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)
   - Winter season has most users
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

   Using drop_first=True during dummy variable creation is important to avoid multicollinearity issues in regression models, this is due to :

   - **Multicollinearity**: When two or more dummy variables are highly correlated (which happens when one variable can be predicted from the others), it can lead to unstable estimates of coefficients in regression models.
   - **Avoiding Redundancy**: By dropping the first dummy variable (often for the baseline category), you ensure each category is represented independently without redundant information, improving model interpretability and stability.
   - **Interpretation**: Without drop_first=True, each category would have its own dummy variable, leading to perfect collinearity (one variable being a linear combination of others), which violates assumptions in regression analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   a. Both temp and atemp seems to have the highest correlation
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   To validate LR I did :

   a. Check Normality distribution of error by plotting error and checking distribution is normal
   b. Multicollinearity check - checked correlation between variables using heatmap
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   a. Temperature (temp)
   b. Year (yr)
   c. Season (winter season to be specific)

General Subjective Questions

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

Linear regression is used to understand the relationship between two variables, here are the steps with explanation:

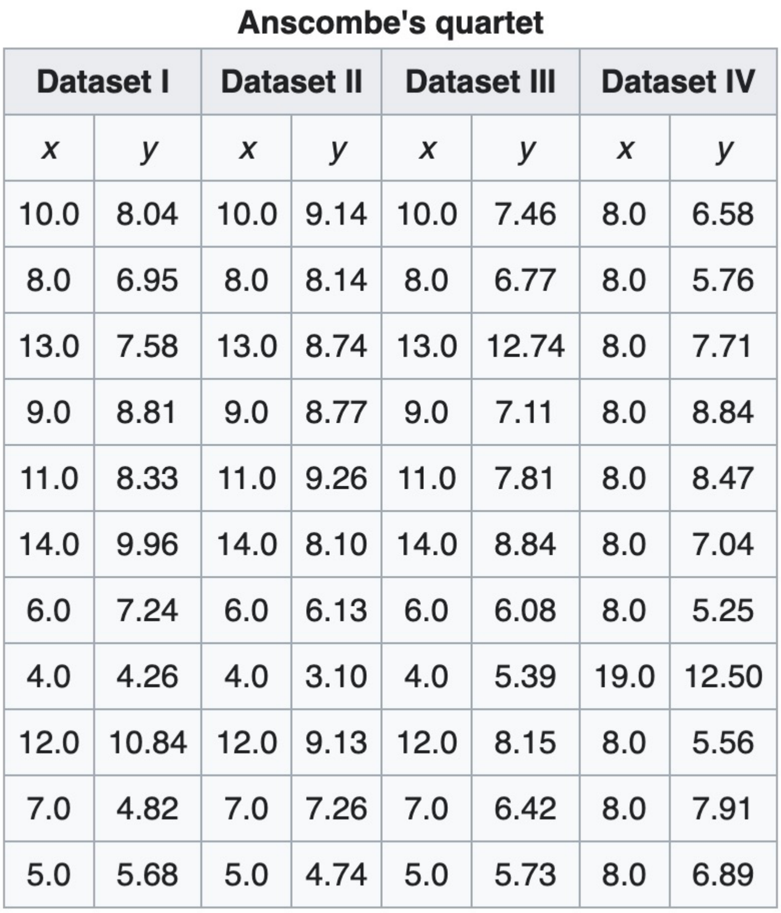
1. **Overview**: Imagine you want to predict a parameter like - person's height based on another parameter like their age. The thing you want to predict is called the "dependent variable"and the thing you use to make the prediction is called the "independent variable"

2. **Steps**:
   - Firstly you collect data, For example, you might have a list of people's ages and their corresponding heights. You can plot this data on a graph, with age on the x-axis (horizontal) and height on the y-axis (vertical).
   - Drawing line: Linear regression tries to draw the best possible straight line through this data. The idea is that this line represents the best guess of the height (y) for any given age (x).
3. **Equation of the Line**: The regression line can be described using a simple equation: y=mx+c
   - y is the height you want to predict.
   - x is the age
   - m is the slope of the line
   - c is the y-intercept, where the line crosses the y-axis
4. **Finding the Best Line**: Linear regression calculates the slope (m) and the y-intercept (b) in such a way that the line is as close as possible to all the data points. This is done using a method called "least squares," which minimizes the sum of the squares of the vertical distances between the data points and the regression line.
5. **Making Predictions**: Once you have your line, you can use it to predict the height for any age by plugging the age into the equation

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

   Anscombe's quartet is a group of four different datasets that have nearly identical statistical properties (mean, variance, correlation, etc.) but look very different when graphed. It was created by the statistician Francis Anscombe to demonstrate the importance of graphing data before analyzing it and to show how relying solely on statistical properties can be misleading.

