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)*

- The lowest demand for bikes is in January, while the highest demand is in September.
- Bike demand is lower on holidays compared to non-holidays
- Bike demand is lowest in the spring season and highest in the fall season.
- Compared to 2018, the demand for bikes increased in 2019.
- The demand for bikes is higher on Mondays, and throughout the week, the demand remains relatively similar.
- Bikes are more in demand on clear weather days compared to days with possible snow or rain when people prefer staying at home or using cars.

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

- The purpose of creating dummy variables is to convert categorical columns into binary format (0s and 1s).
- When creating multiple dummy variables, dropping the first one eliminates redundancy.
- If a certain dummy variable is 1, we can infer that the dropped variable is 0, simplifying interpretation and avoiding multicollinearity issues.

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

The pair-plot analysis among the numerical variables reveals that "temp" and "atemp" exhibit the highest correlation with the target variable "cnt."

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

The assumptions of Linear Regression were validated after building the model on the training set through the following steps:

- Preparation of a test dataset by keeping common columns between the train and test sets.
- Creation of scatter plots for the train and test datasets to check if the points fall on a straight line, indicating linearity.
- Comparison of the R-squared values between the train and test datasets to evaluate the goodness of fit.

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

The top three features explaining the demand of shared bikes are:

1. Year
2. Temperature
3. Weather

General Subjective Questions

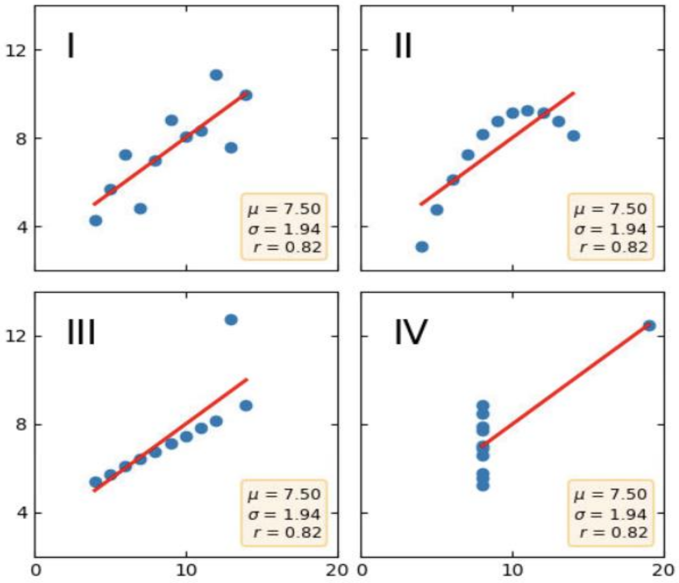
1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning model used to analyze and model the relationship between two variables.

The algorithm follows several steps:

I. Data reading and understanding using the data dictionary.
II. Data visualization through Exploratory Data Analysis (EDA).
III. Conversion of categorical columns into binary format using dummy variables.
IV. Splitting the data into train and test sets.
V. Performing multiple linear regression tasks on the train dataset to obtain the best model with maximum Adjusted R-squared.
VI. Applying the trained model on the test dataset, keeping only the common columns.
VII. Comparing the train and test data using scatter plots to assess linearity.
VIII. Evaluating the R-squared values of the train and test data to determine the best fit.

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



Anscombe’s quartet:

I. Anscombe’s quartet consists of four datasets with similar statistical values but different scatter plot patterns.
II. It emphasizes the importance of data visualization in identifying data abnormalities.
III. Anscombe’s quartet serves as an example of why data visualization is necessary for building accurate models.

3. *What is Pearson’s R? (3 marks)*

- Pearson’s R, also known as Pearson’s correlation coefficient, is a statistical measure of the linear relationship between two variables.
- The coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations.
- The value of Pearson’s R ranges between -1 and 1, indicating the strength and direction of the correlation.

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

- Scaling is a pre-processing step that normalizes categorical independent variables within a specific range.
- It is performed to handle features with varying magnitudes, units, and ranges in the collected data.
- Scaling ensures that the algorithm considers both magnitudes and units, resulting in accurate modeling.

- Normalized scaling brings the data within the range of 0 and 1, while standardized scaling replaces values with their z-scores.

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

- When the VIF value is infinite, it indicates perfect correlation (multicollinearity) between two independent variables.
- This occurs when one variable can be linearly predicted from the other with high accuracy.
- To address this issue, one of the variables causing multicollinearity should be dropped from the dataset.

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

- Q-Q plot What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.is a scatter plot that compares two different quantiles from a sample against each other.
- It is used to determine if two samples are from the same population and to assess similarities in distribution shape and tail behavior.
- Q-Q plots are important in linear regression to test the assumption of normality between the predicted and actual distributions.