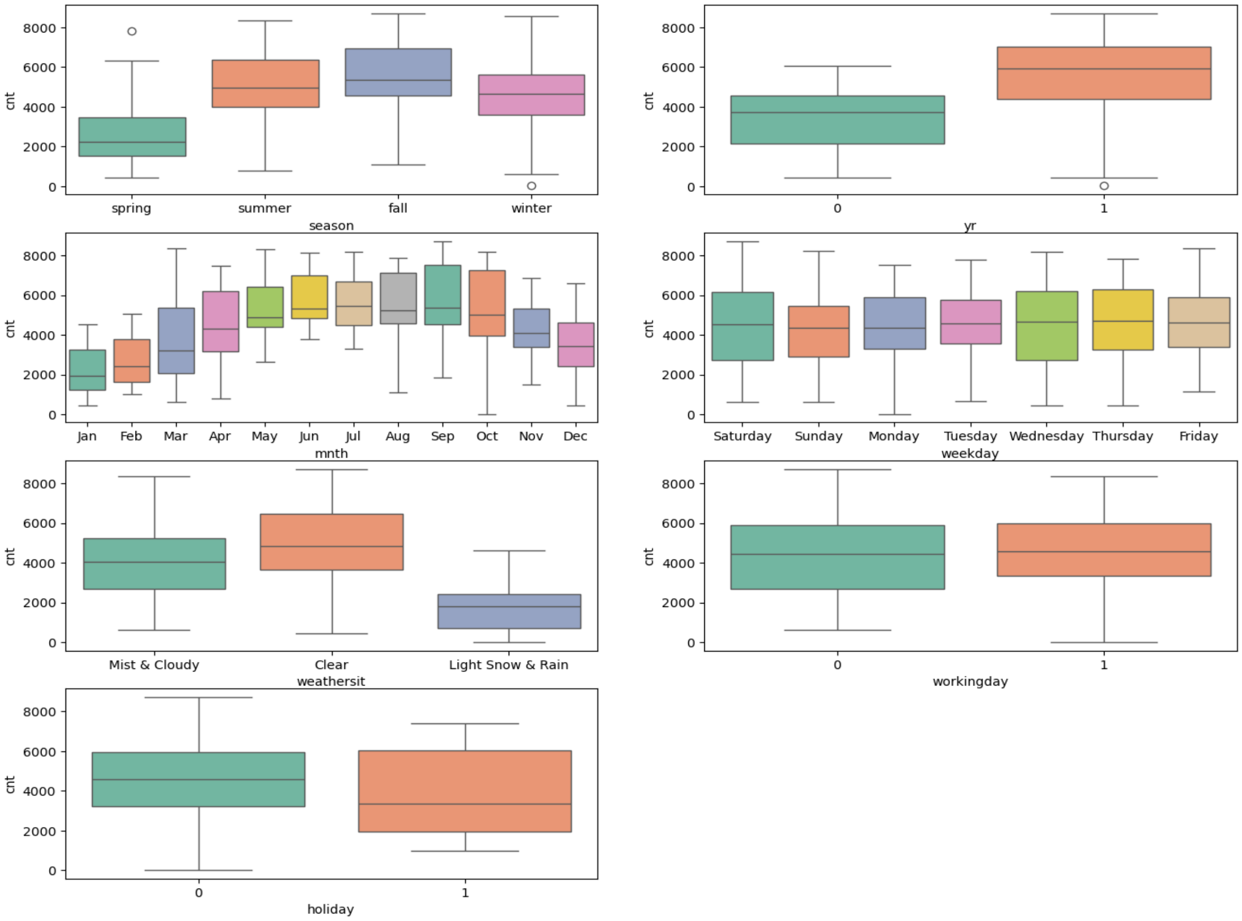
Assignment-based Subjective Questions

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 categorial features are there in the dataset Season, Year, Month, Weekday, Weathersit, Workingday and Holiday.

I have used Bar Plot and Box Plot to see their inference on the dependant variable on cnt column.

- ✓ Season:- Fall has the highest total count of rental bikes compared to other seasons like Spring, Summer, and Winter. The demand was high during this season.

