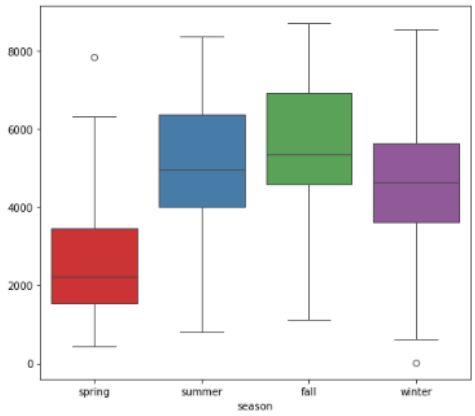
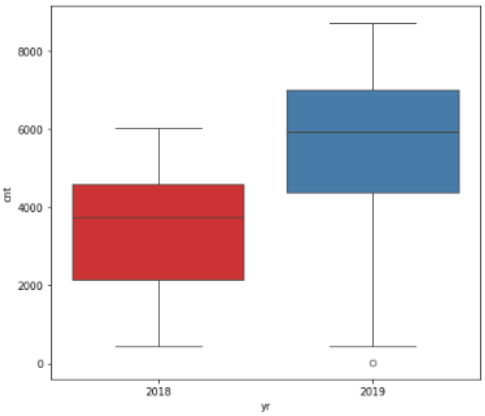
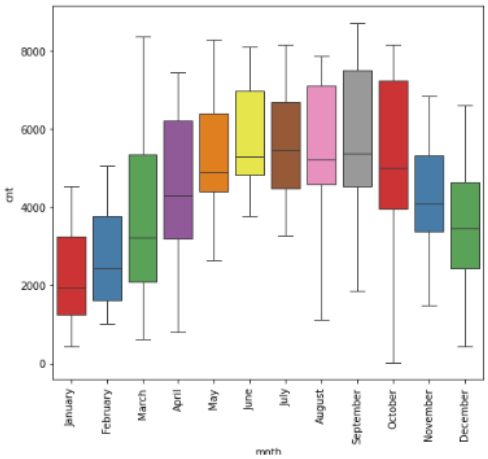
Assignment-based Subjective Questions

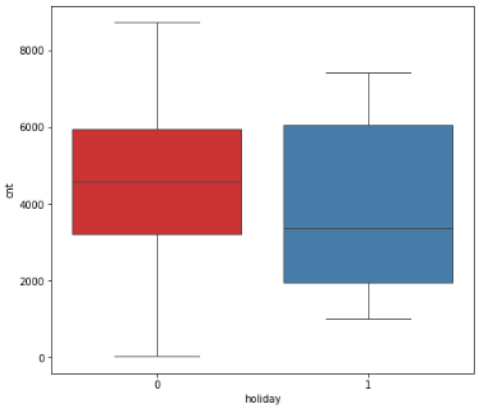
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

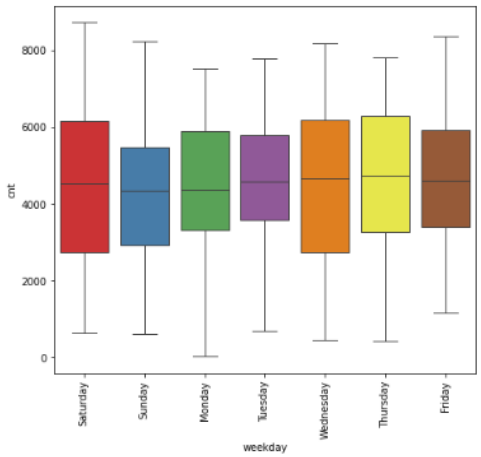
Ans: There're total of **7 categorical variables** within the dataset are namely **season**, **yr**, **mnth**, **holiday**, **weekday**, **workingday** and **weathersit.** The Analysis is summarised below

