Inferential Statistics (IS)

Guided project report

Submitted to

By

Pirangi Charan Teja Goud

In partial Fulfillment of

PGP-DSBA

Table of Contents

List of Tables

List of Figures

List of Equations

1. Probability Formula:

$$P(Event) = \frac{FavorableOutcomes}{Total\ Outcomes}$$

2. Conditional Probability Formula:

$$P(A|B|) = \frac{P(A \cap B)}{P(B)}$$

3. Union of two Events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

4. ANOVA F-Test Formula:

$$F = \frac{Between\ Group\ Variance}{Within\ Group\ Variance}$$

Data Dictionary

Column Name	Data Type	Description
ID	Integer	Unique identifier for each student
Gender	Text	Student’s gender (Male/Female)
Age	Integer	Student’s age in years
Class	Text	Academic class or year
Major	Text	Declared major field of study
Grad Intention	Text	Whether the student intends to graduate (Yes/No)
GPA	Float	Student's Grade Point Average
Employment	Text	Employment status (Full-time, Part-time, Unemployed)
Salary	Float	Monthly salary in dollars
Social Networking	Float	Hours spent on social networking per day
Satisfaction	Integer	Satisfaction score (scale-based)
Spending	Float	Amount of money spent per semester
Computer	Text	Type of computer owned (Laptop/None)
Text Messages	Integer	Number of text messages sent per day

Introduction

This report provides statistical insights into three different problems: Student Demographics and Behavioral Analysis, Moisture Content in ABC Asphalt Shingles and The Relationship Between Salary, Education, and Occupation. This report applies statistical techniques such as probability analysis, hypothesis testing, and ANOVA to address key business problems. Each section follows a structured approach, detailing the problem, methodology, results, and conclusions. The insights derived from this analysis will support organizations in enhancing decision-making processes and optimizing operational efficiency.

Problem 1:

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).

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

Solution: the probability that a randomly selected CMSU student will be male is

Probability = (Number of male students) / (Total number of students)

From the dataset:

- Number of male students = 29
- Total number of students = 62

$$\text{P(Male)} = \frac{62}{29} = 0.468 \text{ (approx.)}$$

Thus, the probability that a randomly selected CMSU student will be male is **0.468** (or **46.8%**).

1.2 What is the probability that a randomly selected CMSU student will be female?

Solution: The probability of selecting a female student is calculated using the formula

Probability = (Number of female students) / (Total number of students)

From the dataset:

- Number of female students = 33
- Total number of students = 62

$$P(\text{Female}) = \frac{33}{62} = 0.532(\text{approx.})$$

Thus, the probability that a randomly selected CMSU student will be female is **0.532 (or 53.2%)**.

1.3What is the conditional probability of different majors among male students in CMSU?

Probability = (Number of male students in a specific major) / (Total number of male students)

- **Management:**

$$P(\text{Management} / \text{Male}) = \frac{6}{29} = 0.207 \textbf{ (approx.)}$$

- **Retailing/Marketing:**

$$P(\text{Management} / \text{Male}) = \frac{5}{29} = 0.172 \textbf{ (approx.)}$$

- **Other:**

$$P(\text{Other} / \text{Male}) = \frac{4}{29} = 0.138 \textbf{ (approx.)}$$

- **Economics/Finance:**

$$P(\text{Economics or Finance} / \text{Male}) = \frac{4}{29} = 0.138 \textbf{ (approx.)}$$

- **Accounting:**

$$P(\text{Accounting} / \text{Male}) = \frac{4}{29} = 0.138 \textbf{ (approx.)}$$

- **Undecided:**

$$P(\text{Undecided} / \text{Male}) = \frac{3}{29} = 0.103 \textbf{ (approx.)}$$

- **International Business:**

$$P(\text{International Business} / \text{Male}) = \frac{2}{29} = 0.069 \textbf{ (approx.)}$$

- **CIS:**

$$P(\text{International Business} / \text{Male}) = \frac{1}{29} = 0.034 \textbf{ (approx.)}$$

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

Solution: The conditional probability of a female student being in a particular major is given by the formula:

Probability = (Number of female students in a specific major) / (Total number of female students)

From the dataset, the total number of female students is 33. Below are the probabilities for different majors among female students:

- **Retailing/Marketing:**

$$P(\text{Retailing or Marketing /Female}) = \frac{9}{33} = 0.273 \textbf{ (approx.)}$$

- **Economics/Finance:**

$$P(\text{Economics or Finance / Female}) = \frac{7}{33} = 0.212 \textbf{ (approx.)}$$

- **Management:**

$$P(\text{Management / Female}) = \frac{4}{33} = 0.121 \textbf{ (approx.)}$$

- **International Business:**

$$P(\text{International Business / Female}) = \frac{4}{33} = 0.121 \textbf{ (approx.)}$$

- **Other:**

$$P(\text{Other / Female}) = \frac{3}{33} = 0.091 \textbf{ (approx.)}$$

- **CIS:**

$$P(\text{CIS / Female}) = \frac{3}{33} = 0.091 \textbf{ (approx.)}$$

- **Accounting:**

$$P(\text{Accounting / Female}) = \frac{3}{33} = 0.091 \textbf{ (approx.)}$$

These probabilities represent the likelihood that a randomly selected female student is in a specific major.

1.5 What is the probability that a randomly chosen student is a male and intends to graduate?

Solution: The probability that a randomly chosen student is male and intends to graduate is calculated using the formula:

Probability = (Number of male students who intend to graduate) / (Total number of students)

From the dataset:

- Number of male students who intend to graduate = 17
- Total number of students = 62

$$P(\text{Male} \cap \text{Graduate Intention}) = \frac{17}{62} = 0.274$$

Thus, the probability that a randomly chosen student is male and intends to graduate is **0.274 (or 27.4%)**.

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

Solution: the probability that a randomly selected student is a female and does NOT have a laptop:

Probability = (Number of female students without a laptop) / (Total number of students)

From the dataset:

- Number of female students who do not have a laptop = 4
- Total number of students = 62

$$P\ (\text{Female} \cap \text{No Laptop}) = \frac{4}{62} = 0.065$$

The probability that a randomly selected student is female and does NOT have a laptop is **0.065 (or 6.5%)**.

