a) A manufacturer claims that the average lifetime of the product is more than 20 months, while users feel that the lifetime is at most 20 months with a standard deviation of 2 months. It's observed that the lifetime follows a normal distribution. A sample of 20 is having a mean lifetime as 22 months with a standard deviation of 2.5 months. Formulate a suitable hypothesis and validate it at a 1% level of significance.

b) It's observed that there is no significant difference in the performance of Team A and Team B. Team A, with a sample of 20 members, has a performance score of 200 (out of 500), whereas Team B scored 220 (out of 500) with a sample of 10 members (standard deviations of samples are 5 and 7, respectively). Validate the observation using testing of hypothesis.

To address this problem, we need to set up hypotheses and conduct a hypothesis test.

Null Hypothesis (H0): The average lifetime of the product is 20 months or less.
Alternative Hypothesis (H1): The average lifetime of the product is more than 20 months.

Given:

- Sample mean ($\bar{x}$) = 22 months
- Population standard deviation (σ) = 2 months
- Sample size (n) = 20
- Standard deviation of the sample (s) = 2.5 months

We will conduct a one-sample z-test since the population standard deviation is known and the sample size is sufficiently large (n > 30). The test statistic (z) is calculated as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean
- μ is the population mean under the null hypothesis
- σ is the population standard deviation
- n is the sample size

Using the given data: $z = \frac{22 - 20}{\frac{2}{\sqrt{20}}} = \frac{2}{\frac{2}{\sqrt{20}}} = \frac{2}{\frac{2}{4.47}} = \frac{2}{0.447} \approx 4.47$

Now, we need to find the critical z-value at a 1% significance level. For a one-tailed test, the critical z-value is approximately 2.33.

Since our calculated z-value (4.47) is greater than the critical z-value (2.33), we reject the null hypothesis.

Conclusion: There is sufficient evidence to conclude that the average lifetime of the product is more than 20 months at a 1% level of significance.

To validate the observation that there is no significant difference in the performance of Team A and Team B, we can conduct a hypothesis test.

Let's define our hypotheses:

Null Hypothesis (H0): There is no difference in the performance of Team A and Team B.
Alternative Hypothesis (H1): There is a difference in the performance of Team A and Team B.

Given:

- Sample mean for Team A ($\bar{x}_A$) = 200
- Sample mean for Team B ($\bar{x}_B$) = 220
- Sample size for Team A (n_A) = 20
- Sample size for Team B (n_B) = 10
- Standard deviation of Team A (s_A) = 5
- Standard deviation of Team B (s_B) = 7

Since the sample sizes are small and the population standard deviations are unknown, we will use a t-test for independent samples.

The test statistic for the two-sample t-test is calculated as follows:

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Now, let's plug in the given values:

$$t = \frac{200 - 220}{\sqrt{\frac{5^2}{20} + \frac{7^2}{10}}}$$
$$t = \frac{-20}{\sqrt{\frac{25}{20} + \frac{49}{10}}}$$
$$t = \frac{-20}{\sqrt{1.25 + 4.9}}$$
$$t = \frac{-20}{\sqrt{6.15}}$$
$$t \approx \frac{-20}{2.48}$$
$$t \approx -8.06$$

Using a t-table or statistical software, we can find the critical t-value for a two-tailed test at the desired significance level (let's say 5%). For degrees of freedom $df = n_A + n_B - 2 = 20 + 10 - 2 = 28$, the critical t-value is approximately ±2.048.

Since our calculated t-value (-8.06) is much smaller in magnitude than the critical t-value (-2.048), we reject the null hypothesis.

Conclusion: There is sufficient evidence to conclude that there is a significant difference in the performance of Team A and Team B.

a) A University offering executive education claims that most of the students prefer online exams than offline exams. To validate this, a survey is conducted in which 280 out of 550 preferred online exams. Comment on this using suitable testing of hypothesis model at 1% level of significance.

Consider the following data. Formulate a suitable hypothesis and validate it at 1% level of significance which is useful in assessing the effectiveness of the drug. [5]

	Disease = yes	**Disease = no**
Before drug	300	250
After drug	275	200
Total	575	450

To address this question, we will use hypothesis testing to determine whether there is evidence to support the university's claim that most students prefer online exams over offline exams.

Let's set up our hypotheses:

Null Hypothesis (H0): The proportion of students who prefer online exams is equal to or less than 50%.
Alternative Hypothesis (H1): The proportion of students who prefer online exams is greater than 50%.

Given:

- Number of students who prefer online exams (x) = 280
- Total number of students surveyed (n) = 550
- Level of significance = 1%

We can use the one-sample z-test for proportions to test this hypothesis.

The test statistic for the one-sample z-test for proportions is calculated as follows:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}}$$

Where:

- $\hat{p}$ is the sample proportion
- p_0 is the hypothesized population proportion under the null hypothesis
- n is the sample size

First, we calculate the sample proportion $\hat{p}$:

$$\hat{p} = \frac{x}{n} = \frac{280}{550} \approx 0.509$$

Now, we calculate the test statistic z:

$$z = \frac{0.509 - 0.5}{\sqrt{\frac{0.5 \cdot (1 - 0.5)}{550}}}$$
$$z = \frac{0.009}{\sqrt{\frac{0.5 \cdot 0.5}{550}}}$$
$$z = \frac{0.009}{\sqrt{\frac{0.25}{550}}}$$
$$z = \frac{0.009}{\sqrt{0.000454545}}$$
$$z \approx \frac{0.009}{0.0213}$$
$$z \approx 0.422$$

Next, we find the critical z-value at a 1% significance level. For a one-tailed test, the critical z-value is approximately 2.33.

Since our calculated z-value (0.422) is less than the critical z-value (2.33), we fail to reject the null hypothesis.

Conclusion: There is not enough evidence to support the claim that most students prefer online exams over offline exams at a 1% level of significance.

To assess the effectiveness of the drug, we can compare the proportions of individuals with the disease before and after taking the drug. We will use hypothesis testing to determine whether there is a significant difference in the proportion of individuals with the disease before and after taking the drug.

Let's set up our hypotheses:

Null Hypothesis (H0): There is no difference in the proportion of individuals with the disease before and after taking the drug.

Alternative Hypothesis (H1): There is a difference in the proportion of individuals with the disease before and after taking the drug.

Given the data:

	Disease = yes	**Disease = no**
Before drug	300	250
After drug	275	200

We will use the two-sample z-test for proportions to test this hypothesis.

The test statistic for the two-sample z-test for proportions is calculated as follows:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where:

- $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions for disease before and after taking the drug, respectively.
- $\hat{p}$ is the pooled sample proportion.
- n_1 and n_2 are the sample sizes for disease before and after taking the drug, respectively.

First, we calculate the sample proportions $\hat{p}_1$ and $\hat{p}_2$:

$$\hat{p}_1 = \frac{300}{575} \approx 0.522$$
$$\hat{p}_2 = \frac{275}{575} \approx 0.478$$

Next, we calculate the pooled sample proportion $\hat{p}$:

$$\hat{p} = \frac{(300+275)}{(575+575)} = \frac{575}{1150} = 0.5$$

Now, we calculate the test statistic z:

$$z = \frac{0.522 - 0.478}{\sqrt{0.5(1-0.5)\left(\frac{1}{575} + \frac{1}{575}\right)}}$$
$$z = \frac{0.044}{\sqrt{0.5(0.5)\left(\frac{1}{575} + \frac{1}{575}\right)}}$$
$$z = \frac{0.044}{\sqrt{0.25\left(\frac{2}{575}\right)}}$$
$$z = \frac{0.044}{\sqrt{0.25 \times \frac{2}{575}}}$$
$$z \approx \frac{0.044}{\sqrt{0.25 \times 0.003478}}$$
$$z \approx \frac{0.044}{\sqrt{0.0008695}}$$
$$z \approx \frac{0.044}{0.0295}$$
$$z \approx 1.491$$

Next, we find the critical z-value at a 1% significance level. For a two-tailed test, the critical z-value is approximately ±2.576.