   For example see this data :



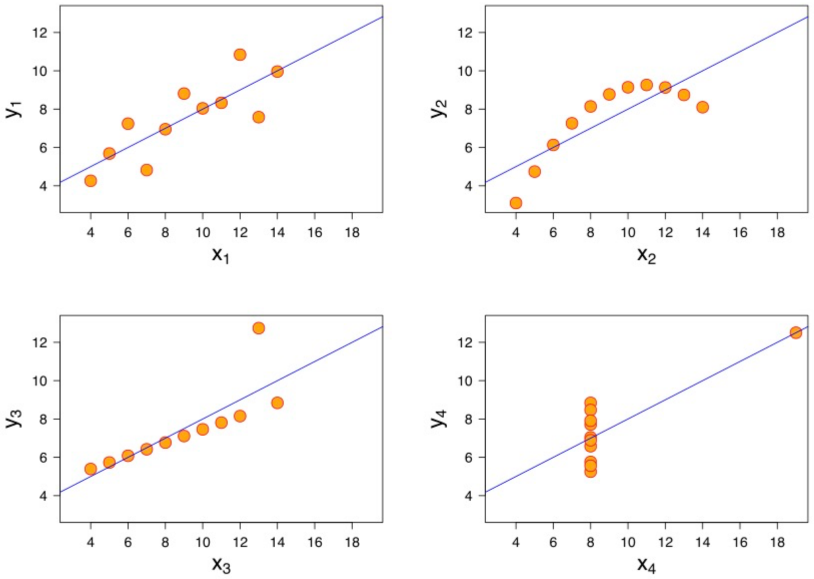
Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset 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

For all four datasets:

Property	**Value**	**Accuracy is**
Mean of *x*	9	exact
Sample variance of *x*: s^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places

But when plotting over graph its :



3. What is Pearson’s R? (3 marks)

Pearson’s R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. Here's a brief explanation:

1. Range: Pearson's R ranges from -1 to 1.
    - 1 indicates a perfect positive linear relationship.
    - -1 indicates a perfect negative linear relationship.
    - 0 indicates no linear relationship.
2. Interpretation:
    - Values closer to 1 or -1 indicate a stronger linear relationship.
    - Values closer to 0 indicate a weaker linear relationship..
3. Use: Pearson’s R is used in statistics to quantify the degree to which two variables are linearly related.
4. Assumptions: It assumes that both variables are continuous and normally distributed, and the relationship between them is linear.

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

Scaling is a process in data preprocessing where the values of a dataset are adjusted to a common scale. It is an important step in many machine learning algorithms because it ensures that all features contribute equally to the model

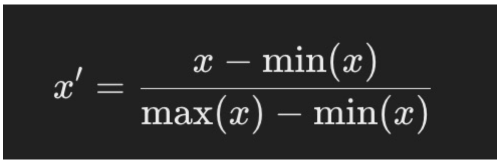
It is performed to :

1. Avoiding Dominance - Features with larger scales can dominate the cost function, scaling prevents this issue by ensuring no single feature dominates due to its scale.

2. Improving Model Performance **-** Many machine learning algorithms, especially those based on distance calculations like k-nearest neighbors (KNN) and support vector machines (SVM), are sensitive to the scale of the data. Scaling helps these algorithms perform better.

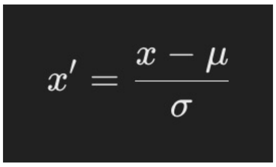
Difference between normalized scaling and standardized scaling is :

- **Normalized Scaling**: Rescales data to a fixed range, typically [0, 1] example – min max
- **Standardized Scaling**: Rescales data to have a mean of 0 and a standard deviation of 1, example Z-score normalization
- **Use Case**: Normalization is best for non-Gaussian data and maintaining relationships, while standardization is ideal for Gaussian data and algorithms assuming normally distributed inputs.
- Formula :
  Min Max :



$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Z-score



$$x' = \frac{x - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 dataset follows a specific theoretical distribution, such as the normal distribution. Here's a brief explanation:

   1. **Purpose**: To compare the quantiles of your dataset with the quantiles of a theoretical distribution.
   2. **How It Works**:
      - Plot the quantiles of your sample data on the y-axis.
      - Plot the corresponding quantiles of the theoretical distribution on the x-axis.
   3. **Interpretation**:

    - **Straight Line**: If the points lie approximately on a straight line, your data likely follows the theoretical distribution.
    - **Deviations**: Significant deviations from the line suggest the data does not follow the theoretical distribution.

A Q-Q plot is important in linear regression for checking the normality of residuals; if the residuals follow a straight line in the Q-Q plot, it indicates they are normally distributed, satisfying a key assumption of linear regression and ensuring reliable model inferences.