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?

Ans - The season, year, and weather situation are significant categorical variables affecting the demand for shared bikes.

The demand is highest during the fall, has grown from 2018 to 2019, and is highest during clear weather.

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

Ans - Using drop_first=True during dummy variable creation helps in avoiding the "dummy variable trap" by preventing multicollinearity in the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Ans - From the pair plot, we can observe that the variables temp and atemp (which represent the actual and "feels like" temperature, respectively)

have a strong positive linear relationship with the target variable cnt.

Among them, both seem to be highly correlated, but temp seems to exhibit a slightly clearer linear trend with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set?

Ans- We validated the assumptions of linear regression by checking the linearity of the model,

assuming the independence of residuals, confirming homoscedasticity through the "Residuals vs. Predicted Values" plot,

and checking the normality of residuals using a histogram.

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

Ans - From the coefficients table we generated earlier, the top three features with the highest absolute coefficients are:

1. temp
2. season_winter
3. season_spring

These features contribute significantly towards explaining the demand for shared bikes.

General Subjective Questions

1. **Linear regression algorithem.**

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation. It's one of the simplest and most widely used algorithms for regression tasks.

Steps involved are:

1. Data Collection
    a. Data set should have two types of data, the dependent variable(target) and one or more independent variables(features).
    b. If we are working with multiple Linear regression then there would be multiple features.
2. Data Preprocessing
    a. This step includes working on the data, making it fit for the model to be trained on.
    b. Handling missing values, scaling, splitting data to train and test
3. Model Representation
    a. In a simple linear regression, model is represented by
    $Y = B_0 + B_1 . X + E$
    This is also called the 'best fit line'

Here

- Y is dependent variable
- X is independent variable
- B_0 is the intercept
- B_1 is the coefficient for the variable.
- E represents the error term

4. Modedl Training
   a. Using the training data the model will be preparing the coefficient for each variable keeping in mind to minimize the SSE (Sum of Squared Error) This is done by using Least square method.
5. Model Evaluation
   a. Test data is used to evaluate the model's performance. MSE is calculated to do a direct comparison of train and test data prediction capacity.
   b. R square can also be calculated to performance matrix.
6. Model Prediction
   a. When the model is trained, the model can be used to predict target variables other that train and test.

Example:

Dataset:

Lets assume we have dataset as:

Hours Studied	**Score**
2	**60**
3	**70**
4	**80**
5	**85**
6	**90**

Model Representation:

We can represent in such a way that – The model will predict the Exam score(Y) based on the number of hours studied.(X)

The model equation would be

$$Y = B_0 + B_1 . X + E$$

Mode Training:

Using the Least square method, we will estimate the value of B_0 and B_1 that minimize the sum of squared error.
After training, we may have data like. $B_0 = 60$ and $B_1 = 5$

Model Evaluation:

We evaluate the model's performance on a test dataset and calculate metrics like MSE or R^2 to assess how well the model fits the data.

Model Prediction

With the trained model, exam scores can be predicted for students who studied different hours.
For example, if a student studied for 7 hours (X = 7) We can predict the exam score as follows.

Y = 60 + 5 . 7
Y = 95

So, the predicted exam score is 95.
Similarly we can fill for any hours of study to get t he projected score.

2. Anscombe's Quartet:

Introduction:

It is used to illustrate a scenario where we have 4 different datasets but it reflects very similar statisticle properties like – mean, variance, R-square, correlations etc.

The differences pops up when we visualize the dataset using charts. So the Objective of Anscombe's Quartet is to not to completely rely on the statisticle values but with visualization.

Below examples can be understood with a specific dataset, its statistical properties and visualization of the same dataset.

