Business Report Structure (Checklist Based)

Title Page

Title: “Business Report: INFERENTIAL STATISTICS On Clear Mountain State University

Author: Harik Charan

Table Of Contents:

1.) Problem Statement

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).

2.) Objective:

Based on the given data, answer the following questions.

1. What is the probability that a randomly selected CMSU student will be male?
2. What is the probability that a randomly selected CMSU student will be female?
3. What is the conditional probability of different majors among the male students in CMSU?
4. What is the conditional probability of different majors among the female students of CMSU?

5. What is the probability That a randomly chosen student is a male and intends to graduate?
6. What is the probability that a randomly selected student is a female and does NOT have a laptop?
7. What is the probability that a randomly chosen student is a male or has full-time employment?
8. What is the conditional probability that given a female student is randomly chosen, she is majoring in international business or management?
9. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?
10. What is the conditional probability that a randomly selected male earns 50 or more?
11. What is the conditional probability that a randomly selected female earns 50 or more?
12. Are the continuous variables in the data normally distributed? Write a note summarizing your conclusions.

3.) SURVEY DATASET:

First 5 rows of dataset:

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages	
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

Last 5 rows of dataset:

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages	
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	2	4	350	Laptop	250
59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55.0	1	4	500	Laptop	500
60	61	Female	23	Senior	Accounting	Yes	3.5	Part-	30.0	2	3	490	Laptop	50

ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages	
								Time						
61	62	Female	23	Senior	Economics/ Finance	No	3.2	Part-Time	70.0	2	3	250	Laptop	0

- There are no missing values in data set

Statistical Summary Of Data:

ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages	
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	31.500000	21.129032	3.129032	48.548387	1.516129	3.741935	482.016129	246.209677
std	18.041619	1.431311	0.377388	12.080912	0.844305	1.213793	221.953805	214.465950
min	1.000000	18.000000	2.300000	25.000000	0.000000	1.000000	100.000000	0.000000
25%	16.250000	20.000000	2.900000	40.000000	1.000000	3.000000	312.500000	100.000000
50%	31.500000	21.000000	3.150000	50.000000	1.000000	4.000000	500.000000	200.000000

ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages	
75%	46.750000	22.000000	3.400000	55.000000	2.000000	4.000000	600.000000	300.000000
max	62.000000	26.000000	3.900000	80.000000	4.000000	6.000000	1400.000000	900.000000

3.)Questions 1

1. What is the probability that a randomly selected CMSU student will be male?

```
2. The probability that a randomly selected CMSU
   student will be a male is: 0     0.000000
3. 1     1.612903
4. 2     1.612903
5. 3     1.612903
6. 4     1.612903
7.          ...
8. 57    0.000000
9. 58    0.000000
10.    59    0.000000
11.    60    0.000000
12.    61    0.000000
Name: Gender, Length: 62, dtype: float64
```

The probability that a randomly selected CMSU student will be a male is: 0 0.000000

2.) What is the probability that a randomly selected CMSU student will be female?

p_female= No_of_female/Total_value

The probability that a randomly selected CMSU student will be a female is: 0 1.612903

1 0.000000

2 0.000000

3 0.000000

4 0.000000

...

57 1.612903

58 1.612903

59 1.612903

60 1.612903

61 1.612903

Name: Gender, Length: 62, dtype: float64

3.)What is the conditional probability of different majors among the male students in CMSU?

Conditional Probability of Different Majors among Male Students in CMSU:

col_0	count
Major	
Management	0.206897
Retailing/Marketing	0.172414
Accounting	0.137931
Economics/Finance	0.137931
Other	0.137931
Undecided	0.103448
International Business	0.068966
CIS	0.034483

4. What is the conditional probability of different majors among the female students of CMSU?

Conditional Probability of Different Majors among Female Students in CMSU:

col_0	count

Major	
Retailing/Marketing	0.272727
Economics/Finance	0.212121
International Business	0.121212
Management	0.121212
Accounting	0.090909
CIS	0.090909
Other	0.090909

5. What is the probability That a randomly chosen student is a male and intends to graduate?

Probability that a randomly chosen student is a male and intends to graduate: 58.620689655172406

6. What is the probability that a randomly selected student is a female and does NOT have a laptop?

Probability that a randomly selected student is a female and does NOT have a laptop: 6.451612903225806

7. What is the probability that a randomly chosen student is a male or has full-time employment?

Probability that a randomly chosen student is a male or has full-time employment: 51.61290322580645

8. What is the conditional probability that given a female student is randomly chosen, she is majoring in international business or management?