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

Solution: The probability that a randomly chosen student is male or has full-time employment is calculated using the formula:

$$P\ (\text{Male} \cup \text{Full-Time Employment}) = P(\text{Male}) + P\ (\text{Full-Time Employment}) - P\ (\text{Male} \cap \text{Full-Time Employment})$$

From the dataset:

- Number of male students = 29
- Number of students with full-time employment = 11
- Number of male students with full-time employment = 6
- Total number of students = 62

Now, calculating the probabilities:

$$P(\text{Male}) = \frac{29}{62} = 0.468$$

$$P\ (\text{Full-Time Employment}) = \frac{11}{62} = 0.177$$

$$P\ (\text{Male} \cap \text{Full-Time Employment}) = \frac{6}{62} = 0.097$$

Applying the formula:

$$P(\text{Male} \cup \text{Full-Time Employment}) = 0.468 + 0.177 - 0.097 = 0.548$$

The probability that a randomly chosen student is male or has full-time employment is **0.548 (or 54.8%)**.

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

Solution: the conditional probability that given a female student is randomly chosen; she is majoring in international business or management:

$$\text{Probability} = (\text{Number of female students majoring in International Business or Management}) / (\text{Total number of female students})$$

$$= \frac{8}{33} = 0.242$$

Thus, the probability that a randomly chosen female student is majoring in International Business or Management is **0.242 (or 24.2%)**.

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

Solution: The probability that a randomly chosen student has a GPA less than 3 is given by:

$$\text{Probability} = (\text{Number of students with GPA} < 3) / (\text{Total number of students})$$

Using the dataset:

- Number of students with GPA < 3 = 17
- Total number of students = 62
-

$$P(\text{GPA} < 3) = \frac{17}{62} = 0.274$$

Thus, the probability that a randomly chosen student has a GPA less than 3 is **0.274 (or 27.4%)**.

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

Solution:

the conditional probability that a randomly selected male earns 50 or more:

$$P(\text{Earning} \geq 50/\ \text{Male}) = \text{Number of males earning} \geq 50 / \text{Total number of males}$$

$$= \frac{14}{29} = 0.4828 \ = \ 48.28\%$$

the conditional probability that a randomly selected male earns 50 or more is **0.4828 or 48.28%**

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

Solution:

the conditional probability that a randomly selected female earns 50 or more:

P (Earning ≥ 50 / Female) = Number of males earning ≥50 / Total number of females

$$= \frac{6}{11} = \ 0.5455 \ or \ 54.55\%$$

the conditional probability that a randomly selected female earns 50 or more is **0.5455 or 54.55%**

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

Solution:

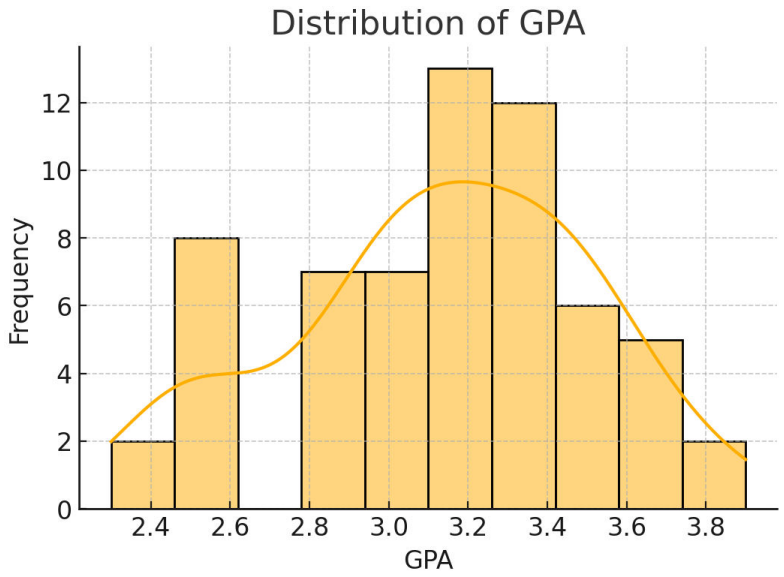

Fig 1: Distribution of GPA

The distribution of **GPA** appears to be approximately normal, with a slight left skew **(skew = - 0.31)**. This suggests that most students have GPAs clustered around the mean, with slightly more values leaning towards the higher end. Since the distribution is close to normal, standard statistical methods can be applied without significant concerns.

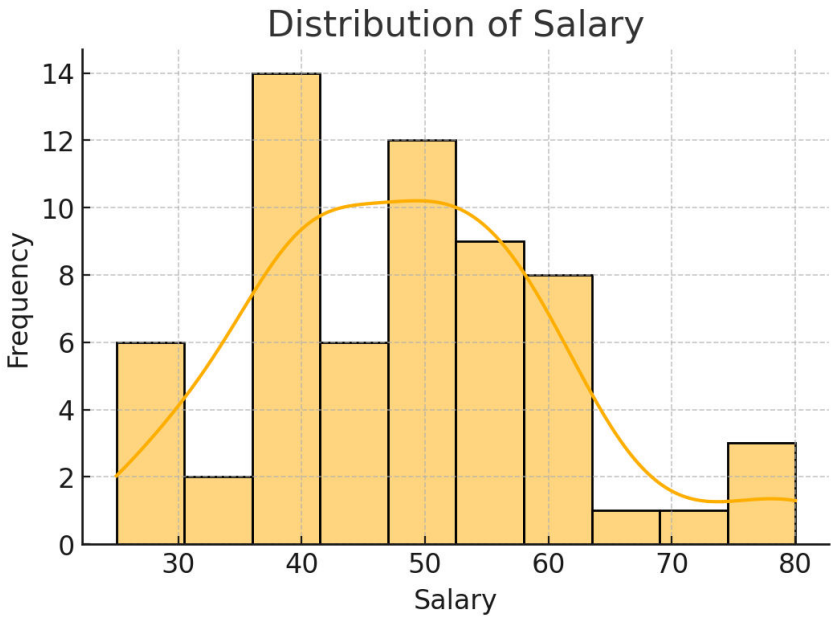

Fig 2: Distribution of Salary

The **Salary** distribution is moderately right-skewed **(skew = 0.53)**. This indicates that most individuals earn within a lower to mid-range salary, but a few earn significantly higher amounts, pulling the distribution to the right. While this skewness is not extreme, it suggests that median-based statistics might provide a better representation of central tendency than the mean.