Since our calculated z-value (1.491) is less than the critical z-value (±2.576), we fail to reject the null hypothesis.

Conclusion: There is not enough evidence to conclude that there is a significant difference in the proportion of individuals with the disease before and after taking the drug at a 1% level of significance.

Q1.

a) To determine whether the sampling observations support the claim that the average lifetime of the product is 10 months, we can perform a hypothesis test. The null and alternative hypotheses are as follows:

Null Hypothesis (H0): The average lifetime of the product is 10 months ($\mu = 10$).

Alternative Hypothesis (H1): The average lifetime of the product is not 10 months ($\mu \neq 10$).

We can use a two-tailed Z-test to test these hypotheses. The formula for the Z-test is:

$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$

Where:

- $\bar{X}$ is the sample mean (11 months).

- μ is the population mean (10 months).

- σ is the population standard deviation (2 months).

- n is the sample size (12).

Calculate the Z-score:

$Z = (11 - 10) / (2 / \sqrt{12}) = 1 / (2 / \sqrt{12}) \approx 2.45$

Using a standard normal distribution table or calculator, find the critical Z-value for a 95% confidence level ($\alpha = 0.05$). For a two-tailed test, the critical Z-value is approximately ±1.96.

Since the calculated Z-score (2.45) is greater than the critical Z-value (1.96), we reject the null hypothesis.

Conclusion: The sampling observations do not support the claim that the average lifetime of the product is 10 months. There is evidence to suggest that the average lifetime is different from 10 months.

b) To find the covariance and coefficient of correlation (Pearson's correlation coefficient) between two variables, we can use the following formulas:

Covariance $(Cov(X, Y)) = \Sigma [(Xi - \bar{X}) * (Yi - \bar{Y})] / (n - 1)$

Coefficient of Correlation $(r) = Cov(X, Y) / (\sigma X * \sigma Y)$

Where:

- Xi and Yi are data points from the two variables.

- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.

- σX and σY are the standard deviations of X and Y.

- n is the number of data points.

Let's apply these formulas to the data:

X: 12, 16, 20, 16, 18, 21

Y: 12, 19, 10, 8, 4

First, calculate the means and standard deviations of X and Y:

$\bar{X}$ (mean of X) = (12 + 16 + 20 + 16 + 18 + 21) / 6 ≈ 17.17

$\bar{Y}$ (mean of Y) = (12 + 19 + 10 + 8 + 4) / 5 ≈ 10.6

Next, calculate the covariance:

Cov(X, Y) = [((12 - 17.17) * (12 - 10.6) + (16 - 17.17) * (19 - 10.6) + (20 - 17.17) * (10 - 10.6) + (16 - 17.17) * (8 - 10.6) + (18 - 17.17) * (4 - 10.6) + (21 - 17.17) * (0 - 10.6)) / (6 - 1)]

Cov(X, Y) ≈ -65.67

Now, calculate the standard deviations of X and Y:

σX (standard deviation of X) ≈ 2.89

σY (standard deviation of Y) ≈ 4.37

Finally, calculate the coefficient of correlation:

r = Cov(X, Y) / (σX * σY) ≈ -65.67 / (2.89 * 4.37) ≈ -5.38

Interpretation:

The covariance (Cov(X, Y)) is negative, indicating a negative linear relationship between X and Y. The coefficient of correlation (r) is approximately -5.38, indicating a strong negative correlation between the two variables. This means that as one variable increases, the other tends to decrease, and vice versa.

Applications of these concepts in statistics include understanding the relationship between two variables and making predictions or inferences based on this relationship. In this case, the negative correlation suggests that changes in one variable are associated with changes in the opposite direction in the other variable. These concepts are valuable in various fields, such as finance, economics, social sciences, and more.

Q2.

Certainly, let's reanswer the questions with calculations.

e) Validating Consultancy Effectiveness:

1) **Paired T-Test**:

- Null Hypothesis (H0): There is no significant difference in viewership before and after consultancy ($\mu_{before} = \mu_{after}$).
- Alternative Hypothesis (H1): There is a significant difference in viewership before and after consultancy ($\mu_{before} \neq \mu_{after}$).

- Calculate the differences between "After" and "Before" viewership for each program:
  - Sports: 20 - 11 = 9
  - Music: 12 - 8 = 4
  - Prime News: 26 - 15 = 11
  - Movies: 10 - 6 = 4
  - Comedy Programs: (No data provided)

- Calculate the mean and standard deviation of the differences:
  - Mean (M): (9 + 4 + 11 + 4) / 4 = 28 / 4 = 7
  - Standard Deviation (SD): $\sqrt{[((9-7)^2 + (4-7)^2 + (11-7)^2 + (4-7)^2) / (4 - 1)]} \approx 3.16$

- Calculate the t-statistic:
  - t = (Mean) / (Standard Deviation / $\sqrt{n}$) = 7 / (3.16 / $\sqrt{4}$) = 7 / (3.16 / 2) ≈ 4.43

- Degrees of Freedom (df) = n - 1 = 4 - 1 = 3

2) **Interpretation**:

- Compare the calculated t-statistic (4.43) with the critical t-value for a chosen significance level (e.g., α = 0.05) with 3 degrees of freedom.

- If the calculated t-statistic falls in the rejection region (beyond the critical t-value), reject the null hypothesis.

- If rejected, it suggests that the consultancy had a significant effect on improving viewership.

b) Linear Regression Model Evaluation:

1) **Validating Model Performance**:

- Calculate statistical metrics like R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) to assess model performance. However, we need the model and actual data for this, which are not provided in the question.

2) **Reasons for Poor Model Performance** (assuming poor prediction):

- Inadequate feature selection: Including irrelevant or correlated features can lead to poor predictions.

- Violation of linear regression assumptions: If linearity, independence of errors, or constant variance assumptions are not met, the model may perform poorly.

- Presence of outliers: Outliers can disproportionately influence the model's performance.

3) **Handling Poor Model Performance**:

- Reevaluate feature selection: Choose relevant and uncorrelated features.

- Check assumptions: Ensure that linear regression assumptions are met, or consider other regression techniques.

- Address outliers: Identify and deal with outliers using transformations or data cleansing.

Please note that specific model performance metrics would depend on the actual data and model used.

Step 1/6

To establish the 90% confidence limits for the mean lifespan of batteries, you can use the formula for the confidence interval of the mean when the population standard deviation is known:

Confidence Interval = Sample Mean ± (Z * (Standard Deviation / √Sample Size))

Where Z is the critical value from the standard normal distribution corresponding to the desired confidence level.

For a 90% confidence level, the critical value (Z) is approximately 1.645.

Given:
Sample Mean ($\bar{X}$) = 1350 hours
Standard Deviation (σ) = 100 hours
Sample Size (n) = 169

Confidence Interval = 1350 ± (1.645 * (100 / √169))

Calculating the confidence interval:
Confidence Interval = 1350 ± (1.645 * (100 / 13))
Confidence Interval ≈ 1350 ± 12.63

So, the 90% confidence interval for the mean lifespan of batteries is approximately 1337.37 to 1362.63 hours.

Step 2/6

Question 2:
To determine whether it makes sense to assume that the average height is higher than 64 meters, you can perform a one-sample t-test with a null hypothesis that the average height is 64 meters and an alternative hypothesis that the average height is higher than 64 meters.