Conditional probability that given a female student is randomly chosen, she is majoring in international business or management: 0.2424242424242243

9. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

Probability that a randomly chosen student has a GPA less than 3: 27.419354838709676

10. What is the conditional probability that a randomly selected male earns 50 or more?

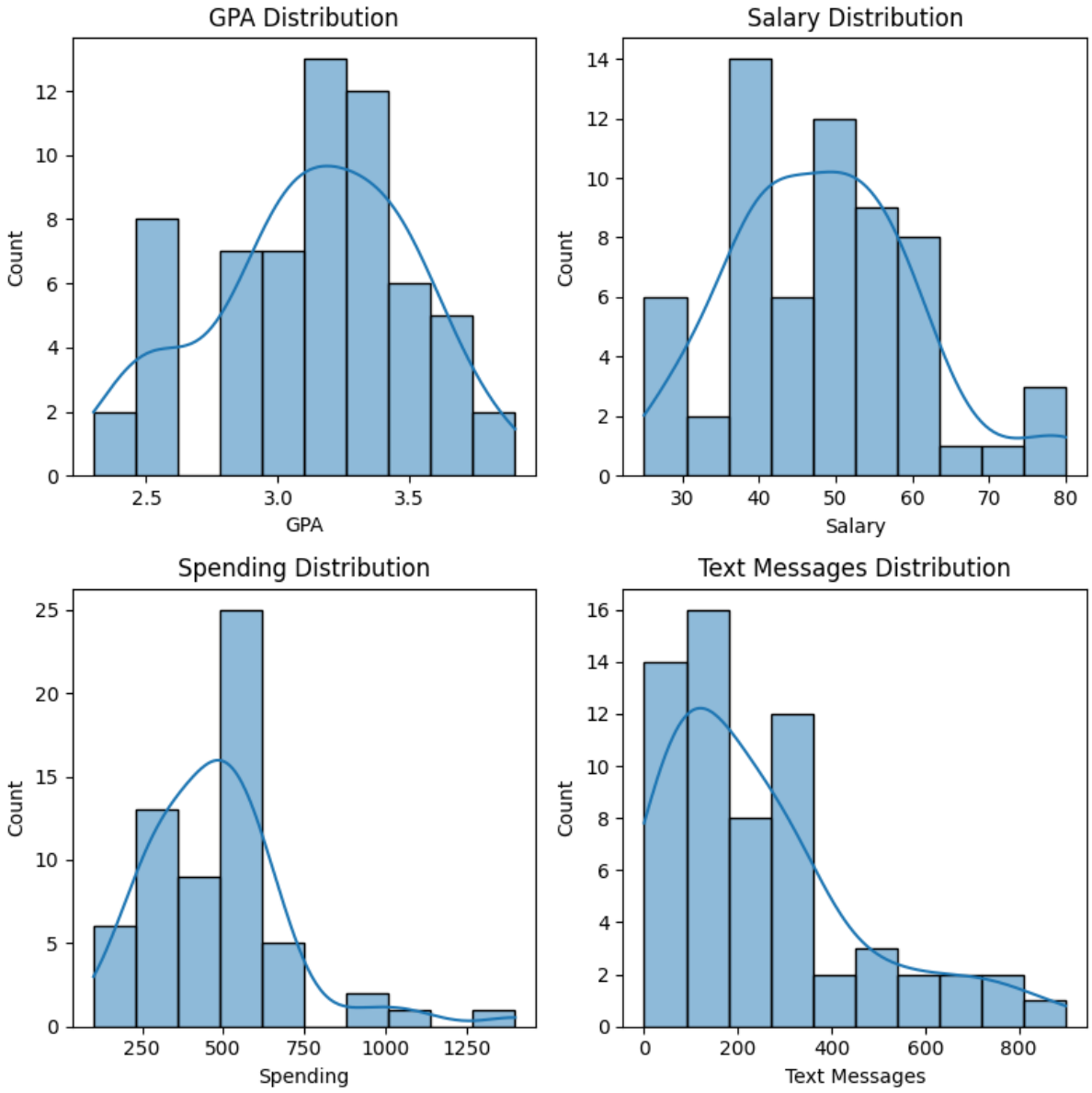
Conditional probability that a randomly selected male earns 50 or more: 48.275862068965516