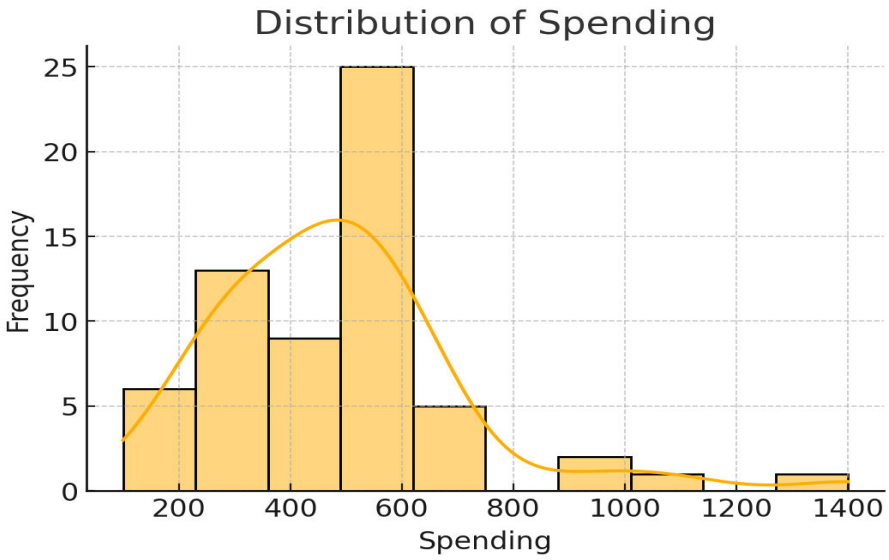

Fig 3: Distribution of Spending

Spending is highly right-skewed **(skew=1.59),** meaning that most individuals spend relatively low amounts, while a few outliers have significantly higher spending levels. This skewness suggests that a small percentage of people drive up the average spending. A log transformation may help normalize this variable for better statistical analysis.

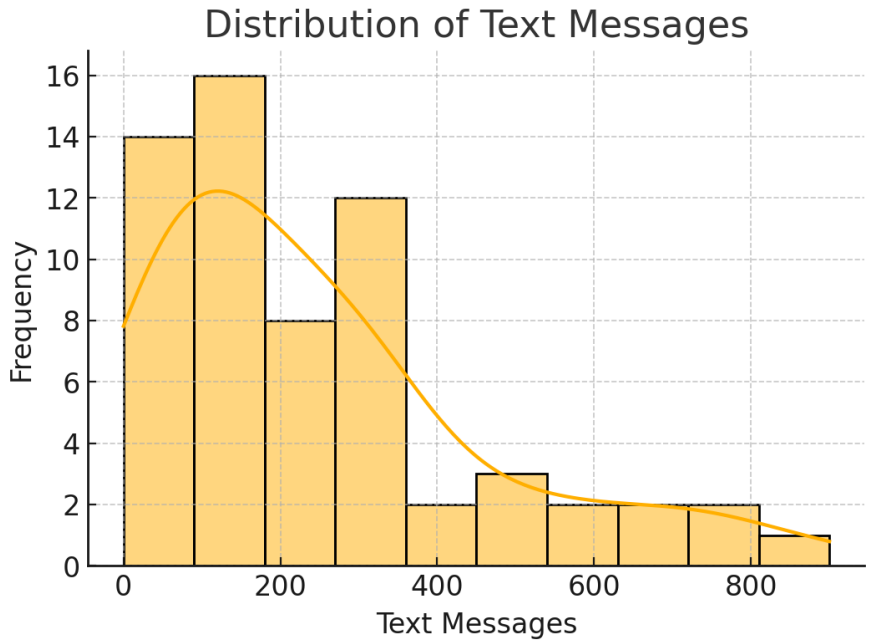

Fig 4: Distribution of Text messages

Similarly, **Text Messages** show a highly right-skewed distribution **(skew=1.30)**, indicating that most individuals send relatively few messages, while a smaller group sends an exceptionally high number. This skewed nature suggests that using median-based measures or transformation techniques could be beneficial in further analysis.

Conclusion: In conclusion, while **GPA** is nearly normal, **Salary, Spending, and Text Messages** exhibit right-skewed distributions due to high-value outliers. If normality is required for analysis, transformations such

logarithmic scaling or non-parametric tests should be considered.

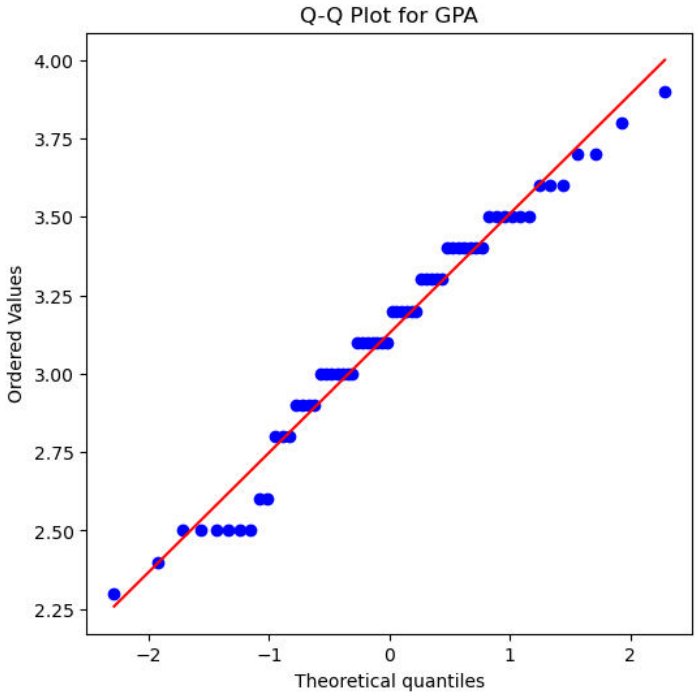

Fig 5: GPA Data: Theoretical vs. Observed Quantiles

- If the points follow the diagonal line closely, the **Salary** data is normally distributed.
- If the points deviate significantly, especially at the higher or lower ends, it suggests that the **Salary** data may have a skewed distribution or outliers.