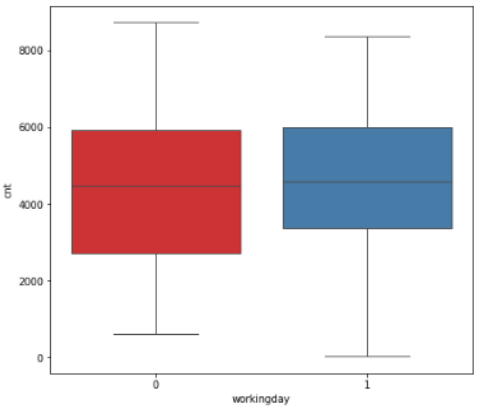


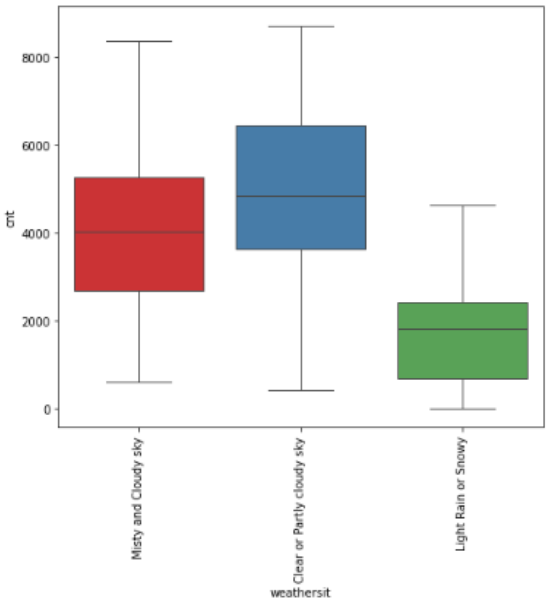












From the analysis we can see that

- ✓ **Maximum demand of bikes is in the fall and minimum in spring season**.
- ✓ **Bikes demand is more in the year 2019 as compared to 2018.**
- ✓ **Bikes display most demand during the months of May till October with highest in September.**
- ✓ **Bikes are slightly more in demand on holidays than on working days.**
- ✓ **Bikes are majorly more in demand on Saturday, Thursday, Wednesday showing that they have been used for mixed purposes namely office and travel.**
- ✓ **Bike demand is almost same irrespective of working day or not.**
- ✓ **Bikes are demanded mostly when the sky is clear or partly cloudy.**

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)
Ans: drop_first=True drops the first column during dummy variable creation in order to prevent multicollinearity and ensure better model performance. It reduces the correlations created among dummy variables.

This way you maintain independence among the dummy variables, which helps prevent overfitting and leads to a better generalization of the model on unseen data. Models often perform better when multicollinearity issues are minimized. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

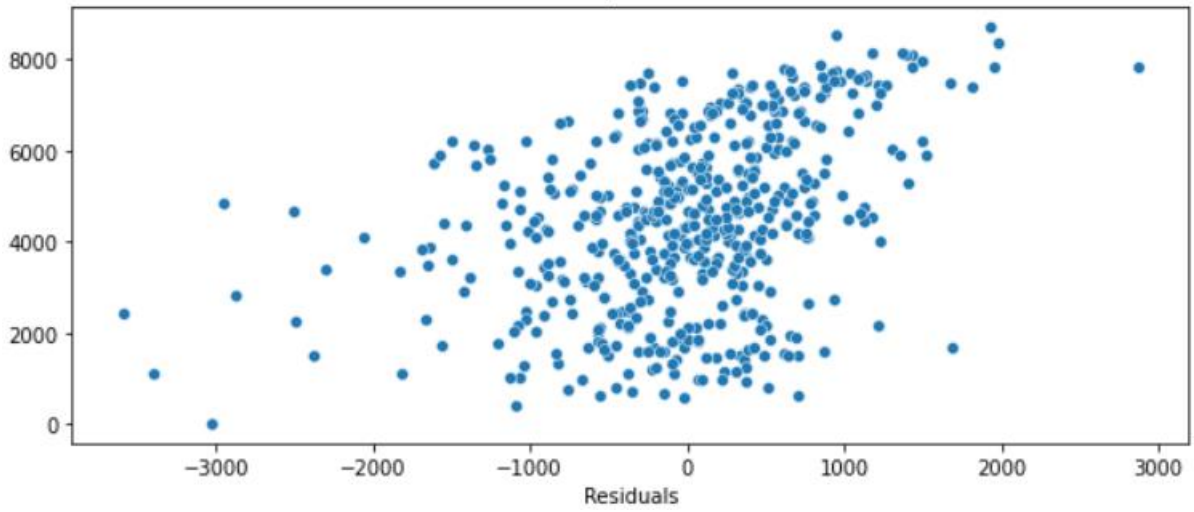
For instance, suppose you have a categorical variable "**Color**" with three categories: **Red, Green, and Blue**. If you create dummy variables without dropping the first category, you'd generate two dummy variables: "**Green**" and "**Blue**." However, having both "**Green**" and "**Blue**" flags implies the absence of "**Red**." By dropping the first category ("**Red**") when creating dummy variables, you ensure independence among the variables, as the absence of both "**Green**" and "**Blue**" inherently implies "Red.