11. What is the conditional probability that a randomly selected female earns 50 or more?

Conditional probability that a randomly selected female earns 50 or more: 54.54545454545454

12. Are the continuous variables in the data normally distributed? Write a note summarizing your conclusions.

Data Distribution And Skew values



1. GPA Distribution (Top Left)

- X-axis: GPA scores (ranging approximately from 2.4 to 3.9)

- Y-axis: Count (number of students)
- Shape: Nearly bell-shaped, resembling a normal distribution, though slightly skewed left.
- Interpretation: Most students have GPA scores between 3.0 and 3.5, with fewer students having very low or very high GPAs.

2. Salary Distribution (Top Right)

- X-axis: Salary (likely in 1000s or some currency unit, ranging from 25 to 80)
- Y-axis: Count
- Shape: Slight left skew, with a peak around 40–50 and a tail extending toward higher salaries.
- Interpretation: Most students earn between 40 and 60, but a few earn significantly higher, up to 80.

3. Spending Distribution (Bottom Left)

- X-axis: Spending (ranging roughly from 100 to 1300)
- Y-axis: Count
- Shape: Right-skewed (positively skewed)
- Interpretation: A large number of students spend between 300 and 600, but some spend significantly more (outliers near 1200–1300).

4. Text Messages Distribution (Bottom Right)

- X-axis: Number of text messages (ranging from 0 to 900+)
- Y-axis: Count
- Shape: Right-skewed
- Interpretation: Most students send fewer than 300 messages, but a small number send very high volumes, possibly outliers.

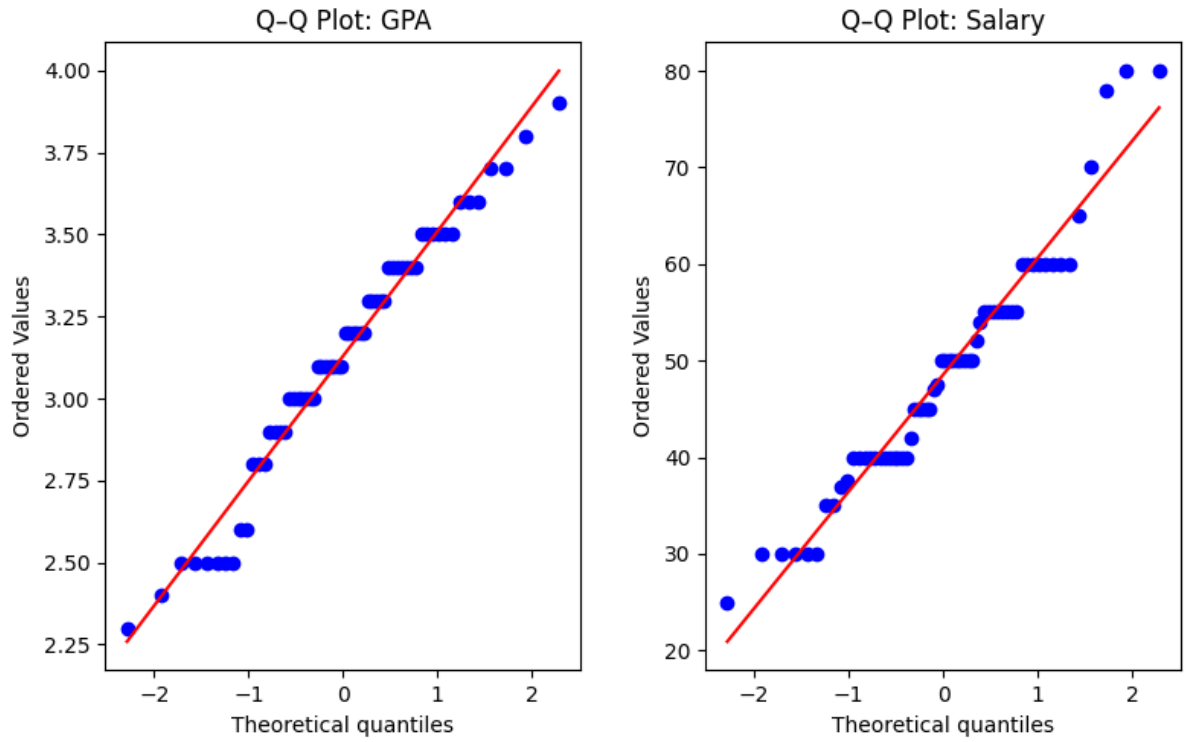
Insights:

- GPA is fairly consistent among students.
- Spending and texting behaviors show significant variation and extreme values,

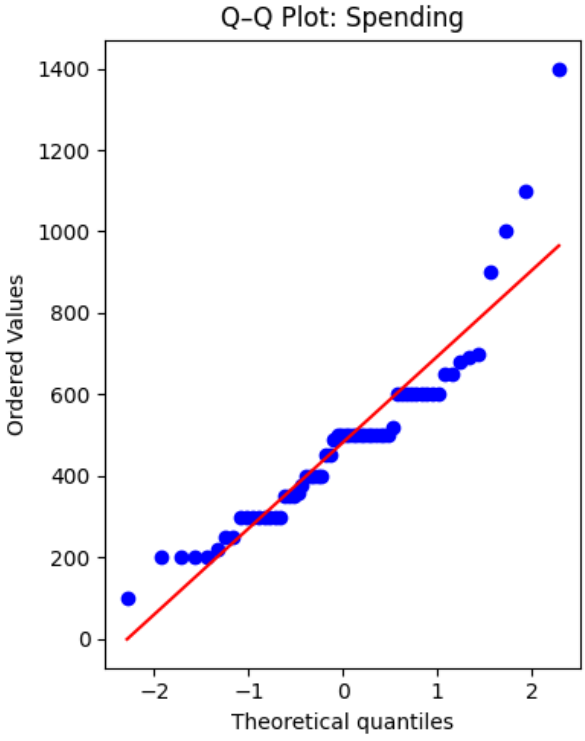
suggesting some students spend or text much more than others.

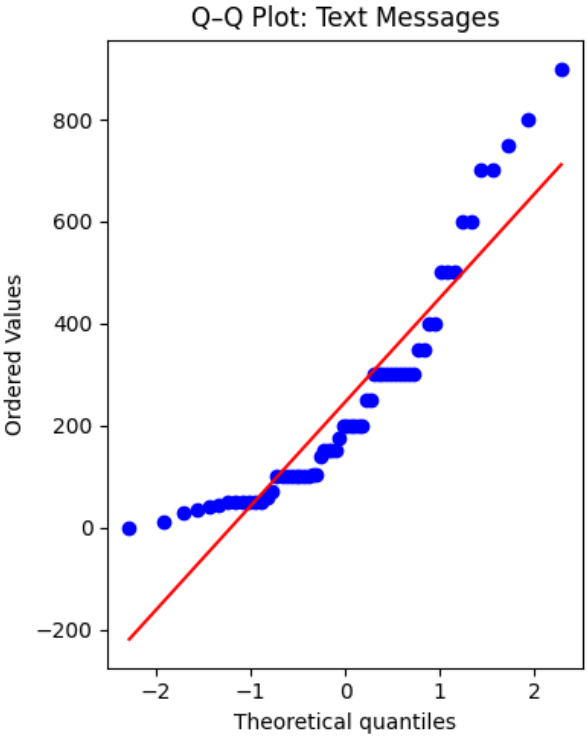
- Salary has some outliers, indicating not all students are in similar financial situations.

