Assigment-based Subjective Questions

Answer 1. The categorical variable ‘weathersit’ can be a used to build a useful regression model as the ‘cnt’ when ‘weathersit’ is ‘bad’, can be easily differentiated from the ‘cnt’ when the weathersit is either ‘good’ or ‘better’ which indicates that the number of users are more on the days when the weather is either ‘good’ or ‘bad’ compared to the number of users on the days when the weather is ‘low’.

Answer 2. ‘drop_first=True’ is essential because it helps to reduce the correlations between the columns created during the dummy variable creation.

Answer 3. The variable ‘registered’ has the highest correlation with ‘cnt’.

Answer 4. I plotted the histogram of the error terms that came out to be normally distributed which is one of the major assumptions of linear regression.

Answer 5. The top three features contributing significantly towards explaining the demand of the shared bikes are: 1. yr, 2. temp and 3. windspeed.

General Subjective Questions

Answer 1. Linear regression is a fundamental supervised machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It is a simple yet powerful technique for making predictions and understanding the relationships between variables. Here's a detailed explanation of the linear regression algorithm:

Key Terminology:
1. Dependent Variable (Target): This is the variable we want to predict or explain. It's often denoted as Y.
2. Independent Variable(s) (Features): These are the variables that are used to predict the dependent variable. If there's only one independent variable, it's a simple linear regression. If there are multiple independent variables, it's a multiple linear regression. These are often denoted as X1, X2, ..., Xn.

3. Linear Equation: The fundamental assumption in linear regression is that the relationship between the independent and dependent variables is linear. The linear equation for a simple linear regression is typically written as:
$Y = \beta 0 + \beta 1 * X + \varepsilon$

Y is the dependent variable.
X is the independent variable.
$\beta 0$ is the y-intercept (constant term).
$\beta 1$ is the slope of the line, representing the change in Y for a unit change in X.
ε represents the error term, accounting for the variability that is not explained by the model.

Model Assumptions:
Linear regression makes several key assumptions, including:
1. Linearity: The relationship between the variables is assumed to be linear.
2. Independence: The errors (residuals) are assumed to be independent of each other.
3. Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables.
4. Normality: The errors are assumed to be normally distributed.

Answer 2. Anscombe's quartet is a famous dataset in statistics that consists of four distinct datasets, each containing 11 data points. What makes the quartet particularly interesting and instructive is that, despite having very different data distributions and characteristics, they share nearly identical summary statistics. Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and not relying solely on summary statistics when analyzing data.

Key Insights from Anscombe's Quartet:
1. Summary Statistics Can Be Deceptive: All four datasets have nearly identical summary statistics (mean, variance, correlation), suggesting a

similar relationship between X and Y. However, when you plot the data, you see very different patterns.

2. Importance of Data Visualization: Anscombe's quartet underscores the importance of data visualization in understanding data. When you visualize the data, you can identify relationships, outliers, and nuances that summary statistics alone cannot reveal.

3. The Need for Robust Analysis: Analysts should not rely solely on summary statistics or simple linear regression. They should use data visualization and more robust analysis methods to gain a complete understanding of their data.

Answer 3. Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is widely used in statistics to assess the degree to which two variables are related to each other. Pearson's r ranges from -1 to 1, where:

If r = 1, it indicates a perfect positive linear relationship. As one variable increases, the other also increases in a linear fashion.
If r = -1, it indicates a perfect negative linear relationship. As one variable increases, the other decreases in a linear fashion.
If r = 0, it suggests no linear relationship between the two variables.

Answer 4. Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the range of variables or features in your dataset. The primary goal of scaling is to bring all variables to a standard range without changing their relationships. Scaling is performed for several reasons, including making the data more suitable for certain algorithms, improving convergence speed, and ensuring that no variable dominates the learning process due to its scale. There are two common methods for scaling: normalized scaling and standardized scaling.

Normalized Scaling (Min-Max Scaling):
Normalized scaling, also known as min-max scaling, is a method for transforming the values of variables to a specific range, usually between 0 and 1.

Standardized Scaling (Z-Score Standardization):
Standardized scaling, also known as z-score standardization, transforms the values of variables to have a mean of 0 and a standard deviation of 1.

Answer 5. The Variance Inflation Factor (VIF) can sometimes become infinite when there is perfect multicollinearity in your dataset. Perfect multicollinearity occurs when two or more independent variables in a regression model are perfectly correlated, meaning one can be exactly predicted from the others using a linear combination. In other words, there is a linear relationship among these variables that has no variation or error.

Answer 6. A Quantile-Quantile (Q-Q) plot is a graphical tool used in linear regression and statistics to assess the normality of a dataset or the goodness of fit of a statistical model, such as linear regression.

If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are approximately normally distributed, which is a good sign for the validity of the linear regression model.

If the points deviate from a straight line in the Q-Q plot, it indicates deviations from normality in the residuals. In this case, you might need to consider data transformations or assess whether the regression model assumptions are met.

If the points curve upward or downward in the Q-Q plot, it may suggest skewness or heavy-tailed distributions in the residuals, which could affect the model's performance.