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**.**

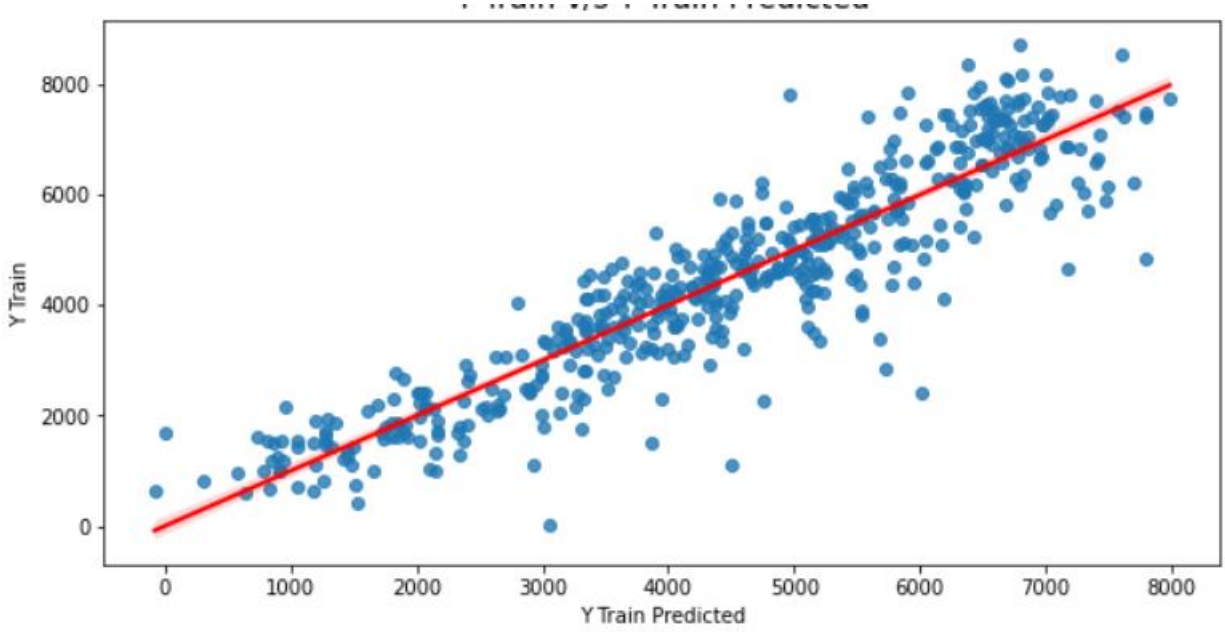
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: After building the model on the training set, I validated its assumptions using various diagnostic techniques