Probability Plots:


Q–Q Plot: GPA
Ordered Values
Theoretical quantiles
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
−2
−1
0
1
2

Q–Q Plot: Salary
Ordered Values
Theoretical quantiles
80
70
60
50
40
30
20
−2
−1
0
1
2


Q–Q Plot: Spending
Ordered Values
Theoretical quantiles
1400
1200
1000
800
600
400
200
0
−2
−1
0
1
2


Q–Q Plot: Text Messages
Ordered Values
Theoretical quantiles
800
600
400
200
0
−200
−2
−1
0
1
2

How to Interpret Q-Q Plots:

- Red Line = the theoretical normal distribution line.
- Blue Dots = actual data.
- If the points fall close to the red line, data is approximately normally distributed.
- If the points curve away from the line, data is not normally distributed (indicating skewness or heavy tails).

Q-Q Plot Analysis:

1. GPA

- Points lie almost perfectly along the red line.
- Interpretation: GPA is very close to normally distributed.
- This confirms what we saw in the histogram earlier (bell-shaped).

2. Salary

- Mostly linear but slight deviations at the lower and upper ends.

Interpretation: Salary is approximately normal, but slightly skewed or has mild outliers.

- Consistent with the histogram's mild left skew.

3. Spending

- Noticeable curve upward at the end, indicating heavy right tail.
- Interpretation: Spending is not normally distributed — it is positively skewed.
- There are outliers or high spenders pulling the tail upward.

4. Text Messages

- Large deviations from the line, especially in the tail.
- Interpretation: Text messages are not normally distributed, with strong positive skew.

- Most students send few messages, but some send very high numbers (as seen in the histogram).

What You Can Conclude:

- If you're considering parametric tests (e.g., t-tests, ANOVA), GPA and Salary are relatively safe to use.
- For Spending and Text Messages, consider using:
  - Log transformations
  - Non-parametric tests (like Mann-Whitney U or Kruskal-Wallis)
  - Or treat outliers carefully.

Variable-wise Assessment:

1. GPA

- Histogram: Bell-shaped, fairly symmetrical.
- Q-Q Plot: Points lie almost perfectly along the diagonal line.
- Conclusion: GPA is normally distributed.

2. Salary

- Histogram: Slight left skew with more values between 40–60.
- Q-Q Plot: Mostly linear with small deviations at the ends.
- Conclusion: Salary is approximately normal, but may have minor skewness or mild outliers. It may still be used in parametric tests.

3. Spending

- Histogram: Right-skewed with a long tail and some large values.
- Q-Q Plot: Noticeable curve at the top end; values deviate upward.
- Conclusion: Spending is not normally distributed — shows strong positive skew and outliers. Consider log transformation or non-parametric tests.

4. Text Messages

- Histogram: Right-skewed with a few very high values.
- Q-Q Plot: Clear deviations from normality, especially in upper tail.
- Conclusion: Text Messages are not normally distributed — also positively skewed with outliers. Use log/sqrt transformation or non-parametric tests.

Final Note:

Among the four continuous variables:

- GPA is normally distributed.
- Salary is close to normal, usable with caution.
- Spending and Text Messages are not normal, and require transformation or non-parametric approaches for accurate analysis

4.)Problem Statement 2

Context:

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

5.)Objective2:

Based on the above context, the manufacturer wants to understand the following:

1. Is there any evidence that mean moisture content in both types of shingles are within the permissible limits?
2. Is the population mean for shingles A and B are equal?

Use the relevant statistical tests to answer the above questions and state your conclusions along with all necessary steps.

Hint:

- Use the test for equality of means for the second question

Data set:A+&+B+shingles.csv

First 5 rows of data:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

Last 5 rows of Data:

	A	B
31	**0.40**	**NaN**
32	**0.29**	**NaN**
33	**0.43**	**NaN**
34	**0.34**	**NaN**
35	**0.37**	**NaN**

Missing Values in Data set:

0

A 0

B 5

There is **0** missing values in column A
There is **5** missing values in column B

Statistical Summary Of Data:

	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

6.)Question2:

1. Is there any evidence that mean moisture content in both types of shingles are within the permissible limits?

t-statistic: -1.473504625382782

p-value (one-tailed): 0.0747763314490751 3

Hypotheses:

- **Null hypothesis (H_0):** $\mu = 0.35$ (the population mean is 0.35)
- **Alternative hypothesis (H_1):** $\mu < 0.35$ (you expect the mean is less than 0.35 because it's a one-tailed test)

Interpretation:

- A **one-tailed p-value = 0.0748**
- Using a common **significance level (α) = 0.05**:
  - Since **0.0748 > 0.05**, you **fail to reject** the null hypothesis.
  - This means there is **not enough evidence** to conclude that the mean of column A is significantly less than 0.35.