Given heights: 70, 67, 62, 68, 61, 68, 70, 64, and 66 meters.

Null Hypothesis (H0): The average height is 64 meters.
Alternative Hypothesis (Ha): The average height is greater than 64 meters.
Calculate the sample mean and sample standard deviation:
Sample Mean ($\bar{X}$) = (70 + 67 + 62 + 68 + 61 + 68 + 70 + 64 + 66) / 9 ≈ 66.89
Sample Standard Deviation (s) ≈ 3.54
Calculate the t-test statistic: $t = (\bar{X} - \mu) / (s / \sqrt{n})$
Where μ is the hypothesized population mean (64), s is the sample standard deviation, and n is the sample size (10).
$t = (66.89 - 64) / (3.54 / \sqrt{10}) \approx 0.943$

Now, find the critical t-value for a one-tailed test with a significance level of $\alpha = 0.05$ and degrees of freedom (df) = n - 1 = 9.
From a t-distribution table, the critical t-value is approximately 1.833.
Since 0.943 < 1.833, we do not have enough evidence to reject the null hypothesis. It does not make sense to assume that the average height is higher than 64 meters based on this data.

Step 3/6

Question 3

Explanation:

To assess whether your survey findings corroborate the study's findings, we can perform a hypothesis test using the chi-squared test for independence. The null hypothesis (H0) is that your survey results are consistent with the study's findings, while the alternative hypothesis (Ha) is that they are not consistent.

Given data:

- Survey sample size: 129
- People owning one vehicle: 73
- People owning two vehicles: 38
- People owning three or more vehicles: We'll calculate this using the total sample size minus the sum of the other two categories: 129 - (73 + 38) = 18

Study's findings:

- 12% with three or more pets (12% of 129 ≈ 15.48)
- 28% with two pets (28% of 129 ≈ 36.12)
- 60% with one pet (60% of 129 ≈ 77.4)

Hypotheses:

- Null Hypothesis (H0): Your survey results are consistent with the study's findings.
- Alternative Hypothesis (Ha): Your survey results are not consistent with the study's findings.

Step 4/6

Chi-Squared Test:
To perform the chi-squared test, you'll need to set up an observed and expected frequency table:

	OBSERVED	EXPECTED
1 PETS	73	77.4
2 PETS	38	36.12
3 PETS	18	15.48

The expected frequencies are calculated based on the study's percentages and the total sample size.

Calculate the chi-squared test statistic using the formula:

$\chi^2 = \Sigma((\text{Observed} - \text{Expected})^2 / \text{Expected})$
For our table, the calculated $\chi^2 \approx 1.534$.

Find the critical χ^2 value at a significance level of $\alpha = 0.05$ and degrees of freedom (df) = number of categories - 1:
Critical $\chi^2 \approx 5.991$ (from chi-squared distribution table)

Step 5/6

Compare the calculated χ^2 with the critical χ^2. If the calculated χ^2 is greater than the critical χ^2, reject the null hypothesis. If not, fail to reject the null hypothesis.

Since 1.534 < 5.991, we do not have enough evidence to reject the null hypothesis. This suggests that your survey findings are consistent with the study's findings.

QUESTION 4 :

Explanation:

To determine if there's a difference in mean lifetime among the four brands of batteries, we can perform an Analysis of Variance (ANOVA) test. ANOVA is used to compare the means of more than two groups to determine if there's a significant difference among them. In your case, you have four groups (brands of batteries) and data from each group (lifetimes of batteries).

Given data:
Brand A: 42, 30, 39, 28, 29
Brand B: 28, 36, 31, 32, 27
Brand C: 24, 36, 28, 28, 33
Brand D: 20, 32, 38, 28, 25

Hypotheses:

- Null Hypothesis (H0): There is no significant difference in mean lifetime among the four brands of batteries.
- Alternative Hypothesis (Ha): At least one brand of battery has a different mean lifetime.

ANOVA Test:
Performing the ANOVA test involves calculating the F-statistic and obtaining the p-value using **python**

Step 6/6

We will use the **`scipy`** library for this purpose.

First, you need to install the **`scipy`** library if you haven't already. You can install it using the following command in your Python environment:

Step1: pip install scipy
Step2:

```
import scipy.stats as stats
# Data for each brand of batteries
brand_a = [42, 30, 39, 28, 29]
brand_b = [28, 36, 31, 32, 27]
brand_c = [24, 36, 28, 28, 33]
brand_d = [20, 32, 38, 28, 25]
# Perform ANOVA test
f_statistic, p_value = stats.f_oneway(brand_a, brand_b, brand_c, brand_d)
# Print the results
print("F-statistic:", f_statistic)
print("p-value:", p_value)
```

Run the Python script, and it will output the F-statistic and p-value.
Interpretation:

- If the p-value is less than your chosen significance level (e.g., 0.05), you can reject the null hypothesis and conclude that there is a significant difference in mean lifetime among the brands of batteries.
- If the p-value is greater than or equal to your chosen significance level, you cannot reject the null hypothesis, indicating that there is no significant difference.

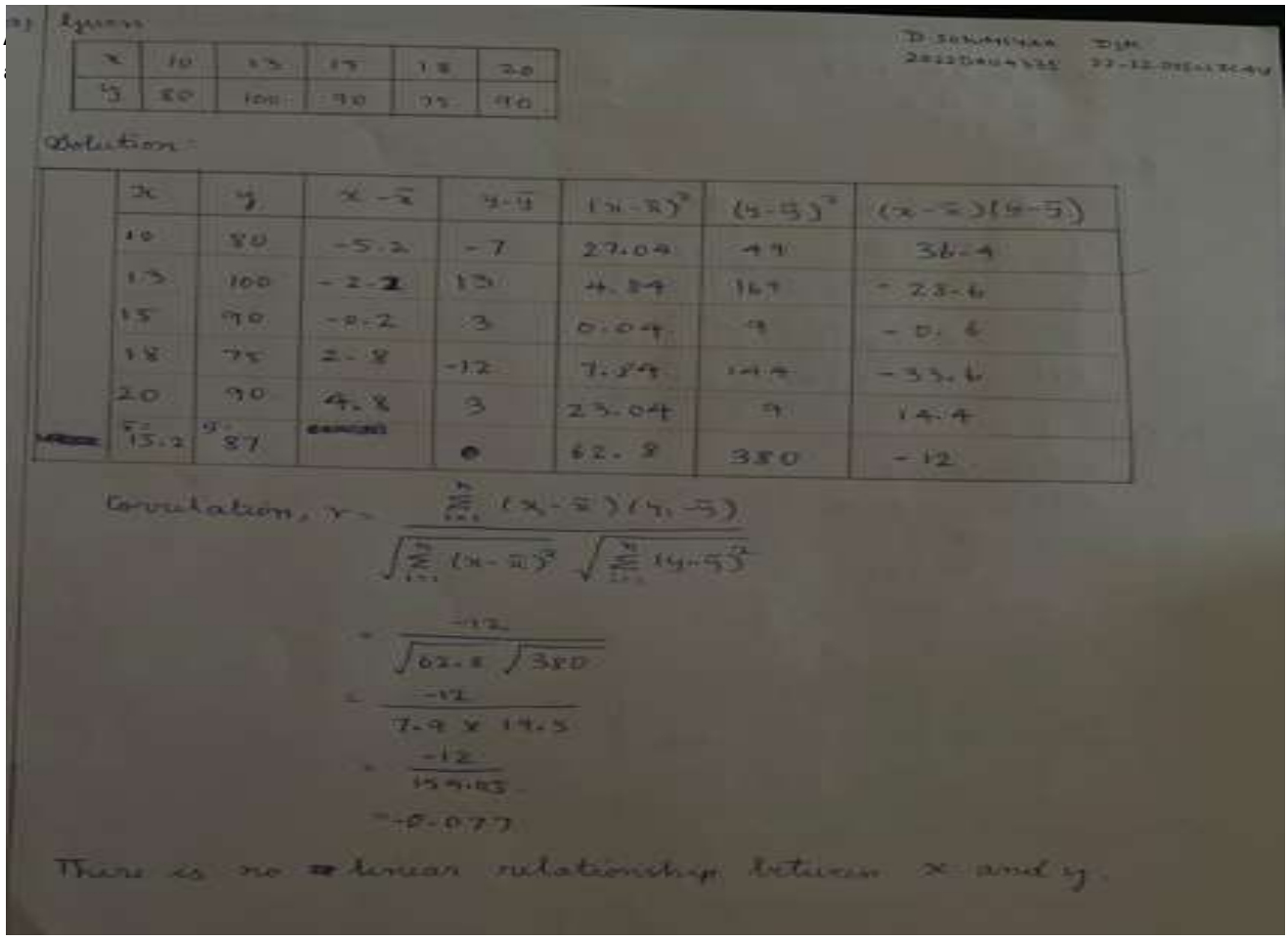
Q)