- ✓ Yr: The total count of rental bikes in 2018 was 12 lakh. After one year, in 2019, the count increased to 20 lakh, indicating a 60% spike in the number of users it is good growth in terms of business.

✓ Mnth: There is a high demand of shared bikes from January till September Starting from October the demand is getting decreased till end of the December month.

✓ Weekday: The bike demand looks same throughout the entire week.

✓ Weathersit: when the wheatersit is Clear there are more number of users are using the shared Bike for their work whereas Mist and Cloudy weather having some users and very few users are using the shared bike with the Light Snow & Rain weather.

✓ Working Day: There are more numbers of users are using the shared bike for their office commute and looks count is same for the 2018 and 2019 years.

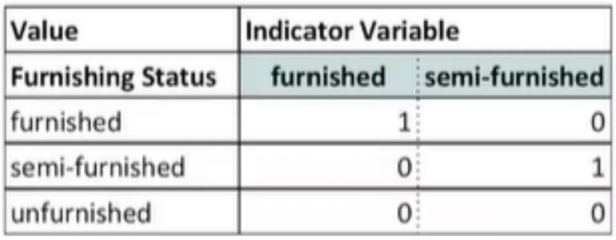
✓ Holiday: There are very few people using the shared bike on holiday.

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

Ans:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.



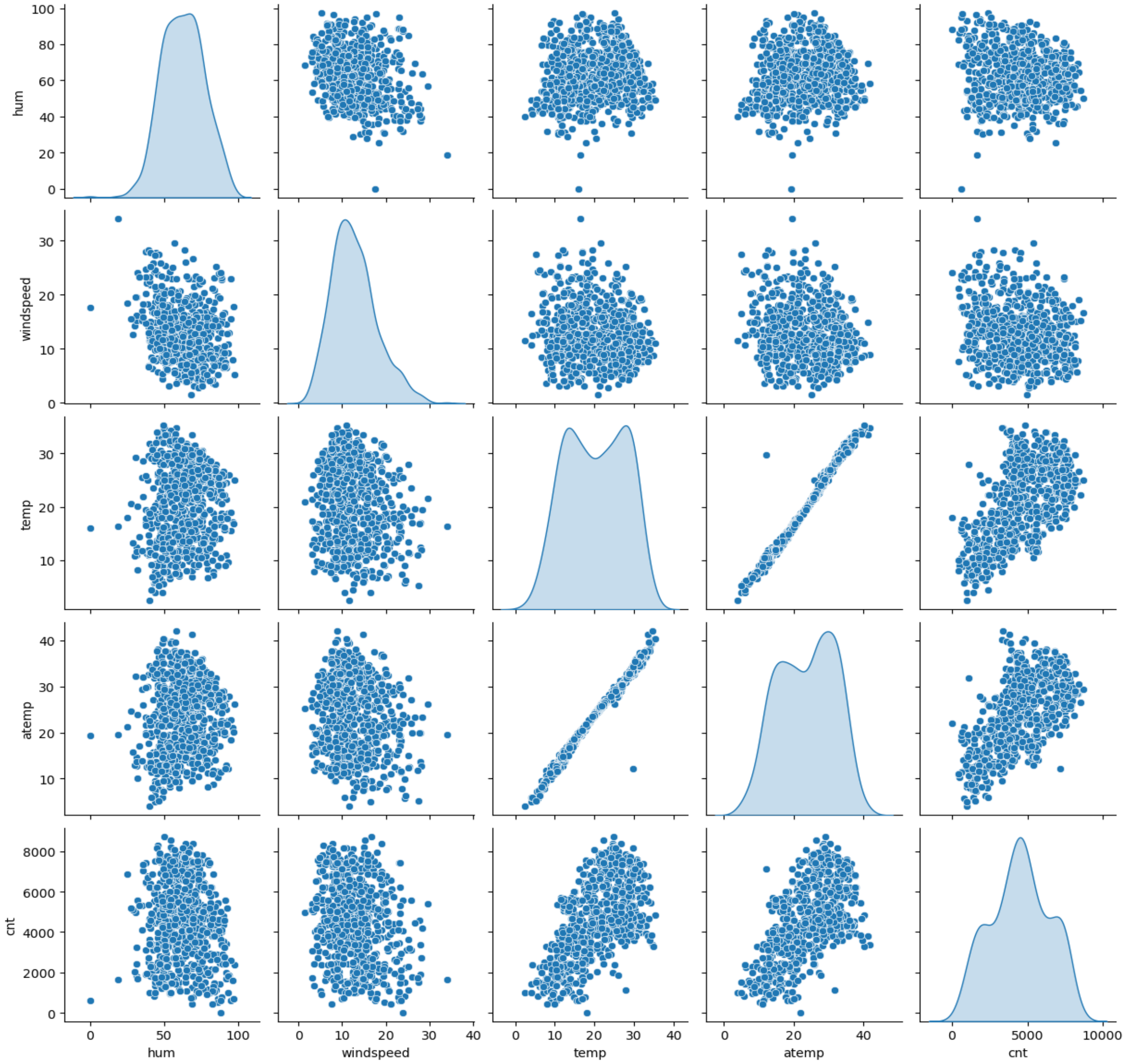
Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the 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:

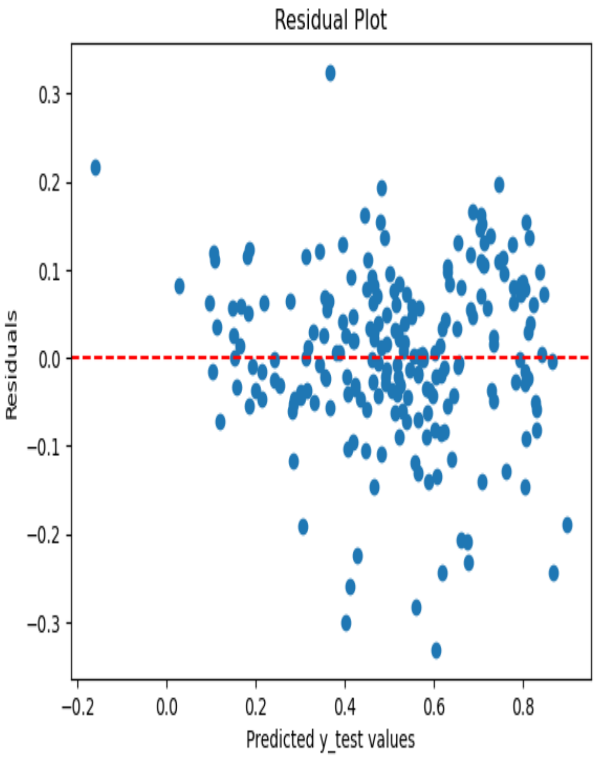
‘temp’ and ‘atemp’ variable has the highest correlation with the target variable (cnt)

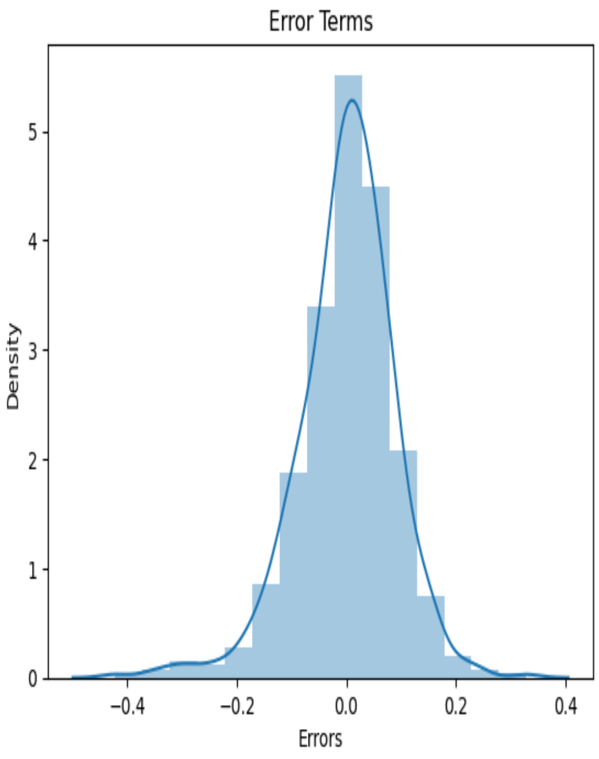


.

