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)**

   From the dataset, categorical variables like season, month, weekday, and weathersit influence the dependent variable cnt (total bike rentals):

   - **Season**: Fall shows the highest bike demand compared to other seasons while spring has the least demand. The number of reservations increased in every season from 2018 to 2019.
   - **Month**: Demand is highest from May to October, indicating seasonal peaks during warmer months.
   - **Weekday**: More reservations occur during the later part of the week, especially Thursday to Sunday.
   - **Holiday:** Bike demand is lower on holidays, likely due to reduced commuting needs.
   - **Weather Situation (weathersit)**: Clear weather results in the highest bike demand, with demand decreasing as weather conditions worsen (e.g., misty or snowy weather).

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

   Dummy variables encode categorical variables into binary (0/1) format. If all categories are represented, it creates **perfect multicollinearity**, where one variable becomes a linear combination of others. For example:

   - season encoded into spring, summer, fall, and winter creates redundancy since knowing any three categories can deduce the fourth.
   - Using drop_first=True removes one category (e.g., winter), reducing redundancy and ensuring the model is stable and interpretable.

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

   Temperature (temp) has the highest positive correlation with the target variable cnt, as seen from the pair-plot and correlation matrix.

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

   - **Linearity**: Verified using scatter plots of predictors against cnt and Component Plus Residual (CCPR) plots.
   - **Normality of Residuals**: Histogram and distribution plot of residuals confirmed that they are approximately normally distributed.
   - **Homoscedasticity**: Residuals vs. Actual Values plot showed no visible pattern, confirming constant variance of errors.
   - **Multicollinearity**: Verified using Variance Inflation Factor (VIF), ensuring all values are below 5.
   - **Independence of Residuals:** The Durbin-Watson statistic (≈ 2.085) confirms no autocorrelation among residuals.

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

From the final model (6th iteration):

1. **Temperature (temp)**: Strong positive impact; warmer weather increases bike rentals.
2. **Year (year)**: Significant growth in demand observed in 2019 compared to 2018.
3. **Windspeed (windspeed)**: Mild negative correlation; high wind speeds deter biking.

These insights suggest focusing on weather-appropriate strategies, like offering promotions during mild weather and marketing heavily in high-demand years.

General Subjective Questions

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

Linear regression is a statistical and machine learning model used to establish a relationship between a dependent variable (target) and one or more independent variables (predictors). It determines how the dependent variable changes when the independent variable(s) change.

Mathematical Representation

The relationship is expressed as:

Y=mX+cY = mX + cY=mX+c

Where:

- YYY: Dependent variable (target) to predict.
- XXX: Independent variable (predictor).
- mmm: Slope of the regression line, representing the rate of change in YYY with respect to XXX.
- ccc: Intercept, the value of YYY when X=0X = 0X=0.

The equation of linear regression is:

$$Y= \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \epsilon$$

Where:

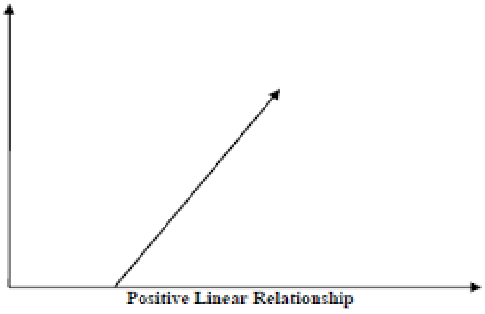
- β0: Intercept
- β1,β2,...,βn : Coefficients (weights) for predictors
- ϵ: Error term (residuals)

Types of Linear Regression

1. Simple Linear Regression:
   Involves one independent variable and one dependent variable.
   Example: Predicting house prices based solely on size.
2. Multiple Linear Regression:
   Involves multiple independent variables to predict one dependent variable.
   Example: Predicting house prices based on size, location, and number of bedrooms.