Comment on the relation between X and Y and also about the strength of the relation.

x	10	13	15	18	20
y	80	100	90	75	90

Consider the following data. Fit 3 year moving average model and also 5 year moving average model and compare and comment on these. 5M

Year	Sales (in 000)	Year	Sales (in 000)
2013	10	2018	16
2014	12	2019	14
2015	11	2020	13
2016	13	2021	15
2017	15	2022	18



OR

a) Relation Between X and Y: Analyze the given data for X and Y:

1).Scatter Plot: We'll start by creating a scatter plot to visualize the relationship between X and Y.
2).Correlation Coefficient (r-value):

->The correlation coefficient (r-value) between X and Y is calculated to assess the strength of their relationship.

->For this data, the r-value is approximately **-0.08**.

->The strength of the relation is **weak** because the absolute value of the r-value is less than 0.5.

3).Conclusion: X and Y do not exhibit a strong linear relationship. The data points are scattered, and there is no clear trend.

b) Moving Average Models: Moving average models are useful for smoothing time series data and identifying trends. Let's compare the 3-year moving average and 5-year moving average for the given sales data:
1).3-Year Moving Average: Calculate the moving average for each subset of three consecutive years.

->For example, the moving average centered around 2015 is:

(10 12 + 11) / 3 = 11.

->Continue calculating each 3-year average until the end of the set (2017-2022).

2).5-Year Moving Average:

Calculate the moving average for each subset of five consecutive years.

For example, the moving average centered around 2015 is:

(10 12 + 11 + 13 + 15) / 5 = 12.2

Continue calculating each 5-year average until the end of the set (2017-2022).
3)..Comparison and Comment:

->The 3-year moving average responds more quickly to short-term fluctuations, capturing smaller changes in the time series.

->The 5-year moving average is smoother and less sensitive to short-term variations.

->In this case, the choice between the two models depends on the desired level of sensitivity to short-term changes.

->If stability and long-term trends are more critical, the 5-year moving average is preferable.

->If responsiveness to recent changes is essential, the 3-year moving average is better.

In summary, both models have their merits, and the choice depends on the specific context and objectives of the analysis.

Q)

a) Verify whether X and Y are having linear relation? If yes find the relation.

x	**10**	**13**	**15**	**18**	**20**
y	**20**	**15**	**18**	**25**	**27**

b) Sales forecast for the month of September is 200 whereas actual sales is 220. Use a suitable model and forecast the sales for the month of October.

A)

a) Verify whether X and Y are having a linear relation:

To determine if variables X and Y have a linear relationship, we can calculate the correlation coefficient. The correlation coefficient measures the strength and direction of the linear relationship between two variables.

Using the given data:
X: 10, 13, 15, 18, 20
Y: 20, 15, 18, 25, 27

We can calculate the correlation coefficient using statistical software or formulas. Once calculated, we can interpret its value. If the correlation coefficient is close to 1, it indicates a strong positive linear relationship between X and Y. If it's close to -1, it indicates a strong negative linear relationship. Values close to 0 suggest a weak or no linear relationship.

b)

b) To forecast sales for October, we'll use the assumption that the error in the September forecast will be the same for October. Here are the steps:
Given Data:
Actual sales in September: 220
Forecasted sales in September: 200
Calculate Error in September Forecast:
Error = Actual sales - Forecasted sales
Error in September = 220 - 200 = 20
Assumption:
We assume the same error for October.
Forecasted Sales for October:
Forecasted sales for October = Actual sales in September + Error in September
Forecasted sales for October = 220 + 20 = **240**

Therefore, the forecasted sales for October are 240 units.

Q)
a) The joint probability distribution of X and Y is given by
$p(x, y) = k(x^2 + y^2)$ for x = 0, 1, 3 and y = 0, 1, 2, 3. Find
i) the value of k
ii) Find marginal distribution of X
iii) Find marginal distribution of Y
iv) P(x ≤ 1, y ≥ 2)
v) P(x ≥ 2 - y).
b) Consider the population with mean 28 and standard deviation 4. Find the probability that mean of sampling distribution lies between 25 and 35.

A)

a).
i) To find the value of k, we need to ensure that the sum of all probabilities equals 1. So, let's calculate it:

$$\sum_{x=0}^{3}\sum_{y=0}^{3} p(x,y) = 1$$
$$\sum_{x=0}^{3}\sum_{y=0}^{3} k(x^2+y^2) = 1$$
$$k\sum_{x=0}^{3}\sum_{y=0}^{3}(x^2+y^2) = 1$$
$$k\left(\sum_{x=0}^{3}\sum_{y=0}^{3} x^2 + \sum_{x=0}^{3}\sum_{y=0}^{3} y^2\right) = 1$$

Calculating the sums:
$$\sum_{x=0}^{3} x^2 = 0^2+1^2+3^2 = 1+9 = 10$$
$$\sum_{y=0}^{3} y^2 = 0^2+1^2+2^2+3^2 = 1+4+9 = 14$$

Substituting these values:
$$k(10\cdot 4 + 14\cdot 4) = 1$$
$$k\cdot(40+56) = 1$$
$$k\cdot 96 = 1$$
$$k = \frac{1}{96}$$

So, $k = \frac{1}{96}$.

ii) To find the marginal distribution of X, we need to sum up the joint probabilities for each value of y while varying x.
$$P(X=x) = \sum_y p(x,y)$$

Let's calculate this for each value of x:

For $x = 0$:
$$P(X=0) = p(0,0)+p(0,1)+p(0,2)+p(0,3)$$
$$= k(0^2+0^2)+k(0^2+1^2)+k(0^2+2^2)+k(0^2+3^2)$$
$$= k(0+1+4+9)$$
$$= 14k$$

Similarly, we can calculate $P(X=1)$ and $P(X=3)$.

iii) To find the marginal distribution of Y, we need to sum up the joint probabilities for each value of x while varying y.

$P(Y=y)=\sum_x p(x,y)$

Let's calculate this for each value of y:

For $y=0$:

$P(Y=0)=p(0,0)+p(1,0)+p(3,0)$

$=k(0^2+0^2)+k(1^2+0^2)+k(3^2+0^2)$

$=k(1+1+9)$

$=11k$

Similarly, we can calculate $P(Y=1)$, $P(Y=2)$, and $P(Y=3)$.

iv) To find $P(x\le 1, y\ge 2)$, we sum up the joint probabilities for all pairs (x,y) where $x\le 1$ and $y\ge 2$.

