Neighbors, SVMs) are sensitive to the scale of the data, and scaling helps improve the model's convergence and accuracy.

Key Difference:

- **Normalization** brings all values into a specific range, usually [0, 1].
- **Standardization** adjusts the values to have a mean of 0 and a standard deviation of 1, which doesn't necessarily bound the values within a fixed range

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

When the **Variance Inflation Factor (VIF)** becomes infinite, it indicates the presence of a **perfect linear relationship** between one predictor variable and the others. The formula for VIF is:

- $VIF = 1/ (1-R^2)$

If the R^2 value of a predictor's regression against other predictors is 1, the VIF becomes infinite, as the denominator equals zero. This scenario is often observed when a predictor variable is perfectly or near-perfectly correlated with other predictors, leading to Perfect Multicollinearity.

Consequences of Infinite VIF:

- **Model Instability**: Regression coefficients can become unstable, with minor changes in data causing large fluctuations.
- **Inaccurate Inferences**: It becomes challenging to interpret the effect of individual predictors due to high multicollinearity.
- **Computational Problems**: Some algorithms may fail to converge or return unreliable results when encountering such multicollinearity.

Managing such multicollinearity is crucial for improving model accuracy and stability.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool that compares the distribution of a dataset against a theoretical distribution, typically a normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

Use in Linear Regression: In linear regression, a Q-Q plot is used to assess whether the residuals (the differences between actual and predicted values) follow a normal distribution, which is a key assumption for the validity of statistical tests and inferences in the model.

Importance in Linear Regression: A Q-Q plot helps verify the normality assumption of residuals. If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are normally distributed. Significant deviations from the line indicate departures from normality, which may suggest issues like skewness or kurtosis that could affect model accuracy and inference reliability.