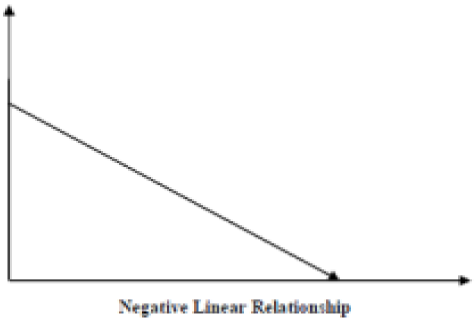
Nature of Linear Relationships

- Positive Linear Relationship:
  Both independent (X) and dependent (Ȳ) variables increase together.
  Example: As temperature (X) increases, ice cream sales (Ȳ) increase.



Graph: An upward-sloping line.

- Negative Linear Relationship:
  As X increases, Ȳ decreases.
  Example: As hours spent partying (X) increase, exam scores (Ȳ) decrease.



Graph: A downward-sloping line.

Linear regression minimizes the sum of squared residuals (errors) to find the best-fitting line. Key assumptions:

1. Linearity: The relationship between predictors and the target is linear.
2. Independence: Residuals are independent.
3. Homoscedasticity: Residuals have constant variance.
4. Normality: Residuals follow a normal distribution.

Steps to Build a Linear Regression Model

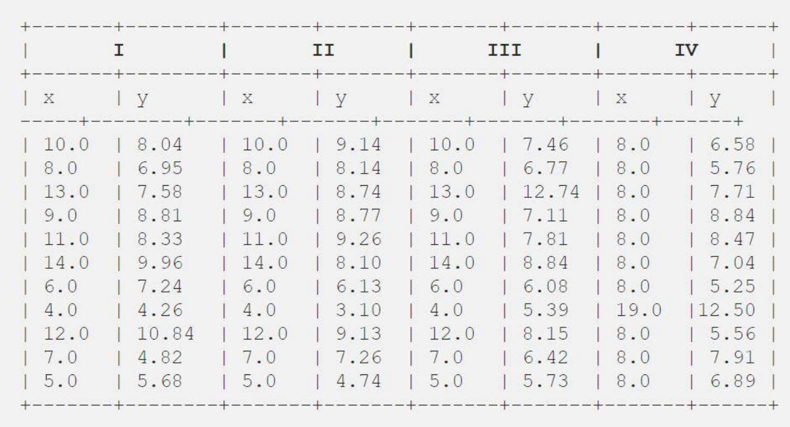
1. Data Preparation:
   - Identify dependent and independent variables.
   - Perform exploratory data analysis (EDA).
   - Handle missing values, outliers, and categorical variables.
2. Model Training:
   - Fit the regression equation to the data to find the best coefficients (mmm, ccc).
3. Model Evaluation:

- Use metrics like R2R^2R2, Adjusted R2R^2R2, Mean Squared Error (MSE), and Residual Plots.

4. Validation:
   - Validate assumptions (e.g., linearity, multicollinearity, homoscedasticity).

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

Anscombe's Quartet, developed by statistician Francis Anscombe, demonstrates the importance of visualizing data during analysis. The quartet comprises four datasets, each containing eleven (x,yx, yx,y) pairs, which share identical descriptive statistics but exhibit significantly different relationships when plotted**.**

