Linear Regression- 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)**

Ans: The categorical variable in the dataset were weather_situation, month, week_day, year,season and working_day which were visualized using a boxplot. These variables had the following effect on our dependent variable 'Count'.

i) From 'Season' Column' we can see that –
- The demand for the bikes is highest during the Fall Season.
- The demand for the bikes is lowest during the Spring Season**.**

ii) For the 'Month' Column we ca see that the demand for bikes in **'**September' is the highest followed by 'October', 'August' and 'June'**.**

iii) The Demand for the bikes was higher during the year 2019 as compared to year 2018.

iv) There is a very slight higher demand on holidays as compared to working days. The demand is almost Same.

v) During the week_days again its observed that the demand for the bikes is higher on Friday, Thursday and Sunday.

vi) The demand for the bikes is higher when the Weather_Situation is Clear, as compared to the days when the Weather_Situation is misty + Cloud or when there is light snow.

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

Ans: Dummy variable creation is used to convert categorical or non-numeric data into numeric data as categorical data cannot be directly used in model building.
It is important to use drop_first = True because it drops the extra column created during dummy variable creation and hence also reduces the correlations between dummy variables.

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

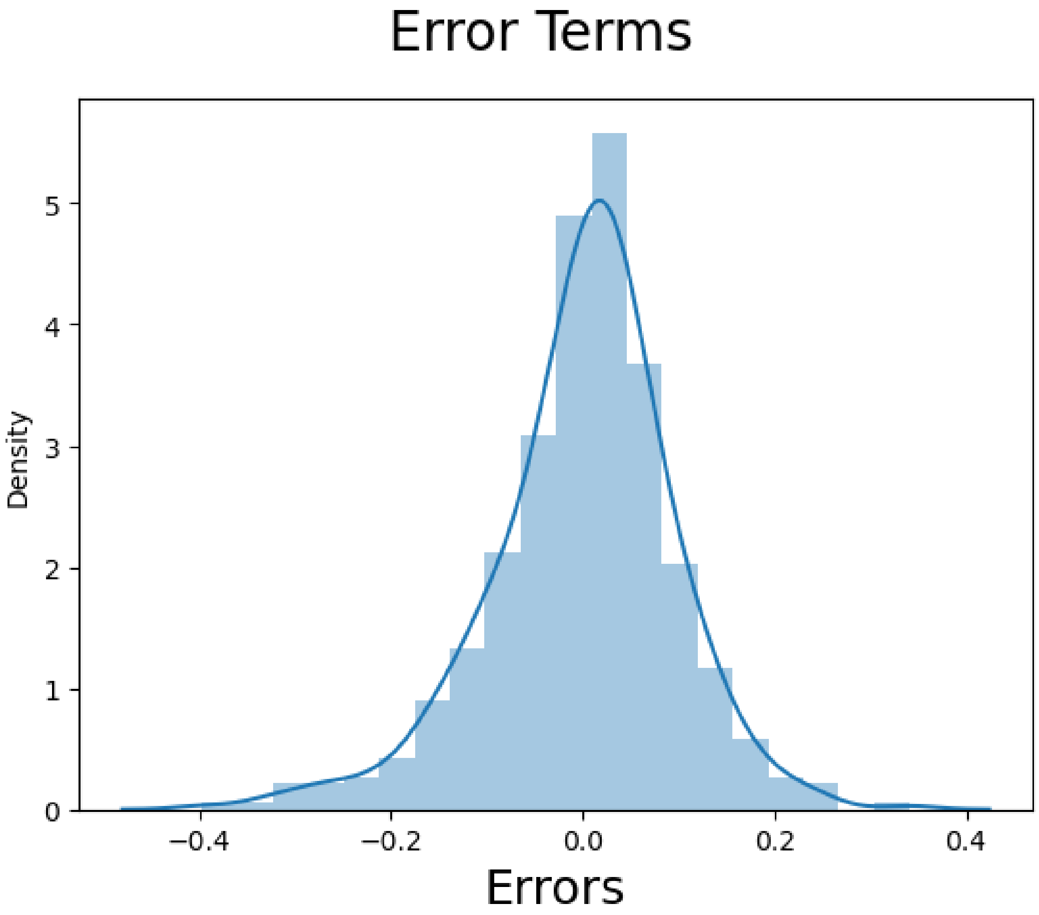
Ans : Looking at the pair plot among variables, Temp and Temp_feel has the highest correlation with the target variable 'Count'.