Summary:

- The **GPA** data appears to be somewhat normally distributed, but there may be slight deviations in the tails, indicating potential skewness or outliers.

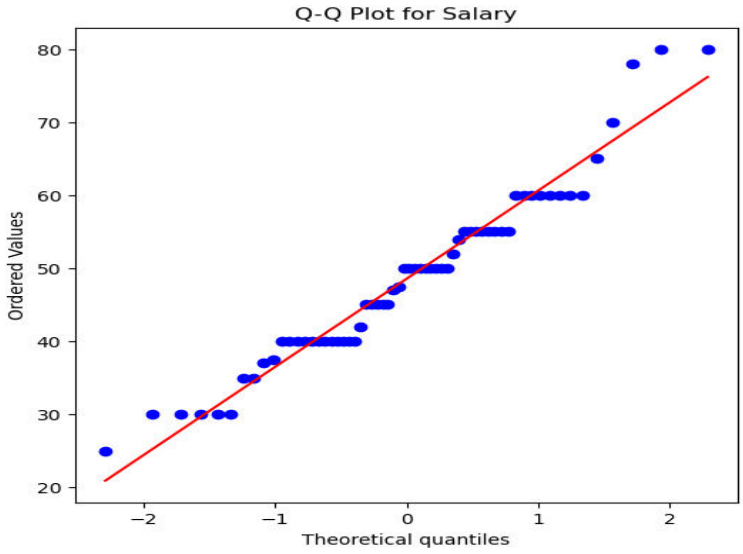

Fig 6: Salary Data: Theoretical vs. Observed Quantiles

- If the points follow the diagonal line closely, the **Salary** data is normally distributed.
- If the points deviate significantly, especially at the higher or lower ends, it suggests that the **Salary** data may have a skewed distribution or outliers.
- **Summary**:
  - The **Salary** data likely deviates from normality, particularly in the tails. This could indicate a right-skewed distribution, where a few individuals have significantly higher salaries compared to the majority.

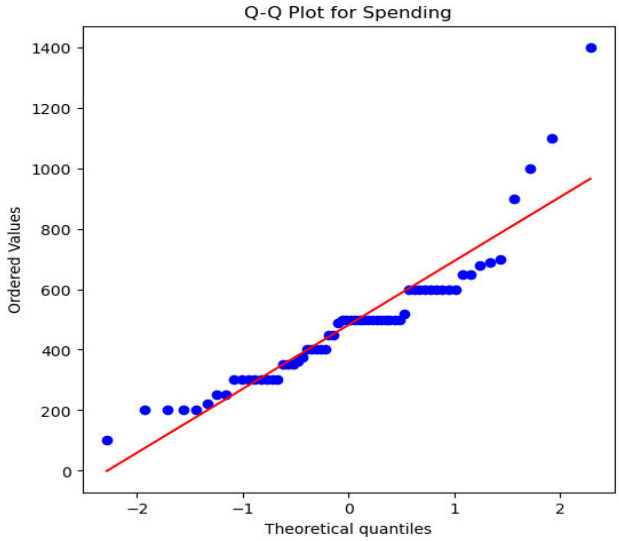

Fig 7: Spending Data: Theoretical vs. Observed Quantiles

- If the points align well with the diagonal line, the **Spending** data is normally distributed.
- Deviations, especially at the higher end (e.g., points curving upward), suggest that the **Spending** data may have a right-skewed distribution, with some individuals spending significantly more than others.
- **Summary**:
    - The **Spending** data appears to be right-skewed, with a few outliers or individuals spending much more than the majority. This is common in spending data, where most people spend within a certain range, but a few spend much more.

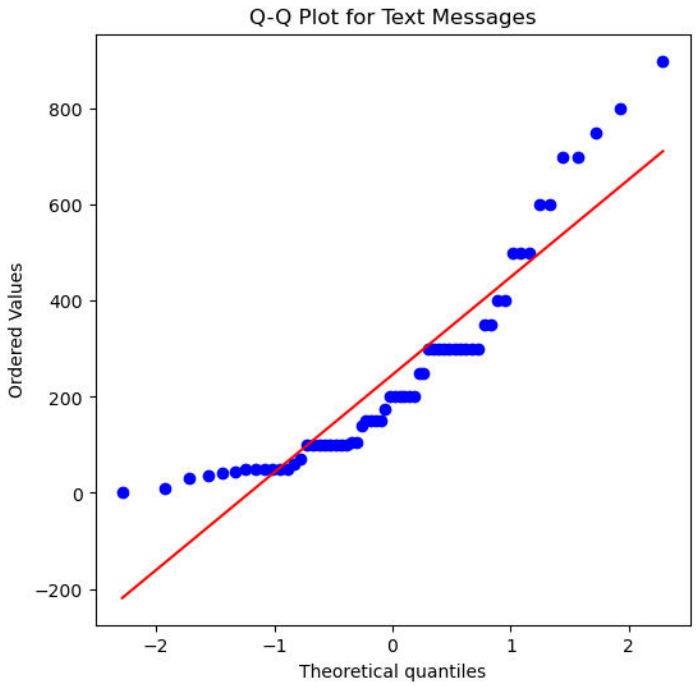

Fig 8: Test Messages Data: Theoretical vs. Ordered Quantiles

- If the points follow the diagonal line, the **Text Messages** data is normally distributed.
- Deviations, especially at the higher end, suggest that the data may have a right-skewed distribution, with some individuals sending significantly more text messages than others.

- **Summary**:
    - The **Text Messages** data is likely right-skewed, with a few individuals sending a much higher number of text messages compared to the majority. This is common in communication data, where most people send a moderate number of messages, but a few are highly active.