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

Ans:





- **Linear Relationship**: The relationship between the independent and dependant variable should be linear.

- **Independence of Residual:** There should not be auto co-relation.

- **Homoscedasticity:** There should be no visible pattern in residual values. Looks same after plotting predicted vs residual analysis.

- **Normality of Residuals**: The residuals should be normally distributed. Attached the image as above.

- **No Multicollinearity:** Independent variables should not be too highly correlated with each other. I have used Variance Inflation Factor (VIF) to check the multicollinearity and drop the column with having high VIF.

- **No endogeneity:** There is no relationship between the errors and the independent variables.

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

Ans:

Below are the 3 features are contributing significantly towards the demand for shared bikes.

- temp
- atemp
- year

General Subjective Questions

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

Ans:

Linear Regression is the supervised Machine Learning model used for predicting a continuous output variable based on one or more predictor variables.

There are two types of Regression.

- Simple:- Only one independent variable and one dependant feature.
- Multiple:- More than one independent variable and one dependant feature.

Simple Linear Regression:

This involves only one independent variable and one dependent variable. The equation for simple linear regression is: $\boldsymbol{y= β0 + β1\, X}$

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression:

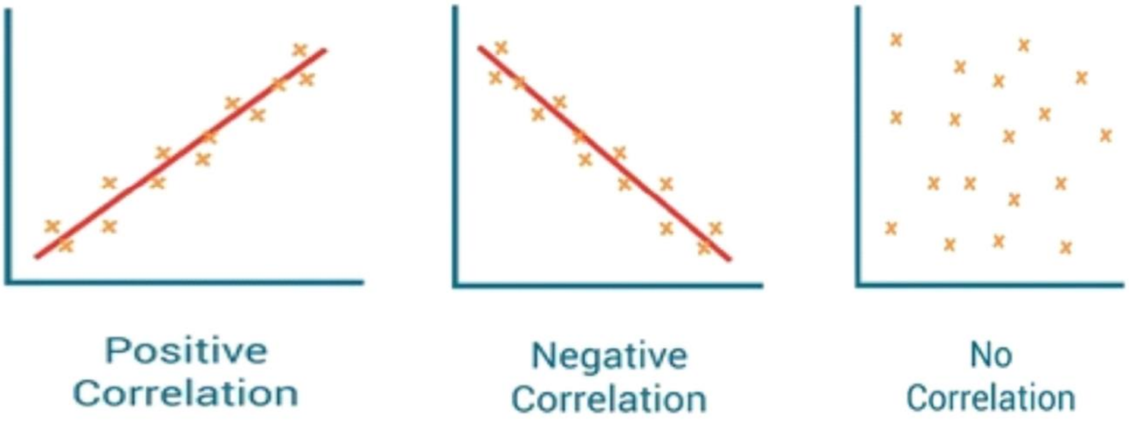
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is: $\boldsymbol{y= β0 + β1X1+ β2X2+\text{.........}\ βn\, Xn}$

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- B0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

There are 3 types of relationships they are described as below.

- **Positive Correlation:** If one variables increases tends to increase the value of other variable.
- **Negative Correlation:** If one variables increases tends to decrease the value of other variable.
- **No Correlation:** If one variable increases, the other variable may or may not increase. It can either increase or decrease.



Assumptions of linear regression include:

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence:** The observations are independent of each other.
3. **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
4. **Normality:** The errors follow a normal distribution.
5. **No multicollinearity:** The independent variables are not highly correlated with each other.
6. **No endogeneity:** There is no relationship between the errors and the independent variables.
7. **Autocorrelation:** There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Auto correlation

R-Squared:

R-square(R^2) is also known as the *coefficient of determination*, It is the proportion of variation in Y explained by the independent variables X. It is the measure of goodness of fit of the model.

Higher the R^2, the more variation is explained by your input variable and hence better is your model

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 → coefficient of determination

RSS → sum of squares of residuals

TSS → Total sum of Squares.

