

Stats Book

Gautam Shaju 21BCE1355

Prem Parikh 21BCE1427

Varun Chaturvedi 21BCE1708

Aran Agarwal 21BCE5443

OVERVIEW:

Simple, Multiple, and Subdivided (Bar Diagrams)

Simple Bar Diagrams

Definition: A simple bar diagram uses single bars at equal spacing to represent data. Each bar's height (in vertical bars) or length (in horizontal bars) corresponds to the value of the category it represents.

Insights: Allows for easy comparison of different categories, highlighting the highest and lowest values. It's straightforward and effective for displaying and comparing data across a small number of categories.

Multiple Bar Diagrams

Definition: These diagrams feature groups of bars, where each group represents a category and each bar within the group represents different sub-categories or data points for comparison.

Insights: Useful for comparing multiple sets of data across the same categories, showing trends over time or differences between groups. It provides a visual display of the variations within each category and across different categories.

Subdivided (Stacked) Bar Diagrams

Definition: Stacked bars divide each bar into segments that represent sub-categories, with the total height or length of the bar representing the aggregated value.

Insights: Offers insights into the proportion of sub-categories within each category, as well as the overall comparison across categories. It's particularly useful for understanding part-to-whole relationships and comparing the total sizes of each category.

Pie Diagram

Definition: A pie diagram (or pie chart) is a circular statistical graphic divided into slices to illustrate numerical proportion. Each slice's angle (and thus its arc length and area) is proportional to the quantity it represents.

Insights: Best for showing the relative proportions of parts to a whole in a single data set. It highlights how each category contributes to the total, making it easy to see the largest and smallest segments at a glance.

Line Diagram

Definition: A line diagram (or line chart) displays information as a series of data points called 'markers' connected by straight line segments. It's often used to visualize data over a continuous time span.

Insights: Ideal for showing trends over time, such as increases or decreases in data values. It can also depict multiple series within the same diagram, allowing for comparison of trends between different data sets.

Mean, Median, Mode, Skewness, Kurtosis

Mean

Definition: The arithmetic average of a set of values, or distribution.

Insights: Provides a central value for the data set, useful for understanding the general tendency or central location of the data.

Median

Definition: The middle value of a data set when the values are arranged in ascending or descending order.

Insights: Offers a better measure of central tendency when the data set contains outliers, as it is not as affected by extremely high or low values.

Mode

Definition: The value that appears most frequently in a data set.

Insights: Useful in understanding the most common or popular value among the data set. It's the only measure of central tendency that can be used with nominal data.

Skewness

Definition: A measure of the asymmetry of the probability distribution of a real-valued random variable.

Insights: Indicates whether the data are spread out more to the left or right of the mean. Positive skewness means a longer tail on the right, while negative skewness means a longer tail on the left.

Kurtosis

Definition: A measure of the "tailedness" of the probability distribution of a real-valued random variable.

Insights: High kurtosis indicates a distribution with heavy tails and a sharp peak, suggesting outliers are more likely. Low kurtosis indicates a distribution with light tails and a flatter peak.

Skewness indicates the asymmetry of the distribution, while Kurtosis highlights the peakedness or flatness of the distribution relative to a normal distribution. Together, they provide comprehensive insights into the distribution's shape, allowing for a nuanced understanding of data characteristics beyond central tendency and variability.

Standard Deviation, Variation, Range

Standard Deviation

Definition: A measure of the amount of variation or dispersion of a set of values.

Insights: Helps understand how spread out the values are in a data set. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates that the values are spread out over a wider range.

Variation

Definition: Often referred to as variance, it's the expectation of the squared deviation of a random variable from its mean, indicating the spread between numbers in a data set.

Insights: Provides a square of the standard deviation, offering a measure of how data points differ from the mean value. Higher variance indicates greater spread.

Range

Definition: The difference between the highest and lowest values in a data set.

Insights: Offers the simplest measure of variability. While easy to calculate, it's sensitive to outliers and may not provide a complete picture of the data's variability.

Coefficient of Variation (CV)

Definition: The Coefficient of Variation is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage and is calculated as the ratio of the standard deviation to the mean.

Insights: The CV is useful for comparing the degree of variation from one data series to another, even if the means are drastically different from each other. It is especially useful in the context of relative variability comparison across different datasets or populations where the scales of measurement are not the same.

Quartile Deviation (QD)

Definition: Quartile Deviation, also known as the semi-interquartile range, is a measure of spread that describes the spread of the middle 50% of the data. It is calculated as the difference between the upper quartile (Q3) and the lower quartile (Q1) divided by 2.

Insights: QD provides a robust measure of variability that is less sensitive to outliers and extreme values than the range or standard deviation. It gives a good indication of the spread of the central portion of the distribution and is particularly useful for skewed distributions.

Box Plot

Definition: A Box Plot, or box-and-whisker plot, is a graphical representation of the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

Insights:

The central box represents the values from Q1 to Q3, providing a visual representation of the interquartile range (IQR).

The line inside the box shows the median of the data.

Whiskers extend from the box to show the range of the data, while points outside of the whiskers can indicate outliers.

Box plots are particularly useful for identifying outliers and for comparing distributions between several groups or datasets.

Scatter Plot

Definition: A Scatter Plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Insights:

Scatter plots are used to observe relationships between variables. If the points are color-coded, an additional variable can be displayed.

They can reveal the distribution trends, concentration areas, and if there's any correlation between the two variables.

By identifying how the variables relate to each other, one can make predictions about future data points.

Coefficient of Correlation - Karl Pearson

Explanation:

The Pearson correlation coefficient, denoted as r , quantifies the degree to which two variables linearly relate to each other. Its calculation involves dividing the covariance of the two variables by the product of their standard deviations. This coefficient is sensitive to outliers, which can significantly affect the correlation value.

Insights:

Interpretation: Values of r near +1 or -1 indicate a strong linear relationship, while values close to 0 suggest a weak or no linear relationship. The sign indicates the direction of the relationship (positive or negative).

Application: Widely used in finance, healthcare, and social sciences to understand and predict relationships between variables. For example, in finance, it may be used to understand the relationship between risk and return for a particular investment.

Coefficient of Correlation - Spearman Rank Correlation

Explanation:

The Spearman Rank Correlation, denoted as ρ or r_s , assesses how well the relationship between two variables can be described by a monotonic function. Unlike Pearson's r , Spearman's ρ does not require the assumption of normal distribution and is less affected by outliers and skewed distributions.

Insights:

Interpretation: Similar to Pearson's r , Spearman's ρ values near +1 or -1 signify a strong monotonic relationship. However, it's crucial in scenarios where the relationship is not linear but consistently increases or decreases.

Application: Useful in ordinal data or when evaluating non-linear relationships, such as ranking preferences in psychological tests or order-based scenarios in marketing research.

Simple Linear Regression and Multiple Linear Regression

Simple Linear Regression:

Simple linear regression not only predicts the dependent variable based on the independent variable but also provides coefficients (slope and intercept) that quantify the relationship. The slope (b) indicates how much Y changes for a one-unit change in X . The intercept (a) represents the predicted value of Y when X is zero.

Multiple Linear Regression:

Multiple linear regression extends the simple linear model to include multiple independent variables. This allows for a more detailed analysis, considering the effect of several factors on the dependent variable simultaneously. It's crucial for models where interactions between different independent variables influence the outcome.

Insights:

Model Fit: The goodness of fit, often assessed by R^2 , indicates how well the regression model explains the variability of the dependent variable. A higher R^2 value suggests a better fit.

Predictive Power: Regression analysis is not just descriptive but also predictive. It enables forecasting future values of the dependent variable based on the known values of the independent variables.

Single Mean - Z test and t-test

Z test:

The Z test's application extends beyond comparing means to proportion tests and variance tests under certain conditions. The normal distribution assumption is crucial, and for real-world applications, the Central Limit Theorem often justifies its use with larger sample sizes.

t-test for a Single Mean:

The t-test adjusts for small sample sizes through the degrees of freedom associated with the t-distribution. The degrees of freedom affect the test's critical values, with smaller sample sizes resulting in a wider distribution and thus a more conservative test.

Insights:

Effect Size: Beyond significance, effect size (e.g., Cohen's d for t-tests) provides information on the magnitude of the difference or relationship, offering practical significance.

Assumptions: Each test comes with assumptions (normality, independence, variance homogeneity). Violations of these assumptions can lead researchers to use alternative statistical methods, such as non-parametric tests.

Two Mean - Z test (Equal Variance) and t-test (Equal Variance)

Two Mean - Z test:

The assumption of equal variances is critical for the accuracy of the Z test. In practice, this assumption is tested using Levene's test or F-test. When variances are unequal, alternative methods or adjustments are needed to avoid misleading conclusions.

Two Mean - t-test (Equal Variance):

The pooled variance estimate is used in this t-test to better estimate the common variance of the two populations. This test is particularly sensitive to the equal variance assumption, and Welch's t-test is often recommended when this assumption is violated.

Insights:

Power Analysis: Conducting a power analysis before the test can determine the sample size needed to detect an effect of a certain size with a given level of confidence. This is crucial for designing studies with adequate sensitivity.

Inter-group Differences: These tests are foundational for experimental designs where the impact of interventions is measured by comparing group means. Understanding how and when to apply these tests directly influences the reliability of conclusions drawn from experimental data.

Paired t-test

Explanation:

The paired t-test accounts for the natural pairing in the data, reducing variability due to extraneous factors. This makes the paired t-test more powerful than two independent samples tests for detecting a difference between the means of the paired samples.

Insights:

Dependency: The test assumes that the pairs are dependent (i.e., there is a meaningful connection between the observations in each pair), which is a significant consideration in its application.

Application Scenarios: Commonly used in crossover studies, before-and-after studies (e.g., pre-test/post-test designs), and matched case-control studies.

Chi-Square Distribution

Definition: The Chi-Square distribution is a statistical distribution widely used in hypothesis testing, particularly in tests of independence and goodness of fit. It is defined by the sum of the squares of k independent standard normal random variables, where k represents the degrees of freedom. The shape of the Chi-Square distribution depends on the degrees of freedom: as the degrees of freedom increase, the distribution becomes more symmetrical.

Applications:

Goodness of Fit Test: Determines how well observed data fit the expected distribution in categorical data. For example, it can test if a dice is fair by comparing the observed frequencies of each face with the expected frequencies.

Test of Independence: Assesses whether there is a significant association between two categorical variables in a contingency table. This is pivotal in fields like market research and epidemiology to explore relationships between categorical variables (e.g., gender and product preference).

Insights:

- The Chi-Square test provides a method to quantify the discrepancy between observed and expected data. A high Chi-Square statistic suggests a significant difference between the observed and expected data, indicating that not all differences can be attributed to chance.
- It's crucial for understanding the relationship dynamics between categorical variables, helping to identify patterns or associations that may not be immediately apparent.

ANOVA (Analysis of Variance)

One-Way ANOVA

Definition: One-Way ANOVA is a statistical test that compares the means of three or more independent (unrelated) groups to determine if at least one group mean is significantly different from the others based on a single independent variable (factor). It generalizes the t-test to more than two groups.

Applications:

- Widely used in experimental designs where the effects of a single factor (e.g., treatment type) on a continuous outcome variable are investigated across multiple groups.
- Essential in fields such as agriculture, where it might be used to compare the yields of different varieties of a crop, or in pharmaceuticals, to assess the effectiveness of different drugs.

Insights:

- One-Way ANOVA helps in determining the impact of a single factor while controlling for the variability within groups. It essentially decomposes the total variance observed in the data into the variance between groups and within groups.
- The F-statistic, derived from the ANOVA, allows researchers to test the null hypothesis that all group means are equal. A significant F-statistic suggests that at least one group mean is significantly different, warranting further investigation.

Two-Way ANOVA

Definition: Two-Way ANOVA extends the One-Way ANOVA to examine the influence of two independent variables (factors) on a dependent variable simultaneously. This method assesses not only the main effects of each factor but also whether there is an interaction between the factors affecting the dependent variable.

Applications:

- Useful in complex experiments where the effects of two factors are of interest. For example, in clinical trials, researchers might investigate the effect of medication dosage and therapy type on patient outcomes.
- Allows for the analysis of interaction effects, which is critical when the effect of one factor depends on the level of another factor.

Insights:

- Two-Way ANOVA provides a comprehensive view of how multiple factors and their interactions influence the outcome variable. It's invaluable for identifying synergistic or antagonistic effects between factors.

- The analysis helps in understanding not only the individual effects of each factor but also how these factors work together, offering deeper insights into complex phenomena that cannot be achieved through simpler analyses.

STATS BOOK

HR DATASET - Unveiling the Secrets of Placement Success: A Data-Driven Analysis of University Student Performance and Employability

Introduction

In today's competitive job market, securing a coveted placement after graduation is a primary concern for university students. Business schools, particularly, strive to equip their students with the knowledge, skills, and experiences necessary to navigate this challenging landscape. This case study delves into the data of a hypothetical cohort from a leading business school to identify key factors influencing placement success. By employing diverse statistical techniques, the case explores the intricate web of relationships between academic performance, extracurricular involvement, skills, and ultimately, landing that crucial first job.

Data Description

The data set under scrutiny encompasses a diverse range of student attributes:

- **Academic Performance:** CGPA (Cumulative Grade Point Average)
- **Extracurricular Involvement:** Internships (Yes/No), Projects (number), SSC_Marks (Secondary School Certificate marks), HSC_Marks (Higher Secondary Certificate marks)
- **Skills:** Aptitude Test Score, Soft Skills Rating
- **Placement Status:** Placed/Not Placed

Methodology

To unearth the secrets of placement success, a multifaceted approach incorporating various statistical techniques will be employed:

1. Descriptive Statistics:

- **Mean, Median, Standard Deviation:** These measures will be calculated for CGPA, Aptitude Test Score, and Soft Skills Rating, both for placed and not-placed students. This provides a basic understanding of the **central tendency (average)** and **dispersion (spread)** of these variables within each group.

Significance: Identifying the **mean CGPA** of placed and not-placed students helps understand the average academic performance associated with placement. The **median** offers an alternative perspective, reflecting the performance of the "middle student" in each group and potentially revealing the presence of outliers not captured by the mean. The **standard deviation** indicates the **variability** in each group, showcasing the range of performance within each category.

2. Inferential Statistics:

- **Coefficient of Correlation:** This statistic measures the **strength and direction** of the relationship between Aptitude Test Score, Soft Skills Rating, and placement status. A positive correlation indicates that higher scores are associated with a higher likelihood of placement, and vice versa.

Significance: The coefficient of correlation helps quantify the **association** between variables. A value close to 1 indicates a strong positive correlation, suggesting that higher scores tend to coincide with placement success. Conversely, a value close to -1 indicates a strong negative correlation, implying that lower scores are more likely to be associated with not securing a placement.

- **t-test:** This test compares the means of **two independent groups** (placed vs. not-placed) for variables like CGPA, SSC_Marks, and HSC_Marks. A statistically significant difference in the means would suggest that these variables **influence placement outcomes**.

Significance: The t-test determines if the observed differences in the means between the two groups are likely due to chance or if a **meaningful relationship** exists. If the t-test shows a statistically significant difference, it suggests that, on average, students in the placed group performed differently on that variable compared to the not-placed group.

3. Predictive Modeling:

- **Simple Linear Regression:** This technique builds a model to **predict placement** based on a **single independent variable**, such as CGPA. This initial

model provides a basic understanding of the relationship between the chosen variable and placement.

Significance: Simple linear regression helps **visualize and quantify** the relationship between a single independent variable and the dependent variable (placement status). It can identify the **direction and strength** of the association, allowing us to see how changes in CGPA, for example, might impact the likelihood of placement.

- **Multiple Linear Regression:** This more complex model incorporates **multiple independent variables** (CGPA, Aptitude Test Score, Soft Skills Rating, etc.) to predict placement status. This model aims to provide a **more comprehensive picture** of the factors influencing placement outcomes.

Significance: Multiple linear regression offers a **more nuanced understanding** of the factors influencing placement by considering the **combined impact** of various variables. It helps assess the **relative importance** of each variable in predicting the dependent variable and provides a more accurate prediction of placement based on a combination of factors.

Connecting to Economics:

The findings of this study have significant **economic implications** for both students and universities.

- **Human Capital Theory:** Placement success can be viewed through the lens of **human capital theory**, which posits that skills and knowledge acquired through education and experience contribute to an individual's earning potential. Higher levels of academic performance, strong test scores, and relevant skills developed through internships and projects can be seen as investments in human capital, potentially leading to higher placement rates and potentially higher future earnings.
- **Labor Market Signaling:** The findings can also be viewed through the lens of **labor market signaling theory**, which suggests that educational credentials and test scores act as signals

STORY ARC :

"Pathways to Success: Navigating the Placement Maze"

Act 1: The Academic Ascent

- Introduction to the vibrant world of a business school, where students embark on their academic journey.
- Meet a diverse group of students, each with a unique academic profile, and learn about their hopes and aspirations.
- Highlight the significance of CGPA and academic achievements in shaping the initial phase of their career paths.

Act 2: Beyond the Classroom Walls

- Introduce the element of extracurriculars – internships, projects, and soft skills.
- Follow students as they navigate through the challenges of balancing academics with real-world experiences.
- Showcase the passion and drive that some students invest in internships and projects, while others focus solely on academic excellence.

Act 3: The Testing Grounds

- Dive into the world of aptitude tests, a critical checkpoint for students on their journey to securing placements.
- Explore how the students prepare for these tests and the impact of their scores on their placement outcomes.
- Examine the correlation between aptitude test scores and overall preparedness for the corporate world.

Act 4: The Training Grounds

- Uncover the role of placement training programs in honing students' skills for the professional arena.
- Witness the transformation of students as they undergo rigorous training to enhance their communication, interview, and presentation skills.
- Reflect on the varying degrees of participation in these programs and their influence on the final lap of the placement race.

Act 5: The Crossroads

- Explore the pivotal moment when students face the dichotomy of placement results.
- Engage with the emotional rollercoaster as some celebrate securing coveted positions, while others grapple with disappointment.

- Investigate the factors that contribute to the divergence in placement outcomes despite similar academic and extracurricular backgrounds.

Act 6: Reflections and Realizations

- Delve into the introspective phase where students reflect on their journey, acknowledging the highs and lows.
- Unearth the lessons learned, both academically and personally, as students contemplate their futures.
- Highlight the resilience and adaptability demonstrated by those who faced setbacks, portraying the human side of the placement process.

Act 7: The Epiphany

- Showcase the students' newfound clarity as they realize the diverse paths to success.
- Explore the realization that success is not a linear journey and that the definition of a successful career varies for each individual.
- Conclude with a sense of optimism as students embark on their post-graduation journeys, equipped with a deeper understanding of their strengths and areas for growth.

Conclusion:

The story arc encapsulates the multifaceted journey of students navigating the complex terrain of placements. It brings to light the intertwining factors of academics, extracurriculars, testing, and training that contribute to their unique narratives. Ultimately, it is a tale of resilience, self-discovery, and the myriad pathways to success in the ever-evolving landscape of career development.

HR DATASET insights -

Insights:

Placement Status:

- The majority of students are placed, as seen in the pie chart of "PlacementStatus." Approximately 61.7% of students are placed, while the remaining 38.3% are not placed.

CGPA and Aptitude Test Score Trends:

- The line chart for "CGPA" and "AptitudeTestScore" provides insights into the distribution and trends among students.
- CGPA varies across students, with some showing consistently high scores, while others have fluctuating scores.
- Aptitude Test Scores also exhibit variations, indicating diverse levels of performance in this area.

Internships and Projects:

- The dataset includes columns for "Internships" and "Projects," which could provide additional insights into students' practical experience.
- Further analysis of the impact of internships and projects on placement status may provide valuable insights.

Soft Skills Rating:

- "SoftSkillsRating" is another column that could be explored for its correlation with placement status.
- Higher soft skills ratings might influence placement outcomes, but a detailed analysis is needed.

Placement Training:

- The presence of "PlacementTraining" as a categorical variable suggests the inclusion of placement training in the dataset.
- Investigating the impact of placement training on placement status could be insightful.

Educational Background:

- The dataset includes "SSC_Marks" and "HSC_Marks," representing secondary and higher secondary education.
- Analyzing the correlation between academic performance and placement status may reveal interesting patterns.

Potential for Further Analysis:

- Further analysis can be conducted to explore the influence of individual factors such as internships, projects, soft skills, and academic performance on placement status.
- Statistical tests or machine learning models could be employed to identify significant predictors of placement.

Placed vs. Not Placed Comparison:

- Comparisons between the profiles of placed and not placed students in terms of CGPA, aptitude test scores, and other factors could reveal characteristics associated with successful placements.

CH 1 - Simple, Multiple and Subdivided (Bar Diagram)

Simple, Multiple and Subdivided Bar Diagram

In this section of the Harvard case study statistics book, we delve into the fundamental yet powerful visualization tools that statisticians and data analysts use to communicate data findings effectively: Simple, Multiple, and Subdivided Bar Diagrams. Each of these tools offers a unique perspective on the data, catering to various analytical needs and storytelling in research and business contexts.

Simple Bar Diagrams: The Foundation of Data Visualization

At its core, a Simple Bar Diagram is a graphical representation of data where individual categories or groups are represented by bars, with the length or height of each bar proportional to its value. This simplicity belies its power; by transforming numerical data into visual form, bar diagrams facilitate immediate comprehension of relationships, differences, and trends.

Application in Case Studies:

Consider a case where a business aims to compare its sales across different quarters of a financial year. A simple bar diagram can vividly showcase the fluctuations in sales, allowing stakeholders to quickly identify periods of growth or decline. This visual clarity is invaluable for making informed decisions, such as allocating resources or adjusting marketing strategies.

Insights:

- Color Coding: Enhancing simple bar diagrams with color coding can further improve readability and impact. By assigning different colors to bars representing different quarters, viewers can more easily follow trends and comparisons.
- Annotations: Adding annotations directly onto the bars, such as percentage changes or significant milestones, can turn a basic diagram into a storytelling tool, providing context and insights at a glance.

Multiple Bar Diagrams: A Comparative Lens

Multiple Bar Diagrams elevate the utility of simple bar diagrams by allowing the comparison of multiple data series within the same category. By placing bars side-by-side within each category, these diagrams enable a direct comparison between different groups, variables, or time periods.

Application in Case Studies:

In a scenario where a corporation wishes to analyze the performance of three different products across several regions, a multiple bar diagram serves as an excellent tool. It not only highlights which products are performing well but also indicates regional market preferences or deficiencies.

Insights:

- Interactive Elements: In digital reports, incorporating interactive elements such as hover text or clickable bars can enrich the user experience, offering more detailed data on demand.
- Clustered vs. Stacked: Opting for clustered bars for direct comparisons while using stacked bars (a variation of multiple bar diagrams) can offer insights into the composition of the data alongside comparisons.

Subdivided (Stacked) Bar Diagrams: Unveiling Composition and Trends

Subdivided Bar Diagrams, or Stacked Bar Diagrams, are a variant where each bar is segmented into sub-categories, showing the total and the composition of the total in a single bar. This method is particularly useful for understanding how various parts contribute to the whole across different categories or over time.

Application in Case Studies:

Imagine a study aimed at analyzing the sources of revenue for a multinational company. A subdivided bar diagram can efficiently display how different revenue streams (such as product sales, services, and licensing fees) contribute to the total revenue across different regions or over sequential quarters.

Insights:

- Gradient Shading: Utilizing gradient shading for the segments of each bar can indicate intensity or concentration, adding depth to the visual representation of data.
- Nested Diagrams: For complex datasets, embedding smaller, detailed bar diagrams within each segment of a subdivided bar can offer readers a micro-view within the macro-analysis, revealing underlying patterns or anomalies.

CASE STUDY 1 - HR

Python

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
data=pd.read_csv("merged.csv")
# Set the aesthetic style of the plots
sns.set_style("whitegrid")

# Simple Bar Diagram - Placement Status
plt.figure(figsize=(8, 4))
sns.countplot(data=data, x='PlacementStatus')
plt.title('Placement Status Count')
plt.xlabel('Placement Status')
plt.ylabel('Count')
```

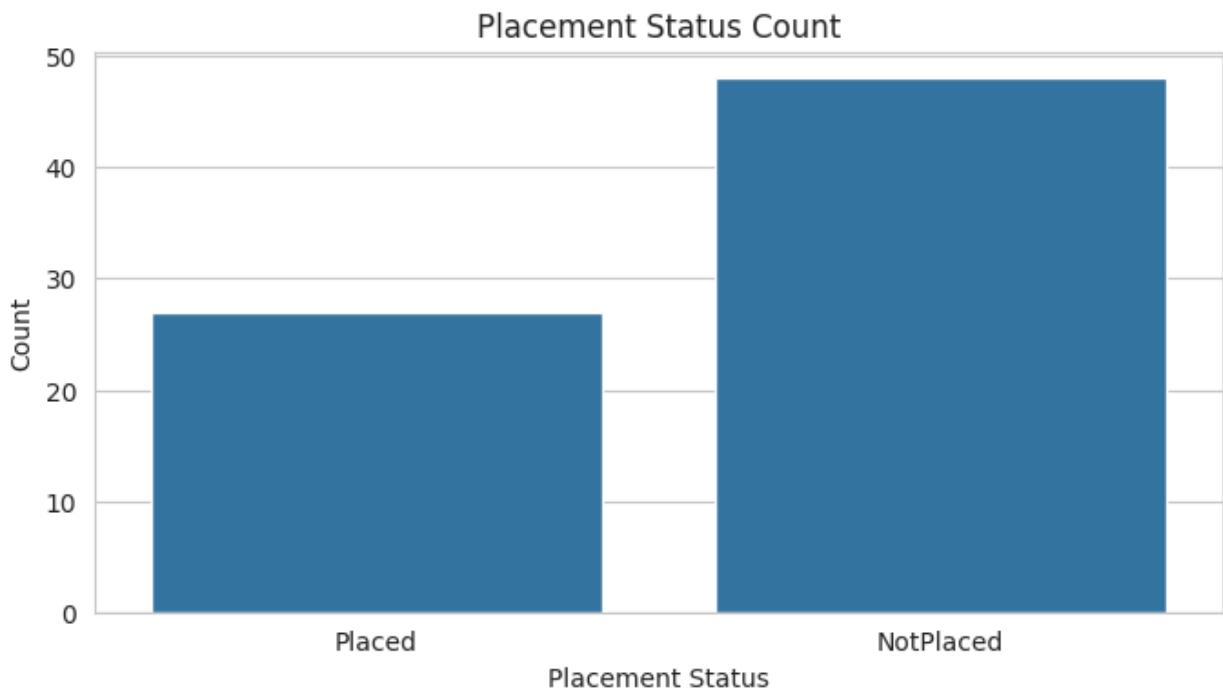
```

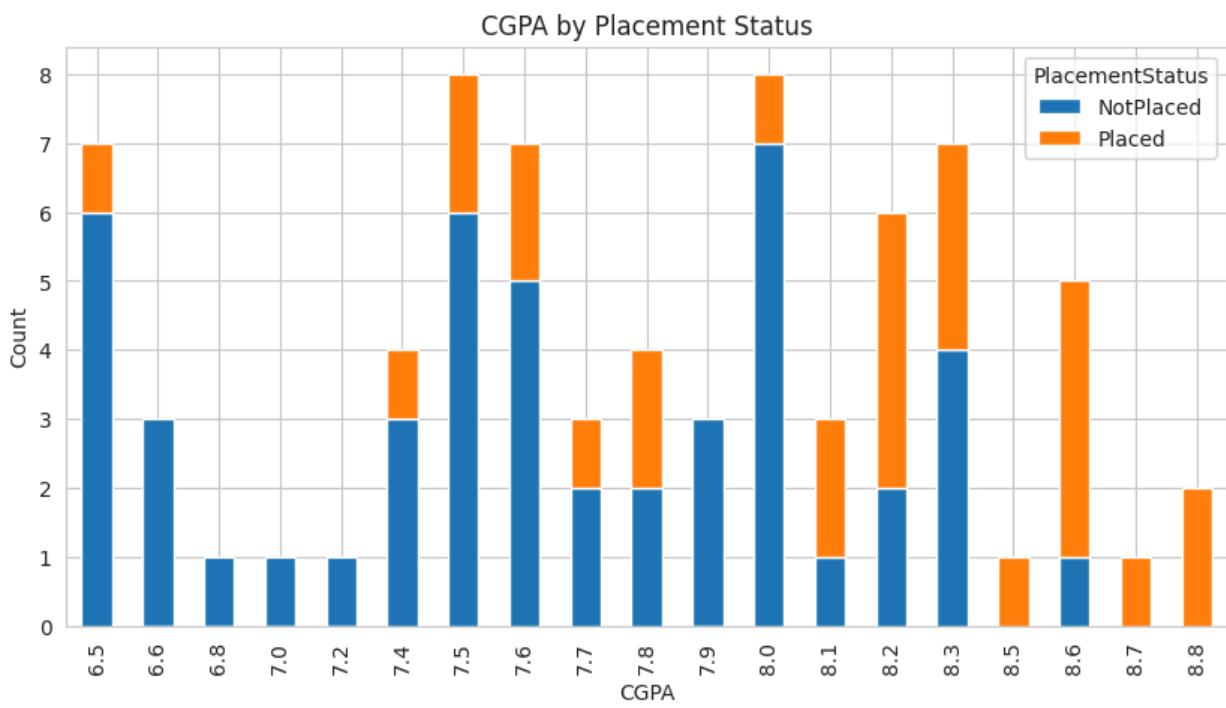
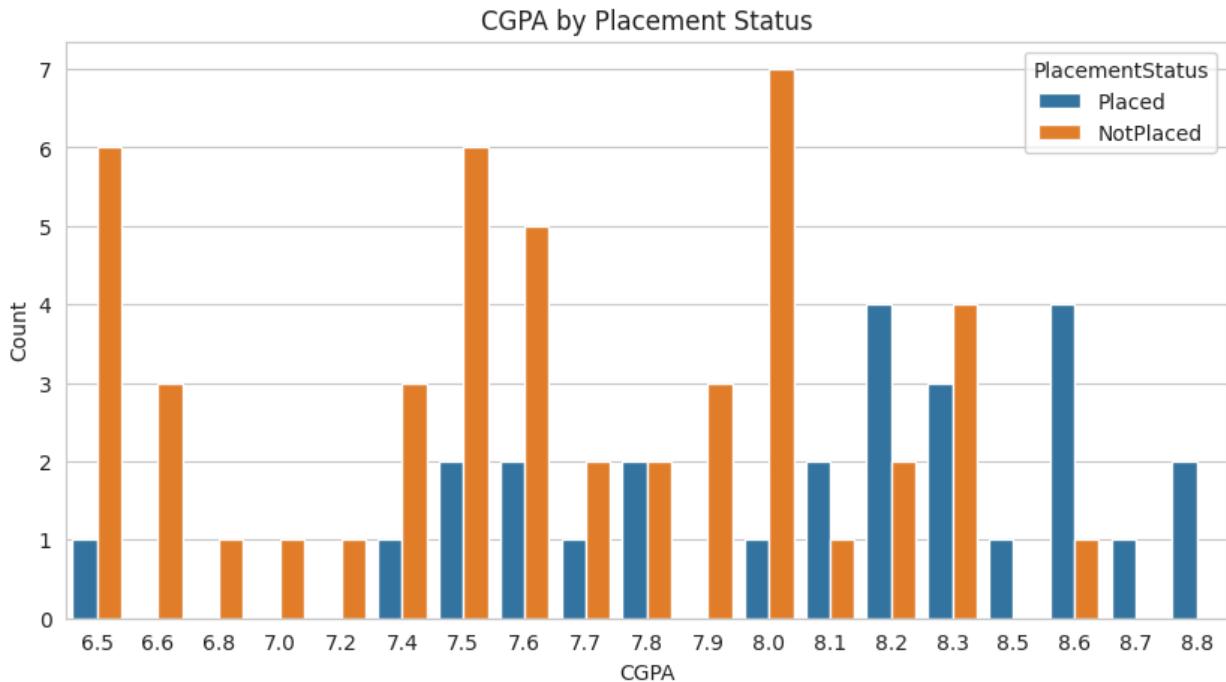
plt.show()