- **Conclusion:**
  - **GPA**: Approximately normally distributed, suitable for parametric methods with minor adjustments.
  - **Salary**: Right-skewed, non-parametric methods or transformations recommended.
  - **Spending**: Right-skewed, non-parametric methods or transformations recommended.
  - **Text Messages**: Right-skewed, non-parametric methods or transformations recommended.
  - For the skewed datasets (Salary, Spending, and Text Messages), consider using **log transformations** or **non-parametric statistical tests** to handle the skewness and outliers. For GPA, parametric methods can be used, but it's important to check for outliers or slight deviations from normality.

Problem 2:

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.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is essential for understanding the distribution, patterns, and potential anomalies within the dataset before conducting statistical tests. In this analysis, we will examine Shingle A and Shingle B individually using summary statistics and visualizations.

Visualization for each shingle type: Histogram for Shingle - A

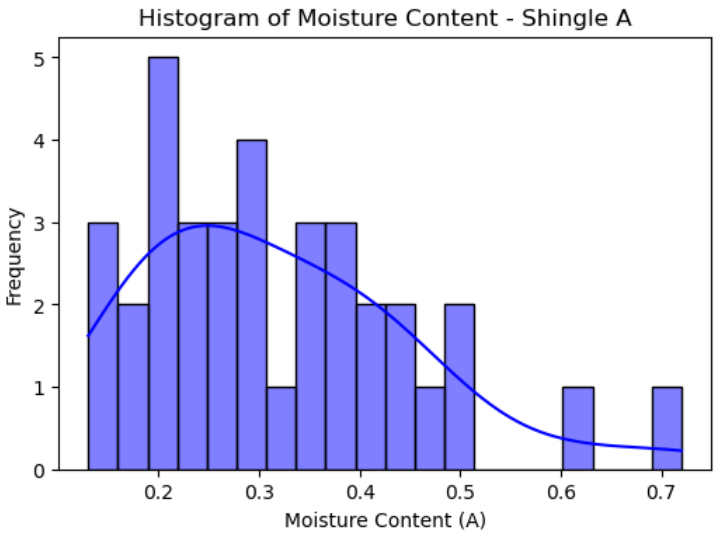

Fig 9: Histogram of Moisture Content – Shingle A

- **Summary:**
  - The histogram shows the distribution of moisture content in Shingle A.
  - If the shape is approximately **normal**, it indicates a well-distributed moisture content.
  - If **skewed**, it suggests uneven moisture retention.

Boxplot of Shingle – B:

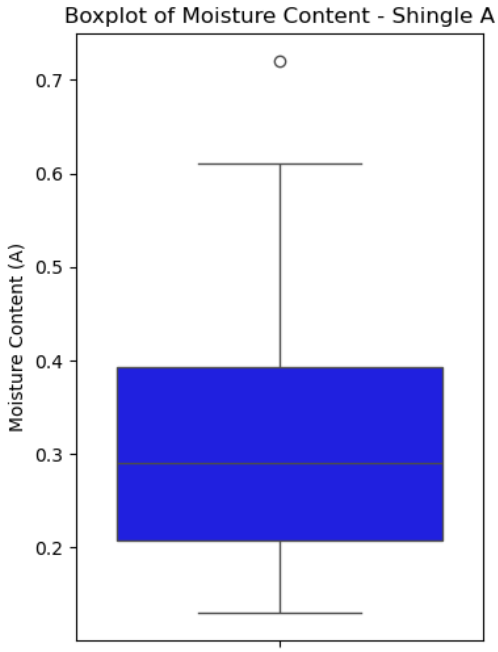

Fig 10: Boxplot of Moisture Content – Shingle A

- **Summary:**
  - The boxplot helps identify **outliers** (values far from the whiskers).
  - If outliers exist, further investigation is needed—whether they are natural variations or errors.

Histogram for Shingle – B

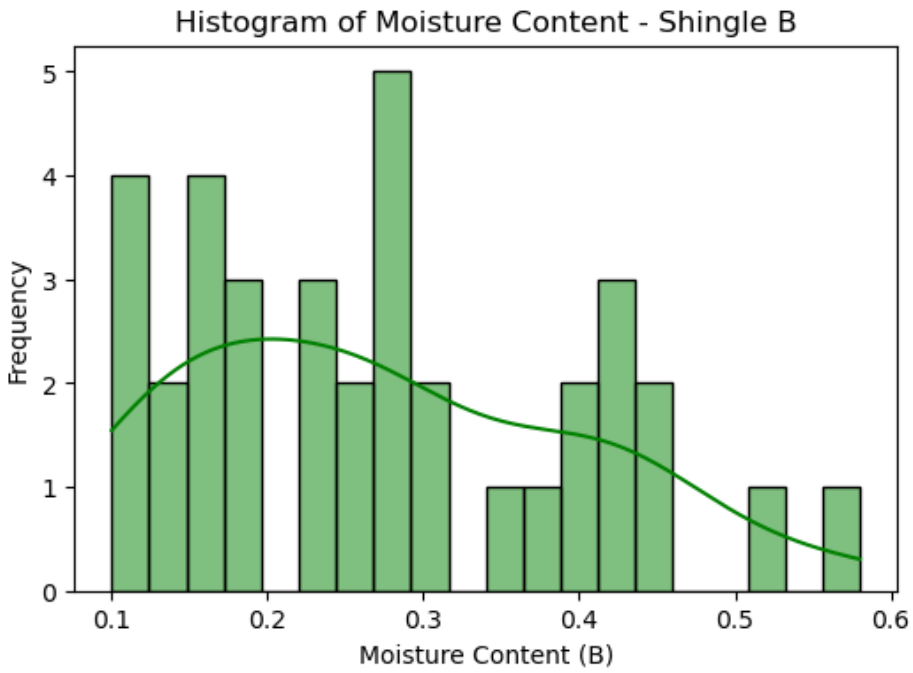

Fig 11: Histogram of Moisture Content – Shingle B

- **Summary**
  - Similar insights as above but specific to Shingle B.
  - A wider spread would indicate **higher variability** in moisture content.