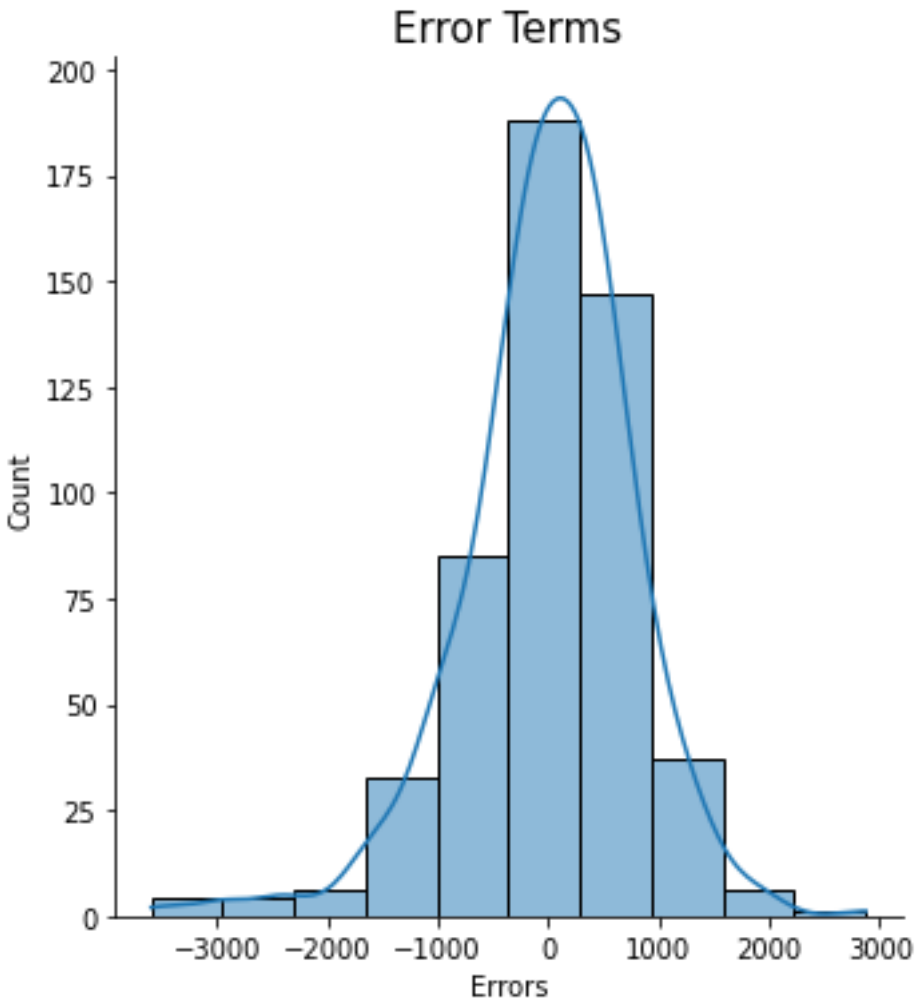
- ✓ **Residual Analysis:** Based upon the analysis of the graph plotted for the **residuals** (the differences between predicted and actual values) against the predicted values, I saw that they were randomly scattered and there was specific trend or pattern.



- ✓ **Homoscedasticity:** There was almost constant variance of predictions thereby validating the concept of Homoscedasticity



- ✓ **Normal Distribution of Errors:** The mean of residuals is **0.00000000000271602527** which is very close to **0.** Thus, the error terms were normally distributed.



- ✓ **Analysis of Variance Inflation Factor (VIF):** The VIF of the final independent variables was less than 5 and was under limit. Thus, we concluded that there was no correlation between them.

	feature	vif
0	hum	1.87
1	Misty and Cloudy sky	1.55
2	temp	1.27
3	winter	1.25
4	Light Rain or Snowy	1.24
5	summer	1.19
6	windspeed	1.18
7	September	1.11
8	yr	1.03
9	Sunday	1.01

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:
My model equation is mentioned below
cnt = 4491.3033 + (998.5631 * **yr**) + *(1125.0534 *** **temp**) - *(213.0869 *** **hum**) - *(276.1097 *** **windspeed**) + *(343.4570 *** **summer**) + *(526.3388 *** **winter**) + *(239.1522 *** **September**) - *(131.7794 *** **Sunday**) - *(349.2905 *** **Light Rain or Snowy**) - *(214.7093 *** **Misty and Cloudy sky**)

based upon it the top 3 features contributing to explain the demand of the shared bikes are