# Multiple Bar Diagram - Projects by Placement Status
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='CGPA', hue='PlacementStatus')
plt.title('CGPA by Placement Status')
plt.xlabel('CGPA')
plt.ylabel('Count')
plt.show()

# Subdivided Bar Diagram - Internships by Placement Training
pd.crosstab(data['CGPA'], data['PlacementStatus']).plot(kind='bar',
stacked=True, figsize=(10, 5))
plt.title('CGPA by Placement Status')
plt.xlabel('CGPA')
plt.ylabel('Count')
plt.show()

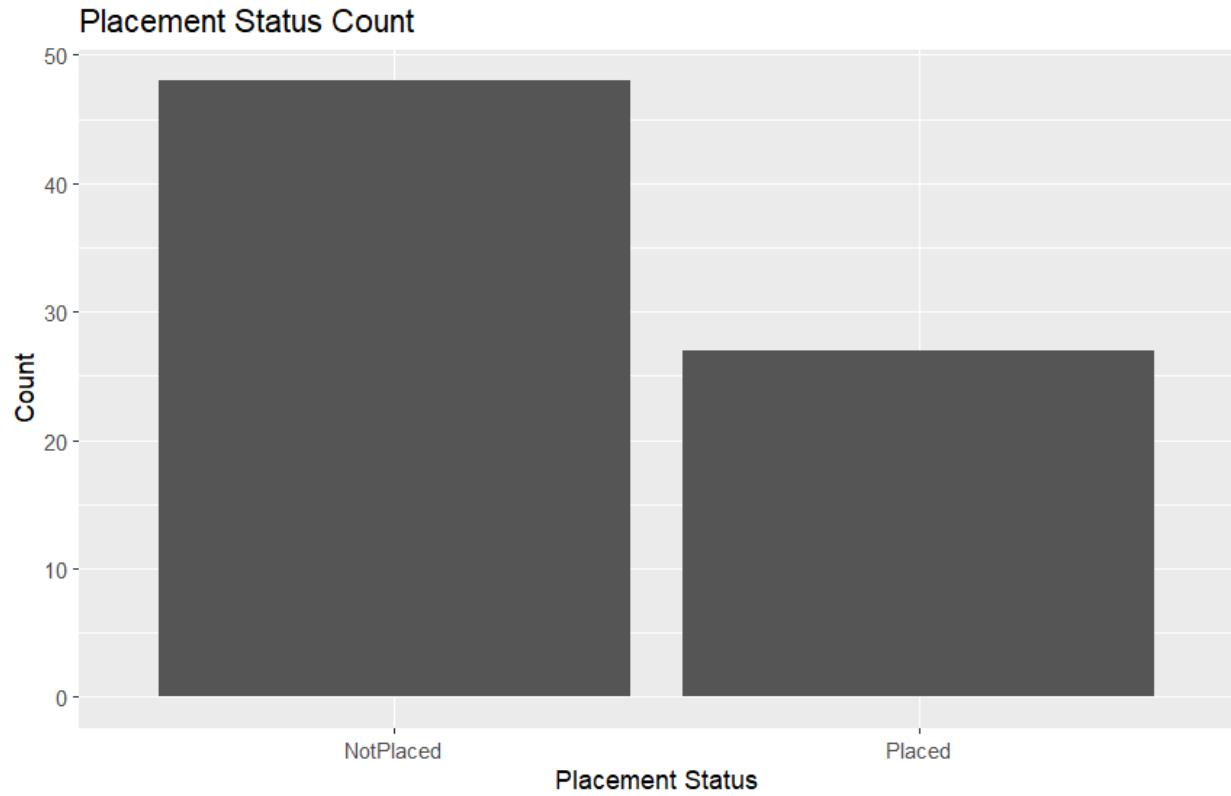
```





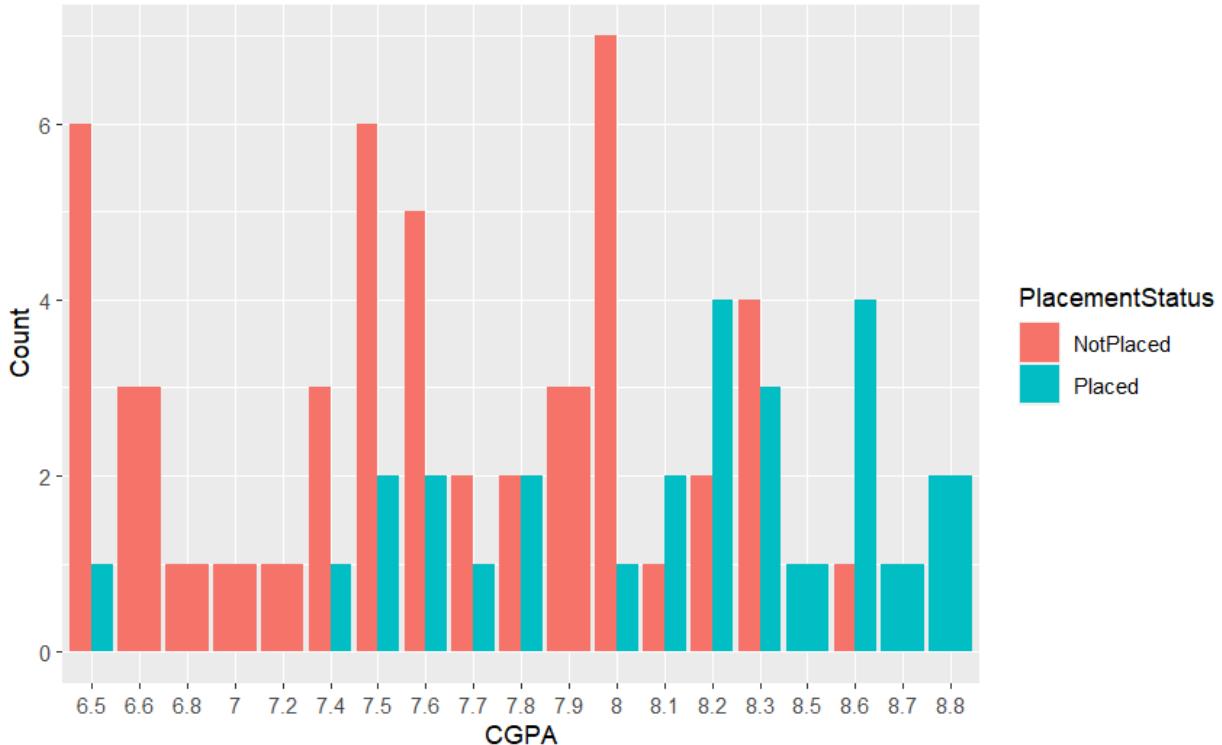
R

```
#----- Visualization(Simple subdivident)
> library(ggplot2)
> ggplot(data, aes(x=factor(PlacementStatus)) + geom_bar() +
  labs(title="Placement Status Count", x="Placement Status", y="Count")
>
> ggplot(data, aes(x=factor(AptitudeTestScore), fill=PlacementStatus)) +
  geom_bar(position="dodge") + labs(title="Internships by Placement Status",
  x="Number of Internships", y="Count")
>
> ggplot(data,aes(x = PlacementStatus, y= AptitudeTestScore)) +
  + geom_bar(stat="identity" )
>   labs(title="Internships by Placement Status", x="Aptitude Score",
  y="Count")
$  
x  
[1] "Aptitude Score"  
  
$y  
[1] "Count"  
  
$title  
[1] "Internships by Placement Status"  
  
attr(),"class")
[1] "labels"
```



```
> ggplot(data, aes(x=factor(CGPA), fill=PlacementStatus)) +  
  geom_bar(position="dodge") + labs(title="CGPA by Placement Status", x="CGPA",  
  y="Count")
```

CGPA by Placement Status



Excel

Bar Diagram

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Column" or "Bar" button in the Charts group.

Step 4: Choose the desired bar chart type from the dropdown menu.

Multiple Bar Graph:

Step 1: Organize your data: Place categories in the first column and data series in adjacent columns. Step 2: Select your data range, including headers. Step 3: Go to the "Insert" tab on the Excel ribbon. Step 4: In the "Charts" group, click on the "Insert Column or Bar Chart" button. Step 5: Select "Clustered Bar" or "Stacked Bar" depending on your preference. Step 6: Excel will create the chart. You can then customize it: Step 7: Click on the chart title to edit it. Step 8: Use the "Chart Elements" button to add or remove elements like legends or data labels. Step 9: Use the "Chart Styles" and "Chart Filters" buttons to change the appearance and data selection.

Subdivided Bar Diagram (using Pivot Chart)

Step 1: Select your data range.

Step 2: Go to the Insert tab and click "PivotTable".

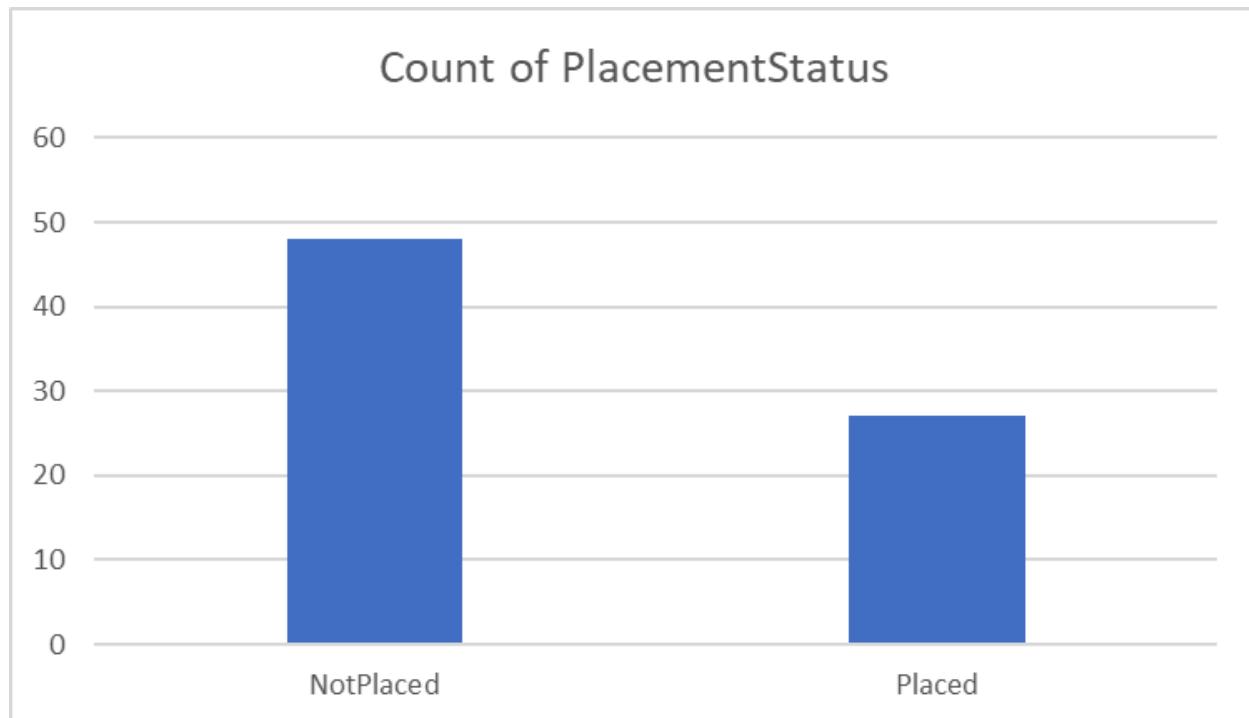
Step 3: In the PivotTable Fields pane, drag your category field to the Rows area.

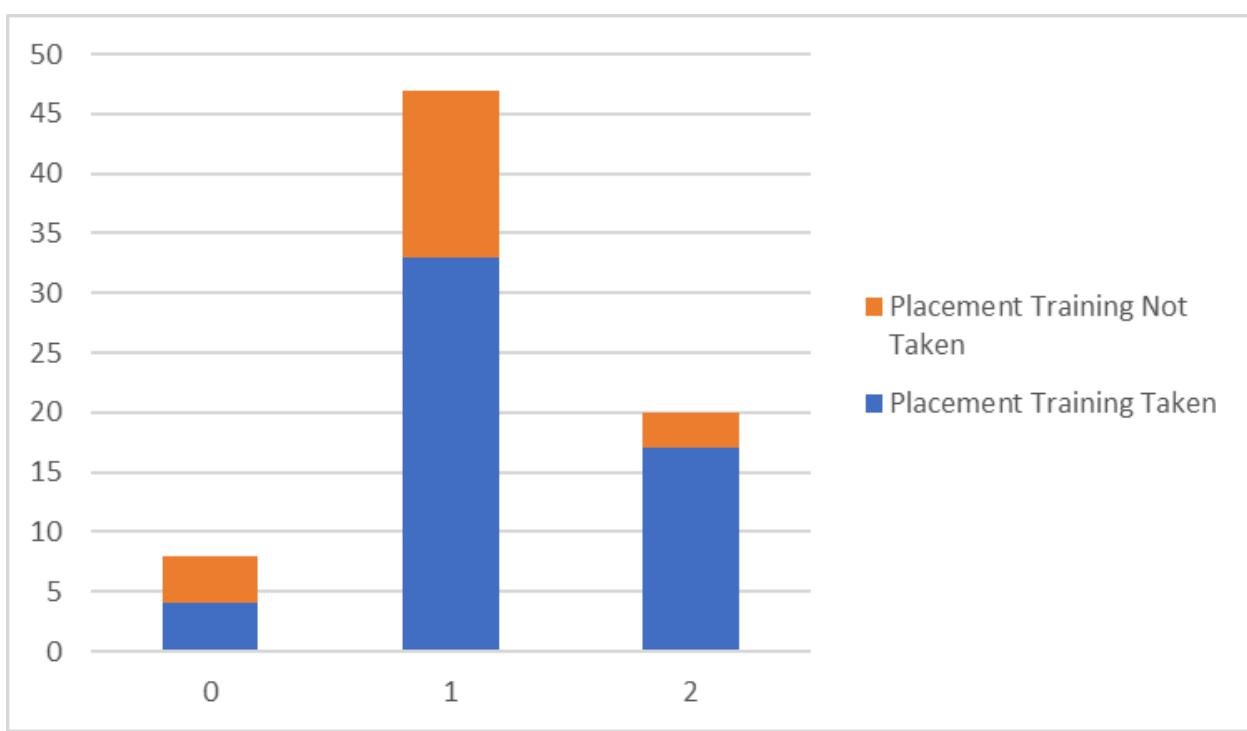
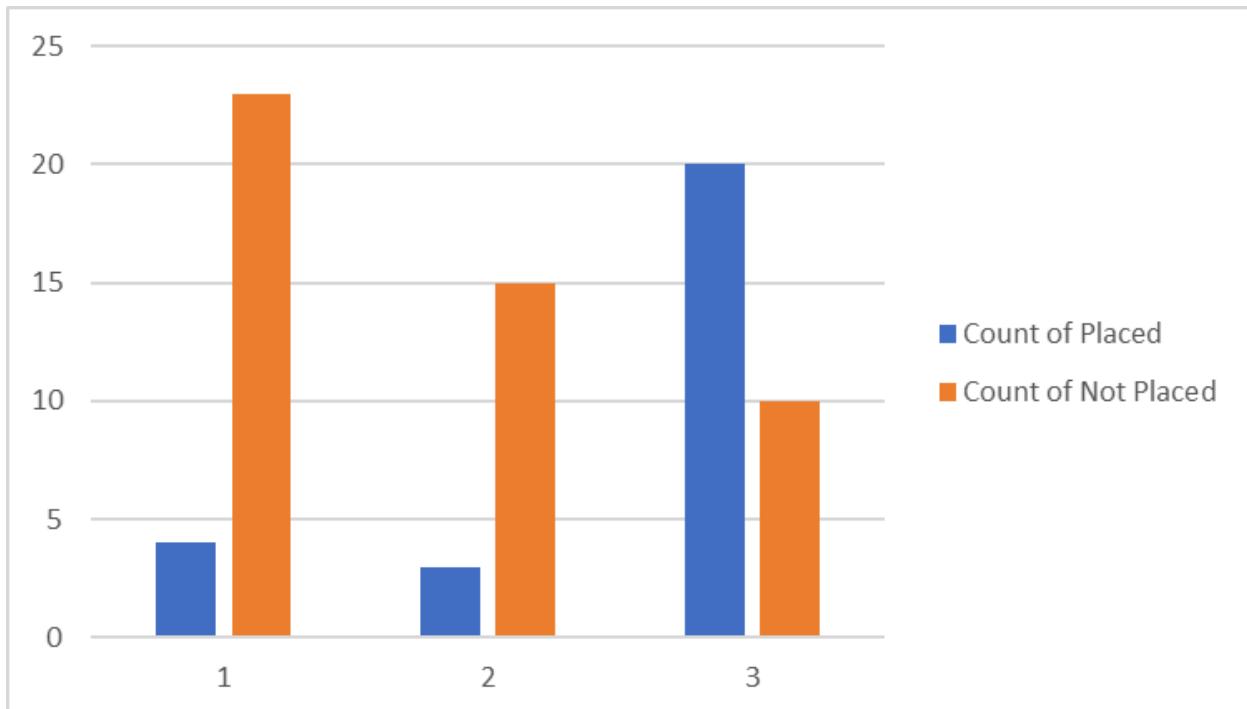
Step 4: Drag your subcategory field to the Columns area.

Step 5: Drag your value field to the Values area.

Step 6: With the PivotTable selected, go to the PivotTable Analyze tab.

Step 7: Click "PivotChart" to create a stacked bar chart.



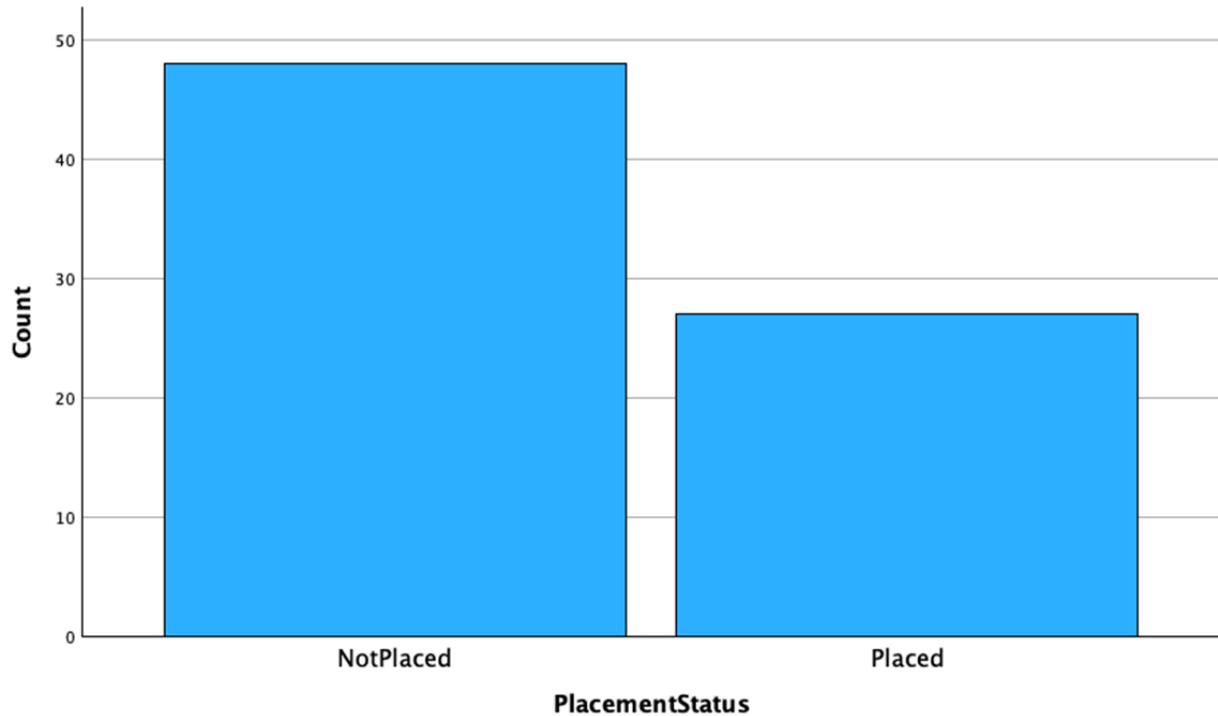


SPSS

Steps

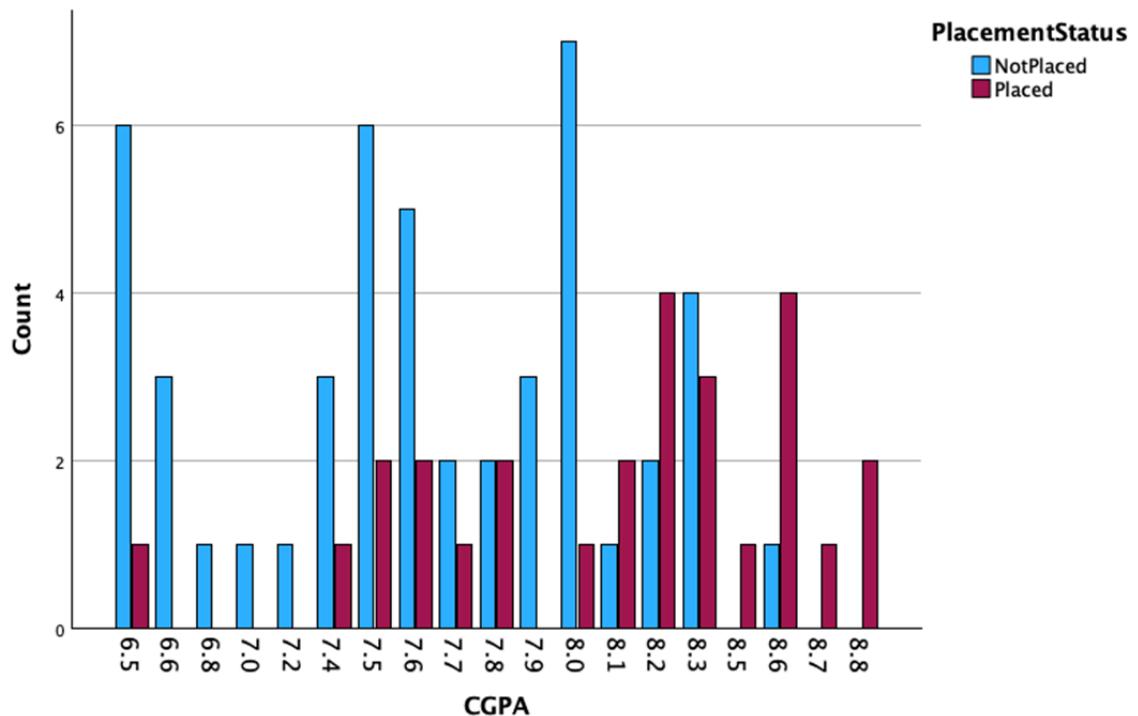
1. Go to Graphs > Chart Builder.

2. Select the Bar option.
3. Drag the appropriate variables to the x-axis (category) and y-axis (value).

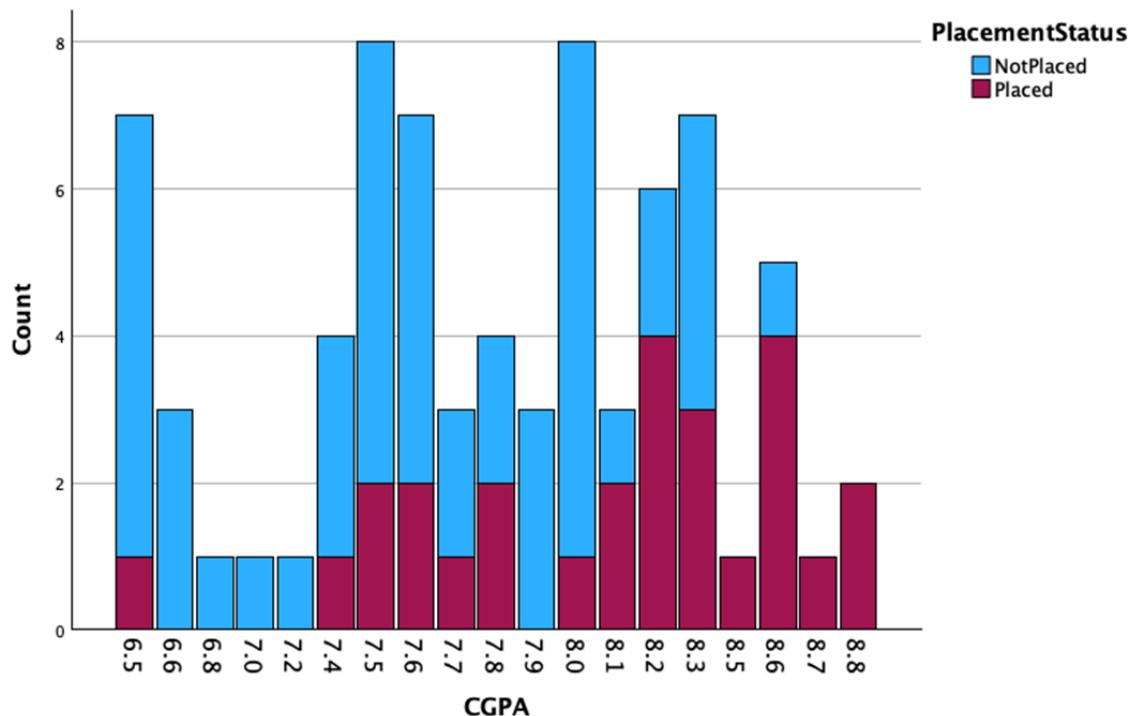


Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a clustered bar chart.
3. Drag the categorical variables to the x-axis and grouping variable to the cluster



1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a stacked bar chart.
3. Drag the categorical variables to the x-axis and sub-categories to the stacking



Practical Output

Interpretation: The college has a higher number of students who are not placed compared to those who are placed. This suggests a need to investigate the factors contributing to the lower placement rate.

Graph 2: Placement Status by CGPA (with finer CGPA intervals)

Low CGPA (6.5-7.0): Majority of students with a CGPA in this range are not placed.

• **Mid CGPA (7.1-8.0):** There is a mix, but still a significant number of students are not placed.

• **High CGPA (8.1-8.8):** More students are placed as the CGPA increases, but there are still some not placed.

Interpretation: Students with higher CGPAs tend to have a better placement rate. However, even among high CGPA students, there are those who are not placed. This suggests that while CGPA is an important factor, there might be other influencing factors such as soft skills, interview performance, and participation in extracurricular activities.

Graph 3: Placement Status by CGPA (with broader CGPA intervals)

This stacked bar chart also shows the placement status by CGPA, but with broader intervals, making it easier to see the overall trends.

• **Low CGPA (6.5-7.4):** A significant majority of students are not placed.

• **Mid CGPA (7.5-8.0):** There is a more balanced distribution between placed and not placed students.

• **High CGPA (8.1-8.8):** The majority of students are placed, with fewer not placed.

Interpretation: This graph reiterates that higher CGPA students have a better chance of being placed. The broader intervals provide a clearer view of the general trend, showing that placement rates improve with higher CGPA.

Recommendations for the HR Department:

1. **Focus on Skill Development:** Given that high CGPA students still face placement challenges, it may be beneficial to offer additional training in soft skills, interview techniques, and practical experience through internships.

2. **Career Counseling:** Implement more robust career counseling services to help students better prepare for placements.

3. **Employer Engagement:** Strengthen relationships with potential employers to understand their requirements and better align student training programs.

4. **Alumni Networking:** Utilize the alumni network to create more placement opportunities and mentorship programs for current students.

Practical Outputs

Graphs: Focus on improving placement rates, especially for students with lower CGPAs. Implement targeted skill development programs.

CASE STUDY 2 - Marketing

Python

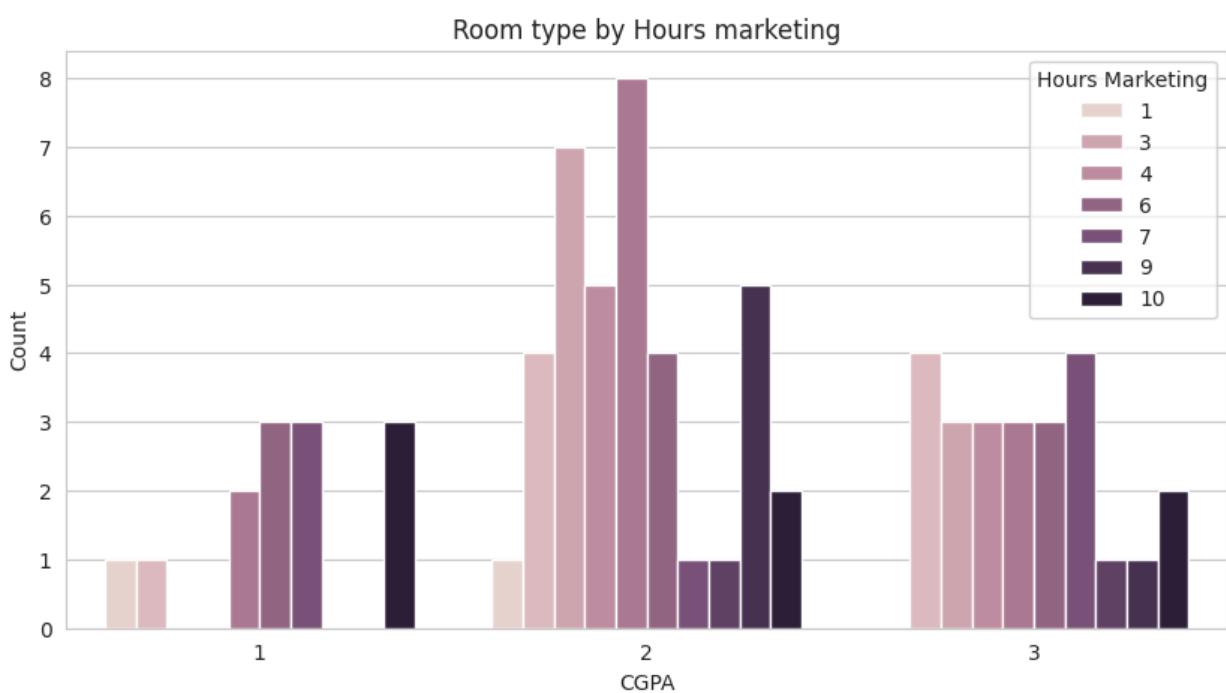
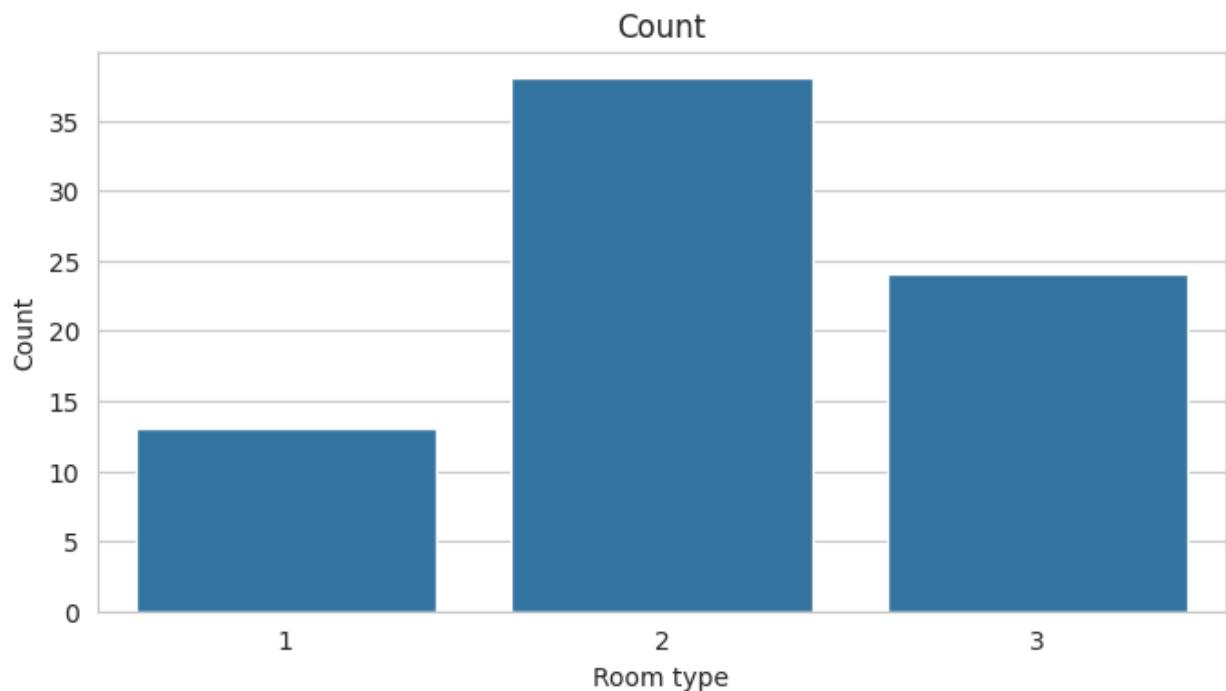
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

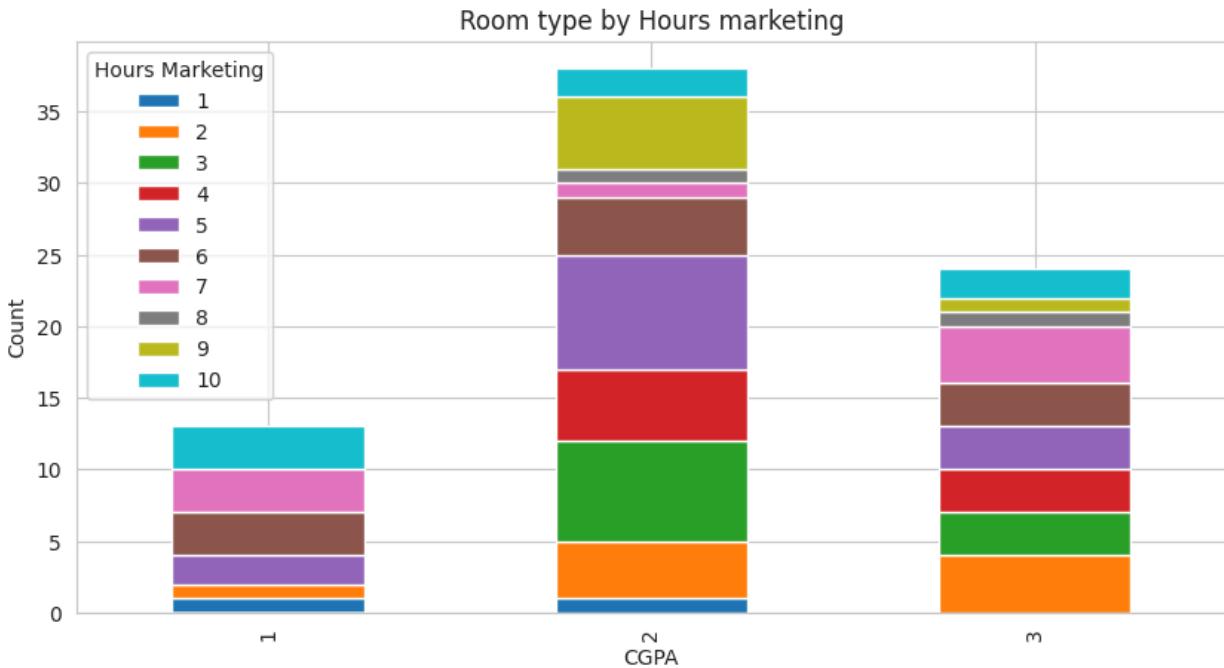
```
data=pd.read_csv("merged.csv")
# Set the aesthetic style of the plots
sns.set_style("whitegrid")