Cost Function For Linear Regression

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

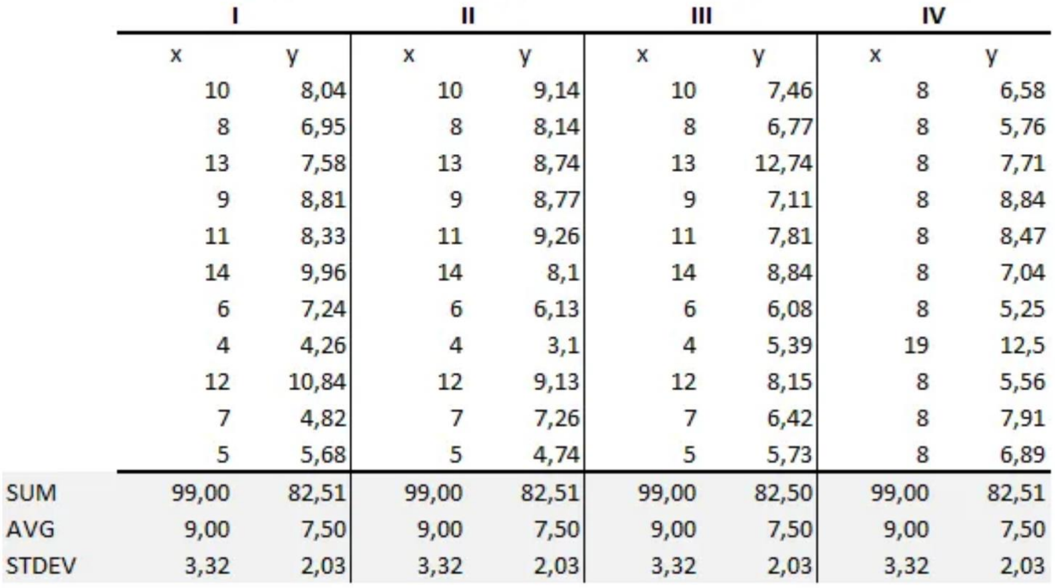
n → number of data points

yi → actual value

$\hat{y}$ → predicted value.

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



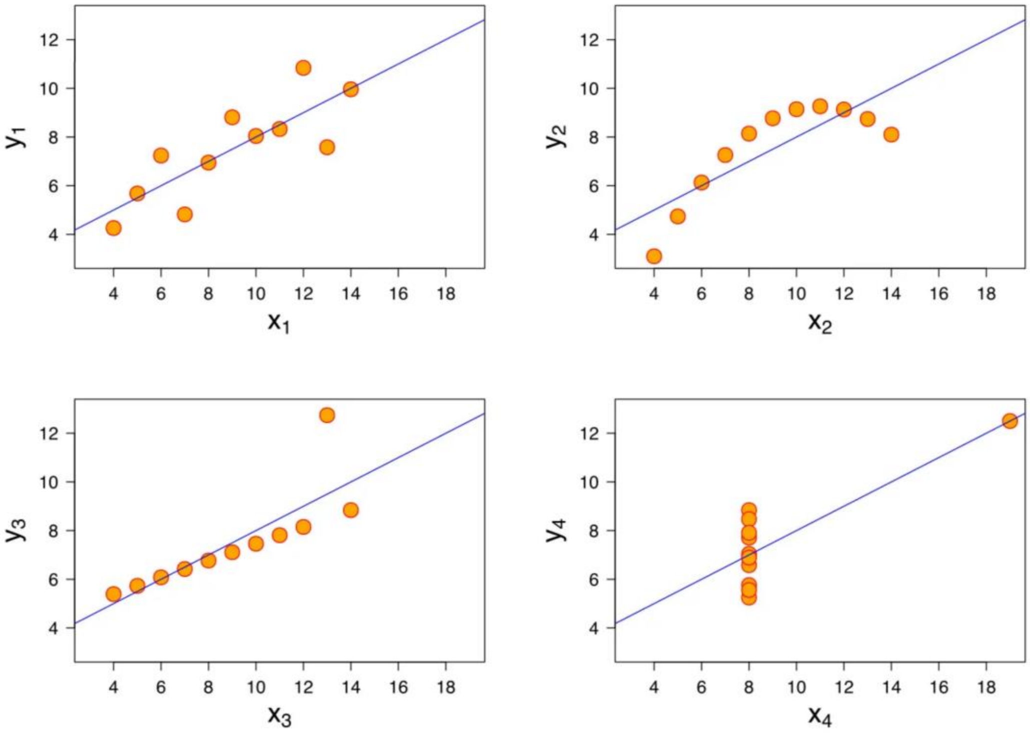
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