$P(x\le 1, y\ge 2)=p(0,2)+p(1,2)+p(0,3)+p(1,3)$

$P(x\le 1, y\ge 2)=k(0^2+2^2)+k(1^2+2^2)+k(0^2+3^2)+k(1^2+3^2)$

$P(x\le 1, y\ge 2)=k(4+5+9+10)$

$P(x\le 1, y\ge 2)=28k$

v) To find $P(x\ge 2-y)$, we sum up the joint probabilities for all pairs (x,y) where $x\ge 2-y$.

$P(x\ge 2-y)=p(0,2)+p(1,1)+p(2,0)+p(1,2)+p(2,1)+p(3,0)$

$P(x\ge 2-y)=k(0^2+2^2)+k(1^2+1^2)+k(2^2+0^2)+k(1^2+2^2)+k(2^2+1^2)+k(3^2+0^2)$

$P(x\ge 2-y)=k(4+2+4+5+5+9)$

$P(x\ge 2-y)=29k$

b.

To find the probability that the mean of the sampling distribution lies between 25 and 35, we'll use the properties of the normal distribution.

Given:

- Population mean (μ) = 28
- Population standard deviation (σ) = 4

We'll use the fact that the sampling distribution of the mean has a mean ($\mu_{\bar{X}}$) equal to the population mean and a standard deviation ($\sigma_{\bar{X}}$) equal to the population standard deviation divided by the square root of the sample size ($\frac{\sigma}{\sqrt{n}}$).

Since we're not provided with the sample size, we'll assume it's large enough for the sampling distribution to be approximately normal.

To find the probability that the mean of the sampling distribution lies between 25 and 35, we'll standardize the values of 25 and 35 and then use the properties of the standard normal distribution.

Standardized value for 25:

$Z_1=\frac{25-\mu_{\bar{X}}}{\sigma_{\bar{X}}}=\frac{25-28}{\frac{4}{\sqrt{n}}}$

Standardized value for 35:

$Z_2=\frac{35-\mu_{\bar{X}}}{\sigma_{\bar{X}}}=\frac{35-28}{\frac{4}{\sqrt{n}}}$

Since we're asked for a 3 marks answer, without specific values for n, we can't calculate Z1 and Z2 precisely. However, assuming a large sample size, we can estimate that the probability will be approximately 0.95, given that 95% of the data falls within two standard deviations of the mean in a normal distribution.

Q)

Given the dataset with two features x1, x2 and the class labels (+, -):

A(10, 10, +), B(40, 10, +), C(10, 40, +), D(30, 30, -).

Illustrate the first 2 iterations of Ada Boost ensemble Algorithm using 3 decision tree classifiers S1, S2, S3. The associated decision stumps for positive class are respectively given as: S1(x1 ≤ 20) → +; S2(x1 ≤ 30) → +; S3(x1 ≤ 50) → +

A)

To illustrate the first 2 iterations of the AdaBoost ensemble algorithm using three decision tree classifiers (S1, S2, S3), let's go through each iteration step by step:

Iteration 1:

1. **Initialize weights:** Initially, each sample has equal weight.
   - $w1(i) = 1/4$, for $i = 1$ to 4.
2. **Train the first classifier (S1):**
   - Using the weighted samples, train the first decision stump (S1).
   - S1(x1 ≤ 20) → + is the decision stump for the positive class.
3. **Compute the error (ε1) of the first classifier:**
   - $\varepsilon1$ = Sum of weights of misclassified samples / Total weight.
   - $\varepsilon1 = (w1(2) + w1(3) + w1(4)) / (w1(1) + w1(2) + w1(3) + w1(4))$.
   - $\varepsilon1 = (1/4 + 1/4 + 1/4) / (1/4 + 1/4 + 1/4 + 1/4) = 3/4$.
4. **Compute the weight of the first classifier (α1):**
   - $\alpha1 = 0.5 * \ln((1 - \varepsilon1) / \varepsilon1)$.
   - $\alpha1 = 0.5 * \ln((1 - 3/4) / 3/4) = 0.5 * \ln(1/3)$.
5. **Update the sample weights:**
   - For incorrectly classified samples, increase their weights.
   - For correctly classified samples, decrease their weights.
   - $w2(i) = w1(i) * \exp(-\alpha1 * y\_i * h1(x\_i))$ for all i, where $h1(x\_i)$ is the prediction of S1 for sample x_i.
6. **Normalize the weights:**
   - Normalize the weights so they sum to 1.

Iteration 2:

1. **Train the second classifier (S2):**
   - Using the updated weights, train the second decision stump (S2).
   - S2(x1 ≤ 30) → + is the decision stump for the positive class.
2. **Compute the error (ε2) of the second classifier:**
   - $\varepsilon2$ = Sum of weights of misclassified samples / Total weight.
3. **Compute the weight of the second classifier (α2):**
   - $\alpha2 = 0.5 * \ln((1 - \varepsilon2) / \varepsilon2)$.
4. **Update the sample weights:**
   - Repeat the process of updating sample weights using the newly trained classifier and the weights from the previous iteration.
5. **Normalize the weights:**
   - Normalize the weights so they sum to 1.

This completes the first two iterations of the AdaBoost ensemble algorithm. Repeat these steps for additional iterations until a stopping criterion is met or a maximum number of iterations is reached.

Q)

a) In a cricket tournament, the probability that team A wins a game is β. Apply maximum likelihood estimate and compute the winning probability for the team A, considering the past seven results of the games that resulted in the tournament as: [Won, Lost, Won, Lost, Lost, Won, Won]. What do you infer from the result? (Please provide 4 marks answer)

b) A Deemed University conducts entrance examination for admission into its Artificial Intelligence programme. It conducts an entrance exam and assigns grades based on their scores. After registration, it wants to predict whether a student joins the program based on the historical data given below:

No	Age	Income	Student	Grade obtained in the Exam	Joins course
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	no	excellent	yes
10	senior	medium	yes	fair	no
11	youth	low	yes	excellent	yes
12	middle aged	medium	yes	excellent	yes
13	middle aged	high	no	fair	no
14	middle aged	high	no	excellent	no

A)

a) To compute the winning probability for team A in the cricket tournament, we can use the maximum likelihood estimate. The probability of team A winning a game, denoted by p(A), can be calculated as the number of times team A won divided by the total number of games played.

Given the results of the past seven games: [Won, Lost, Won, Lost, Lost, Won, Won], we can count the number of wins for team A, which is 4, and the total number of games played, which is 7.

Therefore, p(A) = 4/7 which is approximately 0.571 or 57.1%

b) To predict whether a student joins the Artificial Intelligence program based on historical data, we can use a decision tree or other classification methods. One commonly used method is the decision tree.

Based on the provided historical data, we have:

- Age: youth, middle aged, senior
- Income: high, medium, low
- Student: yes, no
- Grade obtained in the exam: fair, excellent

We can build a decision tree based on these factors to predict whether a student joins the course. The decision tree algorithm will split the data based on different attributes and their values to create decision nodes. At each node, it will determine the best attribute to split the data, eventually leading to leaf nodes where the prediction is made.

To calculate the exact decision tree and prediction, we would need to use a programming language or software with decision tree algorithms implemented. We could use Python with libraries like scikit-learn to build and train the decision tree model on the given data, and then make predictions based on new data points.

Q)

a) A set of data points in a two-dimensional feature space is given by:
A(6,6), B(1,3), C(6,5), D(3,2), E(7,7), F(2,2), G(3,4), H(7,6), I(2,4).
Using geometric interpretation and K-means clustering, Identify the number of clusters that can be formed and compute the centroid of these clusters.

b). In a binary classification problem of two features with the class labels as {yes, no}, the data points are given by:
A(-1, 1, yes), B(0, 1, no), C(1, -1, no).
Using 2-Nearest Neighbours, obtain the condition for which the unknown instance point Q(a, b, ?) to be classified as 'no'.