# Simple Bar Diagram - Placement Status
plt.figure(figsize=(8, 4))
sns.countplot(data=data, x='Room Type')
plt.title('Count')
plt.xlabel('Room type')
plt.ylabel('Count')
plt.show()

# Multiple Bar Diagram - Projects by Placement Status
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='Room Type', hue='Hours Marketing')
plt.title('Room type by Hours marketing')
plt.xlabel('CGPA')
plt.ylabel('Count')
plt.show()

# Subdivided Bar Diagram - Internships by Placement Training
pd.crosstab(data['Room Type'], data['Hours Marketing']).plot(kind='bar',
stacked=True, figsize=(10, 5))
plt.title('Room type by Hours marketing')
plt.xlabel('Room type')
plt.ylabel('Count')
plt.show()
```





R

```

> library(ggplot2)
> library(dplyr)
>
> # Read the data
> data <- read.csv("C:/Users/HP/Downloads/STATS BOOK - Merged (3).csv")
> df<- data.frame(data$Hours.Marketing)
>
> #----- Visualization(Simple subdivident)
> library(ggplot2)
> ggplot(data, aes(x=factor('Room Type'))) + geom_bar() + labs(title="Room type
Count", x="Room Type", y="Count")
>
> ggplot(data, aes(x=factor('Room Type'), fill=Hours.Marketing)) +
geom_bar(position="dodge") + labs(title="Room Type by hours marketed", x="Room
Type", y="Count")

```

Room Type by hours marketed



Excel

Bar Diagram

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Column" or "Bar" button in the Charts group.

Step 4: Choose the desired bar chart type from the dropdown menu.

Multiple Bar Graph:

Step 1: Organize your data: Place categories in the first column and data series in adjacent columns. Step 2: Select your data range, including headers. Step 3: Go to the "Insert" tab on the Excel ribbon. Step 4: In the "Charts" group, click on the "Insert Column or Bar Chart" button. Step 5: Select "Clustered Bar" or "Stacked Bar" depending on your preference. Step 6: Excel will create the chart. You can then customize it: Step 7: Click on the chart title to edit it. Step 8: Use the "Chart Elements" button to add or remove elements like legends or data labels. Step 9: Use the "Chart Styles" and "Chart Filters" buttons to change the appearance and data selection.

Subdivided Bar Diagram (using Pivot Chart)

Step 1: Select your data range.

Step 2: Go to the Insert tab and click "PivotTable".

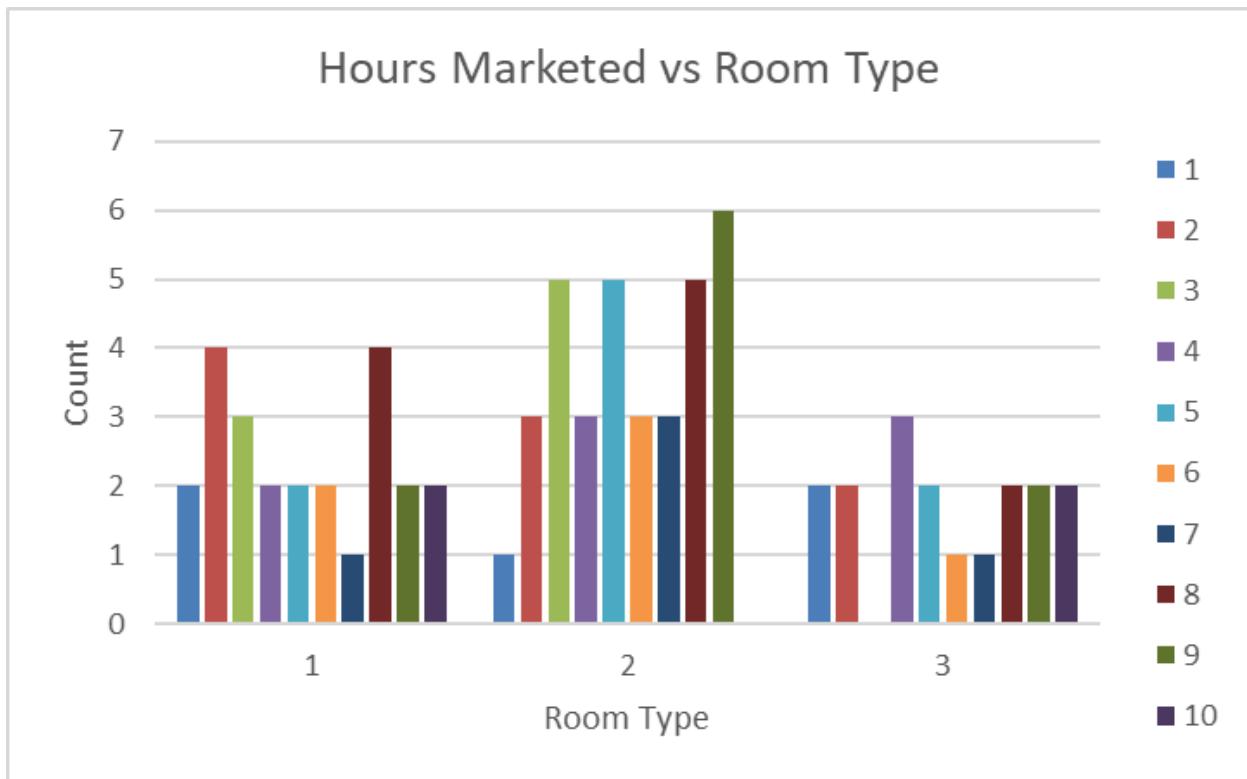
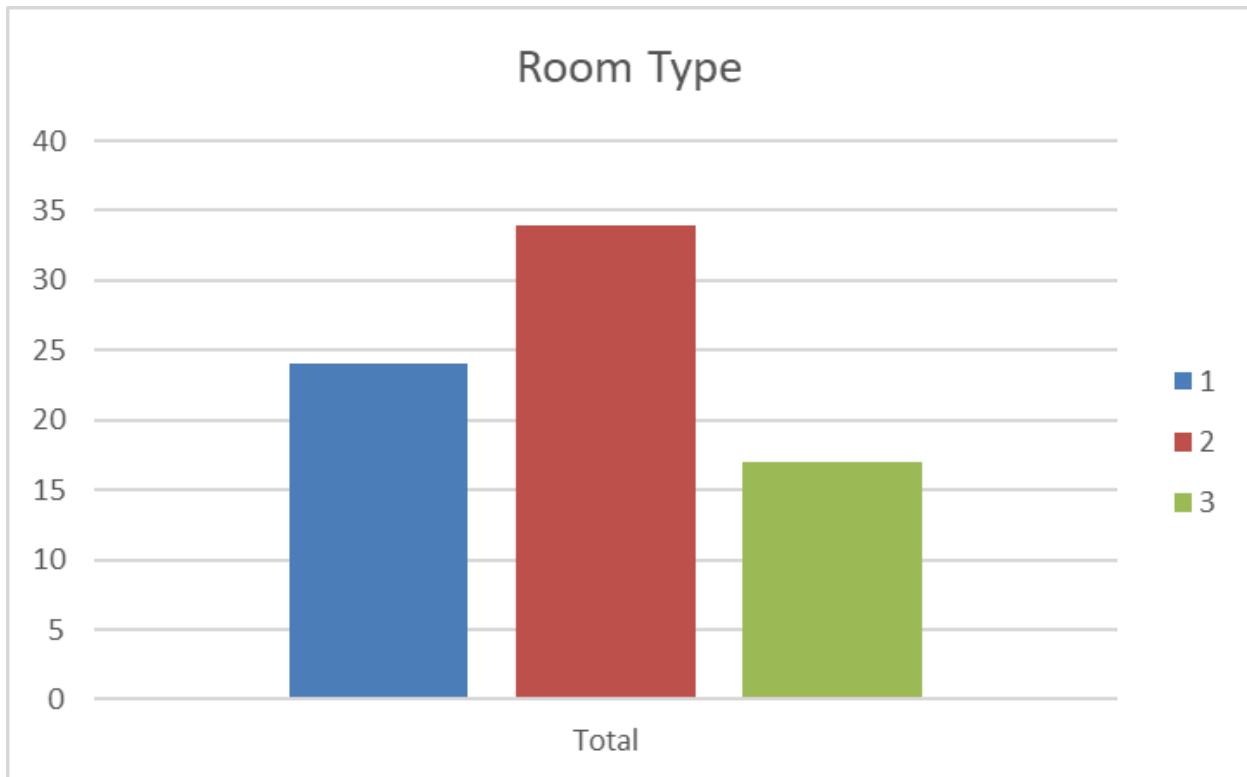
Step 3: In the PivotTable Fields pane, drag your category field to the Rows area.

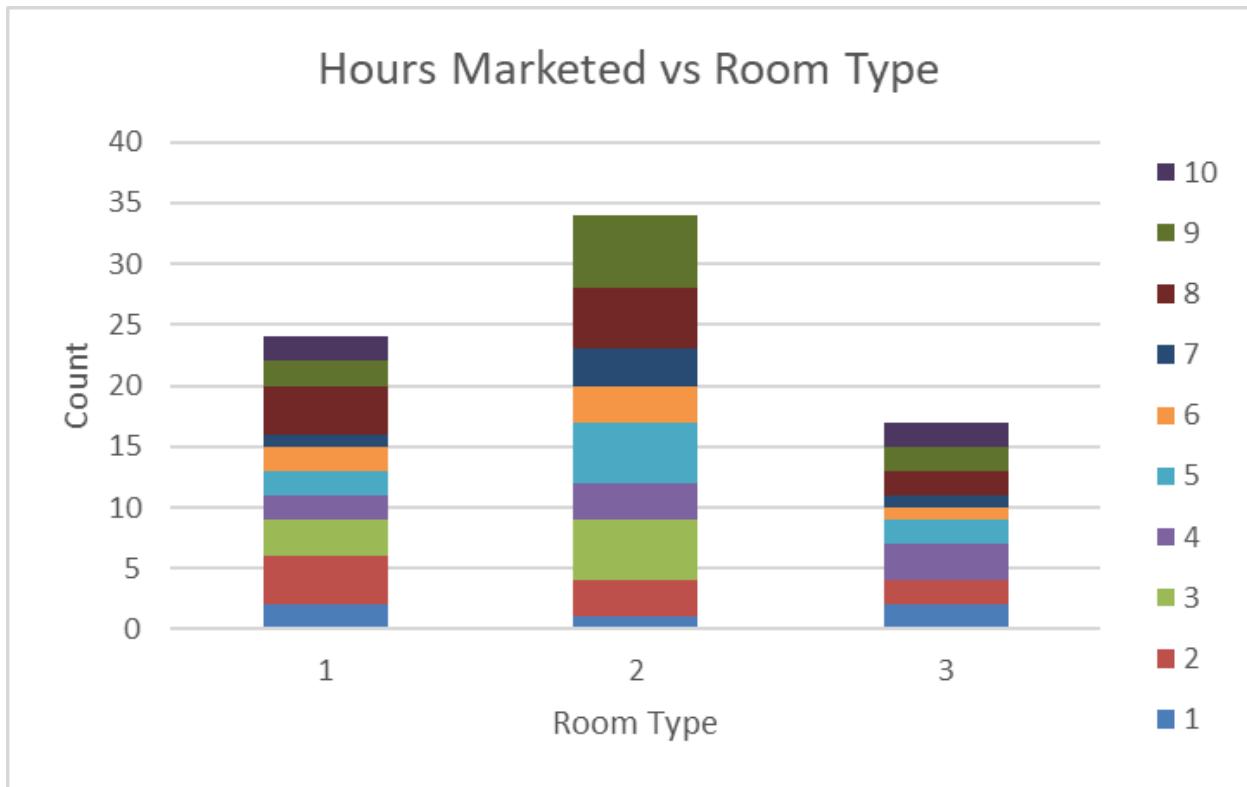
Step 4: Drag your subcategory field to the Columns area.

Step 5: Drag your value field to the Values area.

Step 6: With the PivotTable selected, go to the PivotTable Analyze tab.

Step 7: Click "PivotChart" to create a stacked bar chart.

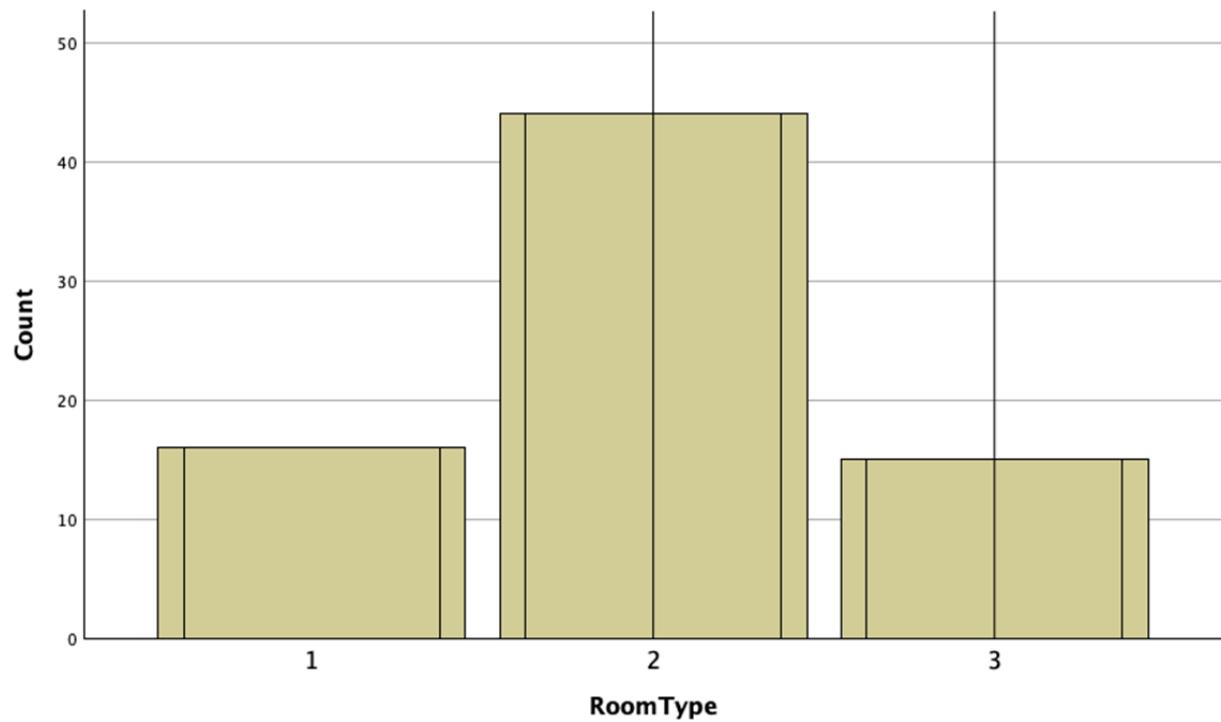




SPSS

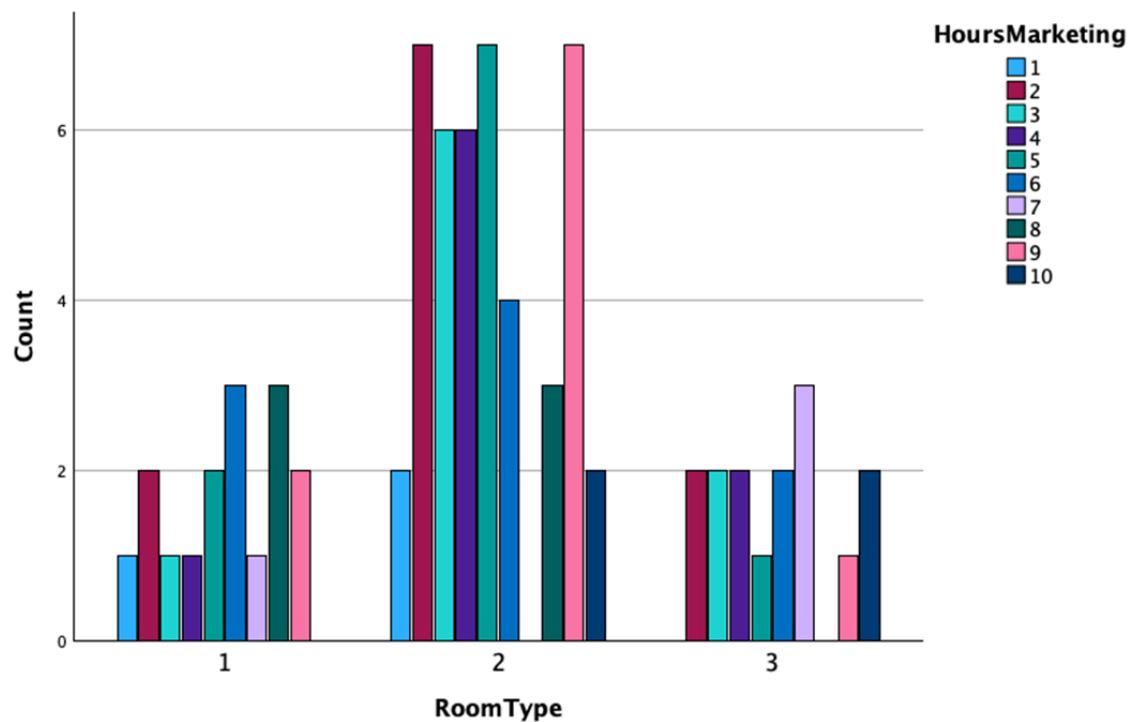
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option.
3. Drag the appropriate variables to the x-axis (category) and y-axis (value).



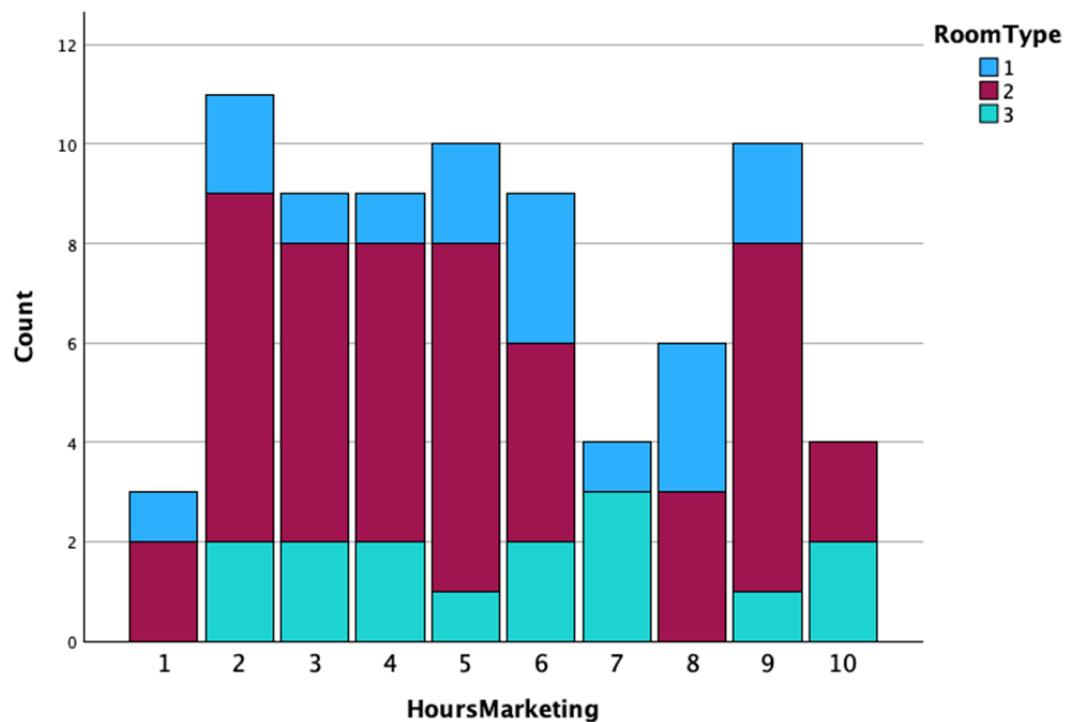
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a clustered bar chart.
3. Drag the categorical variables to the x-axis and grouping variable to the cluster



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a stacked bar chart.
3. Drag the categorical variables to the x-axis and sub-categories to the stacking



Practical Output

1. Room Type Distribution The bar chart shows three room types, with Type 2 being the most common (about 45 counts), followed by Types 1 and 3 (both around 15-20 counts). This suggests the marketing department may have different types of spaces, with Type 2 being the primary workspace.
2. Hours Marketing Distribution (Images 2 and 3): These graphs show the distribution of marketing hours across different room types. Room Type 2 has the highest concentration of marketing activities, with peaks at 2, 5, and 9 hours. Room Types 1 and 3 have more evenly distributed hours, but generally lower counts.
Utilize Room Type 2 as primary workspace for marketing activities. Plan for varied project durations, with common periods of 2, 5, and 9 hours.

CASE STUDY 3 - Operations

Python

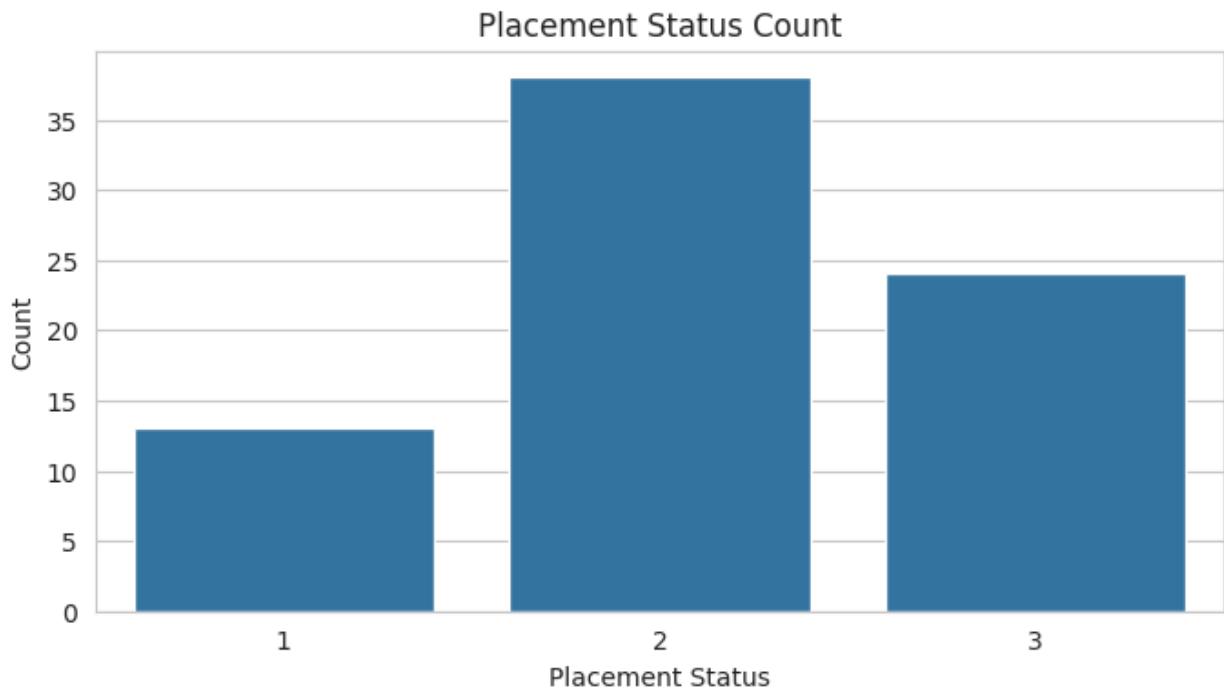
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
data=pd.read_csv("merged.csv")
# Set the aesthetic style of the plots
sns.set_style("whitegrid")

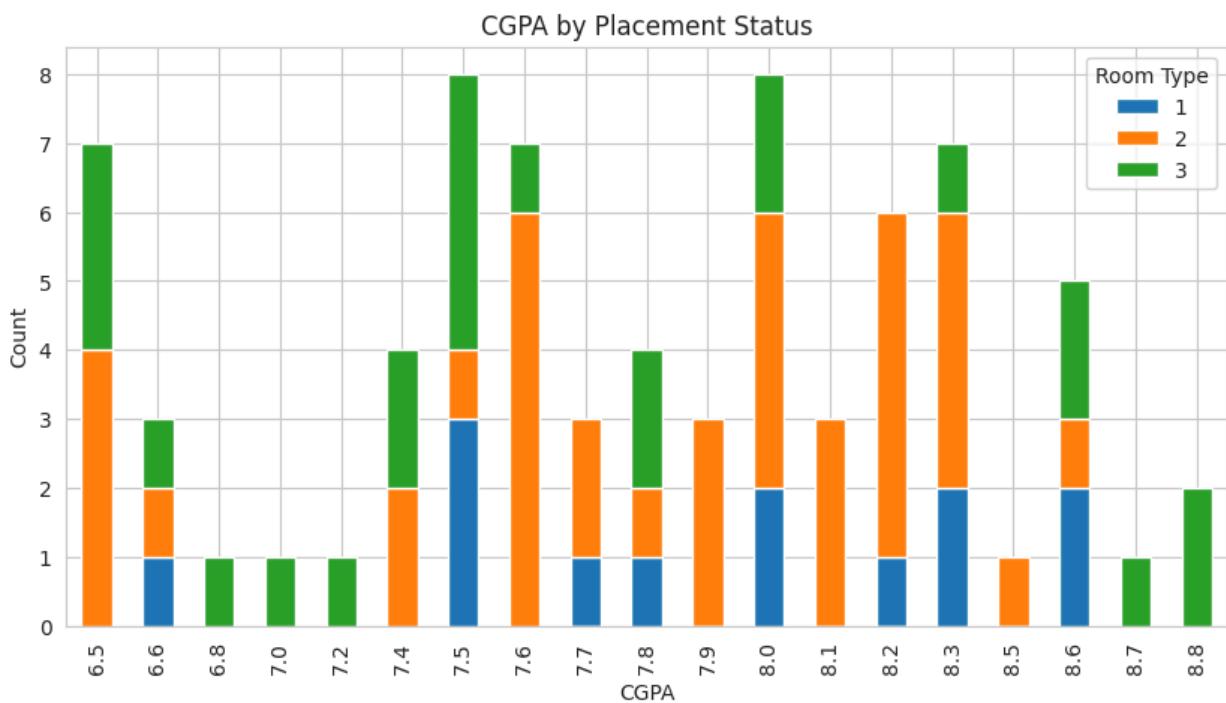
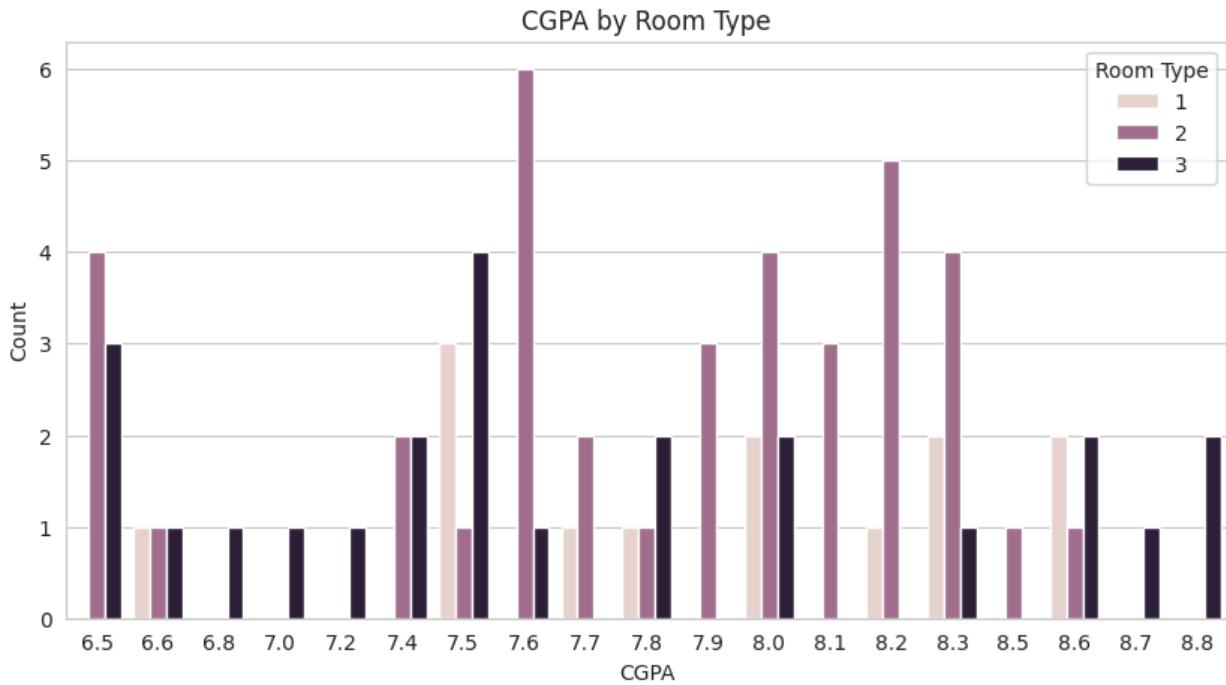
# Simple Bar Diagram - Placement Status
plt.figure(figsize=(8, 4))
sns.countplot(data=data, x='Room Type')
plt.title('Placement Status Count')
plt.xlabel('Placement Status')
plt.ylabel('Count')
plt.show()

# Multiple Bar Diagram - Projects by Placement Status
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='CGPA', hue='Room Type')
plt.title('CGPA by Room Type')
plt.xlabel('CGPA')
plt.ylabel('Count')
plt.show()

# Subdivided Bar Diagram - Internships by Placement Training
```