1. **temp**: it has the highest coefficient values of ***1125.0534*** thus a unit increase in temp will increase the demand by a factor of ***1125.0534***.
2. **yr:** Yearly demand will increase by a factor of **998.5631**
3. **winter:** Winter Season will contribute significantly as demand will increase by a factor of ***526.3388*.**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Ans: Linear regression is a fundamental statistical and supervised machine learning technique used for modelling the relationship between a dependent variable (target) and one or more independent variables (features). It's of two types

- ✓ **Simple Linear Regression**: There's **one target variable** and **only one independent variable**
- ✓ **Multiple Linear Regression**: There's **one target variable** and **more than one independent variable**

It assumes a linear relationship between the predictor variables and the target variable. The goal of linear regression is to find the best-fitting linear equation that predicts the target variable based on the input features.

The linear regression algorithm can be expressed mathematically in a simple form as:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

The linear regression algorithm can be expressed mathematically in a simple form as:
Here

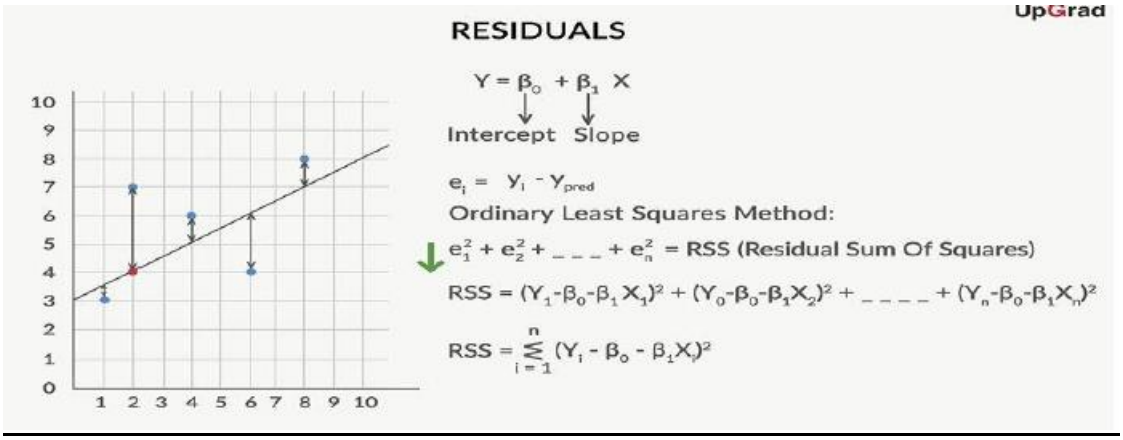
y:	**Dependent (target) variable**
x1, x2,.. xn:	**Independent variables**
b0:	**Intercept**
b1, …, bn :	**Coefficients associated with each independent variable**

Assumptions of Linear Regression:

- ✓ **Linearity**: The relationship between the predictors and the target variable should be linear.
- ✓ **Independence of Errors**: Residuals (errors) should be independent of each other.
- ✓ **Homoscedasticity**: Residuals should have constant variance (homogeneity of variance).
- ✓ **Normality of Residuals**: The residuals should be normally distributed.
- ✓ **No or Minimal Multicollinearity**: The predictors should be independent of each other.

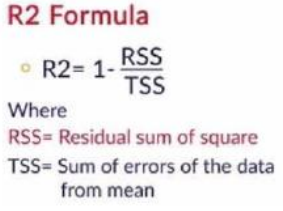
Best Fit Line:

The best-fit line is found by **minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot**. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



Strength of Linear Regression: The strength of the linear regression model can be assessed using 2 metrics:

i. **R^2 or Coefficient of Determination:** Accuracy of the model can be determined by is R2 statistics.
   - ✓ R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1.
   - ✓ In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes.
   - ✓ Overall, the higher the R-squared, the better the model fits your data.
   - ✓ Mathematically, it is represented as: **$R^2 = 1 - (RSS / TSS)$**



ii. **Residual Standard Error (RSE)**

- ✓ **RSS (Residual Sum of Squares):** In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