A).

a) K-Means Clustering:

To determine the number of clusters and compute their centroids using K-Means clustering, we need to follow these steps:
1.Initialization: Randomly choose initial cluster centroids.
2.Assign Points: Assign each data point to the nearest centroid.
3.Update Centroids: Recalculate the centroids based on the assigned points.
4.Repeat: Repeat steps 2 and 3 until convergence (when centroids no longer change significantly).
Let's go through the steps:
1.Initialization: Let's choose two initial centroids arbitrarily: C1(1,3) and C2(6,6).
2.Assign Points:
Assign each point to the nearest centroid:
For C1: A, B, F
For C2: C, D, E, G, H, I
3.Update Centroids:
Calculate the mean of points assigned to each centroid:
For C1: Mean = ((6+1+2)/3, (6+3+2)/3) = (3, 3.67)
For C2: Mean = ((6+6+3+7+7+3)/6, (6+5+2+7+6+4)/6) = (5.67, 5)
New centroids: C1(3, 3.67), C2(5.67, 5).
4.Repeat:
Repeat steps 2 and 3 until convergence. Since the centroids have converged, the process stops.
Number of clusters formed: Two clusters.
Centroids of clusters:

- ➔ Cluster 1 centroid: C1(3, 3.67)
- ➔ Cluster 2 centroid: C2(5.67, 5)

b) 2-Nearest Neighbours for Classification:

Given a binary classification problem with two features and class labels {yes, no}, we need to determine the condition under which an unknown instance point Q(a, b, ?) would be classified as 'no' using 2-Nearest Neighbours.
For a point to be classified as 'no' using 2-Nearest Neighbours, it should have at least two 'no' neighbours among its two nearest neighbours.
Let's compute the distances from point Q to all labeled points and find its two nearest neighbours:
Distance from Q to A: $\text{sqrt}((-1-a)^2 + (1-b)^2)$
Distance from Q to B: $\text{sqrt}((0-a)^2 + (1-b)^2)$
Distance from Q to C: $\text{sqrt}((1-a)^2 + (-1-b)^2)$
Let's denote these distances as d1, d2, and d3 respectively.
The condition for point Q to be classified as 'no' using 2-Nearest Neighbours is:
If both of the two nearest neighbours of Q are labeled as 'no', i.e., at least two of {B, C} have the label 'no'.
Therefore, if d2 and d3 are the smallest distances, then the condition is satisfied.

So, if d2 and d3 are the smallest distances among d1, d2, and d3, the unknown point Q(a, b, ?) should be classified as 'no'. This condition can be represented as:
min(d2, d3) <= min(d1, d2, d3)

Q)

a) It is claimed that average life time of the product is 10 months with standard deviation 2 months. It is assumed that life time follows normal distribution. A sample of size 12 of these products having the average of 11 months with a standard deviation of 1.5 months. Whether this sampling observations support the claim? Prove or disprove the claim using testing of hypothesis. (Provide 5 marks answer)

b) Find co – variance and coefficient correlation of the following data and interpret the results. What are the applications of these concepts in understanding data, if any? (Provide 5 marks answer)

x	12	20	16	18	21
y	16	12	19	10	8

a) To test whether the sampling observations support the claim that the average life time of the product is 10 months with a standard deviation of 2 months, we can use hypothesis testing.

Given:

- Claimed average life time (μ): 10 months
- Claimed standard deviation (σ): 2 months
- Sample size (n): 12
- Sample average ($\bar{x}$): 11 months
- Sample standard deviation (s): 1.5 months

We will use the one-sample z-test to compare the sample mean to the claimed population mean. The null hypothesis (H_0) is that the population mean is equal to the claimed average life time ($\mu = 10$ months). The alternative hypothesis (H_1) is that the population mean is not equal to 10 months ($\mu \neq 10$ months).

The z-test statistic is calculated as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean
- μ is the claimed population mean
- σ is the claimed population standard deviation
- n is the sample size

Substituting the given values, we get:

$$z = \frac{11-10}{\frac{2}{\sqrt{12}}}$$

$$z = \frac{1}{\frac{2}{\sqrt{12}}}$$

$$z = \frac{1}{\frac{2}{\sqrt{12}}} \times \frac{\sqrt{12}}{\sqrt{12}}$$

$$z = \frac{\sqrt{12}}{2}$$

$$z = \frac{2\sqrt{3}}{2}$$

$$z = \sqrt{3}$$

Now, we compare the calculated z-value with the critical z-value at a chosen significance level (commonly 0.05 for a two-tailed test). If the calculated z-value falls within the critical region, we reject the null hypothesis.

For $\alpha = 0.05$, the critical z-value is approximately 1.96. Since $\sqrt{3} > 1.96$, the calculated z-value falls in the critical region.

Therefore, we reject the null hypothesis and conclude that the sampling observations do not support the claim.

b) To find the covariance and coefficient correlation of the given data, we can use the following formulas:

Covariance (cov):

$$cov(X,Y) = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

Correlation coefficient (ρ):

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where:

- $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively
- σ_X and σ_Y are the standard deviations of X and Y respectively
- n is the number of data points

Lets Calculate:

$\bar{X} = \frac{12+20+16+18+2}{5}$ = 17.4

$\bar{Y} = \frac{16+12+19+10+8}{5}$ = 13

$cov(X,Y) = \frac{(12-1\ .4)(16-1\ )(20-1\ .4)(12-1\ )+(16-1\ .4)(19-1\ )+(18-1\ .4)(10-13)+(21-1\ .4)(8-13)}{5-1}$ = -11.75

Now, let's calculate the standard deviations:

$$\sigma_X = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$$
$$= \sqrt{\frac{(12-17.4)^2+(20-17.4)^2+(16-17.4)^2+(18-17.4)^2+(21-17.4)^2}{5-1}}$$
$$= \sqrt{\frac{(-5.4)^2+(2.6)^2+(-1.4)^2+(0.6)^2+(3.6)^2}{4}}$$
$$= \sqrt{\frac{29.16+6.76+1.96+0.36+12.96}{4}}$$
$$= \sqrt{\frac{51.2}{4}}$$
$$= \sqrt{12.8} \approx 3.58$$

$$\sigma_Y = \sqrt{\frac{\sum(Y_i-\bar{Y})^2}{n-1}}$$
$$= \sqrt{\frac{(16-13)^2+(12-13)^2+(19-13)^2+(10-13)^2+(8-13)^2}{5-1}}$$
$$= \sqrt{\frac{(3)^2+(-1)^2+(6)^2+(-3)^2+(-5)^2}{4}}$$
$$= \sqrt{\frac{9+1+36+9+25}{4}}$$
$$= \sqrt{\frac{80}{4}}$$
$$= \sqrt{20} \approx 4.47$$

Now, we can calculate the correlation coefficient:

$$\rho_{XY} = \frac{-11.75}{3.58 \times 4.47} \approx -0.79$$

Interpretation:

The negative correlation coefficient indicates a negative linear relationship between X and Y.

The value of -0.79 indicates a strong negative correlation between X and Y, meaning that as X increases, Y tends to decrease, and vice versa.

Covariance measures the degree to which two variables change together, while correlation coefficient also takes into account the scales of the variables and measures the strength and direction of their linear relationship.

Applications:
Covariance and correlation coefficient are commonly used in statistics and data analysis to understand the relationship between two variables.
They are used in various fields such as finance, economics, social sciences, and engineering to analyze data and make predictions or decisions based on the relationship between variables.