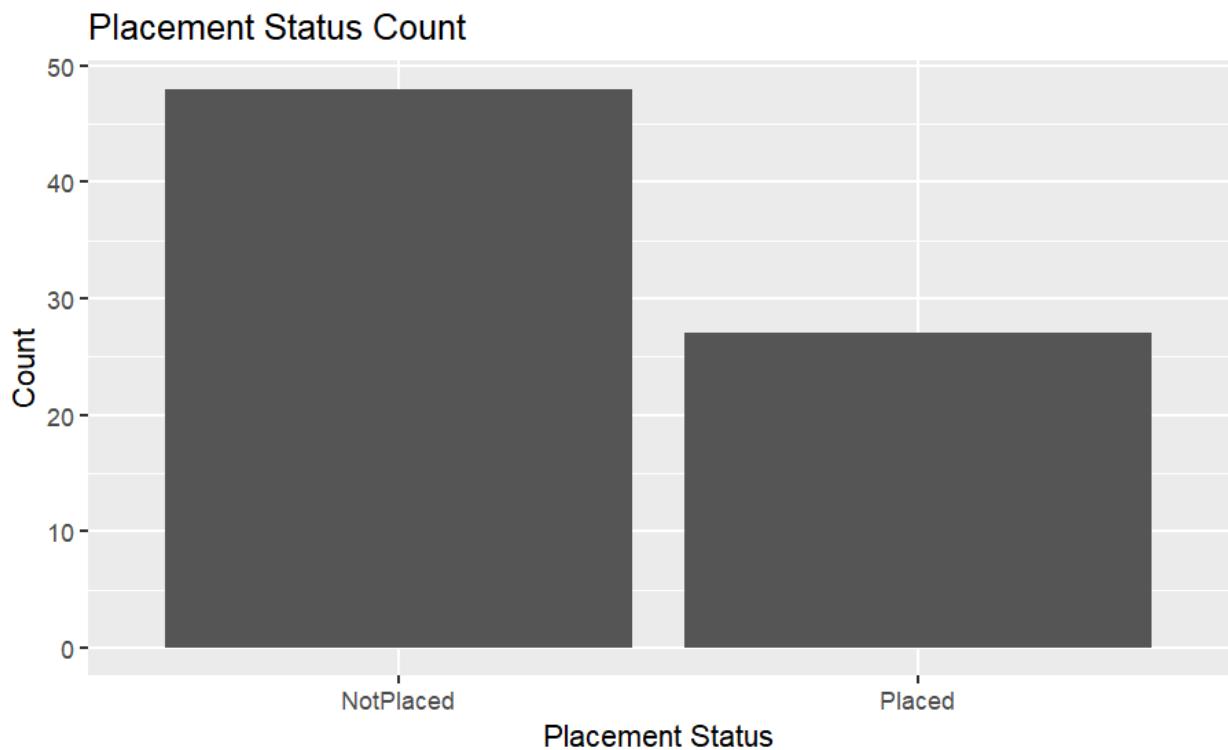
```
pd.crosstab(data['CGPA'], data['Room Type']).plot(kind='bar',
stacked=True, figsize=(10, 5))
plt.title('CGPA by Placement Status')
plt.xlabel('CGPA')
plt.ylabel('Count')
plt.show()
```

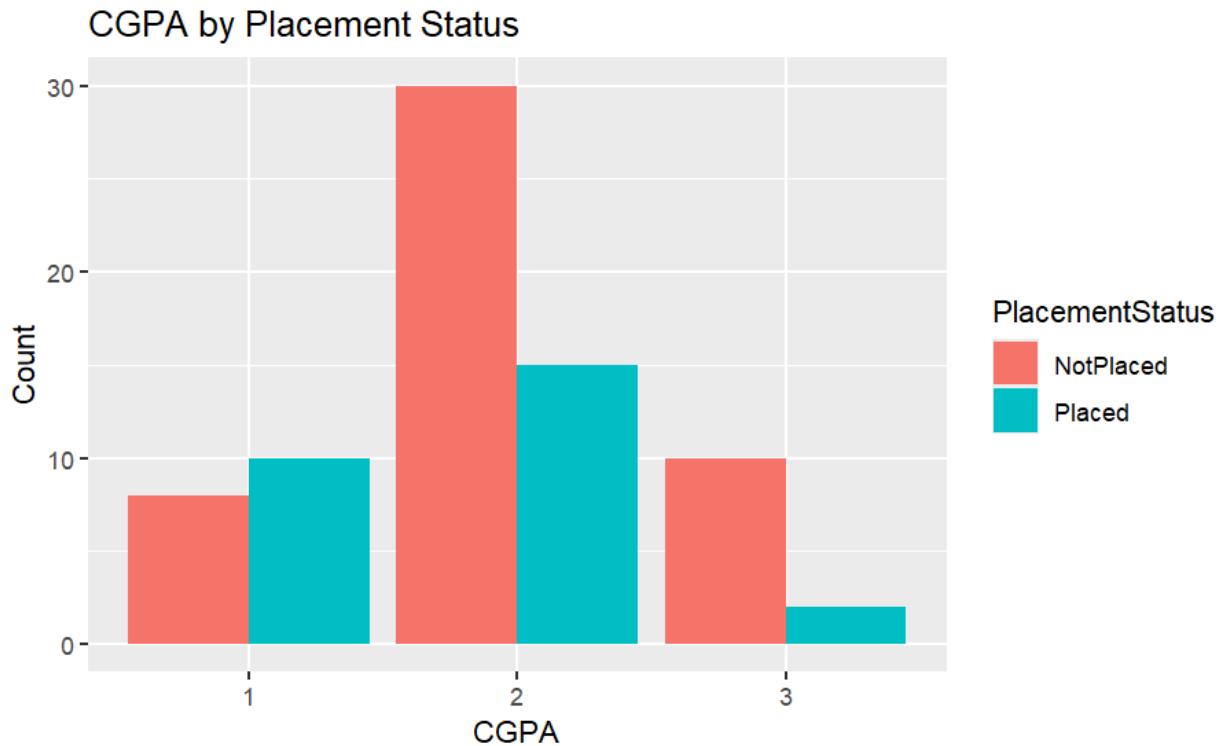




R

```
> library(ggplot2)
> library(dplyr)
>
> # Read the data
> data <- read.csv("C:/Users/HP/Downloads/STATS BOOK - Merged (3).csv")
> df<- data.frame(data$PlacementStatus)
>
> #----- Visualization(Simple subdivident)
> library(ggplot2)
> ggplot(data, aes(x=factor(PlacementStatus))) + geom_bar() +
  labs(title="Placement Status Count", x="Placement Status", y="Count")
>
> ggplot(data, aes(x=factor(data$Room.Type), fill=PlacementStatus)) +
  geom_bar(position="dodge") + labs(title="CGPA by Placement Status", x="CGPA",
  y="Count")
```





Excel

Bar Diagram

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Column" or "Bar" button in the Charts group.

Step 4: Choose the desired bar chart type from the dropdown menu.

Multiple Bar Graph:

Step 1: Organize your data: Place categories in the first column and data series in adjacent columns. Step 2: Select your data range, including headers. Step 3: Go to the "Insert" tab on the Excel ribbon. Step 4: In the "Charts" group, click on the "Insert Column or Bar Chart" button. Step 5: Select "Clustered Bar" or "Stacked Bar" depending on your preference. Step 6: Excel will create the chart. You can then customize it: Step 7: Click on the chart title to edit it. Step 8: Use the "Chart Elements" button to add or remove elements like legends or data labels. Step 9: Use the "Chart Styles" and "Chart Filters" buttons to change the appearance and data selection.

Subdivided Bar Diagram (using Pivot Chart)

Step 1: Select your data range.

Step 2: Go to the Insert tab and click "PivotTable".

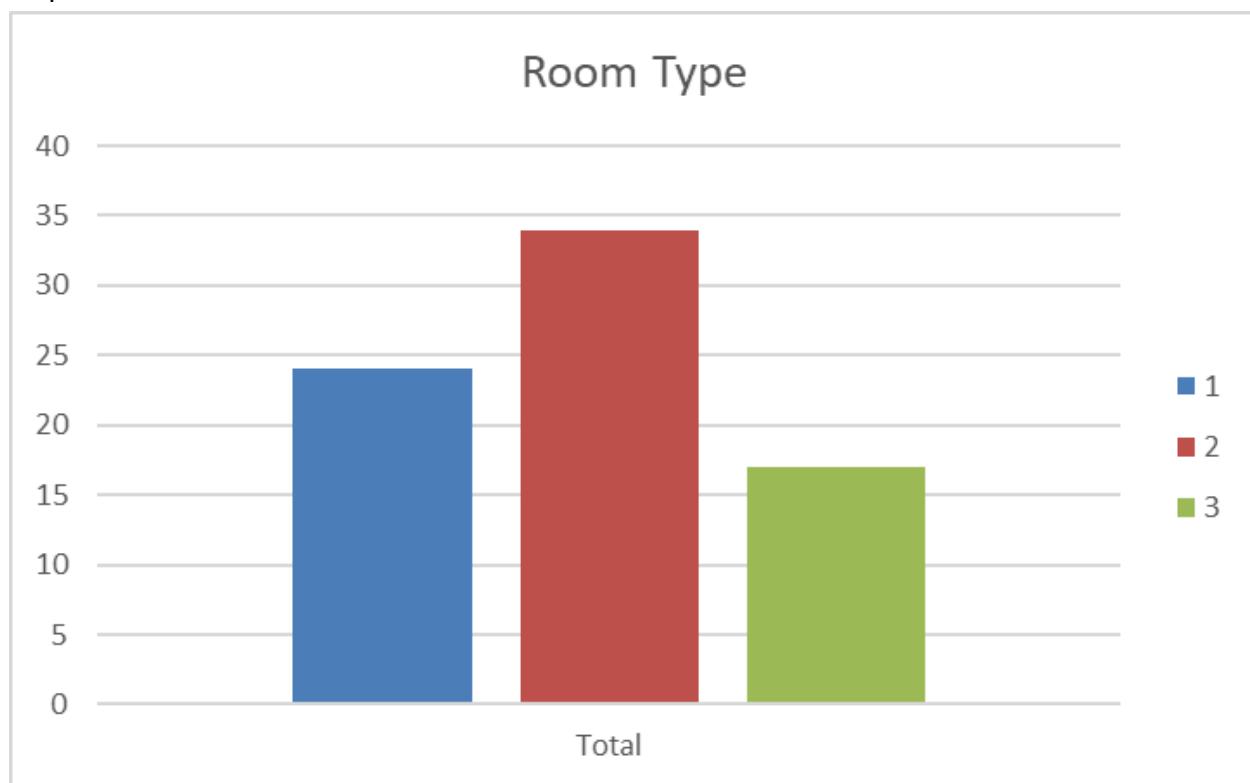
Step 3: In the PivotTable Fields pane, drag your category field to the Rows area.

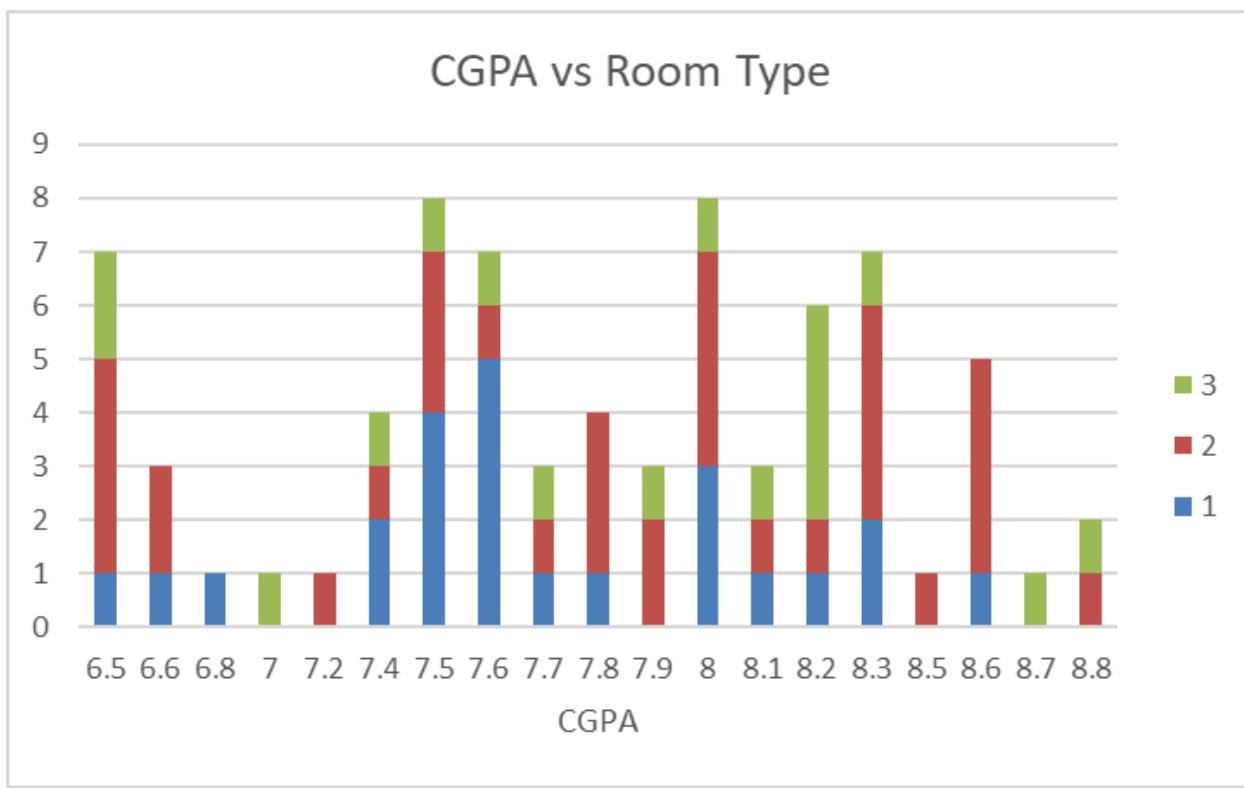
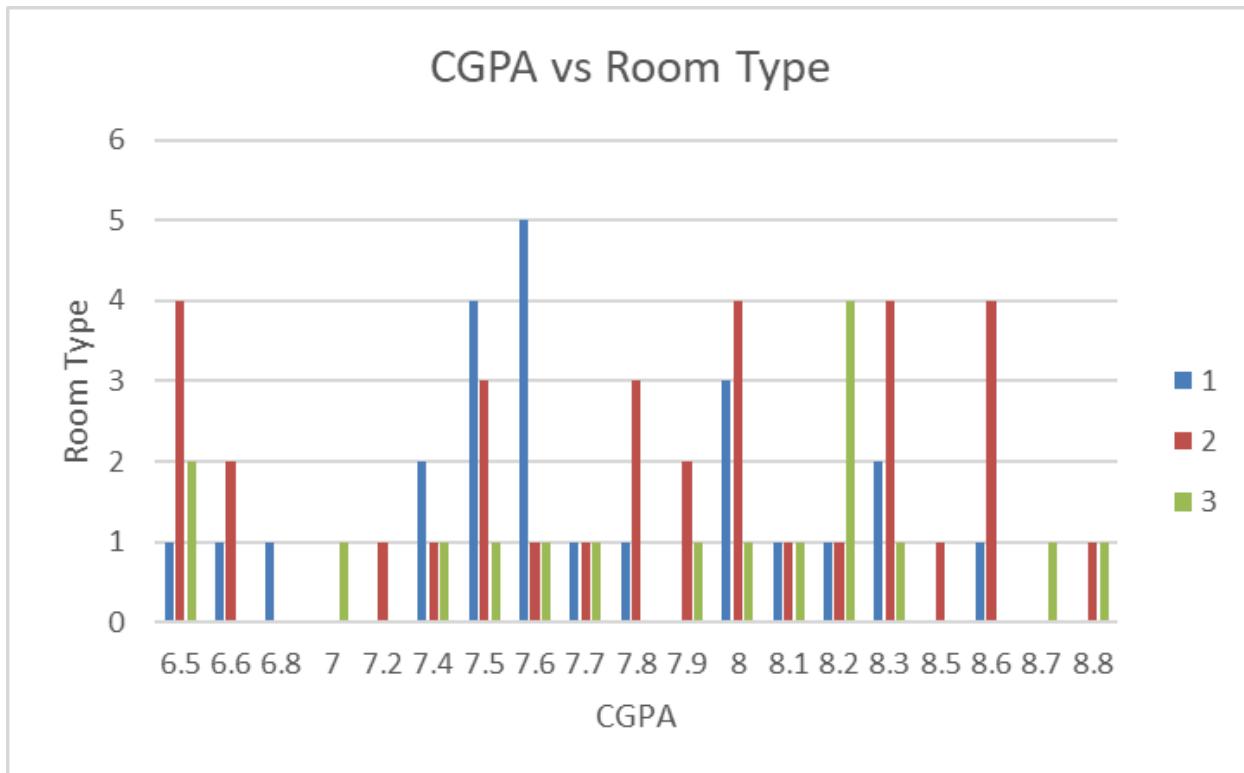
Step 4: Drag your subcategory field to the Columns area.

Step 5: Drag your value field to the Values area.

Step 6: With the PivotTable selected, go to the PivotTable Analyze tab.

Step 7: Click "PivotChart" to create a stacked bar chart.

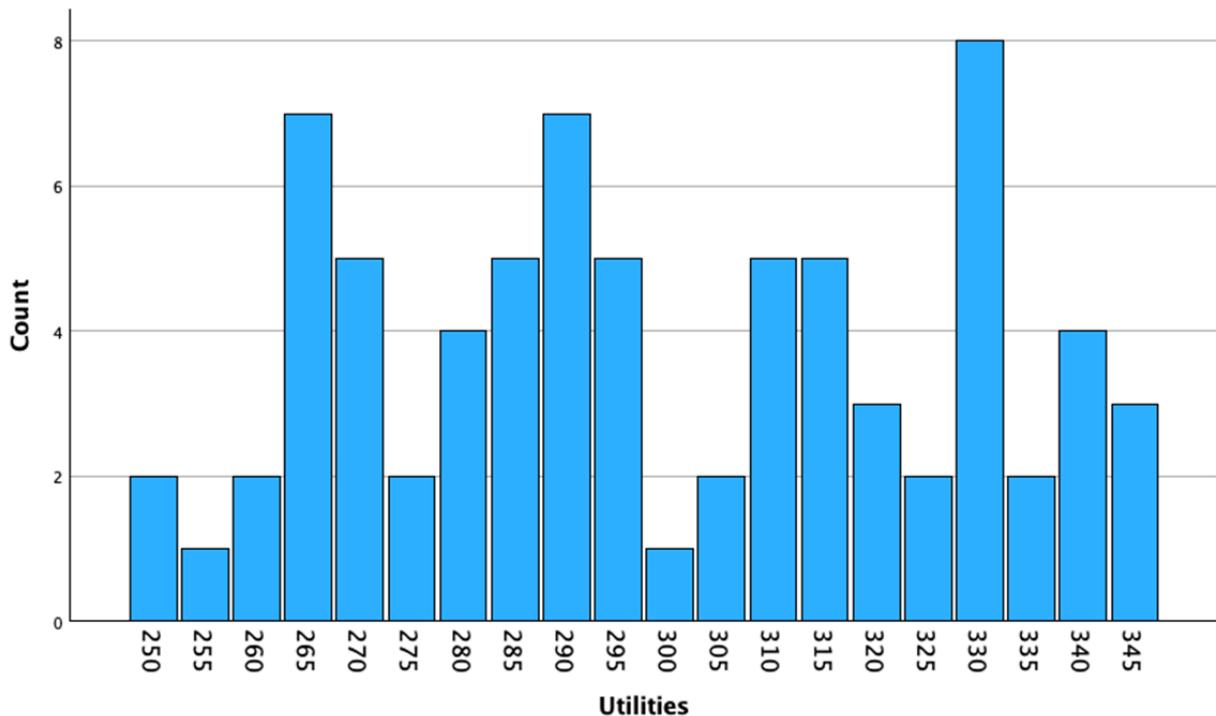




SPSS

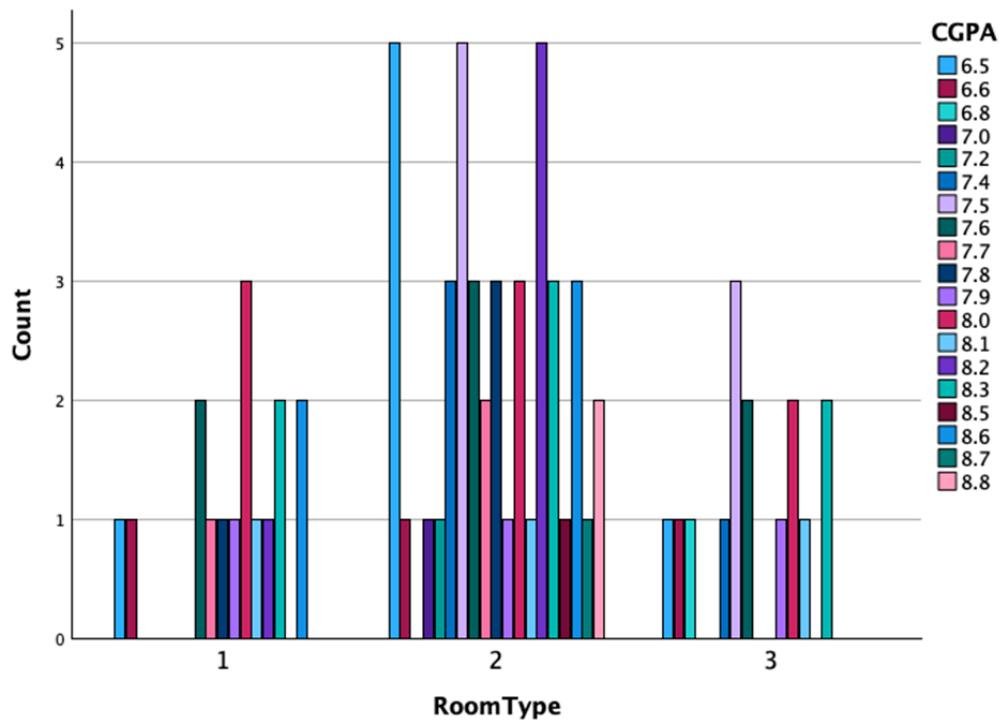
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option.
3. Drag the appropriate variables to the x-axis (category) and y-axis (value).



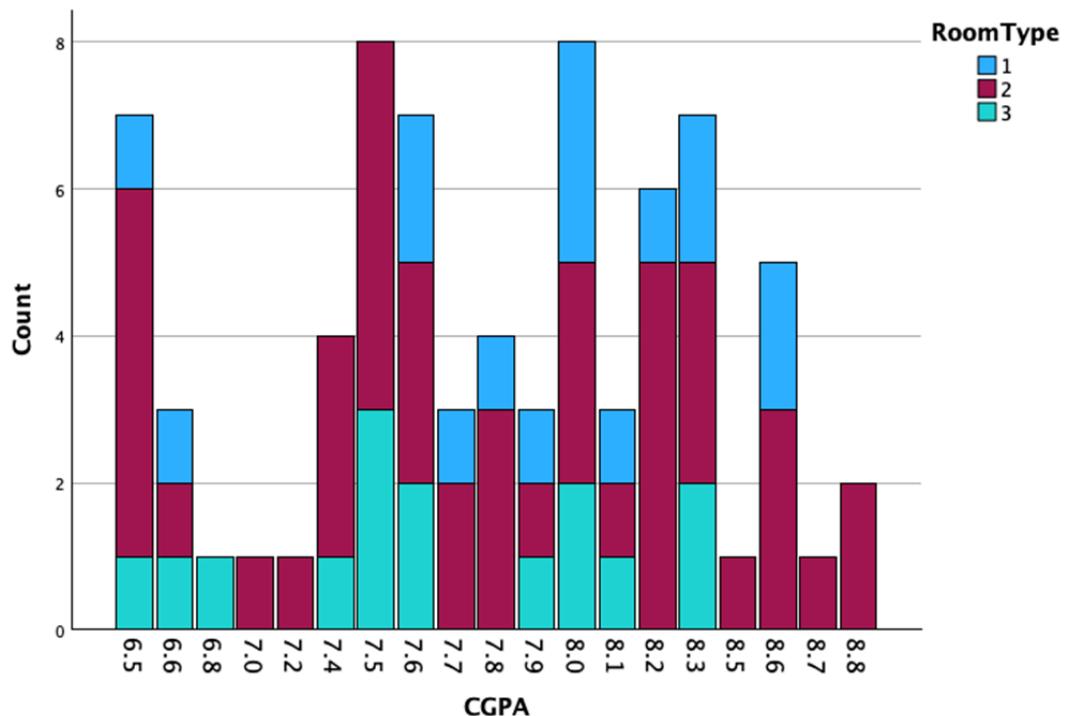
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a clustered bar chart.
3. Drag the categorical variables to the x-axis and grouping variable to the cluster



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a stacked bar chart.
3. Drag the categorical variables to the x-axis and sub-categories to the stacking



Practical Output

Chart 1: Utilities Usage Distribution

This chart shows the count of different utility values. The x-axis represents various utility values (presumably in some units), ranging from 250 to 345. The y-axis shows the count of occurrences for each utility value. Key observations:

- Utility values 265 and 325 are the most frequent, each occurring 8 times.
- There is variability in utility usage, with some values appearing more often than others.
- This information could be used by the operations department to identify common utility usage patterns and potentially optimize resource allocation.

Chart 2: Room Type and CGPA Distribution

This chart compares the count of students with different CGPAs across three room types. The x-axis represents room types (1, 2, and 3), while the y-axis shows the count of students.

Different colors within the bars represent different CGPAs, ranging from 6.5 to 8.8. Key observations:

- Room Type 2 has the highest concentration of students with various CGPAs.
- Room Type 1 has a more evenly distributed count across CGPAs compared to the other room types.
- Room Type 3 has fewer students with higher CGPAs.
- This could help the operations department understand the academic distribution across different living arrangements, potentially influencing housing policies or support services.

Chart 3: CGPA Distribution by Room Type

This chart shows the distribution of CGPAs among students in different room types. The x-axis represents CGPAs, while the y-axis shows the count of students. Different colors within the bars indicate different room types. Key observations:

- Lower CGPAs (6.5 to 7.4) are more frequently associated with Room Type 1.
- Mid-range CGPAs (7.5 to 8.0) are more evenly distributed among Room Types 1, 2, and 3.
- Higher CGPAs (8.1 to 8.8) are more associated with Room Types 2 and 3.
- This data could be used to analyze any correlation between living conditions and academic performance, helping the operations department in making informed decisions about room assignments.

Overall Interpretation

The charts together provide a comprehensive view of utility usage patterns, the distribution of students' CGPAs across different room types, and how CGPA varies by room type. This information can help the operations department of a college in several ways:

- **Resource Management:** By understanding utility usage patterns, the department can optimize the allocation of resources and reduce wastage.
- **Housing Policies:** Insights into the distribution of students across room types and their academic performance can help in designing housing policies that support student success.
- **Targeted Support:** Identifying which room types are associated with lower academic performance can help in providing targeted support to those students, such as additional academic resources or study areas.