Decide to reject or accept null hypothesis

one-sample t-test p-value= 0.07477633144907513

We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis

We conclude that the moisture content is greater than permissible limit in sample A.

Hypothesis Setup:

You are using a **one-tailed t-test**, where:

- **Null Hypothesis (H_o):** Mean moisture content (μ) = 0.35
- **Alternative Hypothesis (H_1):** Mean moisture content (μ) < 0.35
  (You're testing if moisture content is **less than** 0.35)

Test Results:

- **t-statistic** = -1.4735
- **p-value (one-tailed)** = 0.0748
- **Alpha level (significance threshold)** = 0.05

Interpretation:

- Since **0.0748 > 0.05**, we **do not reject** the null hypothesis.
- That means there's **not enough statistical evidence** to claim that the average moisture content is **less than** the permissible limit of **0.35**.
- Therefore, we assume the null hypothesis is plausible — that the **moisture content is at or above 0.35**.

Define Null and alternate hypothesis for sample B:

What This Means:

You are testing whether the **mean of column A is significantly less than 0.35**, using a one-tailed t-test.

Hypotheses:

- **Null hypothesis (H_0):** $\mu = 0.35$
- **Alternative hypothesis (H_1):** $\mu < 0.35$

Results:

t-statistic = -1.4735

p-value (one-tailed) = 0.0748

Interpretation:

- The **t-statistic** is negative, which is expected if you're testing whether the mean is **less than 0.35**.
- The **p-value = 0.0748** is **greater than α = 0.05**, so:
  - You **do not reject the null hypothesis**.
  - You **do not have enough evidence** to claim that the mean of A is significantly **less than 0.35**.

Conclusion:

Based on the one-sample t-test, we **do not have sufficient evidence** to conclude that the mean of df1['A'] is less than 0.35 at the 5% significance level.

2.Is the population mean for shingles A and B are equal?

Calculate the p - value and test statistic

tstat 0.8445012483270873

P Value 0.4050738703654344

What This Test Does:

The ttest_rel compares **two related (paired) samples** — in your case, columns A and B. This is useful when measurements are taken from the **same subjects** under **two conditions** (e.g., before vs. after, or treatment vs. control).

Hypotheses:

- **Null hypothesis (H_0):** The mean difference between A and B is **zero** (no significant change).
- **Alternative hypothesis (H_1):** There **is** a significant difference between the means of A and B.

Interpretation:

- **p-value = 0.4051 > 0.05 → Fail to reject the null hypothesis**
- Conclusion: There is **no statistically significant difference** between the means of columns **A and B**.

Final Statement:

- Based on the paired t-test, there is not enough evidence to suggest a significant difference between the values in column A and column B. Their means are statistically similar.

Decide to reject or accept null hypothesis

Issue:

The **else block** correctly states:

"We do not have enough evidence to reject the null hypothesis..."

But then incorrectly concludes:

"We conclude that mean for shingles A and singles B are not the same"

That is the **opposite** of what the test result indicates.

Correct Interpretation:

If p-value = 0.4051 > 0.05, you **fail to reject** the null hypothesis, so you **cannot conclude** the means are different.

Summary:

- You did **not** find a significant difference between the means.
- Therefore, you **must not conclude** that the means are different.

7.)Problem Statement 3:

Business Context

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor's, and Doctorate.

Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education–occupation combination.

8.)Objective3:

Based on the above context, we want to understand the following:

1. Is there any significant difference in salaries among different levels of education?
2. Is there any significant difference in salaries among different levels of different occupations?
3. Is there a significant interaction between Education and Occupation on Salary?

