ASSIGNMENT SUBJECTIVE QUESTIONS

1. The ‘cnt’ variable can be represented as :
   cnt = C1*season + C2*yr + C3*weathersit + C4*atemp + C5*casual + C6*registered, where the values of all the coefficients are :
   C1 = 2.429e-17
   C2 = -3.816e-17
   C3 = -2.914e-16
   C4 = --3.331e-16
   C5 = 0.3775
   C6 = 0.7968

   From my analysis the model depends equally on categorical (yr, weathersit, season) and numeric (atemp, casual, registered) variables. None of the dependent variables have VIF factory more than 4 and all have P values < 0.5. As per the analysis we can see that the coefficients C1, C2, C3, C4 are very low. So the ‘cnt’ variable depends more strongly on ‘casual’ and ‘registered’ number of bikers. The factors such as ‘yr’, ‘weathersit’, ‘atemp’ affect ‘cnt’ negatively.

2. 1\. Avoiding Multicollinearity: drop_first=True prevents multicollinearity by excluding one dummy variable, preventing perfect correlation between them.
   2. Redundant Information: Including all dummy variables risks redundancy, as the absence of one category can be inferred from the presence of others.
   3. Interpretability: It enhances the interpretability of the model coefficients by having a clear reference category.
   4. Efficient Representation: Using one less dummy variable maintains model efficiency without sacrificing information.
3. ‘temp’ and ‘atemp’ variables have the highest correlation that equals 0.99
4. Validation points :
   1. The error terms were normally distributed around 0.
   2. Prediction on test data was coming linear
   3. The R squared value on training data comes out to be 1 while on test data its 0.99
   4. The F-statistics value is hugely large to support the idea the model we created in that at least one has a non zero coefficient and we haven’t got the result by chance.

5. Casual, registered, atemp

GENERAL ASSIGNMENT QUESTIONS

1.

1. Model Representation: Linear regression models the relationship between a dependent variable Y and one or more independent variables X using a linear equation:

   Y = β0 + β1X1 +β2X2+...+ϵ.

2. Objective: The goal is to minimize the sum of squared differences between observed and predicted values by adjusting coefficients β0 , β1,....

3. Training: The model is trained by optimizing coefficients using a method like gradient descent to minimize a cost function, often Mean Squared Error (MSE).

4. Cost Function: MSE quantifies the average squared difference between observed and predicted values, helping guide the optimization process.

5. Assumptions: Linear regression assumes linearity, independence, homoscedasticity, normality, and no multicollinearity among variables.

6. Coefficients: β0 is the intercept, β1 represents the change in Y for a one-unit change in X, and ϵ is the error term.

7. Evaluation: Model performance is assessed using metrics like R2 (coefficient of determination) to measure the proportion of variance explained.

8. Extensions: Multiple linear regression involves more than one independent variable; polynomial regression fits a polynomial equation.

9. Use Cases: Linear regression is widely applied in predicting outcomes such as sales, stock prices, and housing prices, offering interpretability and simplicity in modeling relationships.

2.

Dataset Creation: Anscombe's quartet consists of four datasets, each containing 11 (x, y) pairs, designed by statistician Francis Anscombe to illustrate the importance of graphical exploration in statistical analysis.

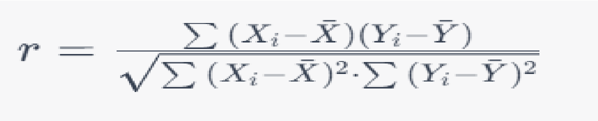
Statistical Properties: Despite having very different distributions and relationships, the quartet's datasets share identical or nearly identical statistical properties, including means, variances, and correlations.
Graphical Exploration: When plotted, the datasets reveal the limitations of relying solely on summary statistics; all four datasets have the same linear regression line and correlation coefficient.
Diverse Relationships: The quartet includes examples of linear relationships, non-linear relationships, and one dataset where the relationship is heavily influenced by an outlier.
Implications: Anscombe's quartet underscores the importance of visualizing data to understand the underlying patterns and challenges assumptions that summary statistics might hide.
Educational Tool: It is frequently used in statistics education to emphasize the concept that datasets with similar summary statistics can exhibit vastly different patterns when graphically examined.
Graphs: The first dataset shows a simple linear relationship, the second a curved relationship, the third a strong outlier, and the fourth a case where the correlation is mainly driven by one point.
Summary Statistics: The datasets have the same mean and standard deviation for both x and y, as well as the same correlation coefficient.

3.

Pearson's correlation coefficient (r) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- 1: Indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally).
- 0: Signifies no linear relationship.
- -1: Suggests a perfect negative linear relationship (as one variable increases, the other decreases proportionally)

The formula for Pearson's correlation coefficient between variables X and Y with n data points is:



$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

Where,

Xi and Yi are individual data points, $\bar{X}$ and $\bar{Y}$ are the means of X and Y, and the summations are over all data points *i.*

4.

Scaling adjusts the values of variables to a standard range. It is performed to ensure that variables with different scales do not dominate in machine learning algorithms, especially in distance-based or gradient-based methods.

Normalized Scaling: Scales values between 0 and 1. Sensitive to outliers, influenced by the range of data.

Standardized Scaling: Standardizes values to have a mean of 0 and a standard deviation of 1. Less sensitive to outliers, influenced by mean and standard deviation.

5. An infinite Variance Inflation Factor (VIF) occurs when perfect collinearity exists among independent variables. This perfect multicollinearity leads to an R2 value of 1 in the VIF formula, causing division by zero.This numerical instability can result from either a perfect linear relationship between variables or computational precision issues, such as rounding errors. Resolving this requires addressing collinearity through variable removal or regularization techniques.

6. A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. In linear regression, Q-Q plots are valuable for checking the normality assumption of residuals. The plot compares the quantiles of the observed residuals against the quantiles of a theoretical normal distribution. If the points lie approximately along a straight line, it suggests that the residuals are normally distributed, validating a key assumption of linear regression. Deviations from linearity indicate departures from normality, prompting further investigation or potential model refinement.