CASE STUDY 4 - Finance

Python

```
import matplotlib.pyplot as plt
```

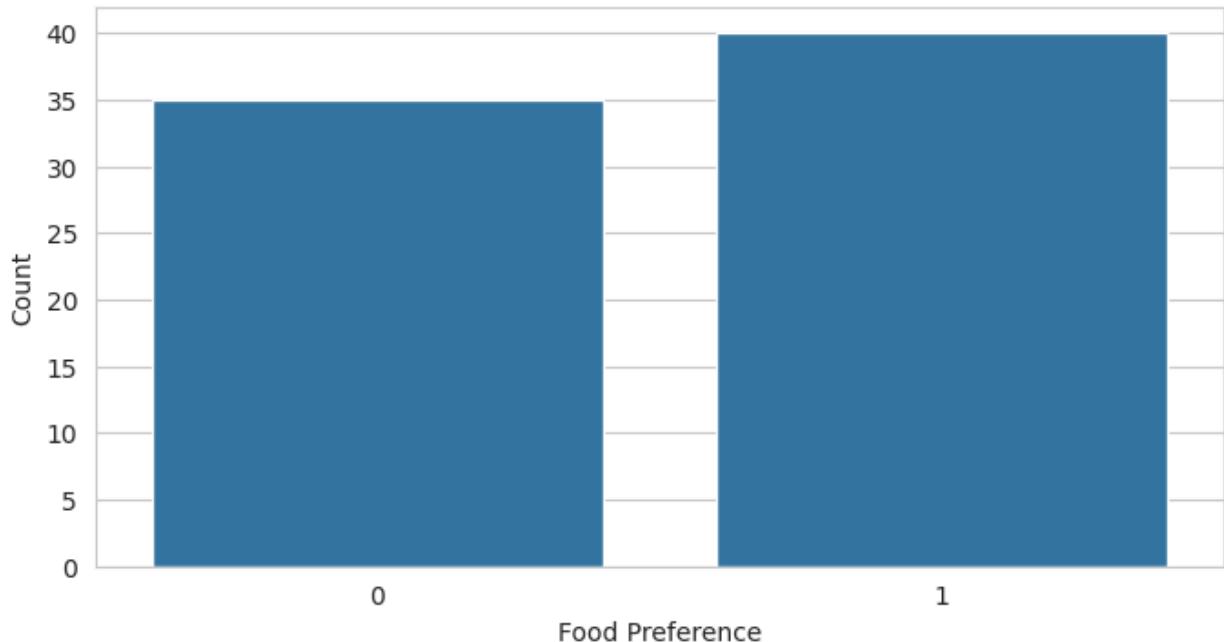
```
import seaborn as sns
import numpy as np
import pandas as pd
data=pd.read_csv("merged.csv")
# Set the aesthetic style of the plots
sns.set_style("whitegrid")

# Simple Bar Diagram - Placement Status
plt.figure(figsize=(8, 4))
sns.countplot(data=data, x='Food Preference')
plt.title('Food Preference Count')
plt.xlabel('Food Preference')
plt.ylabel('Count')
plt.show()

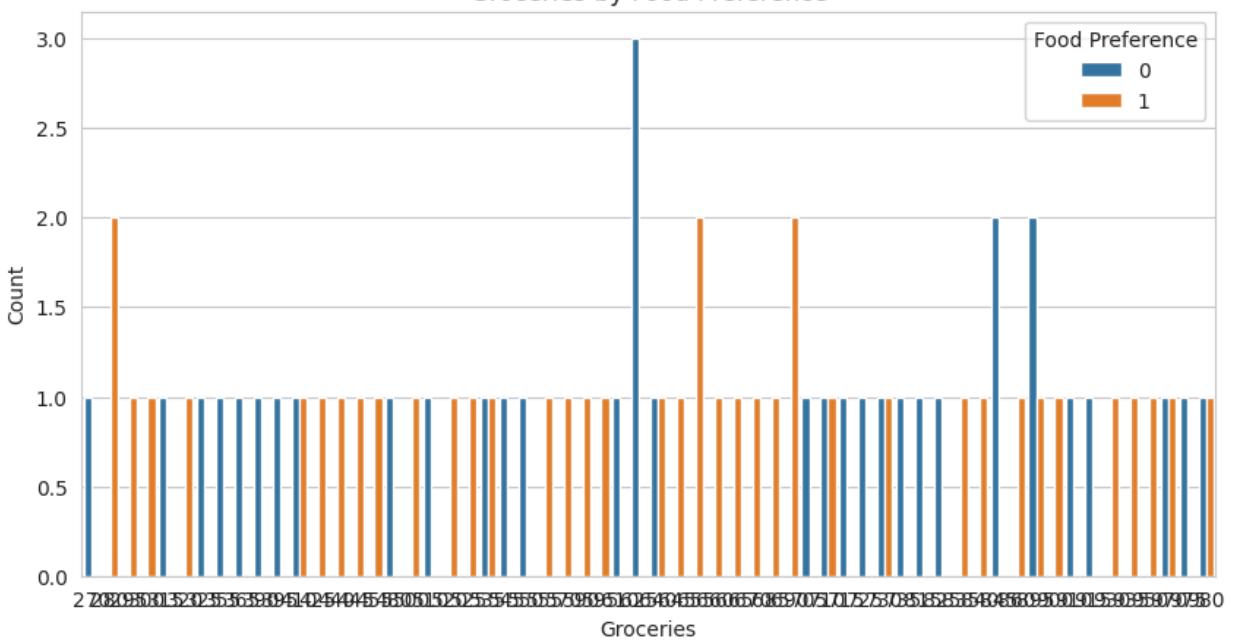
# Multiple Bar Diagram - Projects by Food Preference
plt.figure(figsize=(10, 5))
sns.countplot(data=data, x='Groceries', hue='Food Preference')
plt.title('Groceries by Food Preference')
plt.xlabel('Groceries')
plt.ylabel('Count')
plt.show()

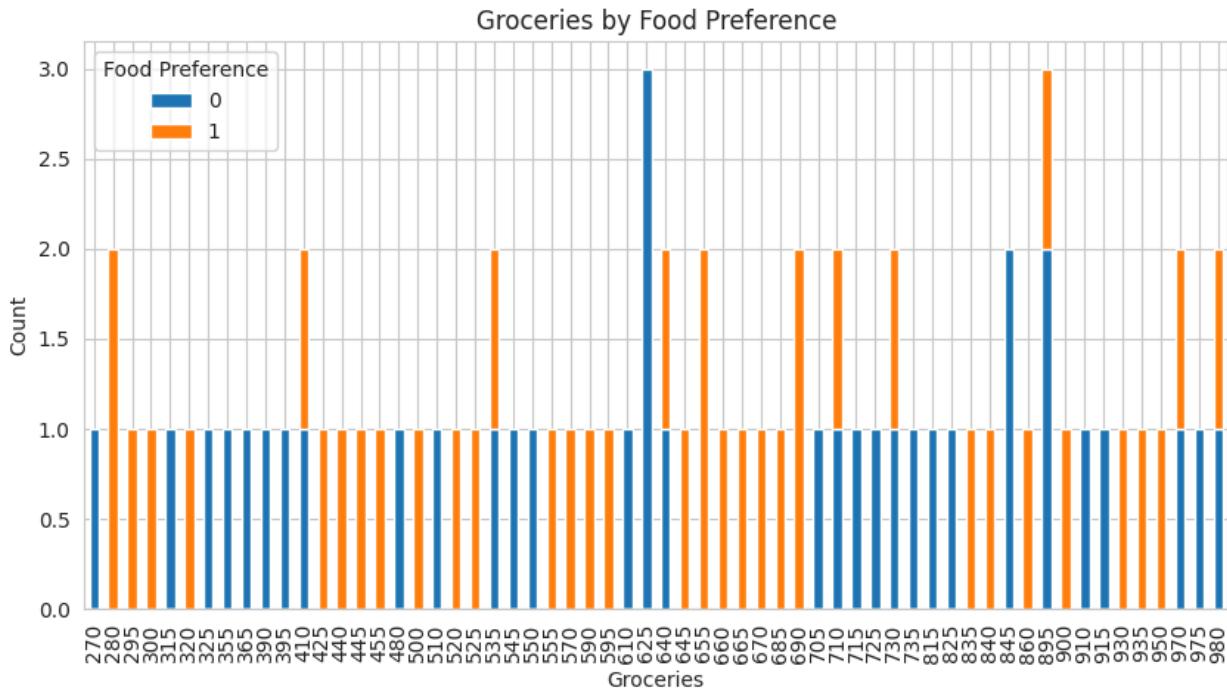
# Subdivided Bar Diagram - Internships by Placement Training
pd.crosstab(data['Groceries'], data['Food Preference']).plot(kind='bar',
stacked=True, figsize=(10, 5))
plt.title('Groceries by Food Preference')
plt.xlabel('Groceries')
plt.ylabel('Count')
plt.show()
```

Food Preference Count



Groceries by Food Preference





R

Excel

Bar Diagram

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Column" or "Bar" button in the Charts group.

Step 4: Choose the desired bar chart type from the dropdown menu.

Multiple Bar Graph:

Step 1: Organize your data: Place categories in the first column and data series in adjacent columns. Step 2: Select your data range, including headers. Step 3: Go to the "Insert" tab on the Excel ribbon. Step 4: In the "Charts" group, click on the "Insert Column or Bar Chart" button.

Step 5: Select "Clustered Bar" or "Stacked Bar" depending on your preference. Step 6: Excel will create the chart. You can then customize it: Step 7: Click on the chart title to edit it. Step 8: Use the "Chart Elements" button to add or remove elements like legends or data labels. Step 9: Use the "Chart Styles" and "Chart Filters" buttons to change the appearance and data selection.

Subdivided Bar Diagram (using Pivot Chart)

Step 1: Select your data range.

Step 2: Go to the Insert tab and click "PivotTable".

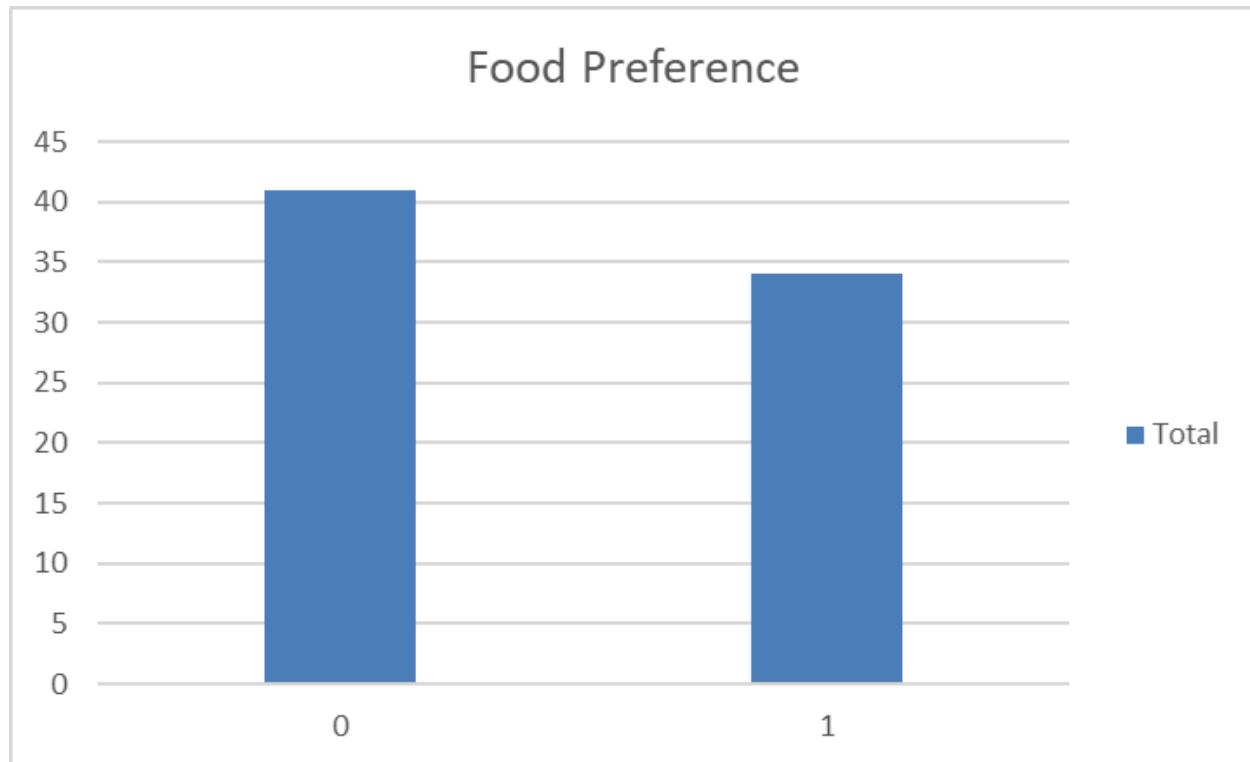
Step 3: In the PivotTable Fields pane, drag your category field to the Rows area.

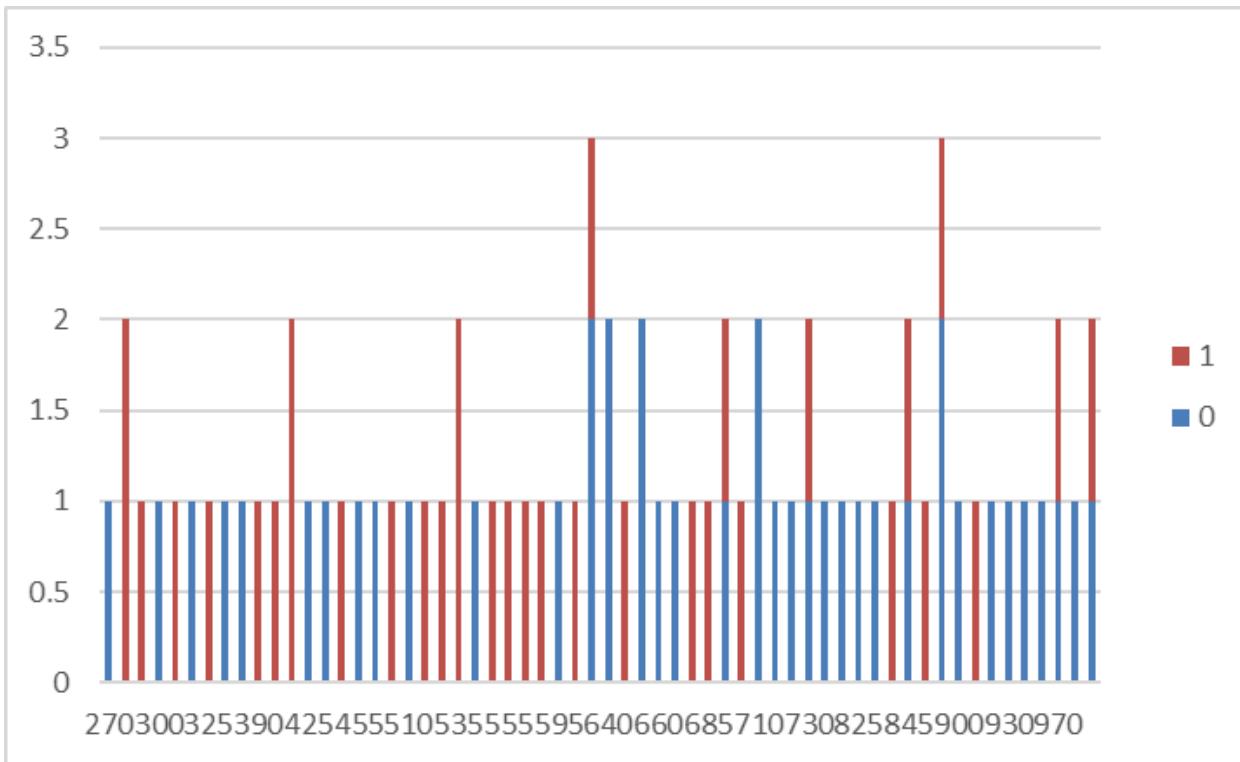
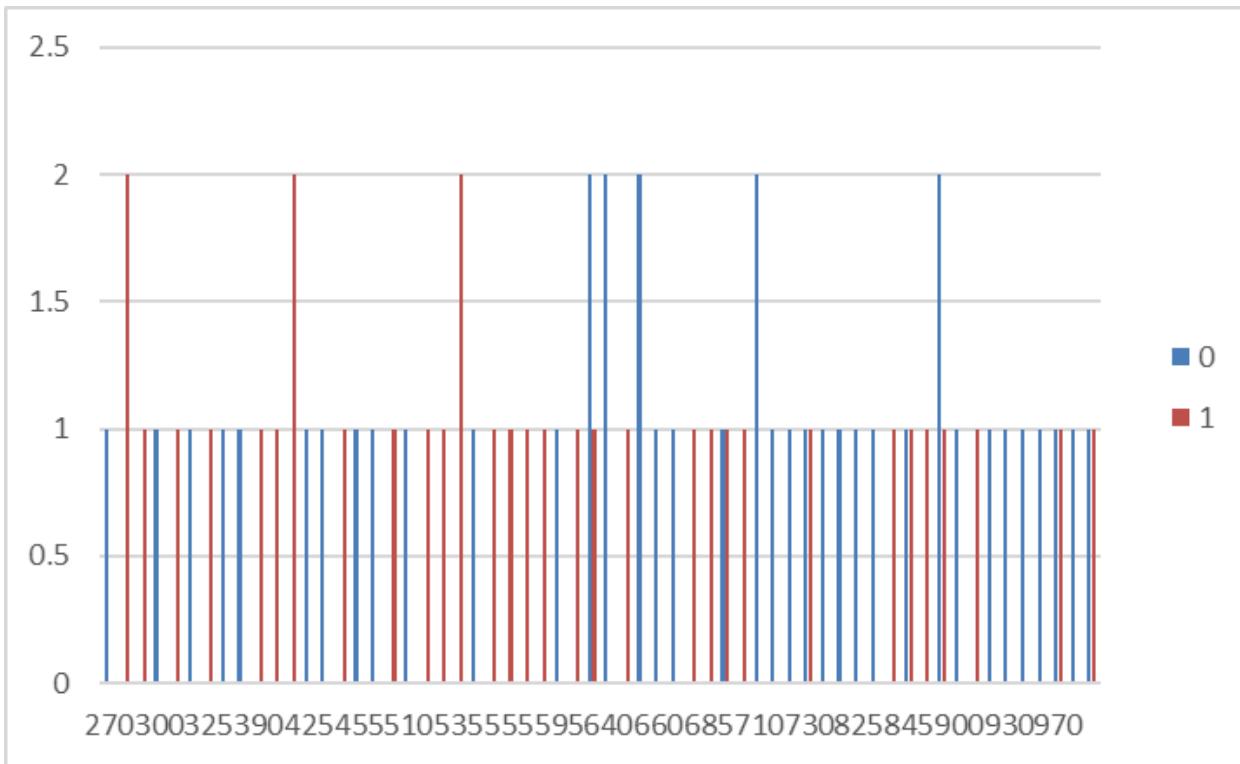
Step 4: Drag your subcategory field to the Columns area.

Step 5: Drag your value field to the Values area.

Step 6: With the PivotTable selected, go to the PivotTable Analyze tab.

Step 7: Click "PivotChart" to create a stacked bar chart.

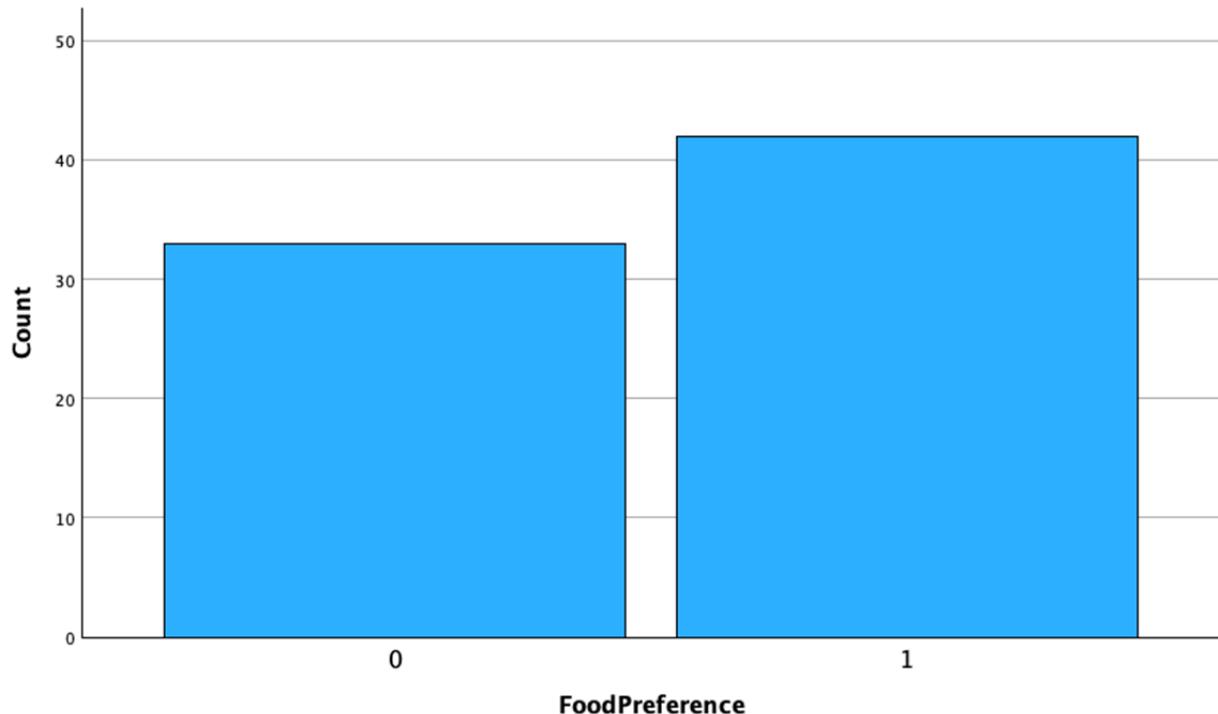




SPSS

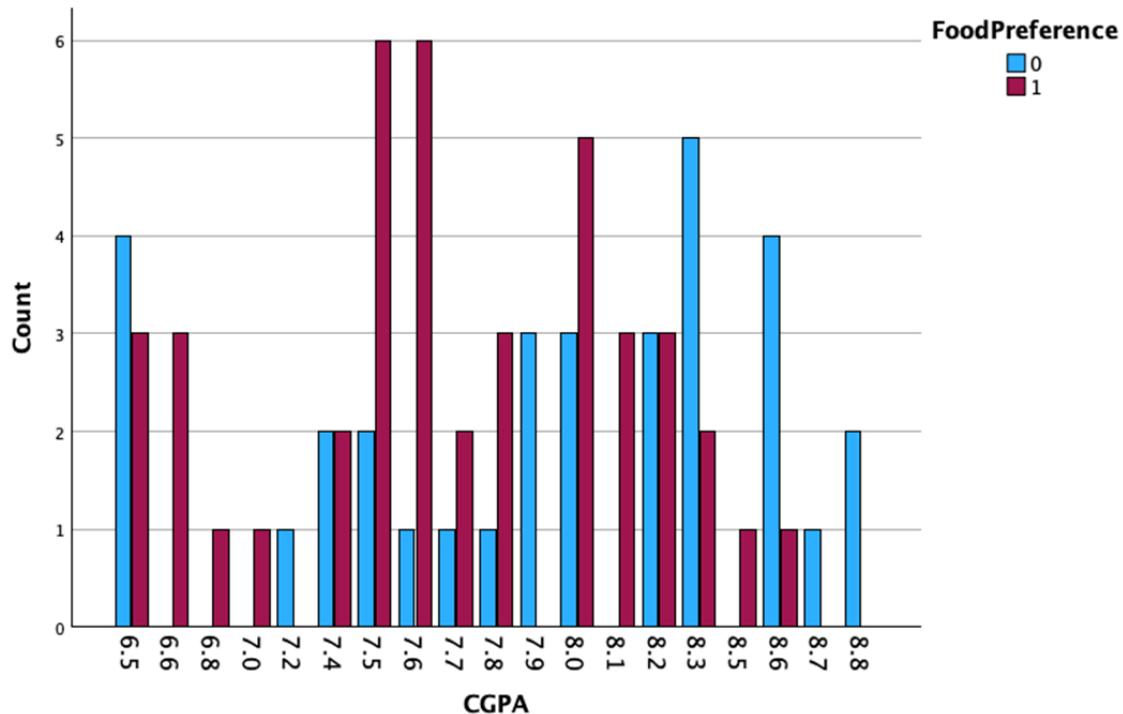
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option.
3. Drag the appropriate variables to the x-axis (category) and y-axis (value).



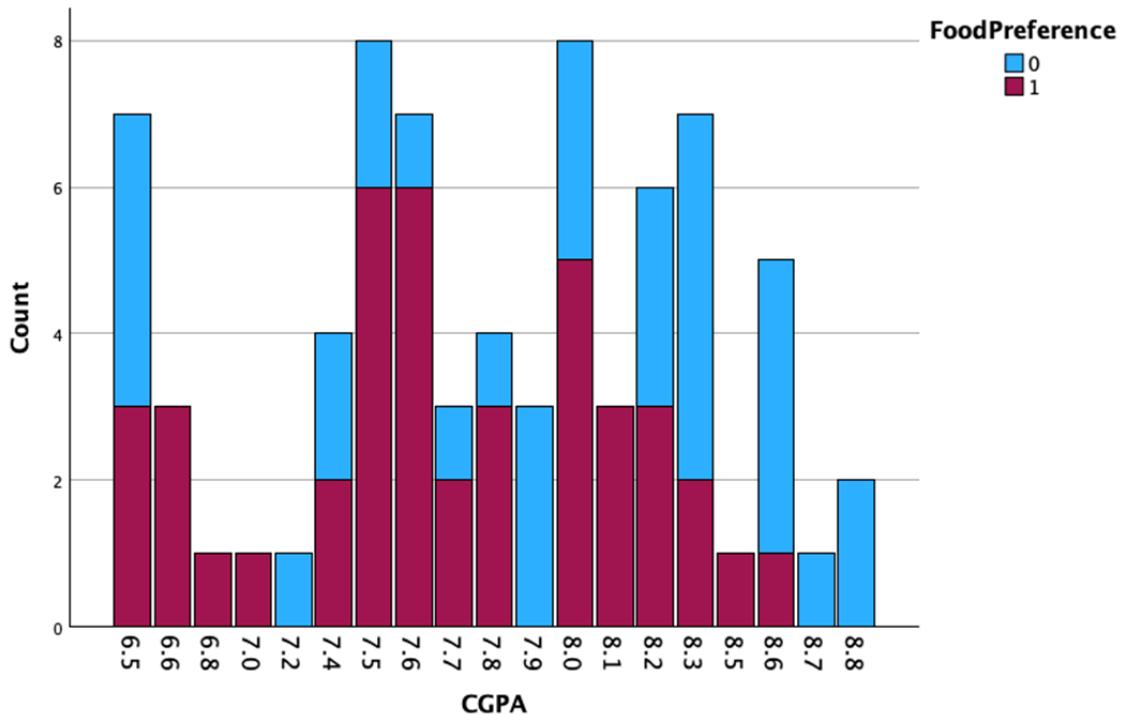
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a clustered bar chart.
3. Drag the categorical variables to the x-axis and grouping variable to the cluster



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Bar option and choose a stacked bar chart.
3. Drag the categorical variables to the x-axis and sub-categories to the stacking



Practical Output

CGPA Distribution show CGPA distributions related to food preferences. There's a wide range of CGPAs represented, from 6.5 to 8.8. The finance department may want to analyze if there's any correlation between academic performance and food choices, which could inform future meal planning and budgeting decisions.

CH 2 - Pie Diagram, Line Diagram

Pie Diagrams: The Art of Proportional Representation

Pie Diagrams, or Pie Charts, serve as a foundational tool in the visual representation of data, transforming numerical proportions into visually intuitive slices of a whole. This section explores the pie diagram's utility, nuances, and innovative applications in conveying complex data sets with clarity and impact.

Conceptual Underpinnings

At its essence, a pie diagram divides a circle into segments or "slices," with each slice's size proportional to the frequency or percentage it represents in the dataset. This method of visual representation excels in illustrating the composition of a whole, making it particularly effective for categorical data where the relationship of parts to a whole is of interest.

Application in Case Studies

Consider the scenario of a multinational corporation analyzing its global sales distribution. A pie diagram can succinctly display the proportion of sales originating from different regions, immediately highlighting which regions are outperforming or underperforming. This visual clarity aids decision-makers in allocating resources efficiently and strategizing market interventions.

Creative Insights

- Interactive Pie Diagrams: Leveraging digital platforms to create interactive pie diagrams can enrich the user's exploration experience. By clicking on a slice, users could reveal additional layers of data, such as subcategories or historical trends, fostering a deeper understanding of the dataset.
- Aesthetic Enhancements: Employing color gradients, textures, and annotations can transform a standard pie diagram into a compelling narrative piece. Such aesthetic enhancements not only draw attention but also facilitate the retention of information by the audience.

Line Diagrams: Tracing Data Through Time

Line Diagrams, or Line Charts, represent one of the most versatile and widely used tools for data visualization, particularly adept at showcasing trends, changes, and comparisons over time. This segment delves into the line diagram's methodology, applications, and creative augmentations for enhanced data storytelling.

Conceptual Underpinnings

Line diagrams plot data points on a two-dimensional plane, connected by lines, to illustrate how a variable changes over time or across categories. The strength of line

diagrams lies in their ability to convey trends and movements within data, offering insights into growth patterns, cyclical behaviors, and potential anomalies.

Application in Case Studies

In an analysis of consumer behavior trends, a line diagram could track the monthly sales volume of a product over several years. The resulting visualization would not only reveal seasonal fluctuations and long-term trends but also pinpoint anomalies that may correlate with external events or interventions, providing a basis for further investigation.

Creative Insights

- Multi-variable Comparison: Incorporating multiple lines representing different variables or categories within the same diagram allows for a direct comparison, revealing correlations or divergences that might not be apparent from isolated data points.
- Enhanced Interactivity: Creating line diagrams with zoom-in capabilities, dynamic time frames, and hover-over data points can transform static charts into interactive exploration tools. This approach encourages users to engage with the data, uncovering insights at their own pace and according to their interests.

Conclusion: Pie and Line Diagrams in Harmonious Analysis

Pie and Line Diagrams, each with its unique strengths, offer complementary perspectives on data. While pie diagrams provide a snapshot of the composition at a specific point in time, line diagrams trace the evolution of data across intervals, together weaving a comprehensive narrative of the dataset. By creatively applying and combining these visualization tools, statisticians, analysts, and decision-makers can unlock deeper insights, driving informed strategies and actions. In the vast landscape of data visualization, Pie and Line Diagrams stand as testament to the power of visual storytelling, transforming raw data into meaningful knowledge that guides the journey from curiosity to clarity.