Boxplot for Shingle – B

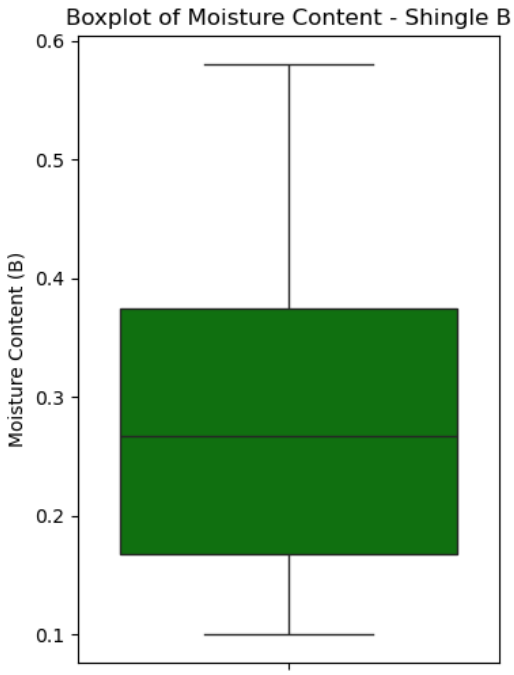

Fig 12: Boxplot of Moisture Content – Shingle – B

- **Summary**
  - Compares the moisture variability with A.
  - If the box (IQR) is larger than A, Shingle B has higher variation in moisture content.

Correlation Heatmap Between Shingles A & B

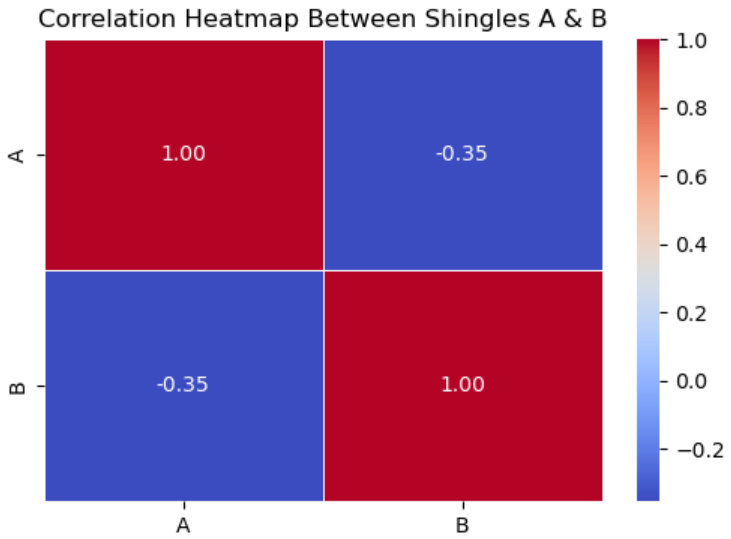

Fig 13: Correlation Heatmap Between Shangles A and B

2.1 Is there any evidence that the mean moisture content in both types of shingles is within the permissible limits?

Solution:

Step 1: State the Null and Alternate Hypotheses

For both Shingle A and Shingle B, we are testing if the mean moisture content is less than the permissible limit (0.35 pounds per 100 square feet).

- **Null Hypothesis (H_0):** The mean moisture content is greater than or equal to 0.35.
  $H_0: \mu \geq 0.35$

- **Alternative Hypothesis (H_1):** The mean moisture content is less than 0.35.
  $H_1: \mu < 0.35$

  This is a **one-tailed t-test** for a **single mean**.

Step 2: Select the Significance Level

- We will use $\alpha = 0.05$ (5%), which means we will reject the null hypothesis if the p-value is less than 0.05.

Step 3: Conduct the Hypothesis Test

We use a one-sample t-test to check if the mean moisture content is significantly less than 0.35.

Results for Shingle A:

- t-statistic: -1.4735
- p-value: 0.0748

Results for Shingle B:

- **t-statistic:** -3.6087
- **p-value:** 0.00048

Step 4: Decision Based on p-value

- For Shingle A: Since p-value (0.0748) > 0.05, we fail to reject H_0.
  - Conclusion: There is not enough statistical evidence to conclude that the moisture content of Shingle A is less than 0.35.
  - This suggests that Shingle A may not meet the permissible moisture limit.

- For Shingle B: Since p-value (0.00048) < 0.05, we reject H_0.
  - Conclusion: There is enough evidence to conclude that the moisture content of Shingle B is less than 0.35.
  - This means Shingle B meets the permissible moisture limit.

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

Solution:

Step 1: Define the Hypotheses

- **Null Hypothesis (H_0):** The mean moisture content of Shingle A and Shingle B are equal.
  Ho: μ A = μ B
- **Alternative Hypothesis (H_1):** The mean moisture content of Shingle A and Shingle B are not equal.
  H1: μ A = μ B

  This is a **two-tailed test**.

Step 2: Choose Significance Level

- The level of significance is **α = 0.05 (5%)**.

Step 3: Identify the Test Statistic

- Since the **population standard deviations are unknown**, and **sample sizes are different**, we use the **independent two-sample t-test (Welch's t-test)**.
- This test assumes **unequal variances** and follows a **t-distribution**.

Step 4: Compute the Test Statistic and p-value

- t-statistic: 1.3912
- p-value: 0.1686

Step 5: Decision Rule

- If p-value < 0.05, we reject H_0 (significant difference in means).
- If p-value ≥ 0.05, we fail to reject H_0 (no significant difference).

Step 6: Conclusion

Since the **p-value (0.1686) is greater than the significance level (0.05)**, we **fail to reject the null hypothesis (H_0)**.

🔷 **Conclusion:** There is not enough statistical evidence to say that the mean moisture content of Shingle A and Shingle B are significantly different. We conclude that their means are statistically similar at the 5% significance level.

Problem 3

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.