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

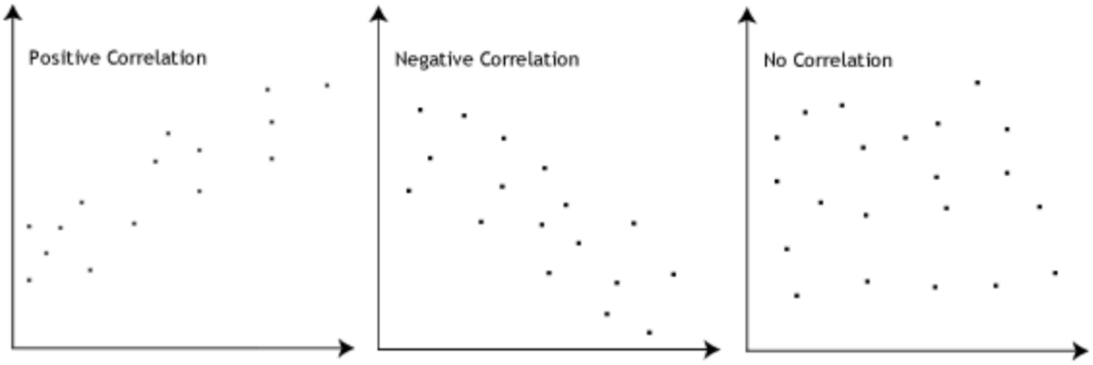
Ans:

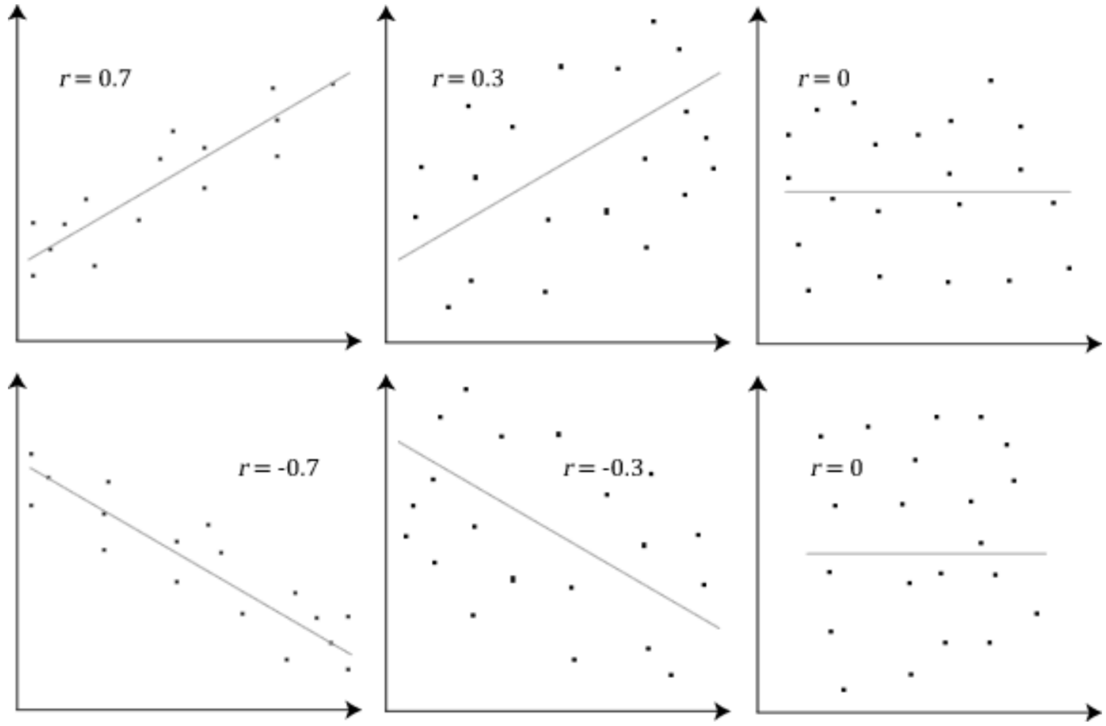
Pearson correlation coefficient, is a measure of the strength of a linear association between two variables and is denoted by *r*

The mathematical representation is below.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:





Their values lies between -1 to 1 they are described below.