CASE STUDY 1 - HR

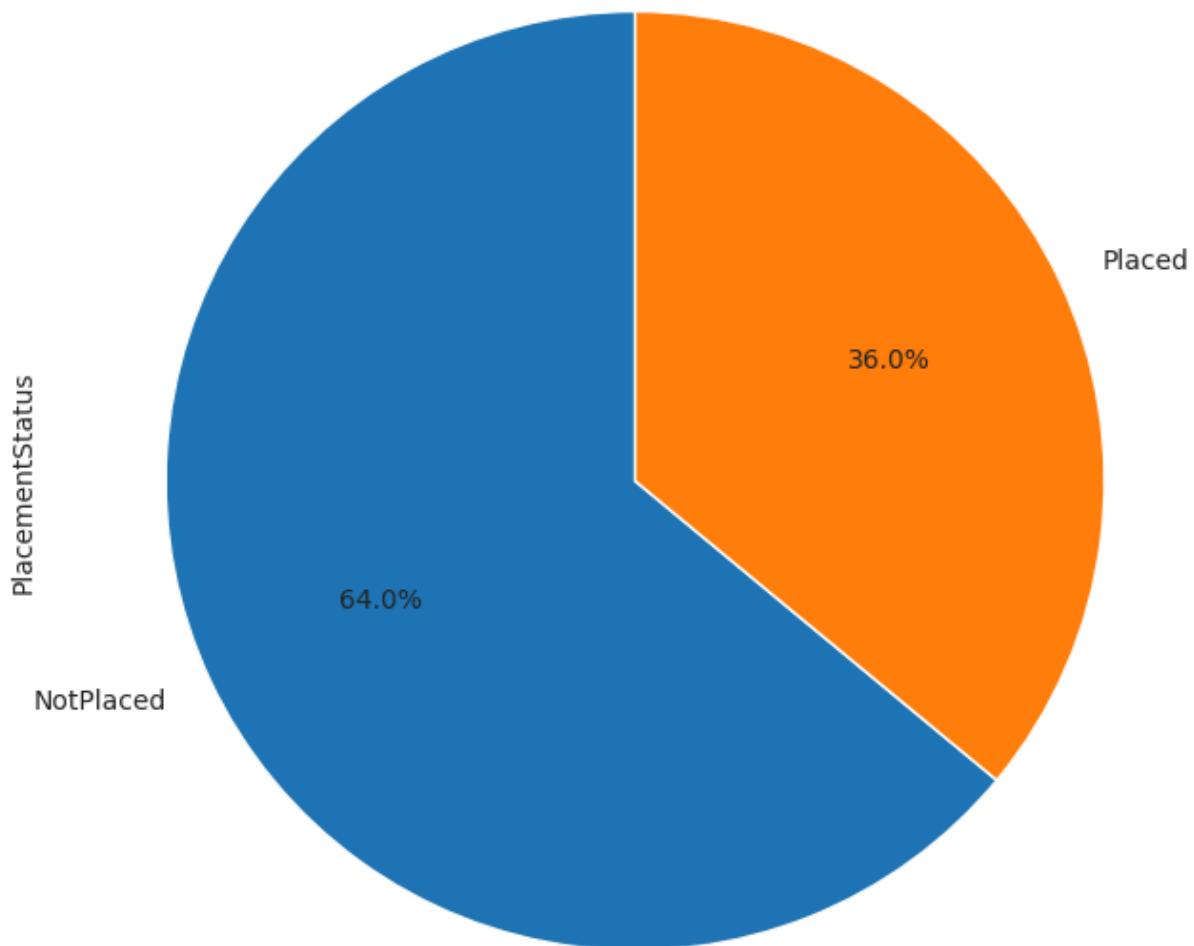
Python

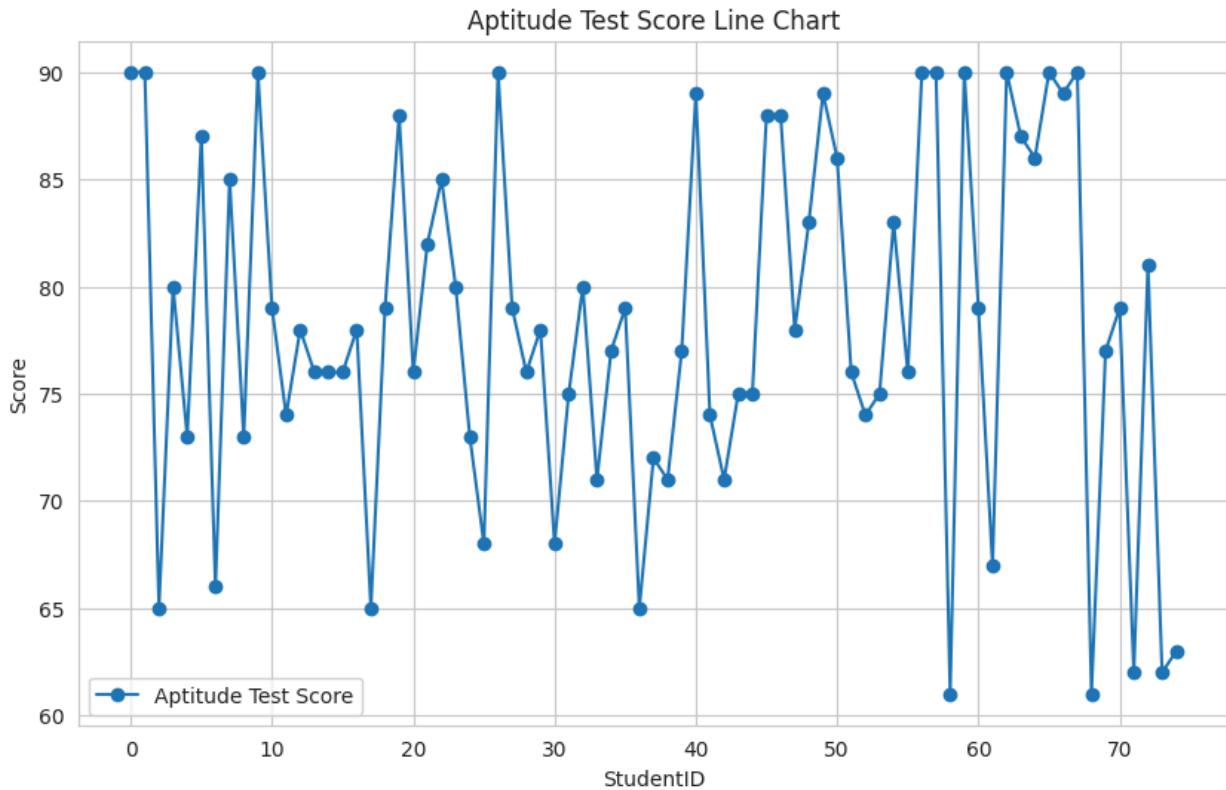
```
import pandas as pd  
import matplotlib.pyplot as plt
```

```
import numpy as np
import seaborn as sns
# Assuming your data is in a DataFrame named 'df'
# Pie Diagram for PlacementStatus
df= pd.read_csv('merged.csv')
plt.figure(figsize=(8, 8))
df['PlacementStatus'].value_counts().plot.pie(autopct='%1.1f%%',
startangle=90)
plt.title('Placement Status Pie Chart')
plt.show()

# Line Diagram for CGPA and AptitudeTestScore
plt.figure(figsize=(10, 6))
plt.plot(df['AptitudeTestScore'], label='Aptitude Test Score', marker='o')
plt.title('Aptitude Test Score Line Chart')
plt.xlabel('StudentID')
plt.ylabel('Score')
plt.legend()
plt.show()
```

Placement Status Pie Chart

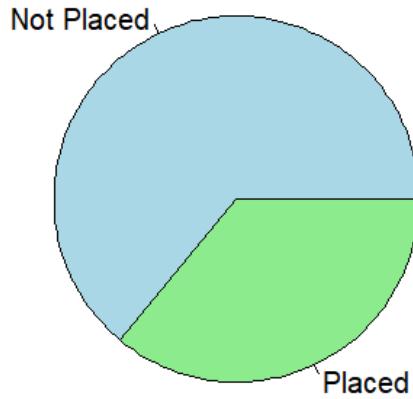




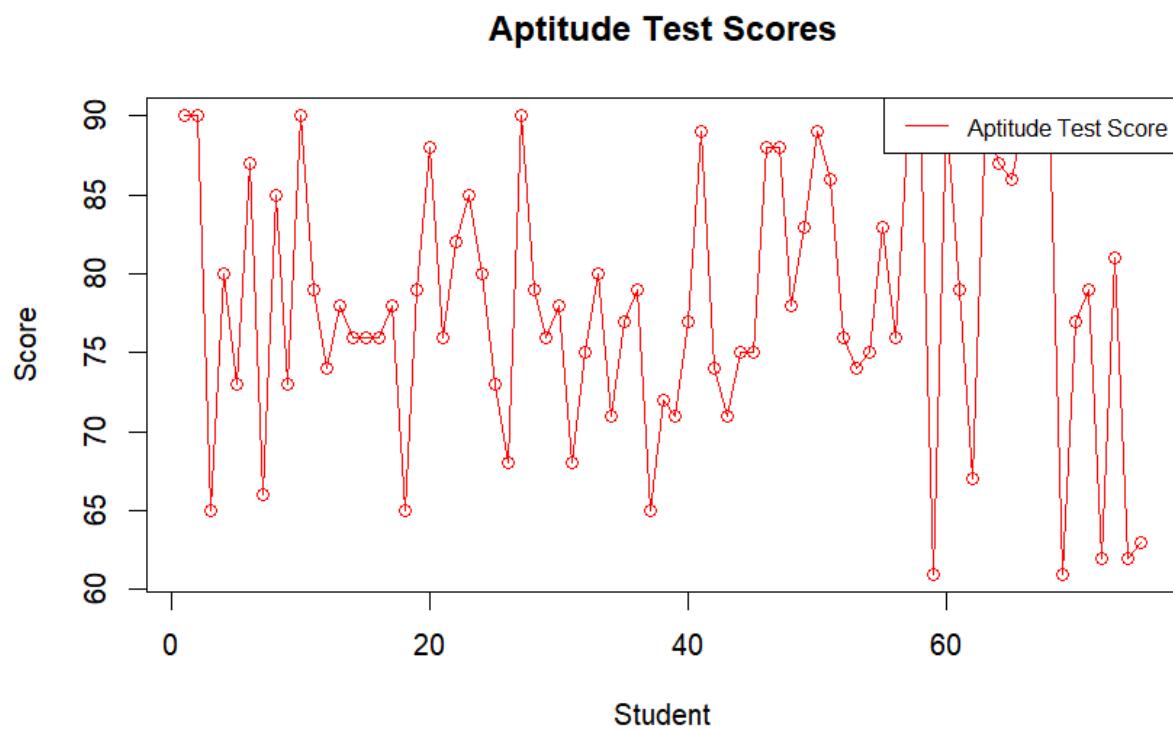
R

```
# Create a pie chart
pie(placement_counts,
    labels = c("Not Placed", "Placed"),
    col = c("lightblue", "lightgreen"),
    main = "Placement Status")
#line graph
CGPA <- data$CGPA
AptitudeTestScore <- data$AptitudeTestScore
```

Placement Status



```
# -----Create a line plot
plot(xlab = "Student", ylab = "Score", main = "Aptitude Test Scores",
AptitudeTestScore) # Plot with data
lines(AptitudeTestScore, type = "o", col = "red") # Add data points with style
legend("topright", legend = c("Aptitude Test Score"), col = "red", lty = 1, cex
= 0.8)
```



Excel

Pie Chart:

Step 1: Organize your data: Place categories in one column and values in an adjacent column.

Step 2: Select your data range, including headers.

Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Pie or Doughnut Chart" button.

Step 5: Select the type of pie chart you want (2D or 3D).

Step 6: Excel will create the chart. You can then customize it:

Line Diagram:

Step 1: Organize your data in columns, with categories (often dates or time periods) in the first column and data series in adjacent columns.

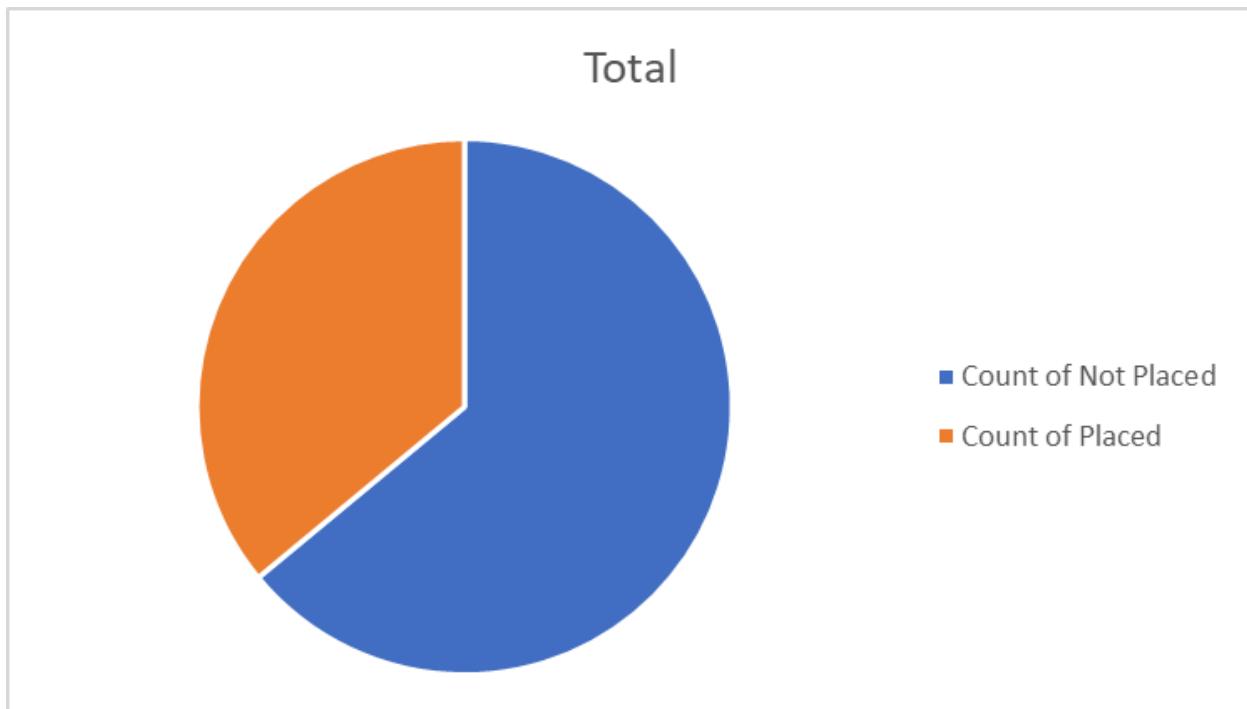
Step 2: Select your data range, including headers.

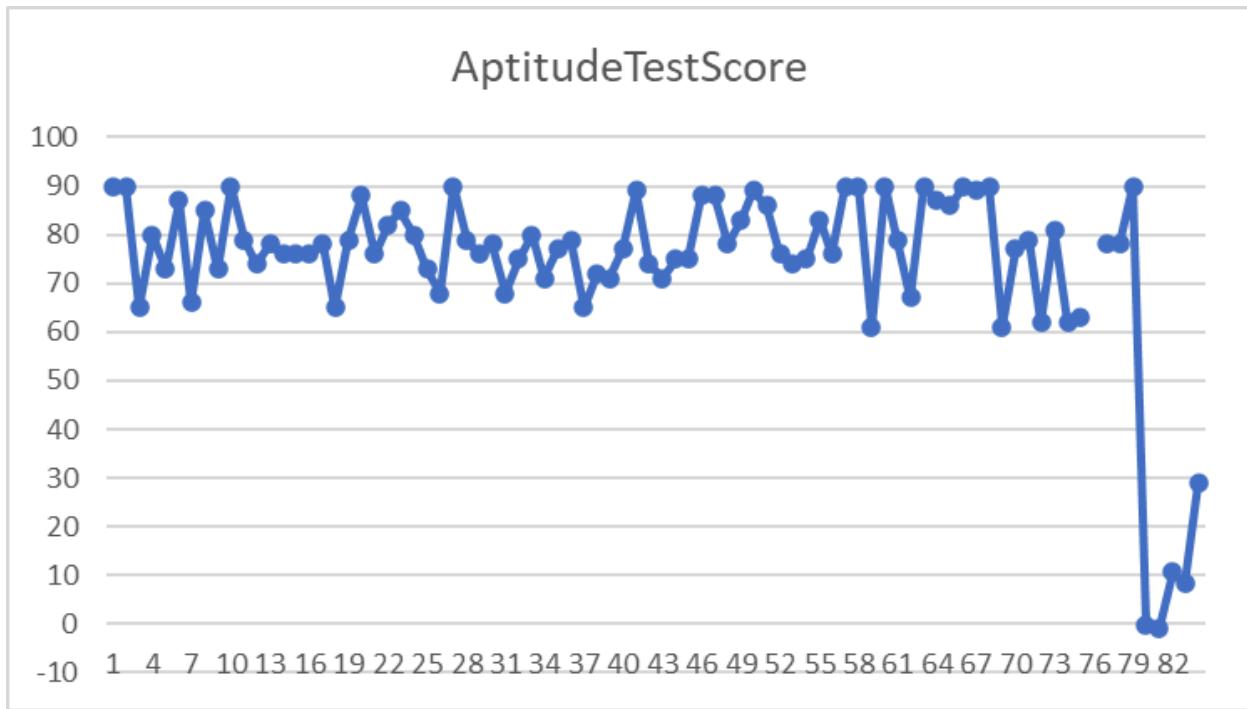
Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Line or Area Chart" button.

Step 5: Select the type of line chart you want (e.g., "Line" for a basic line chart, or "Line with Markers" to show data points).

Step 6: Excel will create the chart. You can then customize it.

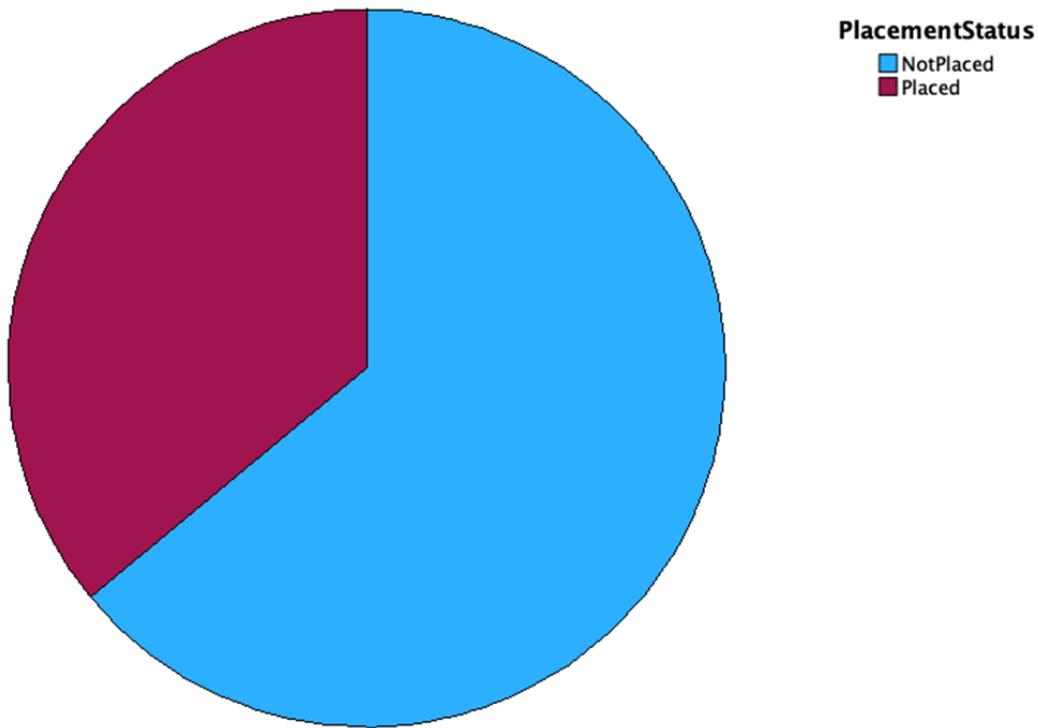




SPSS

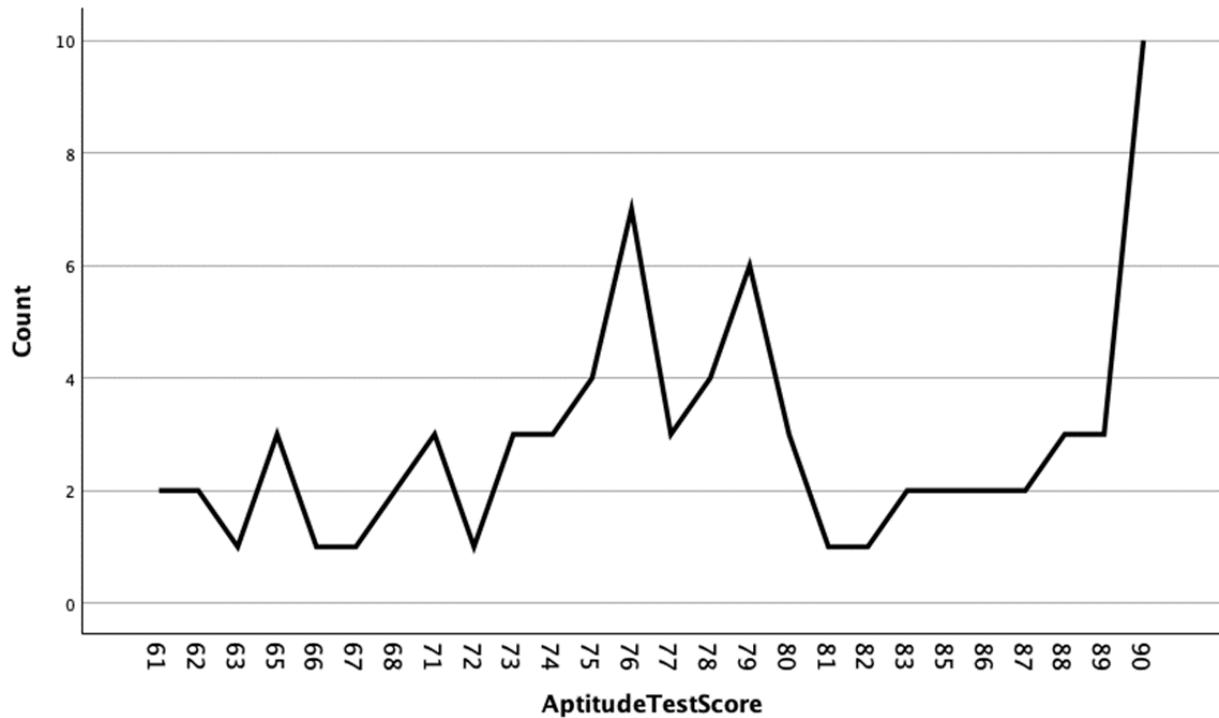
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Pie/Polar option.
3. Drag the categorical variable to the slice by panel.



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Line option.
3. Drag the time variable to the x-axis and the measurement variable to the y-axis.



Practical Output

PIE CHART: The pie chart shows that a majority of students or graduates are not placed (blue section), while a smaller portion are placed (purple section). This suggests the HR department may need to focus on improving job placement rates for students/alumni.

Prioritize improving job placement rates as the majority of students are not placed. Develop strategies to increase placement success.

LINE GRAPH The line graph shows aptitude test scores ranging from approximately 61 to 90. There are peaks at scores around 77 and 80, indicating these are common score ranges. The HR department could use this data to set benchmarks for recruitment or to identify areas where students might need additional support.

CASE STUDY 2 - Marketing

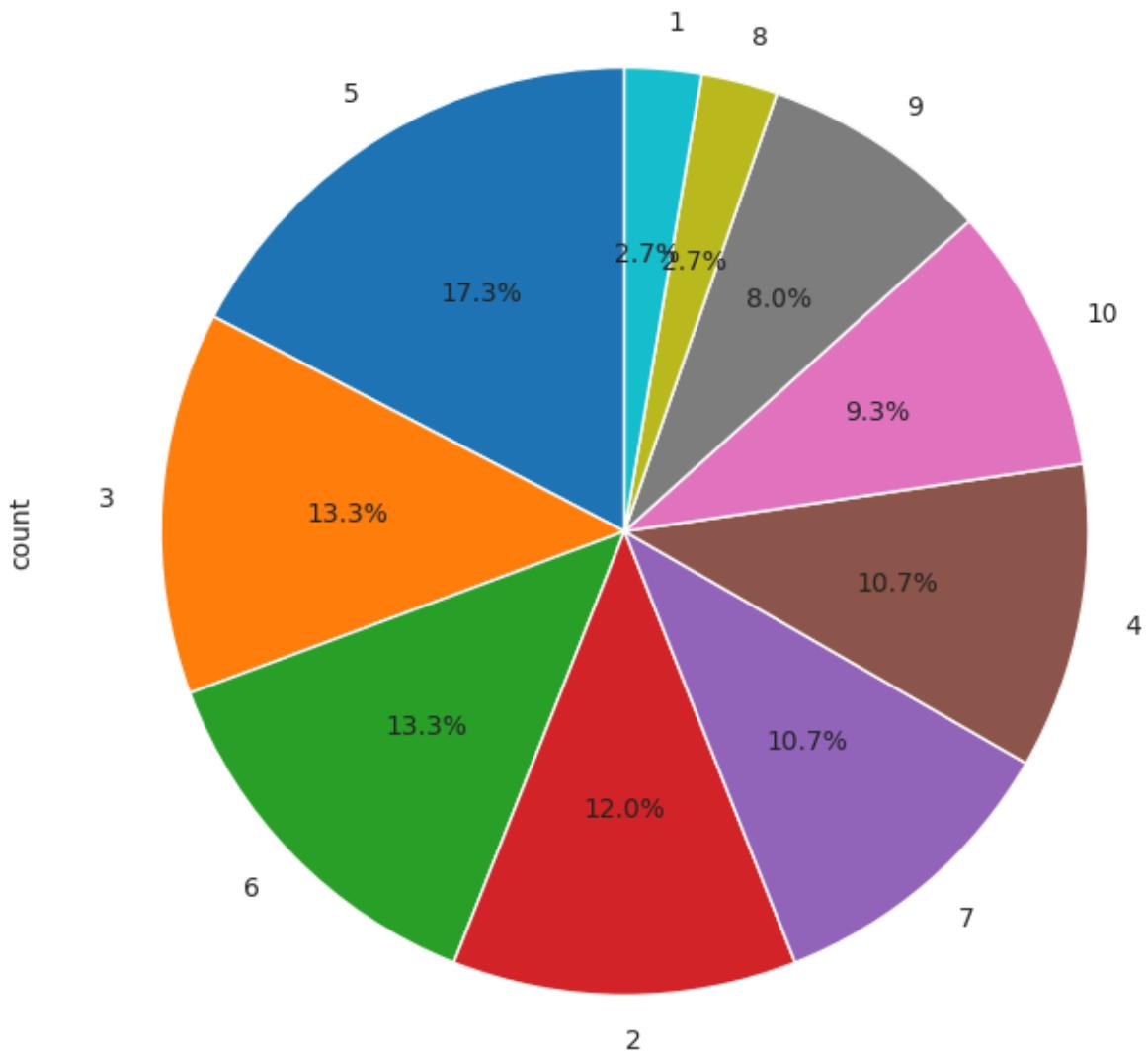
Python

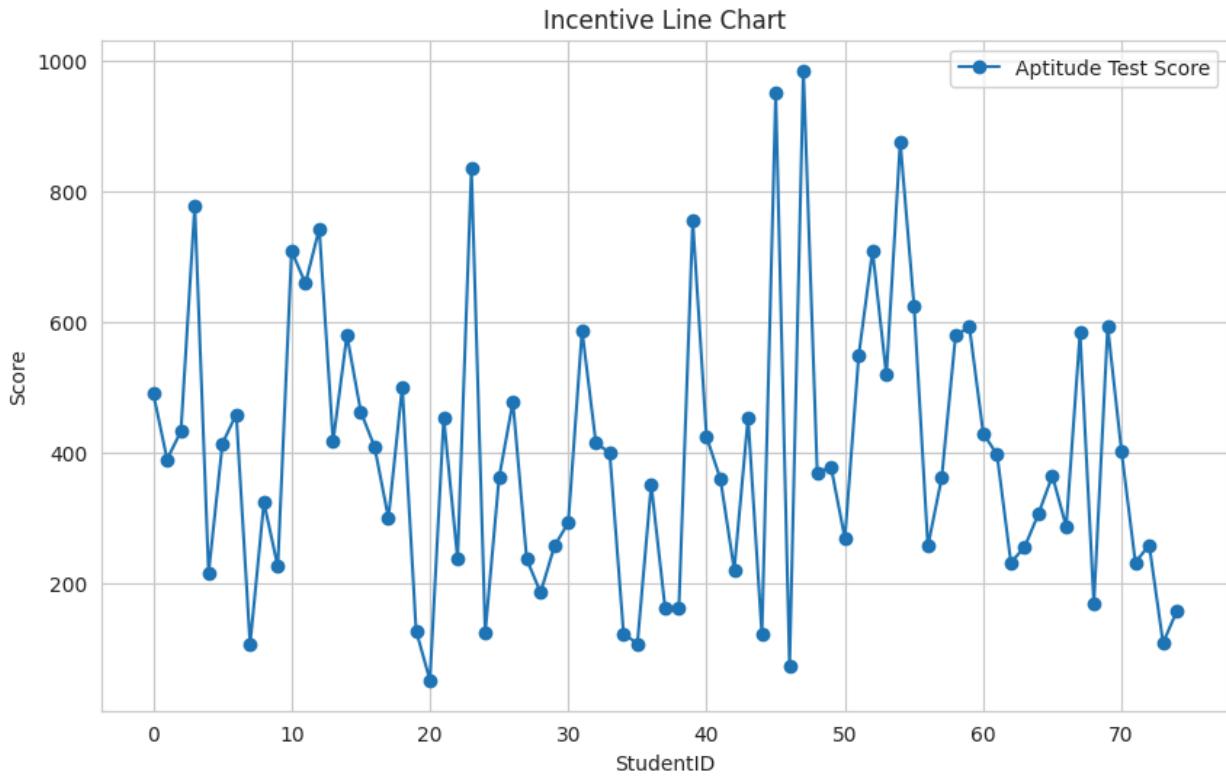
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
# Assuming your data is in a DataFrame named 'df'
# Pie Diagram for PlacementStatus
df= pd.read_csv('merged.csv')
```

```
plt.figure(figsize=(8, 8))
df['Hours Marketing'].value_counts().plot.pie(autopct='%1.1f%%',
startangle=90)
plt.title('Hours Marketing Pie Chart')
plt.show()