Exploratory Data Analysis:

Salary Distribution by Education Level (Boxplot)

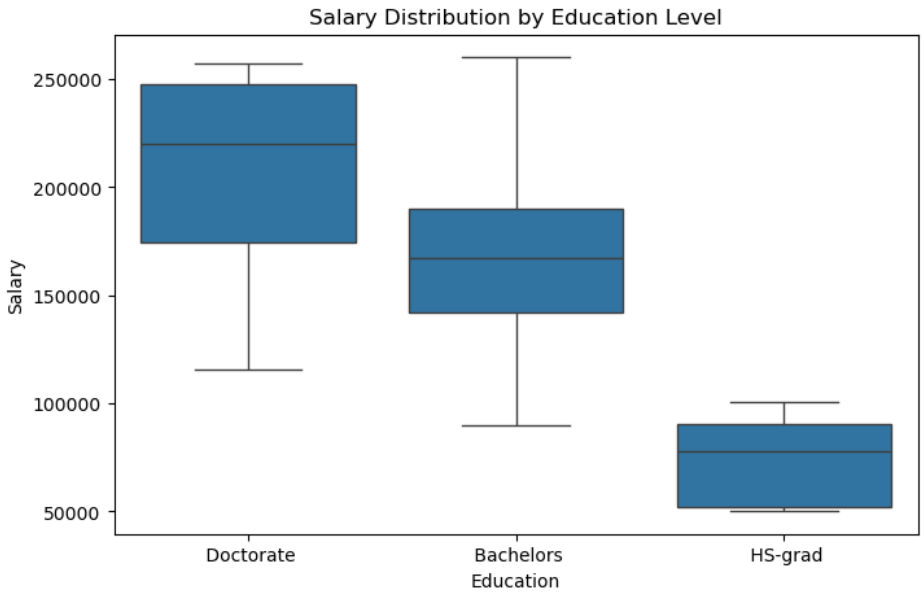

Fig 14: Boxplot – Salary Distribution by Education Level

Summary of Salary Distribution by Education Level (Boxplot Analysis)

The boxplot displays the salary distribution for individuals with different education levels (High School Graduate, Bachelor's, and Doctorate). The key insights are:

Median Salary Comparison

- Doctorate holders have the highest median salary, reflecting their advanced qualifications.
- Bachelor's degree holders earn more than High School graduates but less than Doctorate holders.
- High School graduates (HS-grad) have the lowest median salary.

Salary Variation and Spread

- **Doctorate**: The widest interquartile range (IQR), indicating substantial variation in salaries.
- Bachelor's: Moderate salary variation, with a slightly smaller IQR than Doctorates.
- High School Graduates: The smallest salary spread, indicating relatively consistent earnings.

Presence of Outliers

- No extreme outliers are observed in any education category.

Skewness and Distribution Shape

- The distributions for Doctorate and bachelor's degree holders are more spread out, suggesting a broader salary range.
- The salary distribution for HS Graduates is more compact, implying lower variability in earnings.

Conclusion

The analysis suggests that salary levels tend to increase with higher education levels. Doctorate holders not only earn the highest median salary but also show the most variation in salaries. Bachelor's degree holders follow a similar trend but with lower earnings. High School graduates earn the least, with more stable salaries. These patterns indicate potentially significant differences in salaries across education levels, warranting further statistical analysis (such as ANOVA) to verify the significance of these differences.

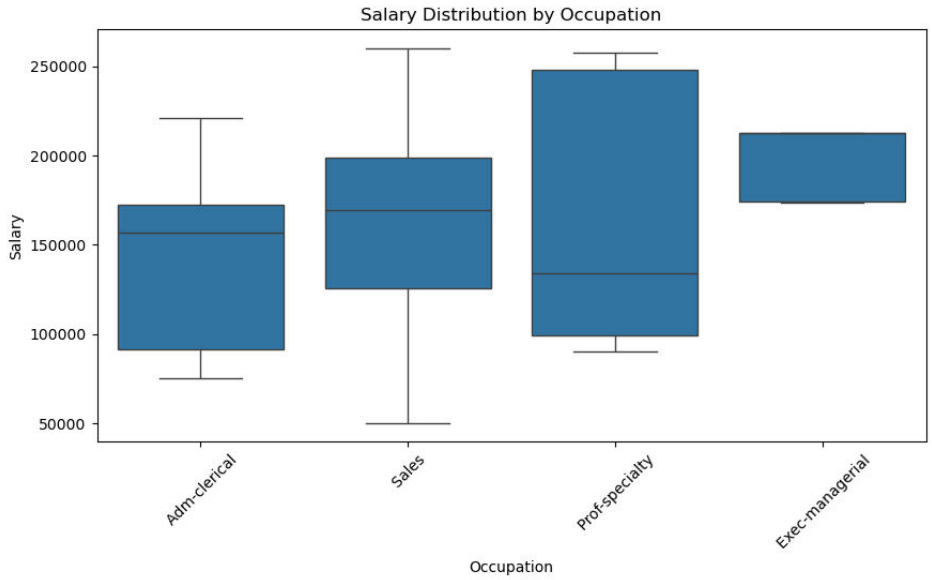

Fig 15: Boxplot – Salary Distribution by Occupation

Summary of Salary Distribution by Occupation (Boxplot Analysis)

The boxplot illustrates salary distributions across various occupational categories: Administrative & Clerical, Sales, Professional Specialty, and Executive & Managerial. Key observations include:

Median Salary Comparison

- Executive & Managerial roles have the highest median salary, indicating that individuals in these positions typically earn more than those in other occupations.
- Sales roles exhibit a moderately high median salary, though with a more dispersed distribution.
- Administrative & Clerical roles have a relatively lower median salary.
- Professional Specialty roles display a highly variable median salary, suggesting significant disparities in earnings within this category.

Salary Variability and Dispersion

- Professional Specialty roles demonstrate the largest salary spread, reflecting substantial variability in earnings.
- Executive & Managerial positions show the least variation, suggesting more consistency in salaries within this category.
- Sales roles exhibit moderate salary dispersion.

Outliers and Distribution Characteristics

- No extreme outliers are present in any occupational category.
- Professional Specialty roles have a broader salary range, potentially indicating the presence of highly paid specialists.
- Administrative & Clerical positions show a more compact salary distribution, signifying less variability.

Conclusion

The salary distribution varies considerably across occupations. Executive & Managerial roles tend to offer the highest and most stable salaries, whereas Professional Specialty roles demonstrate the greatest variation, potentially due to differences in specialization and expertise. These observations underscore the significant influence of occupation on salary levels. To validate these differences statistically, an ANOVA test should be conducted to determine whether they are significant.