Dataset Salary:

First 5 rows of dataset:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

Last 5 Rows Of Dataset:

A		B
31	0.40	NaN
32	0.29	NaN
33	0.43	NaN
34	0.34	NaN
35	0.37	NaN

Missing Values:

0

A 0

B 5

There is **0** missing values in data in A
There is **5** missing values in data in B

Statistical Summary Of DataSet:

	A	B
count	**36.000000**	**31.000000**
mean	**0.316667**	**0.273548**
std	**0.135731**	**0.137296**
min	**0.130000**	**0.100000**
25%	**0.207500**	**0.160000**
50%	**0.290000**	**0.230000**
75%	**0.392500**	**0.400000**

	A	B
max	0.720000	0.580000

9.)Question3:

1. Is there any significant difference in salaries among different levels of education?

Hypotheses:

- Null hypothesis (H_o): Mean salary is the same across all education groups.
- Alternative hypothesis (H_1): At least one group has a different mean salary

Interpretation:

- The p-value from ANOVA ≥ 0.05, so you fail to reject the null hypothesis.
- This means the observed differences in average salaries between Doctorate, Masters, and Bachelors are not statistically significant.
- In other words, education level does not have a significant effect on salary in your sample.

Fail to reject the null hypothesis. There is no significant difference in salaries based on education.

2. Is there any significant difference in salaries among different levels of different occupations?

Hypotheses:

- Null Hypothesis (H_0): All four occupation groups have the same mean salary
- Alternative Hypothesis (H_1): At least one group has a different mean salary

Interpretation:

- The p-value is ≥ 0.05, which means:
  - You do not reject the null hypothesis
  - There is no statistically significant evidence that salaries differ across the four occupation types in your sample.

Possible Reasons:

1. Small group sizes may reduce statistical power.
2. High within-group salary variability can mask differences.
3. The actual salary means might be too close to detect a meaningful difference.

3. Is there a significant interaction between Education and Occupation on Salary?

Reject the null hypothesis. There is a significant interaction between Education and Occupation on Salary.

Then it means:

- The **effect of education on salary depends on occupation**, and vice versa.
- For example, having a Master's might result in higher salary **only in managerial roles**, not others.
- In such cases, you **should not interpret main effects separately** without considering interaction.

10.)Recommendations:

1. GPA (Academic Performance)

Insight: GPA is normally distributed and tightly clustered.

Business Recommendation:

- Use GPA as a predictable metric for academic success and potential.
- Incorporate GPA thresholds into recruitment screening or scholarship eligibility processes, since it’s a stable and reliable indicator.

2. Salary

Insight:

- No significant difference in salary based on education or occupation individually.
- However, interaction between education and occupation is significant, meaning salary outcomes depend on the combination of the two.

Business Recommendation:

- Design compensation strategies that factor in both education level and job role together, rather than evaluating them in isolation.
- For example, offer higher incentives to employees with higher education only if they are in roles where education adds value (e.g., managerial or specialist roles).

3. Spending

Insight:

- Spending behavior is highly skewed with some heavy spenders.
- Not normally distributed.

Business Recommendation:

- Use segmentation strategies:

  Identify high spenders and create VIP loyalty programs.

  Offer budget options or tiered services for low-to-medium spenders.
- Consider spending habits as a proxy for lifestyle profiling and personalized marketing.

4. Text Messages

Insight:

- Very positively skewed. Some users send a very high number of messages.
- Not normally distributed.

Business Recommendation:

- Identify users with very high text usage and offer unlimited texting plans to retain them.
- For users with low usage, bundle messaging with data or app packages to increase overall usage.
- Use text behavior as a predictor of engagement or communication style, which can be useful in CRM or customer support design.

5. Education

Insight:

- Education alone does not significantly affect salary, but does matter in combination with occupation.

Business Recommendation:

- Reevaluate HR policies that prioritize degrees without considering job relevance.
- Provide incentive-based training or certifications aligned with job roles rather than general degrees.

6. Occupation

Insight:

- No significant difference in salary between occupations alone, but there is an interaction effect with education.

Business Recommendation:

- Don’t base pay solely on job title—align compensation with both skill level and educational background.
- Consider cross-functional training to allow employees to transition between roles where their education can yield better compensation outcomes.

General Strategic Recommendations:

- Use multi-factor analysis (like interaction between education and occupation) for decision-making rather than relying on single metrics.
- Employ data-driven segmentation to tailor offers, salaries, or marketing to specific behavior patterns (e.g., high spenders, frequent texters).
- Perform periodic salary audits to ensure fairness across educational and functional roles.