- ✓ **TSS (Total sum of squares):** It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Industry Wide Examples of Linear Regression:

i. **Sports Analytics:**
   - ✓ **Player Performance Prediction**: Analysing player statistics to predict performance metrics such as points scored, goals, or batting averages.
   - ✓ **Team Performance Prediction**: Predicting team success based on factors like player stats, coaching strategies, and historical performance.

ii. **Real Estate:**
   - ✓ **Housing Price Prediction**: Estimating property prices using features like location, square footage, number of bedrooms, etc.
   - ✓ **Rental Price Prediction**: Predicting rental prices based on property characteristics and market trends.

iii. **Marketing and Sales:**
   - ✓ **Sales Forecasting**: Predicting future sales based on advertising expenditure, market size, pricing, etc.
   - ✓ **Customer Lifetime Value**: Estimating the potential value of customers over their lifetime based on historical purchase behaviour

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

Ans: Anscombe's quartet is used to illustrate the importance of Exploratory Data Analysis(EDA) 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.

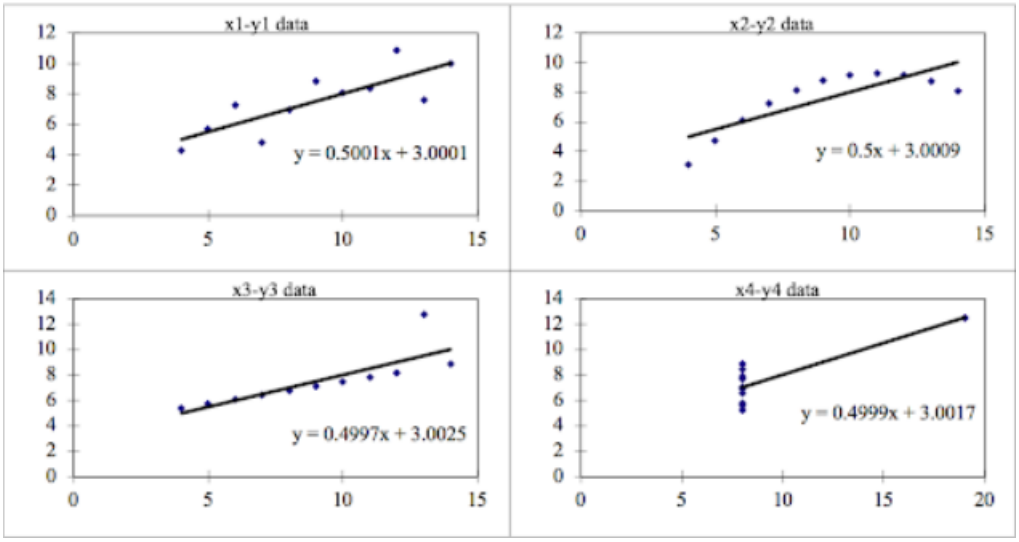
Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines

but having different representations when we scatter plot on graph. The datasets were 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.

The four datasets that make up Anscombe's quartet each include 11 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 Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

i. **Data Set 1:** fits the linear regression model pretty well.
ii. **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
iii. **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
iv. **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Thus, we see that Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Ans: Pearson correlation coefficient, also known as **Pearson R statistical test**, measures the strength between the different variables and their relationships.

Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analysing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

The Pearson's correlation coefficient varies between **-1** and **+1** where:

- ✓ **r = 1** means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- ✓ **r = -1** means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- ✓ **r = 0** means there is no linear association
- ✓ **r > 0 < 5** means there is a weak association
- ✓ **r > 5 < 8** means there is a moderate association
- ✓ **r > 8** means there is a strong association

<u>When to use the Pearson correlation coefficient:</u>

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

- ✓ **Both variables are quantitative**: You will need to use a different method if either of the variables is qualitative.