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

Ans: Assumptions of Linear Regression are-

I. Linear Relationship between X and y.
II. Error terms are normally distributed with mean zero.
III. Error terms are independent of each other.
IV. Error terms have constant variance(homoscedasticity)

After building the model on the training set we can validate the assumption that error terms should normally distributed with mean zero by plotting the distplot of residuals and see if the error terms are normally distributed or not.



The above diagrams shows the residual distribution of the training set. It shows that error terms are normally distributed with mean 0.

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 3 features contributing significantly towards explaining the demand of the shared bikes are-

   i. Temp – Coefficient : 0.447
   ii. Year_2019 – Coefficient: 0.230
   iii. Humidity – Coefficient: -0.239

General Subjective Questions

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

Ans: **Definition :** “Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values”.

- Linear regression is the most basic form of regression analysis.
- Regression is the most commonly used predictive analysis for model.
- Linear regression is based on the popular equation **y = mx + c**.

- ✓ Linear Regression assumes that there should be linear relationship between the dependent variable(y) and independent or predictors variable(x).
- ✓ Regression if performed when dependent variable is of continuous and data type and predictors or dependent variable is of any kind of data type like continuous , categorical/non-numeric,etc.
- ✓ So by calculating the best fit line which describes the relationship between the dependent and independent variables.

Assumptions of Linear Regression:

I. Linear Relationship between X and y.
II. Error terms are normally distributed with mean zero.
III. Error terms are independent of each other.
IV. Error terms have constant variance(homoscedasticity)

Regression is broadly divided into two types-

1) **Simple Linear Regression**: Simple Linear Regression is used when the dependent variable is predicted using only one independent variable.
2) **Multiple Linear Regression**: Multiple Linear Regression is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR is given by:

Yi=β0+β1Xi1+β2Xi2+ β3Xi3...

where, for i=n observations:

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

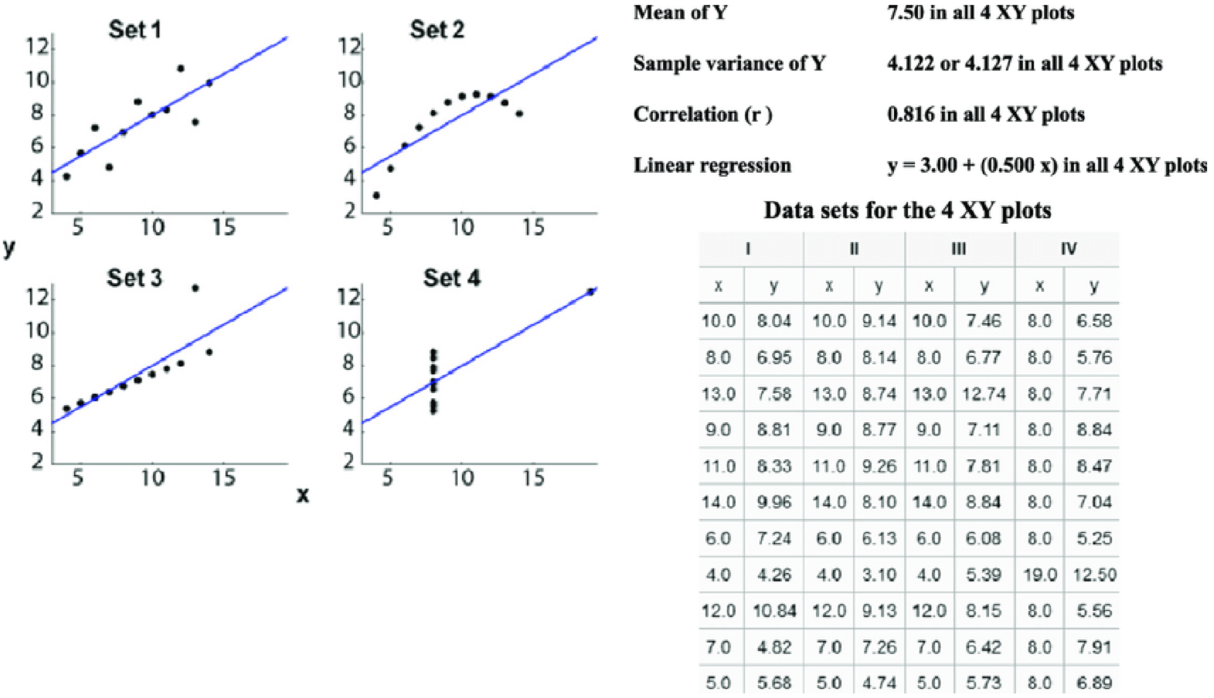
β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

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

Ans: Anscombe’s quartet is a set of four datasets that have nearly identical summary statistics, but very different visual patterns.