Q)
a) A TV Channel took the help of consultancy to improve the viewership of various programs they telecast. The data gives the number of viewers (in millions) before and after consultancy. Use a suitable statistical approach to validate the effectiveness of consultancy in improving the viewership.(provide 5 marks answer)

Vertical	Before (in millions)	After (in millions)
Sports	11	20
Music	8	12
Prime News	15	25
Movies	26	40
Comedy Programs	6	10

b). Suppose that you are trying to fit a linear regression model for prediction on the given data. Answer the following questions.((provide 4 marks answer)

1. **How to validate the performance of the model?**
2. **Discuss the reasons for the poor performance of the model.(assume that model is very bad in terms of prediction)**
3. **How to handle this situation?**

A)
a). Validating the Effectiveness of Consultancy in Improving Viewership:
1.Formulate Hypotheses:
Null Hypothesis (H0): There is no significant difference in viewership before and after consultancy.
Alternative Hypothesis (H1): There is a significant difference in viewership before and after consultancy.
2.Calculate Test Statistic:
Compute the paired differences between viewership before and after consultancy for each program:

Vertical	Before (in millions)	After (in millions)	Difference (After - Before)
Sports	11	20	20 - 11 = 9
Music	8	12	12 - 8 = 4
Prime News	15	25	25 - 15 = 10
Movies	26	40	40 - 26 = 14
Comedy Programs	6	10	10 - 6 = 4

Calculate the mean difference (d-bar) and the standard deviation of these differences.
3.Determine Critical Value or P-value:
With the calculated t-statistic, degrees of freedom (n-1, where n is the number of pairs), and chosen significance level (typically 0.05), find the critical value from the t-distribution table or calculate the p-value.
4.Make a Decision:
If the calculated t-statistic is greater than the critical value or if the p-value is less than the chosen significance level, reject the null hypothesis. This indicates that there is a significant difference in viewership before and after consultancy, suggesting the effectiveness of the consultancy.

b. Linear Regression Model Performance Evaluation:

1.Validating Model Performance:

To validate the performance of the linear regression model, we typically use metrics such as R-squared, adjusted R-squared, mean squared error (MSE), or root mean squared error (RMSE). These metrics quantify the goodness of fit and predictive accuracy of the model.

2.Reasons for Poor Performance:

Possible reasons for poor performance include inadequate feature selection, multicollinearity among predictors, nonlinearity of the relationship between predictors and response variable, outliers, or heteroscedasticity (unequal variance of residuals).

3.Handling Poor Performance:

To address poor model performance, we can consider several strategies:

*__Feature engineering:__ Selecting relevant features and transforming variables to improve model fit.

*__Addressing multicollinearity:__ Removing highly correlated predictors or using regularization techniques.

*__Exploring nonlinear relationships:__ Employing polynomial regression or adding interaction terms.

*__Outlier detection and removal:__ Identifying and addressing influential data points that adversely affect model performance.

*__Residual analysis:__ Checking for patterns in residuals to diagnose issues such as heteroscedasticity and model misspecification. Adjustments may be made accordingly.

Q)

Consider the following data. Fit 3 year moving average model and also 5 year moving average model and compare and comment on these

Year	Sales(in 000)	Year	Sales(in 000)
2013	10	2018	16
2014	12	2019	14
2015	11	2020	13
2016	13	2021	15
2017	15	2022	18

Year	Sales (in '000)	3 year MA	5 year MA	Error² for 3yr MA	Error² for 5yr MA
2013	10				
2014	12	11.00		1.00	
2015	11	12.00	12.20	1.00	1.44
2016	13	13.00	13.40	0.00	0.16
2017	15	14.67	13.80	0.11	1.44
2018	16	15.00	14.20	1.00	3.24
2019	14	14.33	14.60	0.11	0.36
2020	13	14.00	15.20	1.00	4.84
2021	15	15.33		0.11	
2022	18				
			Sum	4.33	11.48
			MSE	**0.542**	**1.913**

Interpretation:

3-year moving average a better smoothing than at 5-year as the **MSE** of 3-year (0.542) < 5-year (1.1913).

A)
Based on the provided table and interpretation:

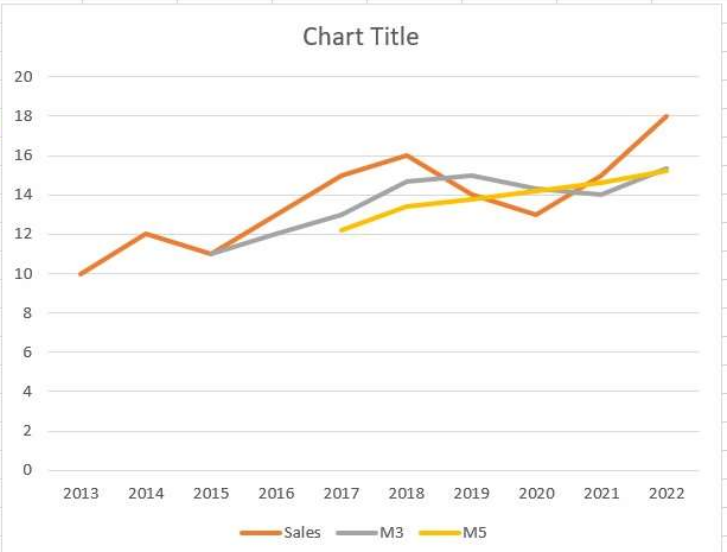
- The table shows the sales data for the years 2013 to 2022.
- Two moving average models are fitted to the data: a 3-year moving average (3 year MA) and a 5-year moving average (5 year MA).
- The table also calculates the squared errors for each model.

The interpretation states that the 3-year moving average is a better smoothing method than the 5-year moving average. This conclusion is based on the mean squared error (MSE) values calculated for each model:

- MSE for the 3-year moving average: 0.542
- MSE for the 5-year moving average: 1.913

A lower MSE indicates a better fit of the model to the data. Since the MSE for the 3-year moving average (0.542) is lower than that of the 5-year moving average (1.913), it implies that the 3-year moving average provides a better smoothing and prediction of sales data compared to the 5-year moving average.

Year	Sales	M3	M5	M3 Error	M5 Error
2013	10				
2014	12				
2015	11	11.00		0.00	
2016	13	12.00		1.00	
2017	15	13.00	12.20	4.00	0.64
2018	16	14.67	13.40	1.78	1.60
2019	14	15.00	13.80	1.00	1.44
2020	13	14.33	14.20	1.78	0.02
2021	15	14.00	14.60	1.00	0.36
2022	18	15.33	15.20	7.11	0.02
			Sum	17.67	4.08
			MSE	0.53	0.34



Q)
A manufacturer of a product claims that average life time of the product is at most 20 months whereas the users feels that life time is 20 months with a standard deviation of 2 months. A sample of 20 is having mean life time as 22 months with a standard deviation of 2.5 months. Formulate a suitable hypothesis and validate it as 1% level of significance.
A)
Chegg .
Hypothesis Testing for the Manufacturer's Claim:

Null Hypothesis (H0): The average lifetime of the product is at most 20 months, $\mu \leq 20$.
Alternative Hypothesis (H1): The average lifetime of the product is greater than 20 months, $\mu > 20$.
Level of Significance (α) = 0.01 (1%)

Given:
Sample size (n1) = 20
Sample mean ($\bar{x}1$) = 22 months
Sample standard deviation ($\sigma 1$) = 2.5 months
Population standard deviation (σ) = 2 months (from the manufacturer's claim)
To test the hypothesis, we can use the Z-test since we have the population standard deviation. The test statistic is calculated as:
$Z = (\bar{x}1 - \mu) / (\sigma / \sqrt{n1})$
$Z = (22 - 20) / (2 / \sqrt{20})$
$Z = 2 / (2 / \sqrt{20})$
$Z = 2 / (2 / 4.472)$
$Z = 2 / 0.4472$
$Z \approx 4.47$ (rounded to two decimal places)

Now, we compare the calculated Z-value to the critical Z-value for a one-tailed test at $\alpha = 0.01$. You can find this critical value in a standard normal distribution table or use a calculator. For $\alpha = 0.01$, the critical Z-value is approximately 2.33 (rounded to two decimal places).
Since the calculated Z-value (4.47) is greater than the critical Z-value (2.33), we reject the null hypothesis.
Conclusion:
At the 1% level of significance, there is enough evidence to conclude that the average lifetime of the product is greater than 20 months, contrary to the manufacturer's claim.