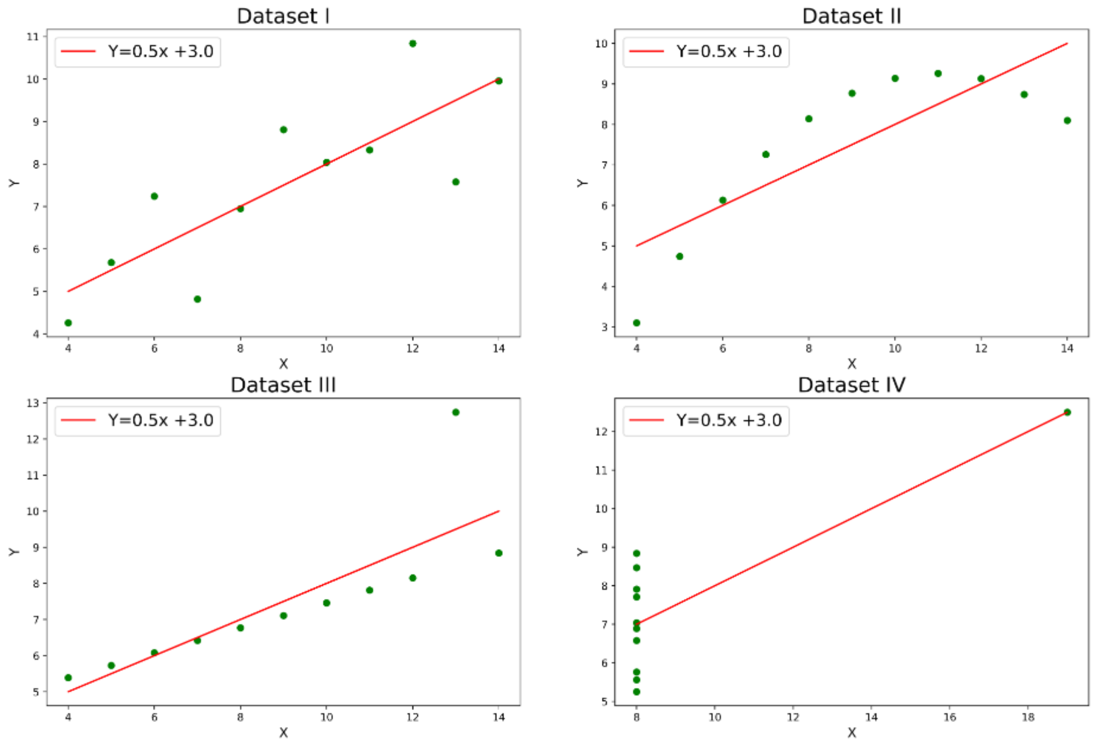


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Insights from Visualization

1. Dataset I: Clean Linear Model
   - This dataset follows a clear linear relationship between xxx and yyy, fitting the regression line well.
   - The linear regression model is valid, and the correlation reflects the relationship accurately.
2. Dataset II: Non-linear Relationship
   - This dataset shows a curved (non-linear) relationship, even though the regression line is the same as in Dataset I.
   - The correlation coefficient is misleading here, as it suggests a strong linear relationship where none exists.
3. Dataset III: Linear with Outlier
   - The dataset follows a linear trend, but an outlier significantly skews the regression line.
   - This highlights how a single point can distort regression results and overstate the strength of the relationship.
4. Dataset IV: Influential Outlier

- In this dataset, there is no meaningful relationship between xxx and yyy, except for a single outlier that drives the high correlation coefficient.
- This demonstrates how even one extreme value can lead to misleading statistical results.



Importance of Visualization in Data Analysis

Anscombe's Quartet underscores the critical role of visualization in data analysis. While summary statistics provide a general overview of the data, they fail to reveal:

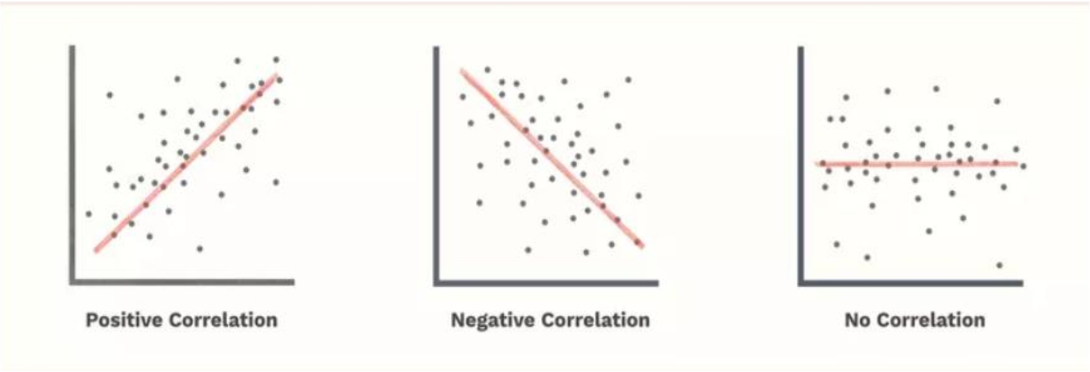
- Non-linear relationships.
- The presence and impact of outliers.
- Patterns or deviations in the dataset.