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

Solution:

Step 1: State the Hypotheses

- **H_0: μ Doctorate = μ Bachelors = μ HS-grad**
- **H_1: At least one μ is different**

Step 2: Check the Assumptions of ANOVA

To perform a **One-Way ANOVA**, we need to check the following assumptions:

1. **Normality**: Salaries within each education level should be approximately normally distributed. We test this using the Shapiro-Wilk test.
2. **Homogeneity of Variance**: The variance in salaries across education levels should be similar. We test this using **Levene's test.**

Step 3: Conduct the Hypothesis Test (One-Way ANOVA)

We perform a **One-Way ANOVA** to determine whether at least one education level has a significantly different salary compared to others.

These are the results after running code:

```
Unique education levels: [' Doctorate' ' Bachelors' ' HS-grad']

Normality Test (Shapiro-Wilk Test):
Education Level:  Doctorate | Test Statistic: 0.8953 | p-value: 0.0676
Education Level:  Bachelors | Test Statistic: 0.9607 | p-value: 0.7051
Education Level:  HS-grad | Test Statistic: 0.8853 | p-value: 0.1783

Levene's Test for Homogeneity of Variance:
Test Statistic: 1.8801, p-value: 0.1669

One-Way ANOVA Results:
F-statistic: 30.9563, p-value: 0.0000
Conclusion: Reject the null hypothesis. There is a significant difference in salaries among education
levels.
```

Step 4: Conclusion from the Results

1. **Normality Test (Shapiro-Wilk Test)**:
    a. Doctorate: p-value = 0.0676 (normal)
    b. Bachelors: p-value = 0.7051 (normal)
    c. HS-grad: p-value = 0.1783 (normal)
    d. Since all p-values > 0.05, the normality assumption holds.
2. **Homogeneity of Variance (Levene's Test)**:
    a. Test Statistic: 1.8801, p-value = 0.1669
    b. Since p-value > 0.05, variances are equal, so we can proceed with ANOVA.
3. **One-Way ANOVA Results**:
    a. F-statistic: 30.9563, p-value < 0.0001
    b. Since p-value < 0.05, we reject H_0 and conclude that salaries differ significantly among education levels.
4. **Post-hoc Analysis (Tukey's HSD Test)** (if applicable):
    a. If ANOVA is significant, Tukey's HSD test helps identify which specific education levels have significantly different salaries.

Conclusion

- Based on One-Way ANOVA, there is a statistically significant difference in salaries among different education levels.
- Since ANOVA only tells us that at least one group is different, a post-hoc test (Tukey's HSD) can be conducted to determine which specific education levels differ from each other.
- This finding suggests that higher education levels are associated with significantly different salary distributions.

3.2 IIs there any significant difference in salaries among different levels of different occupations?

Solution:

Step 1:

Null Hypothesis (H_0)

There is no significant difference in mean salaries among different levels of education.

H0: μ Doctorate = μ Bachelors = μ HS-grad

Alternative Hypothesis (H_1)

At least one education level has a significantly different mean salary compared to others.

H_1: At least one μ is different

Step 2: Assumption Checks

Normality (Shapiro-Wilk Test): Salaries within each education level should be approximately normally distributed.
Homogeneity of Variance (Levene's Test): The variance in salaries across education levels should be similar.

Step 3: Normality Test (Shapiro-Wilk Test)

Education Level	p-value	Normality Assumption
Doctorate	0.0676	Normal
Bachelors	0.7051	Normal
HS-grad	0.1783	Normal

Since all p-values > 0.05, the normality assumption holds.

Step 4: Homogeneity of Variance (Levene's Test)

- Test Statistic: 1.8801
- p-value: 0.1669

Since p-value > 0.05, variances are equal, so we can proceed with ANOVA.

Step 4: Conduct the Hypothesis Test (One-Way ANOVA)

- F-statistic: 30.9563
- p-value: < 0.0001

Since p-value < 0.05, we reject the null hypothesis (H_0) and conclude that salaries differ significantly among education levels.

Conclusion

- Based on One-Way ANOVA, there is a statistically significant difference in salaries among different education levels.
- Since ANOVA only tells us that at least one group is different, a post-hoc test (Tukey's HSD) can be conducted to determine which specific education levels differ from each other.
- This finding suggests that higher education levels are associated with significantly different salary distributions.

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

Solution:

Step1: Null Hypothesis (H_0):

H0: Effect of Education × Effect of Occupation=0

Alternative Hypothesis (H_1):

H1: Effect of Education × Effect of Occupation ≠0

Step 2: Check the Assumptions

1. **Normality Check (Shapiro-Wilk Test)**
    a. Ideally, salaries within each Education-Occupation group should be normally distributed.
    b. However, some groups have very few observations (e.g., Exec-managerial for HS-grad = 0), making normality difficult to assess.

2.Homogeneity of Variance (Levene's Test)

- This should be conducted to ensure similar variance across Education-Occupation groups.
- The assumption might be violated due to the unbalanced design.

Step 3: Conduct the Hypothesis Test (Two-Way ANOVA)

ANOVA Table

Source	Sum of Squares	df	F-Statistic	p-value
Education	1.94e+11	2	136.33	1.76e-12
Occupation	4.08e+08	3	0.19	0.8270
Education × Occupation	4.23e+10	6	9.91	**1.32e-05**
Residual	2.06e+10	29	-	-

- **Education**: The p-value (*1.76e-12*) is very small (< 0.05), indicating that Education has a significant effect on Salary.
- **Occupation:** The p-value (*0.8270*) is greater than 0.05, meaning that Occupation does not have a significant effect on Salary.
- **Education × Occupation Interaction:** The p-value (*1.32e-05*) is less than 0.05, meaning that there is a significant interaction between Education and Occupation on Salary.

Step 4: Conclusion

- Since the interaction effect (Education × Occupation) is significant ($p < 0.05$), we reject the null hypothesis.
- This means that the effect of Education on Salary depends on Occupation, and vice versa.
- However, due to the unbalanced group sizes and empty cells (Exec-managerial for HS-grad = 0), the results should be interpreted with caution.