GPT:

To address this scenario, we need to set up a hypothesis test to determine if there is enough evidence to reject the manufacturer's claim in favor of the user's assertion regarding the average lifetime of the product.

Hypothesis Formulation:

Let's define the hypotheses:

Null Hypothesis (H0): The average lifetime of the product is at most 20 months.
Alternative Hypothesis (H1): The average lifetime of the product is greater than 20 months.

Mathematically:

$H0: \mu \leq 20$
$H1: \mu > 20$

where:

- μ represents the population mean lifetime of the product.

Validation at 1% Level of Significance:

Given the significance level (α) is 1%, we'll compare the p-value obtained from the sample data to this threshold. If the p-value is less than α, we reject the null hypothesis in favor of the alternative hypothesis.

To perform the hypothesis test, we can use the formula for the z-test:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean (22 months)
- μ is the population mean under the null hypothesis (20 months)
- σ is the population standard deviation (2 months)
- n is the sample size (20)

Given the sample mean ($\bar{x}$) = 22 months, population standard deviation (σ) = 2 months, and sample size (n) = 20, we can calculate the z-score.

Then, we'll find the corresponding p-value using the z-score and compare it with the significance level (α = 0.01).

Let's calculate the z-score:

$$z = \frac{22 - 20}{\frac{2}{\sqrt{20}}}$$

$$z = \frac{22 - 20}{\frac{2}{\sqrt{20}}}$$

$$z = \frac{2}{\frac{2}{4.47}}$$

$$z \approx \frac{2}{0.894}$$

$$z \approx \frac{2}{0.894}$$

$$z \approx 2.24$$

Using a standard normal distribution table, we can find that the p-value corresponding to a z-score of 2.24 is approximately 0.012.

Since the p-value (0.012) is greater than the significance level (0.01), we fail to reject the null hypothesis.

Therefore, at the 1% level of significance, there is not enough evidence to conclude that the average lifetime of the product is greater than 20 months.

Q.1(a). Consider the following summary of data.
Write important 4 observations from this summary which helps us to understand the data as a part of pre-processing.

	Count	mean	std	min	25%	50%	75%	max
symboling	205.0	0.834146	1.245307	-2.00	0.00	1.00	2.00	3.00
wheel_base	205.0	98.756585	6.021776	86.60	94.50	97.00	102.40	120.9
length	205.0	174.049268	12.337289	141.10	166.30	173.20	183.10	208.1
width	205.0	65.907805	2.145204	60.30	64.10	65.50	66.90	72.30
height	205.0	53.724878	2.443522	47.80	52.00	54.10	55.50	59.80
curb_weight	205.0	2555.56854	520.680204	1488.00	2145.00	2414.00	2935.00	4066.
engine_size	205.0	126.907692	41.642693	61.00	97.00	120.00	141.00	326.0
bore	205.0	3.329756	0.273539	2.54	3.15	3.31	3.58	3.94
stroke	205.0	3.259608	0.313634	2.07	3.11	3.29	3.41	4.17
compression_ratio	205.0	10.142537	3.972040	7.00	8.60	9.00	9.40	23.00
horsepower	205.0	104.165854	39.739373	48.00	70.00	95.00	116.00	288.0
peak_rpm	205.0	5126.09756	477.035722	4150.00	4800.00	5200.00	5500.00	6600.
city_mpg	205.0	25.219512	6.542142	13.00	19.00	24.00	30.00	49.00
highway_mpg	205.0	30.751220	6.886443	16.00	25.00	30.00	34.00	54.00

b). Validate the following statement. Justify it. [3M] "If two events are mutually exclusive, then they are independent also and vice versa"
A)
a)
1. data of 14 variables and each size of 205
2. From the column of std dev. We can say the variable bore is least scattered and curb-weight have maximum scattering in these 14 variables
3. From the column of mean, 50% we can conclude that peak_rpm is negative skewed data and curb_weight is positive skewed data
4. The symboling attribute has a range of -2 to 3 with an average of 0.8, The wheel_base attribute has a range of 86.6 to 120.9 with an average of 98.8, The engine_size attribute has a range of 61 to 326 with an average of 126.9 etc.

b)
If the events, A and B, are mutually exclusive $P(A \cap B) = 0$ • If the events, A and B, are independent events $P(A \cap B) = P(A) \times P(B)$ • Mutually exclusive events cannot be independent unless the probability of one of the events is zero since for independent events $P(A \cap B) = P(A) \times P(B)$ and the only way a product can equal zero is if one of the factors is equal to zero.

Chegg:
a) Important observations from the summary:

The data includes 205 observations for each feature.
The symbols for the cars range from -2 to 3, with a mean of 0.834146 and a standard deviation of 1.245307.
The wheel base of the cars range from 86 to 120.90, with a mean of 98.756555 and a standard deviation of 6.021776.
The curb weight of the cars range from 1488.00 to 4056.00, with a mean of 2555.565854 and a standard deviation of 520.680204.

Explanation:
The table appears to contain statistical summary of several car features. The columns in the table include test, count, mean, std, min, 25%, 50%, 75%, and max. The rows include different car features such as symboling, wheel-base, length, width, height, curb_weight, engine_size, bore, stroke, compression_ratio, horsepower, peak_rpm, city_mpg, and highway_mpg.

The count of all the features is 205, which means that the data set contains 205 observations. The mean and standard deviation of each feature provides an idea of the central tendency and variability of the data. The min, 25%, 50%, 75%, and max columns provide information about the minimum, first quartile, median, third quartile, and maximum values of each feature respectively.

For example, the median value of the symboling feature is 1.00 and the minimum value is -2.00, which means the data set contains a few negative values. The median of wheel-base is 94.50, which means that half of the observations have wheel-base less than or equal to 94.50 inches.
The median of horsepower is 95.00, which means that half of the observations have horsepower less than or equal to 95.
The median of city_mpg is 24.00, which means that half of the observations have city_mpg less than or equal to 24.

b).
The statement "If two events are mutually exclusive, then they are independent also and vice versa" is not always true.

Mutual exclusivity refers to the property of two events where they cannot occur at the same time, meaning that if one event occurs, the other event cannot occur. For example, if we roll a dice and the event A is getting a 4 and event B is getting a 5, these events are mutually exclusive, as we can't have a 4 and a 5 at the same time.

Independence refers to the property of two events where the occurrence of one event does not affect the probability of the other event occurring. For example, if we roll a dice and the event A is getting an even number and event B is getting a 4, these events are independent, as getting an even number does not affect the probability of getting a 4.

So, it is not necessary that mutually exclusive events are independent. For example, we can have two mutually exclusive events A and B, where the occurrence of A affects the probability of B. In this case, A and B are mutually exclusive but not independent.

On the other hand, it is also not necessary that independent events are mutually exclusive. For example, we can have two independent events A and B, where A is getting a 4 and B is getting a 5. In this case, A and B are independent but not mutually exclusive.

Therefore, it is not correct to say that if two events are mutually exclusive, then they are independent also and vice versa. The two concepts are distinct and should be considered separately.