By graphing data, analysts can better understand the structure, relationships, and issues (like outliers or non-linearity), enabling more accurate and meaningful insights.

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

Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables.

- Value range: −1,1-1, 1−1,1.
    - R=1R = 1R=1: Perfect positive correlation.
    - R=−1R = -1R=−1: Perfect negative correlation.
    - R=0R = 0R=0: No linear correlation. It is calculated as the covariance of two variables divided by the product of their standard deviations.

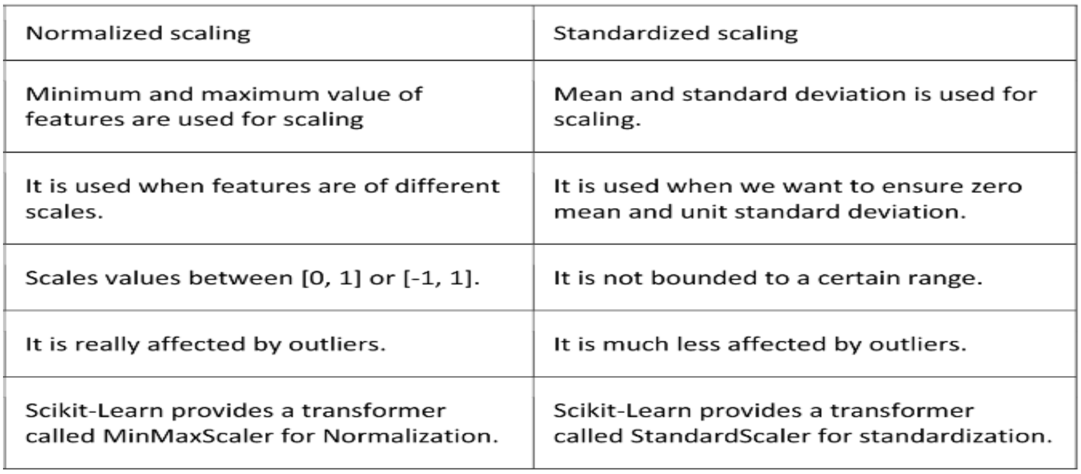


4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling: Adjusts the range of data to ensure that all features contribute equally to the model, especially for algorithms sensitive to magnitude differences (e.g., linear regression, k-NN).

Why Scaling?:

- Prevents features with larger ranges from dominating the model.
- Improves model convergence during training.



Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity, meaning one predictor is an exact linear combination of another. This makes it impossible to estimate unique regression coefficients for the variables.

Key Causes of Infinite VIF

1. Perfect Multicollinearity:
   - Occurs when one variable is fully dependent on another, e.g., X3=2X1+3X2X_3 = 2X_1 + 3X_2X3=2X1+3X2.
2. Duplicate or Highly Correlated Variables:

    - Variables that are identical or strongly correlated provide redundant information. For example, temperature in Celsius and Fahrenheit are duplicates.

3. Dummy Variable Trap:
    - Happens when dummy variables for all categories of a feature are included, creating linear dependency.
        - Example: If you create dummy variables for Red, Blue, Green, their sum equals 1, causing redundancy.
        - Solution: Use drop_first=True to drop one dummy variable.

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

VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is an exact linear combination of another. For example:

- Two variables are duplicates or highly correlated.
- Dummy variable trap occurs (e.g., failing to use drop_first=True). This can be resolved by removing redundant variables.

Example in the Bike Dataset

- A Q-Q plot was used to validate the normality of residuals from the linear regression model predicting cnt.
- The residuals followed the diagonal line closely, indicating they were approximately normally distributed. This confirmed that the assumption of normality was valid and supported the reliability of the model's predictions.