# Line Diagram for CGPA and AptitudeTestScore
plt.figure(figsize=(10, 6))
plt.plot(df['Incentive Received'], label='Aptitude Test Score',
marker='o')
plt.title('Incentive Line Chart')
plt.xlabel('StudentID')
plt.ylabel('Score')
plt.legend()
plt.show()
```

Placement Status Pie Chart





R

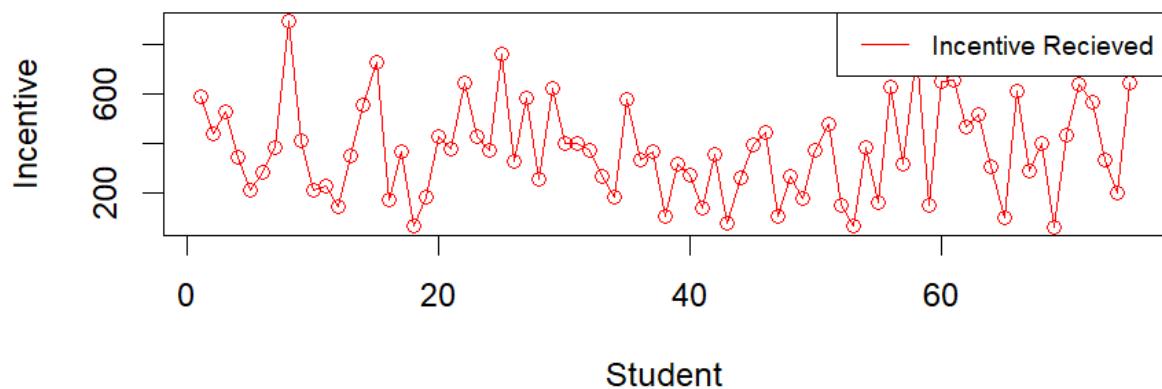
```
> #-----pie and line
> marketing <- table(data$Hours.Marketing)
>
> # Create a pie chart
> pie(marketing,
+     labels = c("10","2","3","4","5","6","7","9","8"),
+     col = c("lightblue",
"lightgreen","red","green","blue","yellow","cyan","maroon","purple"),
+     main = "Hours.Marketing")
> #line graph
> market <- data$Hours.Marketing
> Incentive<- data$Incentive.Received
>
> # -----Create a line plot
> plot(xlab = "Student", ylab = "Incentive", main = "Incentive Recieved",
Incentive) # Plot with data
> lines(Incentive, type = "o", col = "red") # Add data points with style
> legend("topright", legend = c("Incentive Recieved"), col = "red", lty = 1,
cex = 0.8)
```

>

Hours.Marketting



Incentive Recieved



Excel

Pie Chart:

Step 1: Organize your data: Place categories in one column and values in an adjacent column.

Step 2: Select your data range, including headers.

Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Pie or Doughnut Chart" button.

Step 5: Select the type of pie chart you want (2D or 3D).

Step 6: Excel will create the chart. You can then customize it:

Line Diagram:

Step 1: Organize your data in columns, with categories (often dates or time periods) in the first column and data series in adjacent columns.

Step 2: Select your data range, including headers.

Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Line or Area Chart" button.

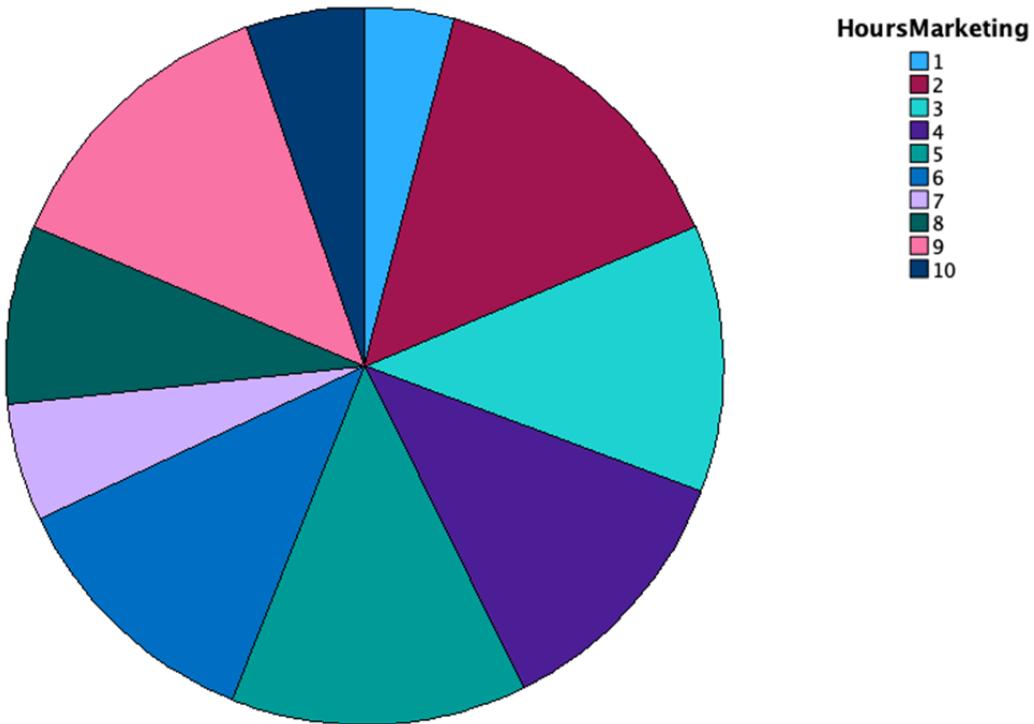
Step 5: Select the type of line chart you want (e.g., "Line" for a basic line chart, or "Line with Markers" to show data points).

Step 6: Excel will create the chart. You can then customize it.

SPSS

Steps:

1. Go to Graphs > Chart Builder.
2. Select the Pie/Polar option.
3. Drag the categorical variable to the slice by panel.



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Line option.
3. Drag the time variable to the x-axis and the measurement variable to the y-axis.



Practical Output

Overall Hours Marketing Distribution (Image 4): The pie chart represents the overall distribution of marketing hours. It shows a relatively even distribution across different hour categories, indicating varied time commitments for marketing activities.

PO

Develop marketing strategies that accommodate diverse time commitments, as hours are relatively evenly distributed.

LINE CHART

Hours Marketing Trend (Image 5): This line graph shows fluctuations in marketing hours, with peaks at 2, 5, and 9 hours. This suggests that marketing activities tend to cluster around these durations.

CASE STUDY 3 - Operations

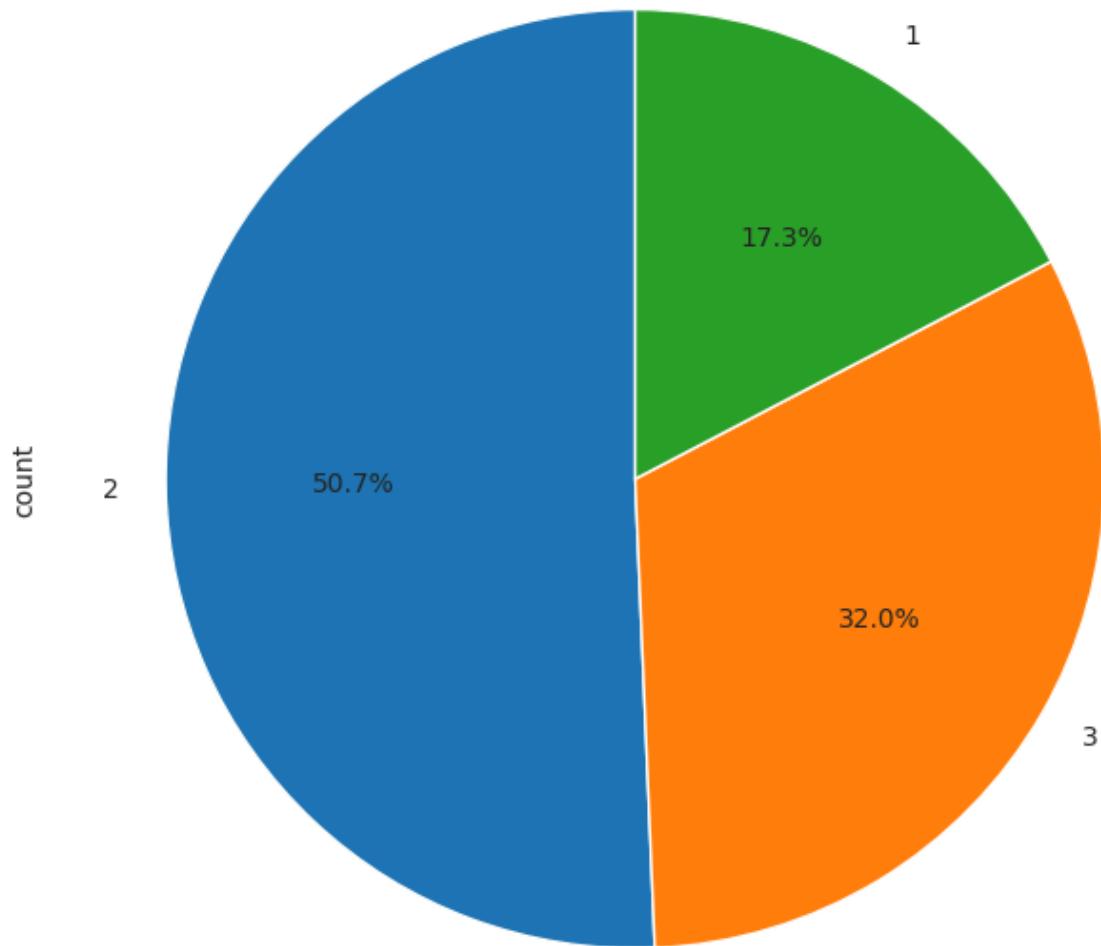
Python

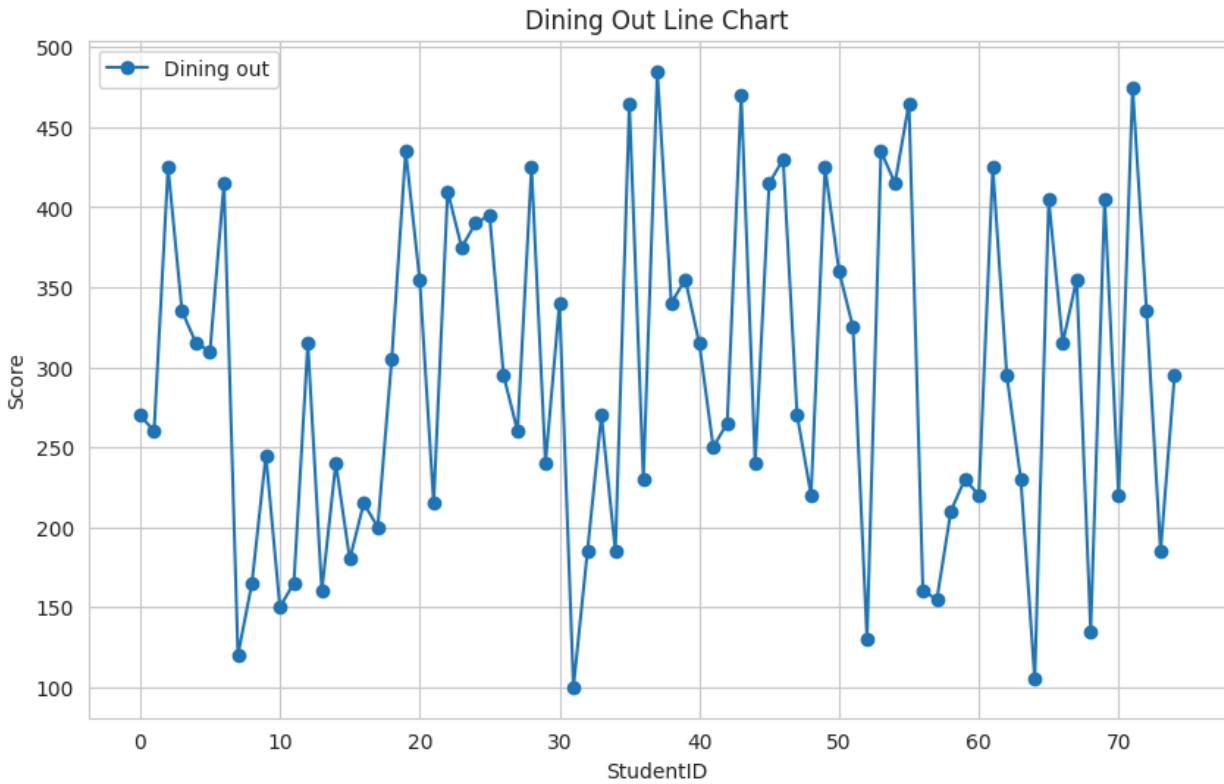
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
# Assuming your data is in a DataFrame named 'df'
```

```
# Pie Diagram for Room Type
df= pd.read_csv('merged.csv')
plt.figure(figsize=(8, 8))
df['Room Type'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90)
plt.title('Room Type Pie Chart')
plt.show()

# Line Diagram for CGPA and Dining Out
plt.figure(figsize=(10, 6))
plt.plot(df['Dining Out'], label='Dining out', marker='o')
plt.title('Dining Out Line Chart')
plt.xlabel('StudentID')
plt.ylabel('Score')
plt.legend()
plt.show()
```

Room Type Pie Chart





R

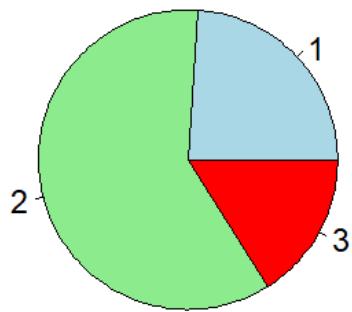
```

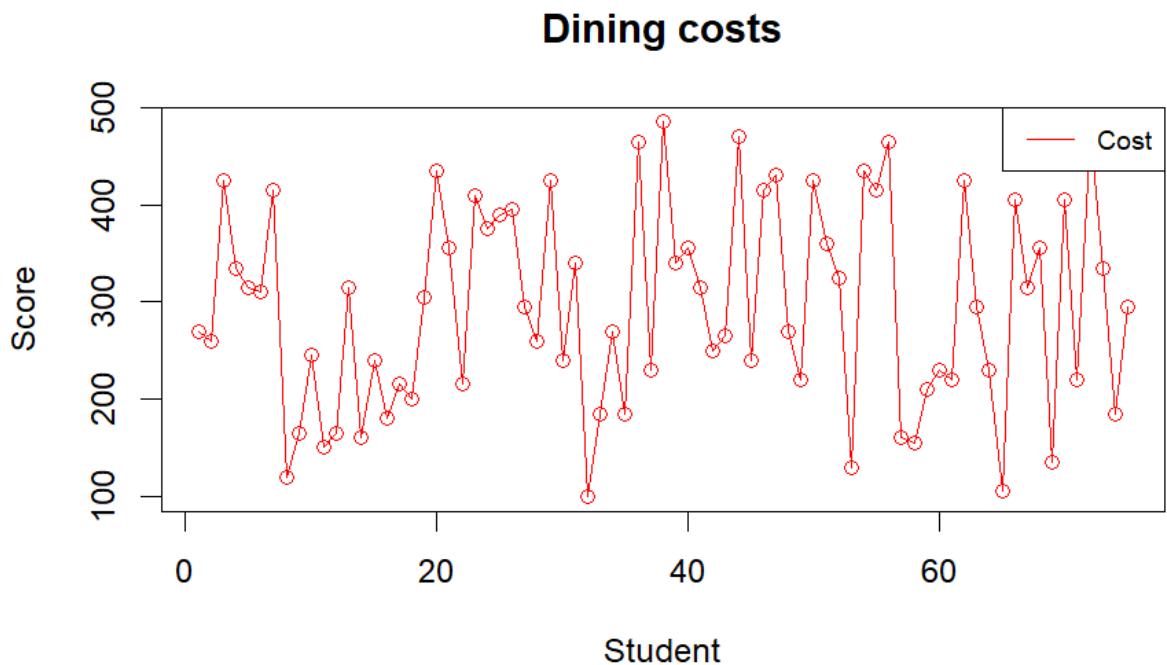
> #-----pie and line
> room_counts <- table(data$Room.Type)
>
> # Create a pie chart
> pie(room_counts,
+     labels = c("1", "2","3"),
+     col = c("lightblue", "lightgreen","red"),
+     main = "Rooms")
> #line graph
> dining <- data$Dining.Out
> Room <- data$Room.Type
> cgpa <- data$CGPA
>
> # -----Create a line plot
> plot(xlab = "Student", ylab = "Score", main = "Dining costs", dining) # Plot with data
> lines(dining, type = "o", col = "red") # Add data points with style
> legend("topright", legend = c("Cost"), col = "red", lty = 1, cex = 0.8)

```

>

Rooms





Excel

Pie Chart:

Step 1: Organize your data: Place categories in one column and values in an adjacent column.

Step 2: Select your data range, including headers.

Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Pie or Doughnut Chart" button.

Step 5: Select the type of pie chart you want (2D or 3D).

Step 6: Excel will create the chart. You can then customize it:

Line Diagram:

Step 1: Organize your data in columns, with categories (often dates or time periods) in the first column and data series in adjacent columns.

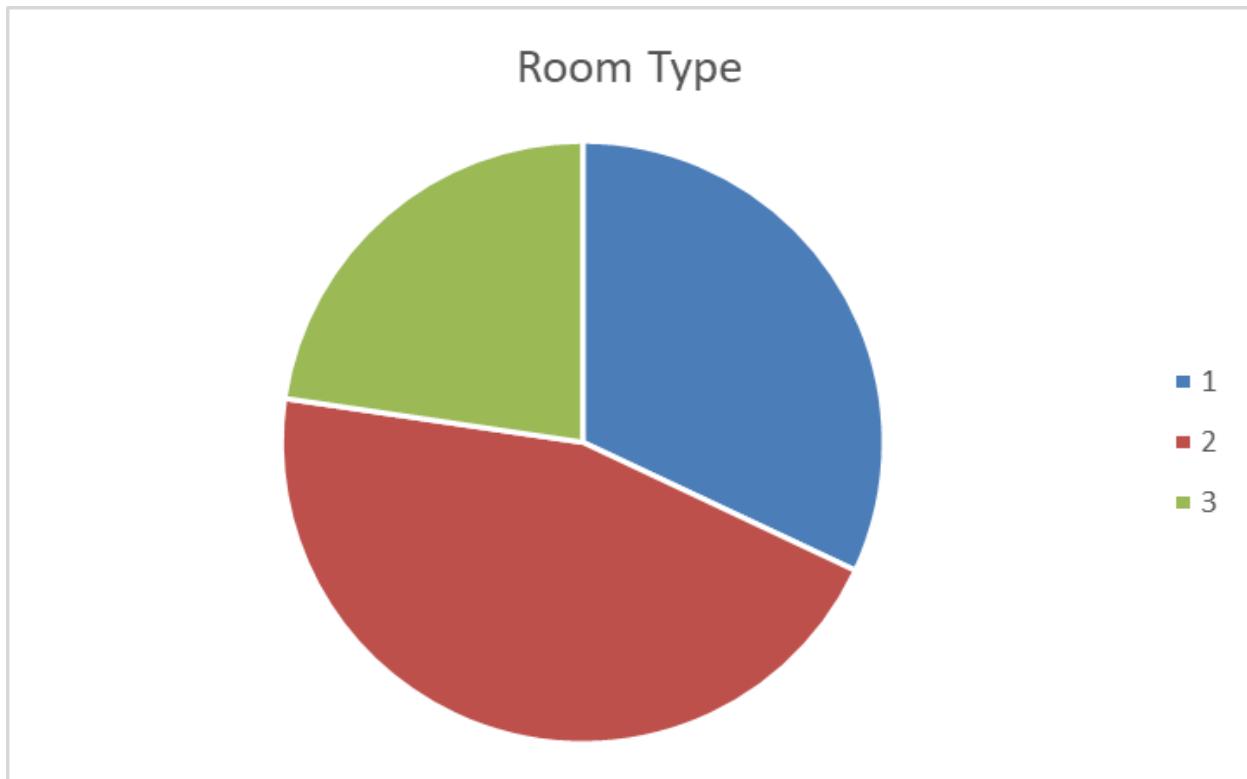
Step 2: Select your data range, including headers.

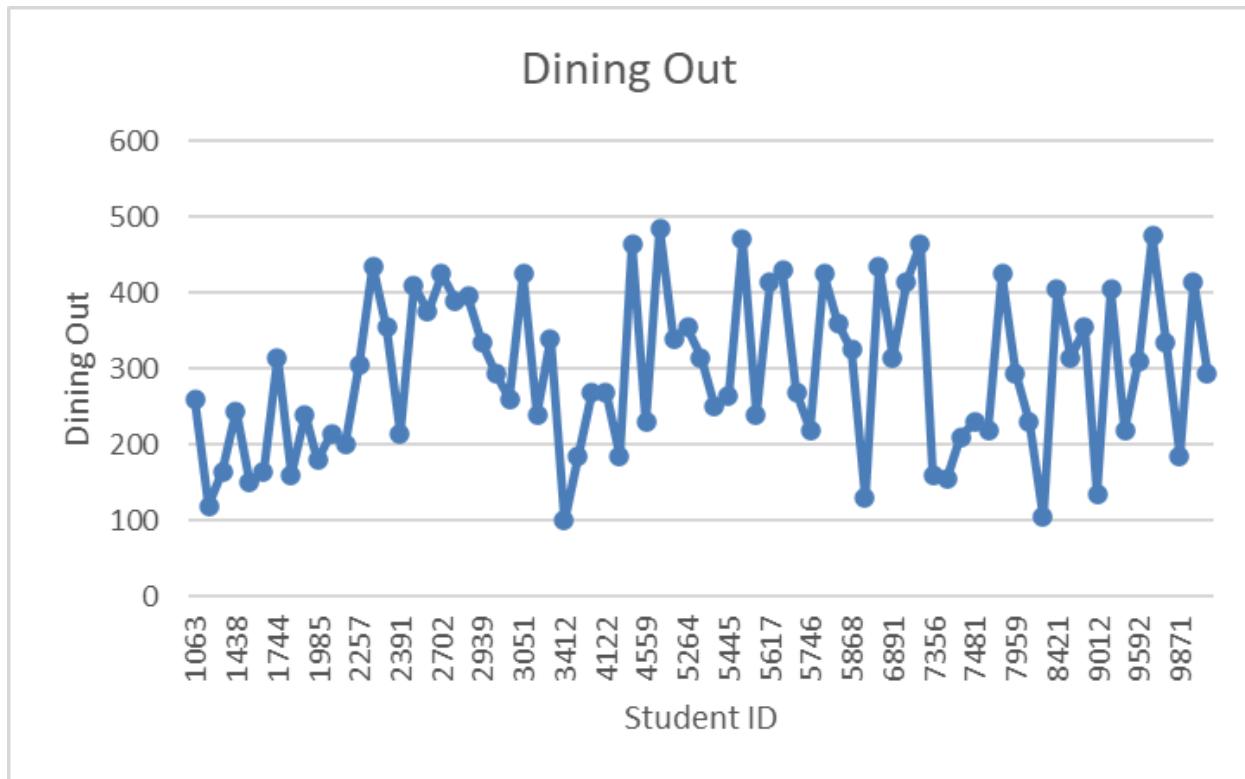
Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Line or Area Chart" button.

Step 5: Select the type of line chart you want (e.g., "Line" for a basic line chart, or "Line with Markers" to show data points).

Step 6: Excel will create the chart. You can then customize it.

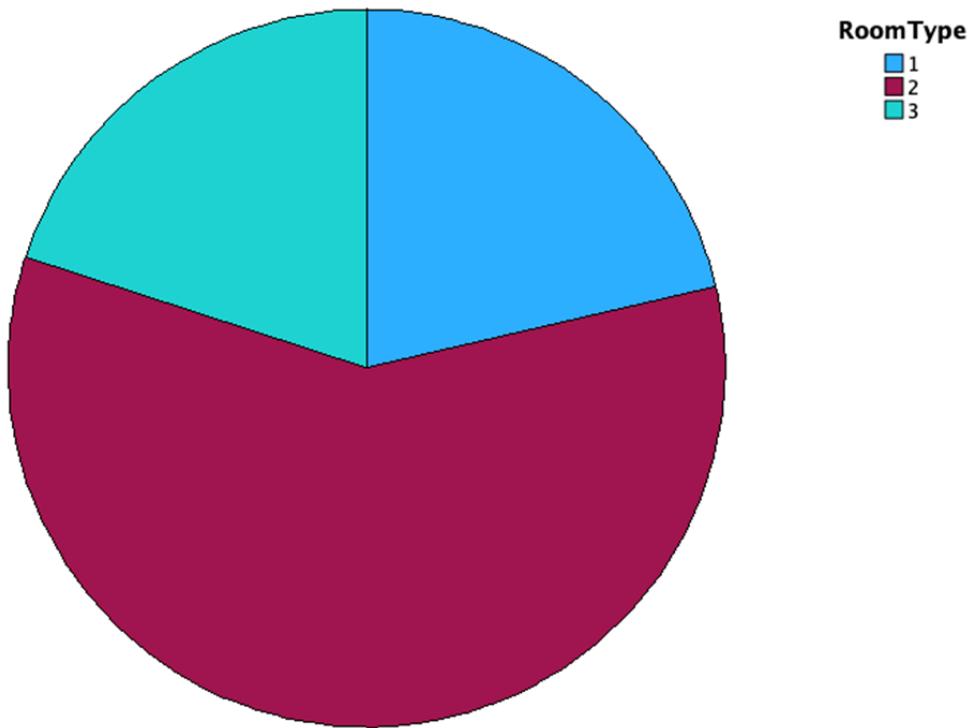




SPSS

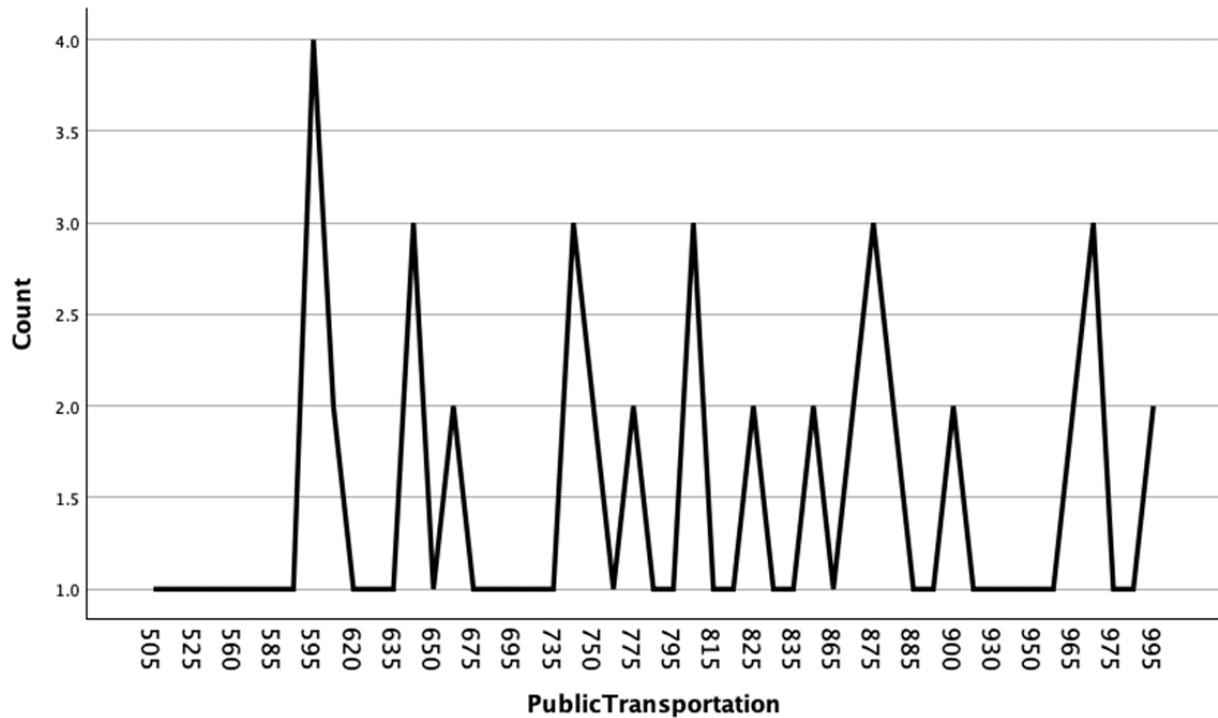
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Pie/Polar option.
3. Drag the categorical variable to the slice by panel.



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Line option.
3. Drag the time variable to the x-axis and the measurement variable to the y-axis.



Practical Output

This graph represents the amount of money students spend on public transportation. The x-axis shows dollar amounts spent on transportation, ranging from \$505 to \$995. The y-axis indicates the frequency or count of students spending each amount.

Key observations:

1. Spending varies widely, with amounts ranging from about \$505 to \$995.
2. There are several peaks in the graph, suggesting common spending amounts:
 - o The highest peak is at \$620, indicating this is the most common amount spent.
 - o Other notable peaks occur around \$655, \$730, \$815, and \$965.
3. There are also periods of lower spending between these peaks.
4. The pattern is irregular, suggesting various factors influence student transportation costs.

Interpretation for college operations:

1. Cost variability: The wide range of spending (\$505 to \$995) indicates significant differences in students' transportation needs or choices.
2. Common expenditures: The peaks represent the most frequent spending amounts, which could correspond to different types of transportation passes or common commute distances.
 - o Negotiate with local transit authorities for student discounts or passes.
 - o Plan for parking needs on campus.
 - o Consider implementing a college shuttle service for common routes.
 - o Develop targeted financial aid or subsidies for transportation costs.
3. Planning and support: The operations department can use this data to:
 - o Negotiate with local transit authorities for student discounts or passes.
 - o Plan for parking needs on campus.
 - o Consider implementing a college shuttle service for common routes.
 - o Develop targeted financial aid or subsidies for transportation costs.
4. Budgeting assistance: Help students budget for transportation expenses, especially focusing on the most common amounts spent.

Plan for varied student transportation needs, potentially negotiating discounts or implementing shuttle services for common routes.

PIE CHART

The pie chart shows the distribution of room types, possibly classrooms or office spaces. There are three types, with Type 2 (burgundy) being the most common, followed by Type 3 (teal), and Type 1 (blue) being the least common. This information could help the operations department in space allocation and maintenance planning.

Optimize resource allocation based on utility usage patterns. Consider room assignment strategies to support academic performance.