- It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
- Each dataset consists of eleven (x,y) pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.
- Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
- Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.
- It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.



Mean of Y	**7.50 in all 4 XY plots**
Sample variance of Y	**4.122 or 4.127 in all 4 XY plots**
Correlation (r)	**0.816 in all 4 XY plots**
Linear regression	**y = 3.00 + (0.500 x) in all 4 XY plots**

Data sets for the 4 XY plots

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

3. What is Pearson’s R? (3 marks)

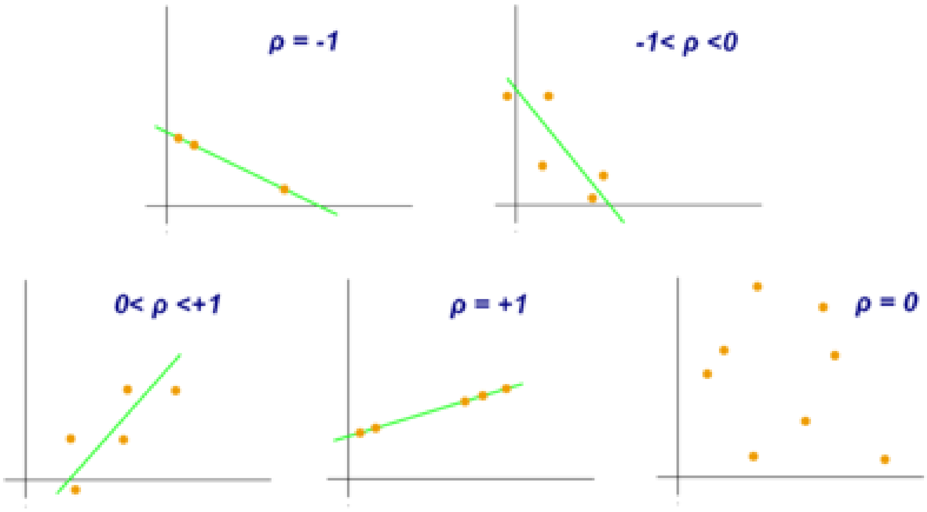
Ans: Pearson's R is a numerical summary of the strength of the linear association between the variables.

- It value ranges between -1 to +1. It shows the linear relationship between two sets of data.
- In simple terms, it tells us can we draw a line graph to represent the data? R = 1 means the data is perfectly linear with a positive slope R = -1 means the data is perfectly linear with a negative slope R = 0 means there is no linear association
- The formula for calculating Pearson’s R is as follows:
- r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2))

- where x and y are the two variables, n is the number of observations, Σ represents the sum of the values, and sqrt represents the square root function.

- Pearson’s R is widely used in statistics, machine learning, and data science to identify the strength and direction of the relationship between two variables.

- It is also used to identify the presence of multicollinearity between independent variables in regression analysis.



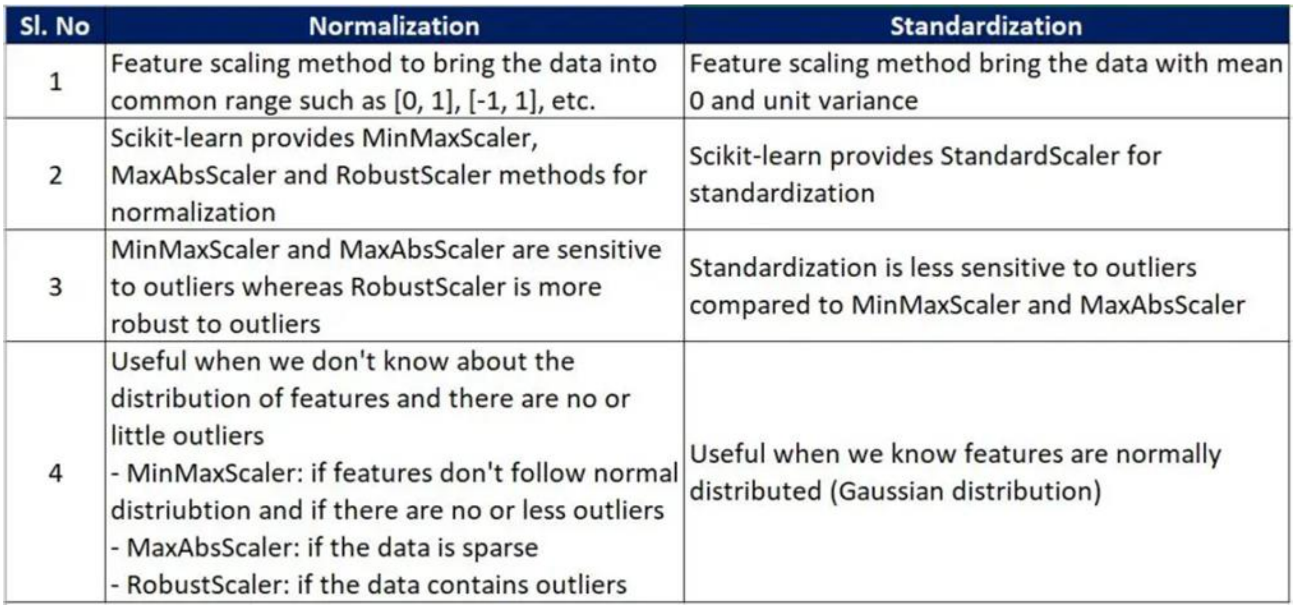
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Ans: Feature scaling is a method used to normalize or standardize the range of independent variables or features of data.