- R=1 → perfect positive corelation
- R is between 0 to 1 → positive corelation
- R is 0 → No corelation
- R is -1 to 0 → negative corelation
- R is -1 → perfect negative corelation

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

Data scaling is the process of transforming the values of the features of a dataset till they are within a specific range, e.g. 0 to 1 or -1 to 1. This is to ensure that no single feature dominates the distance calculations in an algorithm, and can help to improve the performance of the algorithm

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7	$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$	$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$

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

Ans:

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i{}^2}$$

When we calculate the VIF for one independent variable using all the other independent variables, if the R^2 value we get equal to 1 then VIF will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables. It denotes perfect correlation in variables

A rule of thumb for interpreting the Variance Inflation Factor:

- VIF is 1 = not correlated.
- VIF Between 1 and 5 = moderately correlated.
- VIF Greater than 5 = highly correlated.

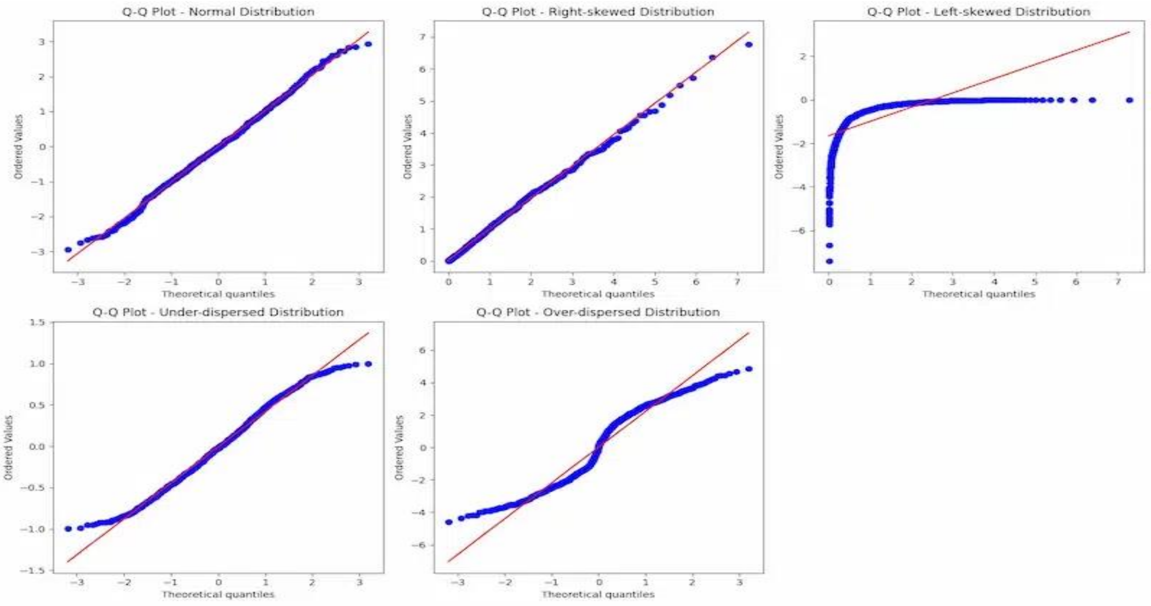
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

Below is the different distribution using QQ plot


Q-Q Plot - Normal Distribution
Q-Q Plot - Right-skewed Distribution
Q-Q Plot - Left-skewed Distribution
Q-Q Plot - Under-dispersed Distribution
Q-Q Plot - Over-dispersed Distribution
Ordered Values
Theoretical quantiles