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)

We have multiple categorical variables:

- **Season**: Compared to the reference season (likely summer or winter), being in spring is associated with a decrease in cnt by approximately 1331.21 units. This effect is statistically significant
- **Weather**: Light snow is associated with a decrease in cnt by approximately 2469.57 units compared to the reference weather situation. This effect is statistically significant. Mist is associated with a decrease in cnt by approximately 667.22 units compared to the reference weather situation. This effect is statistically significant.
- **Weekday**: Being on weekday 6 (likely Saturday) is associated with an increase in cnt by approximately 502.99 units compared to the reference weekday. This effect is statistically significant.
- **Months**: March is associated with an increase in cnt by approximately 409.23 units compared to the reference month. This effect is statistically significant. June is not statistically significant, indicating it may not have a clear effect compared to the reference month. July is associated with a decrease in cnt by approximately 714.16 units compared to the reference month. This effect is statistically significant. August is associated with a decrease in cnt by approximately 414.70 units compared to the reference month. This effect is statistically significant. October is associated with an increase in cnt by approximately 524.32 units compared to the reference month. This effect is statistically significant.

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

Using drop_first=True when creating dummy variables:

- **Prevents Multicollinearity**: Ensures that the dummy variables are not perfectly collinear with each other and the constant term.
- **Allows Model Fit**: Ensures that the regression model can be computed correctly and that coefficients are meaningful and interpretable.
- **Facilitates Interpretation**: Provides a baseline or reference category, making it easier to interpret the effects of other categories relative to this baseline.

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

To determine which numerical variable has the highest correlation with the target variable by examining a pair-plot, you need to look at the scatter plots in the pair-plot that involve the target variable and each numerical predictor. **We see that 'temp' has the highest correlation with target variable.**

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

By performing below diagnostic checks, you can validate the assumptions of linear regression and ensure that the model's estimates and predictions are reliable.

1. Linearity: The relationship between the predictors and the target variable is linear. Plotted the residuals (errors) against the fitted values (predicted values). Look for a random scatter of points around zero. A clear pattern (e.g., a curve) suggests a violation of the linearity assumption.

2. Independence: The residuals (errors) are independent of each other.

3. Homoscedasticity: The residuals have constant variance across all levels of the predictor variables.

4. Normality of Residuals: The residuals are normally distributed. Created Q-Q plot of the residuals to check for normality.

5. No Multicollinearity: The predictor variables are not too highly correlated with each other. Calculated the VIF for each predictor variable. VIF values greater than 10 indicate high multicollinearity. VIF was < 5

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

Based on the final model, the top 3 features significantly contributing to explaining the demand for shared bikes are:

1. Apparent Temperature (atemp): Positive impact, highest coefficient.
2. Weather Situation: Light Snow (weathersit_Light Snow): Strong negative impact.
3. Season: Spring (season_spring): Strong negative impact.

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 (target) and one or more independent variables (predictors). The goal is to find the best-fit line (or hyperplane in multiple dimensions) that predicts the dependent variable based on the independent variables.

The general formula for linear regression is

The general formula for linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where:

- y is the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- β_0 is the y-intercept of the line.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (slopes) of the independent variables.
- ϵ is the error term (residual), representing the difference between the predicted and actual values.

2.

Objectives

The main objectives of linear regression are to:

1. **Estimate the Coefficients** (β0,β1,...,βn\beta_0, \beta_1, \ldots, \beta_nβ0,β1,...,βn): Determine the values that minimize the difference between predicted and actual values.
2. **Predict the Dependent Variable**: Use the estimated coefficients to make predictions based on new data.
3. **Understand Relationships**: Assess how changes in independent variables influence the dependent variable.

3. Assumptions

Linear regression relies on several assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.

2. **Independence**: The residuals (errors) are independent of each other.
3. **Homoscedasticity**: The residuals have constant variance at all levels of the independent variables.
4. **Normality of Residuals**: The residuals are normally distributed.
5. **No Multicollinearity**: The independent variables are not too highly correlated with each other.

4. **Model Fitting**
   Ordinary Least Squares (OLS) is the most common method for fitting a linear regression model. The objective is to minimize the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- y_i is the actual value of the dependent variable.
- $\hat{y}_i$ is the predicted value from the model.

OLS estimates the coefficients (β) by minimizing the RSS.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but different distributions and relationships between variables. Created by Francis Anscombe in 1973, it illustrates the importance of graphing data before analyzing it.
Each dataset in the quartet consists of the same number of observations (11), with identical means, variances, correlation coefficients, and linear regression lines. However, their scatter plots reveal distinct patterns:

1. **Dataset I**: A standard linear relationship.
2. **Dataset II**: A nonlinear relationship with a curve.
3. **Dataset III**: A linear relationship with a single outlier.
4. **Dataset IV**: A vertical line, indicating a constant value for the dependent variable.

Anscombe's quartet emphasizes that relying solely on summary statistics can be misleading and highlights the need for visual inspection to understand the data's true nature.

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

Pearson's RRR, or Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted as rrr and ranges from -1 to 1.

- **r=1r = 1r=1**: Perfect positive linear relationship; as one variable increases, the other variable also increases proportionally.
- **r=−1r = -1r=−1**: Perfect negative linear relationship; as one variable increases, the other variable decreases proportionally.
- **r=0r = 0r=0**: No linear relationship between the variables; changes in one variable do not predict changes in the other.

Calculation

Pearson's rrr is calculated using the formula:

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- $\text{cov}(X, Y)$ is the covariance between the variables X and Y.
- σ_X and σ_Y are the standard deviations of X and Y, respectively.

Interpretation

- **Positive rrr**: Indicates a positive linear relationship; both variables move in the same direction.
- **Negative rrr**: Indicates a negative linear relationship; one variable moves in the opposite direction of the other.
- **Magnitude**: The absolute value of rrr indicates the strength of the linear relationship, with values closer to 1 or -1 indicating a stronger relationship.

Pearson's rrr assumes a linear relationship and requires the data to be normally distributed and free from outliers to provide accurate results.

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

Scaling adjusts the range of feature values to a common scale, improving the performance of machine learning algorithms and ensuring that features contribute equally.

Why Scaling is Performed:

1. Ensures uniform contribution of features.
2. Enhances convergence speed in optimization algorithms.
3. Improves the performance of distance-based algorithms (e.g., k-NN, SVM).

Normalized Scaling:

- Rescales data to a specific range, usually [0, 1].
- Formula: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$.

Standardized Scaling:

- Centers data around the mean with unit variance.
- Formula: $x' = \frac{x - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation.

Normalization is useful for bounded ranges, while standardization is better for algorithms assuming normally distributed data.

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

An infinite VIF indicates that the predictor's variance is not unique and is entirely explained by other predictors, making it impossible to estimate its regression coefficient accurately. This often necessitates removing or combining predictors to address the multicollinearity issue

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

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess whether a dataset follows a theoretical distribution, such as the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

Use in Linear Regression:

1. **Normality of Residuals**: A Q-Q plot helps check if the residuals of a linear regression model are normally distributed. For linear regression to be valid, residuals should be approximately normal.
2. **Model Validity**: If the points on the Q-Q plot lie on or close to the reference line (45-degree line), it suggests that the residuals are normally distributed, supporting the model's assumptions.

Importance:

- **Assumption Checking**: Validates the normality assumption of residuals, crucial for reliable hypothesis testing and confidence intervals.
- **Model Diagnostics**: Identifies deviations from normality, which may indicate problems with the model or the need for transformation of variables.