- ✓ **The variables are normally distributed**: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- ✓ **The data have no outliers**: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- ✓ **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Pearson's correlation coefficient is calculated using the following formula:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where

r = Pearson Coefficient

n = number of pairs of the stock

∑xy = sum of products of the paired stocks

∑x = sum of the x scores

∑y= sum of the y scores

∑x2 = sum of the squared x scores

∑y2 = sum of the squared y scores

<u>Key points about Pearson's correlation coefficient:</u>

- ✓ **Strength of Relationship**: The absolute value of r indicates the strength of the relationship between the variables. Larger absolute values (closer to 1) signify a stronger linear relationship.
- ✓ **Direction of Relationship:** The sign of r (+ or -) indicates the direction of the relationship. A positive value means both variables move in the same direction, while a negative value means they move in opposite directions.
- ✓ **Assumptions:** Pearson's correlation coefficient assumes a linear relationship between the variables and that the data is normally distributed.
- ✓ **Limitations:** It measures only linear relationships and may not capture non-linear relationships. Additionally, it can be sensitive to outliers.

<u>Uses:</u>

Pearson's correlation coefficient is widely used in various fields such as statistics, psychology, economics, and social sciences to assess the strength and direction of relationships between

continuous variables, helping researchers and analysts understand how variables are related numerically.

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

Ans: Scaling in the context of data preprocessing refers to the process of transforming data to a standard scale, often performed on numerical features. It's done to ensure that different features are on a similar scale or magnitude, which can be beneficial for various machine learning algorithms.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So, we need to scale features because of two reasons

i. Ease of interpretation
ii. Faster convergence for gradient descent methods

Difference between Standardized and Normalized Scaling can be done using 2 techniques

i. **Standardized Scaling:** The variables are scaled in such a way that their **mean is 0 and standard deviation is 1.**

$$x = \frac{x - mean(x)}{sd(x)}$$

ii. **Normalized Scaling:** The variables are scaled in such a way that **all the values lie between 0 and 1 using the maximum and the minimum values in the data**.

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

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

Ans: The Variance Inflation Factor (VIF) basically helps explaining the relationship of one independent variable with all the other independent variables. The formula of VIF is given below:

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

VIF measures the severity of multicollinearity in a regression analysis. High VIF values indicate a high correlation between a predictor variable and other predictor variables in the model. VIF quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity.

Sometimes VIF value is calculated as infinite because of perfect multicollinearity among the predictor variables.

Perfect multicollinearity occurs when one predictor variable in a regression model can be exactly predicted from a linear combination of other predictor variables.

Mathematically, it means that there is a perfect linear relationship among the predictors, leading to a situation where one or more variables can be expressed as a perfect linear function of other variables.

In the context of VIF calculation:

- ✓ When perfect multicollinearity exists, one of the variables becomes redundant in the model because it can be precisely predicted from the other variables.
- ✓ In such cases, when calculating the VIF for a variable that is perfectly correlated with other variables, the denominator of the VIF formula becomes very close to zero or zero itself.
- ✓ Division by zero (or near-zero) results in the VIF being computed as infinite.

When encountering infinite VIF values, it's crucial to address the multicollinearity issue by Identifying and removing one of the correlated variables from the model to resolve the perfect multicollinearity problem.

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

Ans: Q-Q plots are also known as **Quantile-Quantile** plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Working of a Q-Q plot:

i. Collect the data for plotting the quantile-quantile plot.
ii. Sort the data in ascending or descending order.
iii. Draw a normal distribution curve.
iv. Find the z-value (cut-off point) for each segment.
v. Plot the dataset values against the normalizing cut-off points.

The use and importance of Q-Q plots in linear regression:

- ✓ **Assumption Checking:** Q-Q plots are useful for checking the assumption of normality in the residuals of a linear regression model.
- ✓ **Normality Assessment:** In linear regression, it is assumed that the residuals are normally distributed with a mean of zero. If this assumption is violated, it can affect the reliability of statistical tests, confidence intervals, and predictions made by the model.

- ✓ **Interpretation**: By examining the Q-Q plot, we can quickly visualize deviations from normality to check for non-normality in data.
- ✓ **Model Improvement:** Identifying non-normality in residuals allows for model improvement.
- ✓ **Decision Making**: Q-Q plots help in making decisions about the appropriateness of assuming normality. Depending on the degree of departure from the diagonal line, analysts can decide whether the assumption of normality is acceptable or if further actions need to be taken.