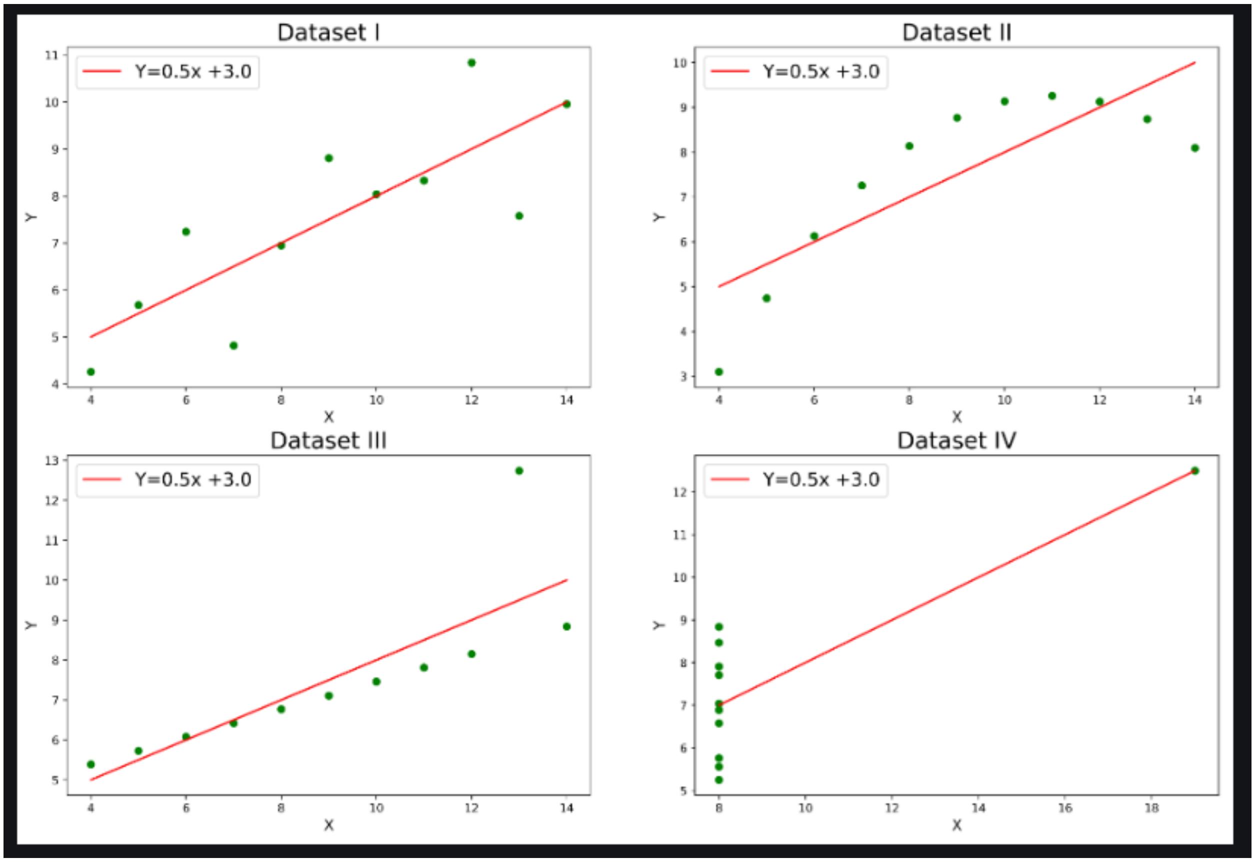
anscombe.csv

DataSet:

Properties for such dataset: I, II, III, IV:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Visualization:



Understanding:

We can clearly infer from the statical properties that the data set must be similar in nature but that is contradicted by looking at the visualization part.(scatter plot)

This example is just to suggest to emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. Pearson's *r*

Introduction:

commonly referred to as the Pearson correlation coefficient, 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:

- r=1 indicates a perfect positive linear relationship.
- r=−1 indicates a perfect negative linear relationship.
- r=0 indicates no linear relationship between the variables.

Understanding & Usability:

It can be used for following insights-

1. Bivariate Relationship: Suitable to get association between to variables at a time. Not fit for multiple variable analysis simultaneously.
2. Assumption of Homoscedasticity: This assumes that the spread of the data point should be roughly constant along the regression line.
3. Linearity: It specifically measures linear relationship, non linear values may not be captured.
4. Non Causality: Itr also assumes that even though the variables are strongly correlated, it does not mean change in one will effect the other.

Calculation:

$$r = Cov(x,y) / Sx . Sy$$

Where

Cov(x,y) is the covariance

Sx and Sy is Standard deviation

Interpretation:

- ***r*** as 1 or -1 as perfect correlation
- ***r*** as 0 shows no correlation
- **$0 < |r| < 1$** shows streangth and direction of the correlation. Colser the r to 1 is stronger correlation

Pearson's r is a valuable tool for understanding linear associations between two continuous variables. However, it's important to remember that it has limitations and should be used in conjunction with other statistical techniques for a comprehensive analysis of relationships in data.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling: Scaling is the process of transforming numerical data to fit within a specific range or scale, usually between 0 and 1 or between -1 and 1 for certain techniques.

Why is it performed?: Scaling is essential for algorithms that are sensitive to the magnitude of variables. It ensures that each variable contributes equally to the computation, preventing variables with larger magnitudes from disproportionately influencing the model's outcome.

Difference:

- **Normalized Scaling**: Often referred to as Min-Max scaling, it transforms features by scaling them to a range of [0, 1].
- **Standardized Scaling**: Also known as Z-score normalization, it transforms features to have a mean of 0 and standard deviation of 1.

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

The Variance Inflation Factor (VIF) measures the inflation in the variances of the parameter estimates due to multicollinearity. The formula for VIF for a particular variable is: where $VIF=1/1-R_{j2}$

is the R-squared value obtained by regressing the j-th predictor on all of the other predictors. When a predictor is a perfect linear combination of other predictors (perfect multicollinearity), ⍰⍰2*Rj*2 becomes 1, making the denominator zero and hence VIF becomes infinite.

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 plot, or Quantile-Quantile plot, is a graphical tool to help assess if a dataset follows a particular theoretical distribution. It plots the quantiles of two distributions against each other.

Use and Importance in Linear Regression: In linear regression, the residuals (errors) are assumed to be normally distributed. A Q-Q plot is used to visually check this assumption. If the data points in the Q-Q plot lie along a straight line (approximately 45-degree angle), it indicates that the residuals are approximately normally distributed. If they deviate significantly, it indicates non-normality, which might suggest model mis-specification or the presence of outliers. Validating the normality of residuals is essential for the reliability of hypothesis tests and confidence intervals in linear regression.