Assignment-based Subjective Questions

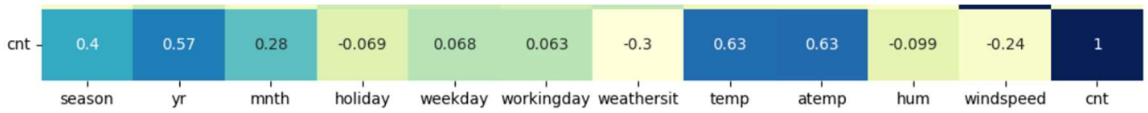
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

- Weather Impact: Weather impacted the pattern of renting bikes. During Clear, Few clouds, partly cloudy, partly cloudy days and on days when its Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. Bikers tend to rent more.
- Seasonal Trends: Bike rentals tend to change depending on the season. During Summer and Spring there has been a decline in renting trends.
- Weekday vs. Weekend: Bike rentals may vary between weekdays (like Monday to Friday) and weekends (Saturday and Sunday). Weekends have seen a negative correlation with renting count.

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

   Using `drop_first=True` during dummy variable creation helps to avoid multicollinearity issues by eliminating one of the dummy variables, preventing perfect predictability of one variable from the others. This enhances model interpretability and efficiency while avoiding the "dummy variable trap."

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



   Column ‘temp’ and ‘atemp’ show highest correlation of 0.63 with the ‘cnt’, Count.

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

   There could be multiple techniques for evaluation like

   - Residual Analysis – By checking the residuals (the differences between observed and predicted values) for randomness and constant variance. By plotting scatterplots.
   - Homoscedasticity Test – Used to check the constant variance of residuals.
   - Normality Test - Assessed the normality of residuals using statistical tests like the Shapiro-Wilk test or visually inspecting histograms and QQ plots.

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

Top three features with positive coefficient are during the months June, September, and October.

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 (features).

Linear regression assumes that the relationship between the independent variables X and the dependent variable y is linear. Given by

$$; y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon,$$

Where, y is target variable, $X_1, X_2, ..., X_n$ are the independent variables, $\beta_0, \beta_1, ..., \beta_n$ are the coefficients, and ϵ is the error term.

During training, the algorithm iteratively adjusts the coefficients to minimize the error (difference between observed and predicted values) using optimization techniques like gradient descent.

The algorithm calculates the gradient of the cost function with respect to each coefficient and updates the coefficients in the direction that reduces the error.

Linear regression is widely used in various fields, including economics, finance, healthcare, and social sciences, for tasks such as sales forecasting, risk assessment, and trend analysis.

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

Anscombe's quartet consists of four datasets that have nearly identical statistical properties, including means, variances, correlations, and linear regression lines. Despite their similarity in summary statistics, the datasets differ significantly when plotted graphically. This illustrates the importance of visualizing data to understand its underlying patterns and relationships. The quartet highlights the limitations of relying solely on summary statistics and emphasizes the value of exploratory data analysis to uncover insights and detect anomalies. It serves as a cautionary example in statistical analysis, reminding researchers to consider the broader context and visualize their data comprehensively.

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

Pearson's correlation coefficient (often denoted as Pearson's \(r \)) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the relationship, ranging from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's 'r' is sensitive to outliers and assumes that the relationship between variables is linear and that the variables are normally distributed. It's widely used in statistics to assess associations between variables in various fields, including psychology, sociology, and economics.

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

Scaling is the process of transforming data to a standardized range or distribution. It's performed to ensure that all variables contribute equally to the analysis and to address issues related to differences in the scale or magnitude of variables.

1. Purpose of Scaling:
    a. Equalize Variables: Scaling ensures that variables with larger magnitudes don't dominate those with smaller magnitudes in algorithms that use distance-based metrics, such as K-means clustering or gradient descent.
    b. Improves Model Performance: Scaling can improve the performance of machine learning algorithms by accelerating convergence and making them more robust to outliers.
    c. Interpretability: Scaling makes the coefficients or feature importance scores in models more interpretable and comparable.

2. Normalized Scaling:
   In normalized scaling, also known as min-max scaling, values are transformed to fit within a specified range, typically between 0 and 1.
   Given by-

$$\frac{x-\min(x)}{\max(x)-\min(x)},$$

3. Standardized Scaling:
   Standardized scaling, also known as z-score scaling or standardization, transforms data to have a mean of 0 and a standard deviation of 1.

$$\frac{x-\text{mean}(x)}{\text{std}(x)}$$

Difference:

- Normalization: scales the values within a fixed range (e.g., 0 to 1), while standardization transforms data to have a mean of 0 and a standard deviation of 1.
- Normalization is sensitive to outliers, as it adjusts the range based on the minimum and maximum values, while standardization is less affected by outliers because it uses the mean and standard deviation.
- Normalization preserves the original distribution shape, while standardization transforms the distribution to be centered at 0 with a spread of 1.

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

The occurrence of infinite values for Variance Inflation Factor (VIF) typically happens when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity arises when one or more independent variables can be exactly predicted from the others. In such cases, the correlation matrix becomes singular, leading to the inability to invert it, resulting in infinite VIF values. This occurs because the computation of VIF involves taking the inverse of the correlation matrix, which is not possible when perfect multicollinearity exists.

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 (quantile-quantile) plot is a graphical tool used to assess whether a set of data follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically a normal distribution.

Use and Importance in Linear Regression:

- Distribution Assumption: Q-Q plots are crucial for checking the assumption of normality in linear regression residuals. If the residuals are normally distributed, the points on the Q-Q plot should fall approximately along a straight line.
- Detecting Non-Normality: Departures from a straight line in the Q-Q plot indicate deviations from normality in the residuals. This could suggest that the linear regression assumptions are violated, leading to biased estimates and incorrect inferences.
- Model Validity: Ensuring that the residuals are normally distributed is important for making valid statistical inferences and predictions from the linear regression model. Q-Q plots help assess the validity of the model and identify any potential issues that need to be addressed.