- It is performed during the data preprocessing stage to deal with varying values in the dataset.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.
- The purpose of scaling is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.
- Scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude.
- There are two common techniques for scaling: **Normalization and Standardization**
- Normalization or Min-Max Scaling is used to transform features to be on a similar scale . The new point is calculated as:
- **X_new = (X - X_min)/ (X_max - X_min)**
- This scales the range to [0, 1] or sometimes [-1, 1] . Normalization is useful when there are no outliers as it cannot cope up with them .

- Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation . This is often called as Z-score.
- **X_new = (X - mean)/Std**
- Standardization can be helpful in cases where the data follows a Gaussian distribution . However, this does not have to be necessarily true.

Difference Between Normalization and Standardization:



Sl. No	Normalization	Standardization
1	Feature scaling method to bring the data into common range such as [0, 1], [-1, 1], etc.	Feature scaling method bring the data with mean 0 and unit variance
2	Scikit-learn provides MinMaxScaler, MaxAbsScaler and RobustScaler methods for normalization	Scikit-learn provides StandardScaler for standardization
3	MinMaxScaler and MaxAbsScaler are sensitive to outliers whereas RobustScaler is more robust to outliers	Standardization is less sensitive to outliers compared to MinMaxScaler and MaxAbsScaler
4	Useful when we don't know about the distribution of features and there are no or little outliers - MinMaxScaler: if features don't follow normal distriubtion and if there are no or less outliers - MaxAbsScaler: if the data is sparse - RobustScaler: if the data contains outliers	Useful when we know features are normally distributed (Gaussian distribution)

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

Ans: The VIF or Variance Inflation Factor is a measure of the degree of multicollinearity in a set of multiple regression variables .

- It is used to identify the correlation between independent variables in a regression model .
- The VIF value ranges from 1 to infinity, where a value of 1 indicates no correlation between the variables, and a value greater than 1 indicates the presence of multicollinearity .

- The VIF becomes infinite when there is a perfect correlation between two independent variables .
- In other words, when two or more independent variables are highly correlated, the VIF value becomes infinite .

- This is because the VIF is calculated as the ratio of the variance of the estimated coefficient in the presence of multicollinearity to the variance of the estimated coefficient in the absence of multicollinearity .
- When there is perfect correlation between two independent variables, the variance of the estimated coefficient in the presence of multicollinearity becomes zero, leading to an infinite VIF value .

- To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity .
- This will help in reducing the correlation between the independent variables and will result in a lower VIF value .

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

Ans: A Q-Q plot or Quantile-Quantile plot is a graphical tool used to compare the distribution of a sample dataset with a theoretical distribution such as a normal distribution .

- It is a scatter plot that compares the quantiles of two datasets and helps to determine if they come from populations with a common distribution .
- The Q-Q plot is a useful tool in linear regression to check the normality assumption of the residuals .

- In linear regression, the Q-Q plot is used to check if the residuals of a model follow a normal distribution .
- If the residuals are normally distributed, the points on the Q-Q plot will roughly form a straight diagonal line .
- If the residuals are not normally distributed, the points on the Q-Q plot will deviate from the straight diagonal line .
- The Q-Q plot is a useful tool to detect non-normality in the residuals of a model .

- The importance of the Q-Q plot in linear regression is that it helps to identify the presence of non-normality in the residuals of a model .
- If the residuals are not normally distributed, the model may not be a good fit for the data, and the results may be unreliable .
- The Q-Q plot is a useful tool to identify the presence of non-normality in the residuals and to determine if the model needs to be improved or not.