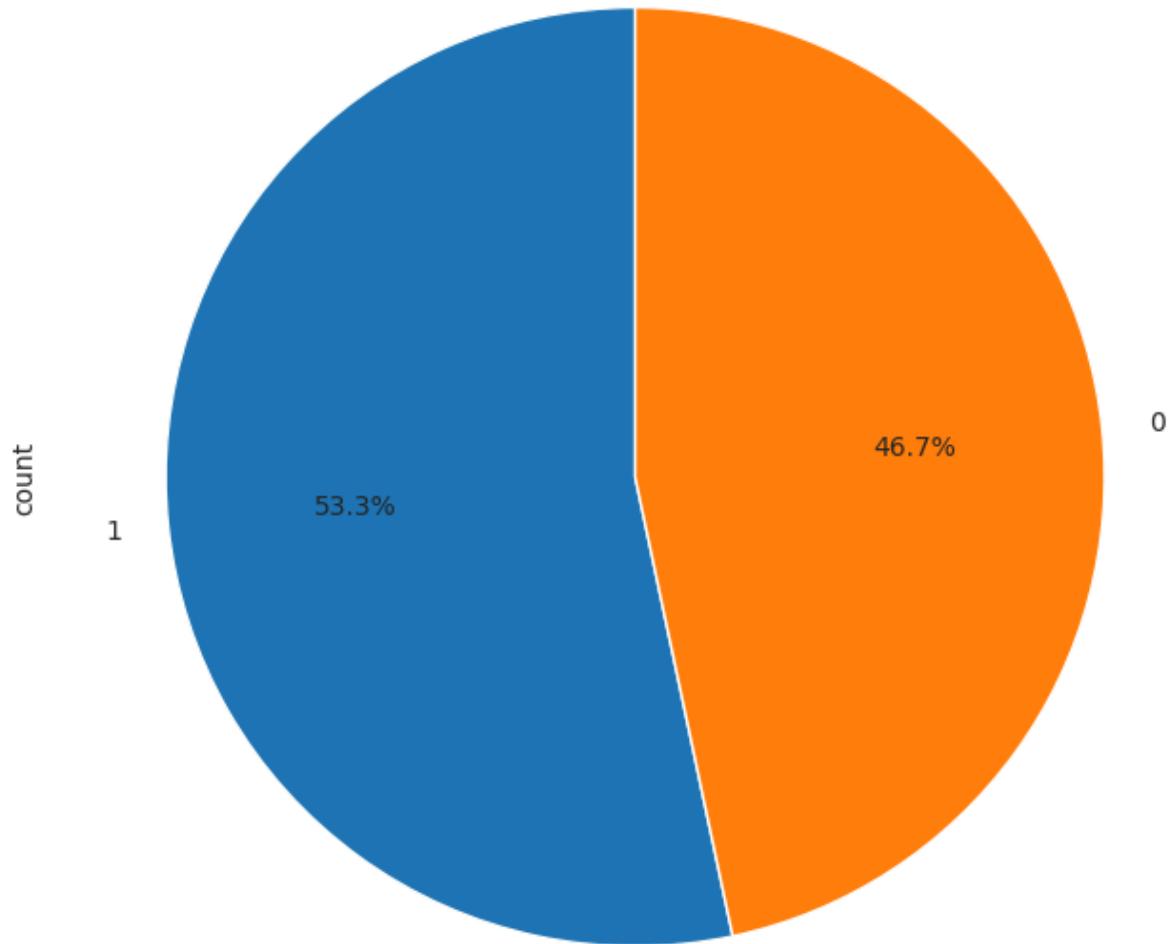
CASE STUDY 4 - Finance

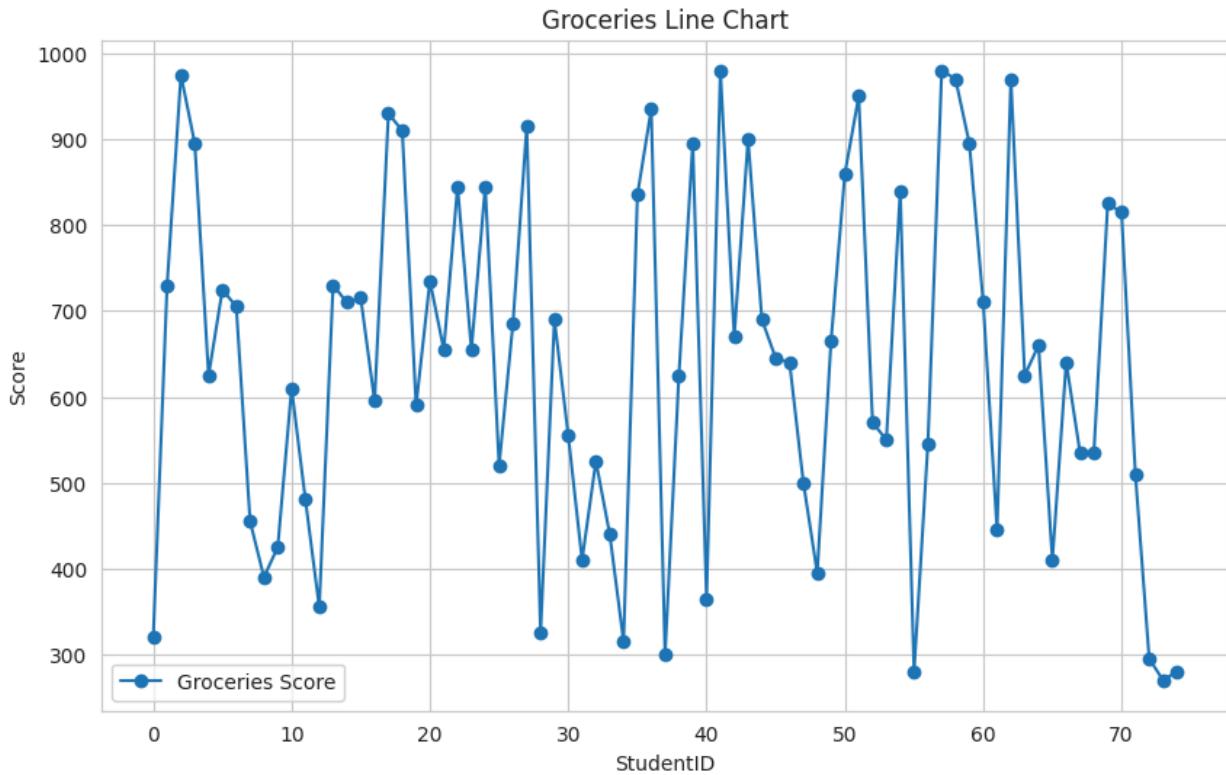
Python

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
# Assuming your data is in a DataFrame named 'df'
# Pie Diagram for Food Preference
df= pd.read_csv('merged.csv')
plt.figure(figsize=(8, 8))
df['Food Preference'].value_counts().plot.pie(autopct='%1.1f%%',
startangle=90)
plt.title('Food Preference Pie Chart')
plt.show()

# Line Diagram for Groceries and Dining out
plt.figure(figsize=(10, 6))
plt.plot(df['Groceries'], label='Groceries Score', marker='o')
plt.title('Groceries Line Chart')
plt.xlabel('StudentID')
plt.ylabel('Score')
plt.legend()
plt.show()
```

Food Preference Pie Chart





R

Excel

Pie Chart:

Step 1: Organize your data: Place categories in one column and values in an adjacent column.

Step 2: Select your data range, including headers.

Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Pie or Doughnut Chart" button.

Step 5: Select the type of pie chart you want (2D or 3D).

Step 6: Excel will create the chart. You can then customize it:

Line Diagram:

Step 1: Organize your data in columns, with categories (often dates or time periods) in the first column and data series in adjacent columns.

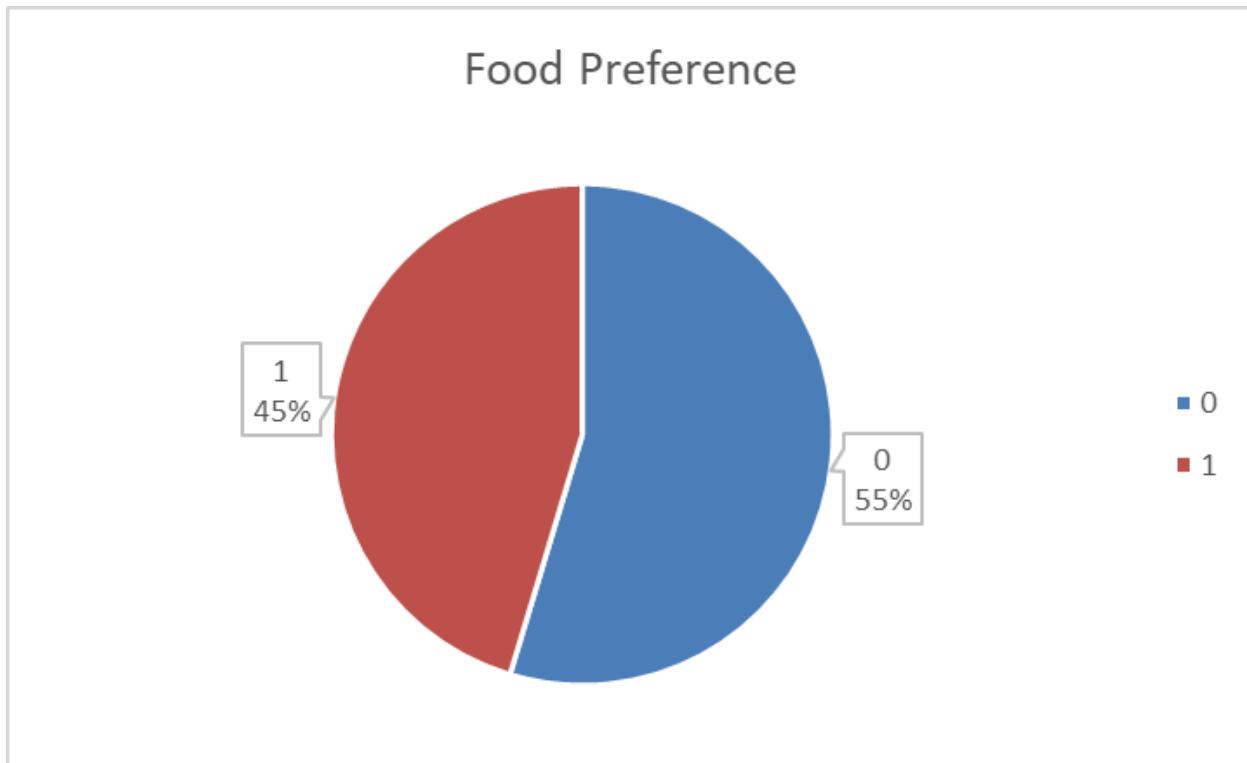
Step 2: Select your data range, including headers.

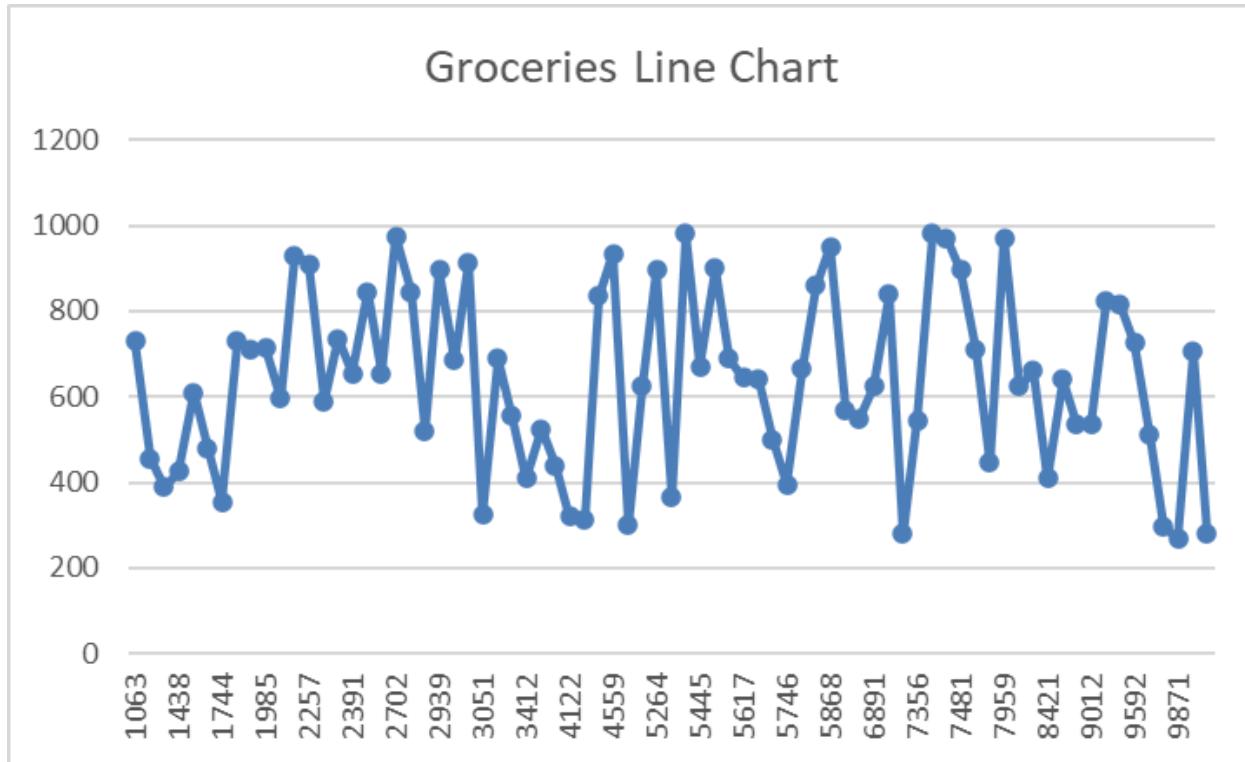
Step 3: Go to the "Insert" tab on the Excel ribbon.

Step 4: In the "Charts" group, click on the "Insert Line or Area Chart" button.

Step 5: Select the type of line chart you want (e.g., "Line" for a basic line chart, or "Line with Markers" to show data points).

Step 6: Excel will create the chart. You can then customize it.

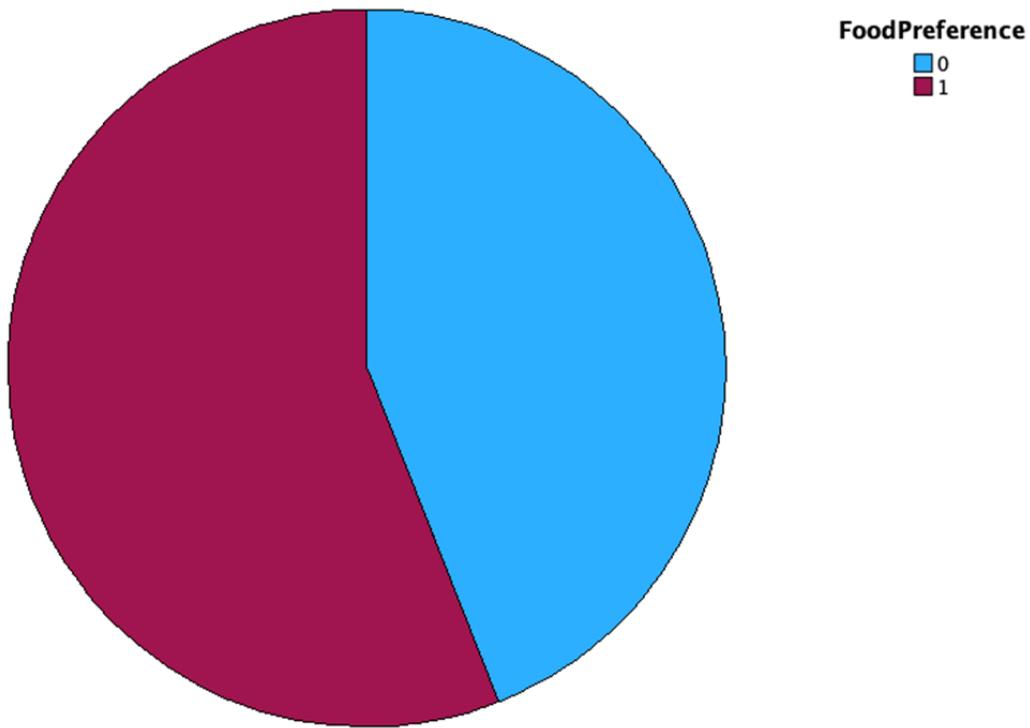




SPSS

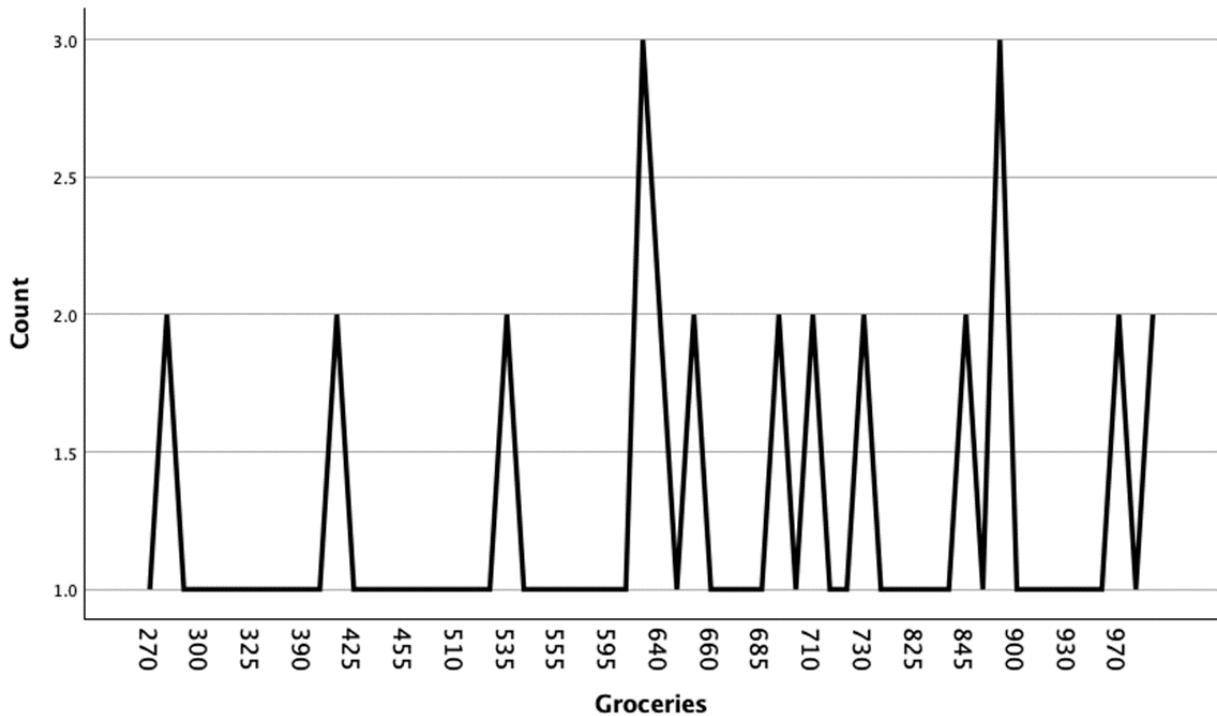
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Pie/Polar option.
3. Drag the categorical variable to the slice by panel.



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Line option.
3. Drag the time variable to the x-axis and the measurement variable to the y-axis.



Practical Output

The pie chart shows a fairly even split between two food preferences, labeled as 0 and 1. This could represent different meal plan options or dietary preferences. The finance department should ensure budget allocations accommodate both options fairly. Ensure budget allocations accommodate both food preference options fairly. Analyze potential correlations between food choices and academic performance.

CH 3- Mean, Median, Mode, Skewness, Kurtosis

Mean, Median, Mode, Skewness, Kurtosis

In the realm of statistical analysis, the quintessential elements that form the backbone of understanding data distributions are Mean, Median, Mode, Skewness, and Kurtosis. Each of these metrics offers a unique lens through which to view and interpret data, providing insights into the central tendency, variability, and shape of data distributions. This section of the Harvard case study statistics book aims to delve deep into these concepts, illustrating their significance and applications with a blend of theoretical rigor and real-world examples.

Mean: The Balancing Point

The Mean, or arithmetic average, is perhaps the most familiar measure of central tendency. It is calculated by summing all the values in a dataset and dividing by the number of values. The Mean serves as a gravitational center of the dataset, offering a single value that summarizes the entire dataset.

Application in Case Studies:

Consider a case study examining the effectiveness of a new teaching method across several schools. The Mean score of student performance across all schools provides a quick snapshot of the method's overall effectiveness, serving as a baseline for further analysis.

Insights:

- Weighted Mean: In scenarios where different data points contribute unequally to the overall outcome, the weighted mean becomes a powerful tool, allowing for a nuanced understanding that acknowledges the varying importance of each data point.

Median: The Middle Ground

The Median represents the middle value in a dataset when the values are arranged in ascending or descending order. It divides the dataset into two equal halves and is unaffected by extreme values or outliers, making it a robust measure of central tendency.

Application in Case Studies:

In analyzing the annual income of residents in a city, the Median income provides a clearer picture of the "typical" income than the Mean, which could be skewed by very high or very low incomes. The Median offers insight into the economic status of the average resident.

Insights:

- Segmented Median Analysis: Breaking down datasets into smaller segments and analyzing the Median of each can uncover underlying patterns and disparities within the data, offering a more granular perspective on the distribution.

Mode: The Common Thread

The Mode is the value that appears most frequently in a dataset. It can be used with any level of data measurement and is particularly useful in identifying the most common or popular choices within a dataset.

Application in Case Studies:

Studying consumer preferences for a particular product feature, the Mode can identify the most preferred feature among consumers. This insight is invaluable for guiding product development and marketing strategies.

Insights:

- Multi-modal Distributions: Identifying datasets with more than one mode can signal significant subgroups within the population, prompting further investigation into the characteristics that define these subgroups.

Skewness: The Asymmetry Indicator

Skewness measures the asymmetry of a distribution around its mean. Positive skew indicates a tail that extends towards more positive values, while negative skew indicates a tail extending towards more negative values. Skewness provides insights into the distribution's deviation from normality.

Application in Case Studies:

Evaluating investment returns, skewness can help in understanding the risk associated with an investment. A positively skewed distribution of returns indicates a higher likelihood of achieving above-average returns, albeit with the risk of extreme negative outliers.

Insights:

- Skewness Adjustment Strategies: In financial modeling, adjusting for skewness can lead to more accurate risk assessments and investment strategies, acknowledging the asymmetrical risks inherent in market returns.

Kurtosis: The Peak and Tail Descriptor

Kurtosis measures the "tailedness" of a distribution, offering insights into the concentration of values around the mean (peak) and the propensity of the distribution to produce outliers (tail). High kurtosis indicates a peaked distribution with fat tails, while low kurtosis indicates a flat distribution.

Application in Case Studies:

Analyzing the distribution of customer satisfaction ratings, kurtosis can indicate the likelihood of extreme satisfaction or dissatisfaction. High kurtosis suggests a greater chance of extreme responses, highlighting areas for potential improvement or innovation.

Insights:

- Tail-Risk Management: In risk management, understanding the kurtosis of loss distributions can be critical. High kurtosis indicates a higher risk of extreme losses, guiding the development of strategies to mitigate tail risks.

CASE STUDY 1 - HR

Python

```
from scipy.stats import skew, kurtosis
from scipy.stats import spearmanr

# Calculate Mean, Median, Mode, Skewness, Kurtosis
statistics = pd.DataFrame(columns=['Mean', 'Median', 'Mode', 'Skewness',
'Kurtosis'])
for column in ['CGPA', 'AptitudeTestScore']:
    statistics.loc[column] = [
        data[column].mean(),
        data[column].median(),
        data[column].mode()[0], # Mode might return multiple values; take
the first one
        skew(data[column]),
        kurtosis(data[column])
    ]

print(statistics)
```

	Mean	Median	Mode	Skewness	Kurtosis
CGPA	7.740000	7.8	7.5	-0.569833	-0.362640
AptitudeTestScore	78.186667	78.0	90.0	-0.231548	-0.769832

R

```
> # -----Basic Statistical Analysis(mean,median,mode)
> summary(data$CGPA)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  6.50   7.50   7.80  7.74   8.20   8.80
> summary(data$AptitudeTestScore)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  61.00  73.50  78.00  78.19   86.00  90.00

> #----- Skewness and Kurtosis
> library(e1071)
> skewness(data$CGPA)
[1] -0.5584748
> kurtosis(data$CGPA)
[1] -0.432501
```

Excel

Mean, Median, and Mode

Step 1: Enter your data into a single column (e.g., column A).

Step 2: For the mean, in an empty cell, type: =AVERAGE(A1:A100)

Step 3: For the median, in another empty cell, type: =MEDIAN(A1:A100)

Step 4: For the mode, in another empty cell, type: =MODE.SNGL(A1:A100)

Note: Replace A1:A100 with your actual data range.

Skewness and Kurtosis

Step 1: Ensure your data is in a single column (e.g., column A).

Step 2: For skewness, in an empty cell, type: =SKEW(A1:A100)

Step 3: For kurtosis, in another empty cell, type: =KURT(A1:A100)

Mean	78.18666667
Median	78
Mode	90
Skewness	-0.236300943
Kurtosis	-0.739257427

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select the statistics you need (mean, median, mode, etc.), and click Continue.

Statistics

		CGPA	DiningOut
N	Valid	75	75
	Missing	0	0
Mean		7.740	295.00
Median		7.800	295.00
Mode		7.5 ^a	315 ^a
Skewness		-.582	.041
Std. Error of Skewness		.277	.277
Kurtosis		-.304	-1.069
Std. Error of Kurtosis		.548	.548

Sum	580.5	22125
-----	-------	-------

- a. Multiple modes exist. The smallest value is shown

Practical Output

- CGPA: Mean: 7.740 Median: 7.800 Mode: 7.5 (multiple modes exist) Slightly negatively skewed (-0.582)
- Dining Out: Mean and Median: 295.00 Mode: 315 Slightly positively skewed (0.041)

CASE STUDY 2 - Marketing

Python

```
from scipy.stats import skew, kurtosis
from scipy.stats import spearmanr

# Calculate Mean, Median, Mode, Skewness, Kurtosis
statistics = pd.DataFrame(columns=['Mean', 'Median', 'Mode', 'Skewness',
'Kurtosis'])
for column in ['Hours Marketing', 'Incentive Received']:
    statistics.loc[column] = [
        data[column].mean(),
        data[column].median(),
        data[column].mode()[0], # Mode might return multiple values; take
the first one
        skew(data[column]),
        kurtosis(data[column])
    ]
print(statistics)
```

	Mean	Median	Mode	Skewness	Kurtosis
Hours Marketing	5.373333	5.0	5.0	0.315333	-0.851997
Incentive Received	396.106667	377.0	162.0	0.682979	0.002179

R

```
> # -----Basic Statistical Analysis (mean,median,mode)
> summary(market)
  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
1.000   3.500  5.000   5.187   7.000 10.000
> summary(Incentive)
  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
61.0   217.5 364.0   373.1 498.5 898.0
> #----- Skewness and Kurtosis
> library(e1071)
> skewness(market)
[1] 0.1903577
> kurtosis(market)
[1] -0.9535828
```

Excel

Mean, Median, and Mode

Step 1: Enter your data into a single column (e.g., column A).

Step 2: For the mean, in an empty cell, type: =AVERAGE(A1:A100)

Step 3: For the median, in another empty cell, type: =MEDIAN(A1:A100)

Step 4: For the mode, in another empty cell, type: =MODE.SNGL(A1:A100)

Note: Replace A1:A100 with your actual data range.

Skewness and Kurtosis

Step 1: Ensure your data is in a single column (e.g., column A).

Step 2: For skewness, in an empty cell, type: =SKEW(A1:A100)

Step 3: For kurtosis, in another empty cell, type: =KURT(A1:A100)

MEAN	414.36
MEDIAN	391
MODE	481
SKEWNESS	0.194471977
KURTOSIS	-0.974635763

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select the statistics you need (mean, median, mode, etc.), and click Continue.

Statistics

	HoursMarketing	IncentiveReceived
N	Valid	75
	Missing	0
Mean	5.31	392.63
Median	5.00	319.00
Mode	2	155 ^a
Skewness	.213	.663
Std. Error of Skewness	.277	.277
Kurtosis	-1.143	-.403
Std. Error of Kurtosis	.548	.548

a. Multiple modes exist. The smallest value is shown

Practical Output

- The mean hours spent on marketing is 5.31, with a median of 5.00 and a mode of 2.
- There's a slight positive skew (.213) in the hours distribution, indicating some longer marketing sessions.

- The standard deviation of 2.661 suggests moderate variability in marketing hours.
 - The range of marketing hours is 9, spanning from 1 to 10 hours.
- Interpretation for College Marketing Department:
- The department likely has a main workspace (Room Type 2) where most marketing activities occur.
 - Marketing activities vary in duration, with common periods being 2, 5, and 9 hours.
 - There's a wide range of time spent on marketing (1-10 hours), indicating diverse project needs.
 - The average marketing session is about 5.31 hours, but there's considerable variation.
 - The department uses different incentives, with an average of \$392.63 and a median of \$319.00.
 - The positive skew in incentives suggests some higher-value incentives are occasionally used.

CASE STUDY 3 - Operations

Python

```
from scipy.stats import skew, kurtosis
from scipy.stats import spearmanr

# Calculate Mean, Median, Mode, Skewness, Kurtosis
statistics = pd.DataFrame(columns=['Mean', 'Median', 'Mode', 'Skewness',
'Kurtosis'])
for column in ['CGPA', 'Dining Out']:
    statistics.loc[column] = [
        data[column].mean(),
        data[column].median(),
        data[column].mode()[0], # Mode might return multiple values; take
        the first one
        skew(data[column]),
        kurtosis(data[column])
    ]
print(statistics)
```

	Mean	Median	Mode	Skewness	Kurtosis
CGPA	7.74	7.8	7.5	-0.569833	-0.362640
Dining Out	295.00	295.0	315.0	0.039821	-1.077681

R

```
> # -----Basic Statistical Analysis (mean,median,mode)
> summary(dining)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
100.0    217.5  295.0    295.0  392.5    485.0
> summary(CGPA)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
6.50    7.50    7.80    7.74    8.20    8.80
>
> #----- Skewness and Kurtosis
> library(e1071)
> skewness(data$CGPA)
[1] -0.5584748
> kurtosis(data$CGPA)
[1] -0.432501
```

Excel

Mean, Median, and Mode

Step 1: Enter your data into a single column (e.g., column A).
 Step 2: For the mean, in an empty cell, type: =AVERAGE(A1:A100)
 Step 3: For the median, in another empty cell, type: =MEDIAN(A1:A100)
 Step 4: For the mode, in another empty cell, type: =MODE.SNGL(A1:A100)
 Note: Replace A1:A100 with your actual data range.

Skewness and Kurtosis

Step 1: Ensure your data is in a single column (e.g., column A).
 Step 2: For skewness, in an empty cell, type: =SKEW(A1:A100)
 Step 3: For kurtosis, in another empty cell, type: =KURT(A1:A100)

MEAN	7.74
MEDIAN	7.8
MODE	7.5

SKEWNESS	-0.58152923
KURTOSIS	-0.30355569
	3

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select the statistics you need (mean, median, mode, etc.), and click Continue.

Statistics

DiningOut

N	Valid	75
	Missing	0
Mean		295.00
Median		295.00
Mode		315 ^a
Skewness		.041
Std. Error of Skewness		.277
Kurtosis		-1.069
Std. Error of Kurtosis		.548

a. Multiple modes exist. The smallest value is shown

Practical Output

- There are 75 valid data points with no missing values.
- The mean and median spending are both \$295.

- The mode (most frequent value) is \$315, but multiple modes exist.
- The standard deviation is \$104.526, indicating significant variation in dining expenses.
- The range of expenses is \$385, suggesting a wide spread of dining costs.

CASE STUDY 4 - Finance

Python

```
from scipy.stats import skew, kurtosis
from scipy.stats import spearmanr

# Calculate Mean, Median, Mode, Skewness, Kurtosis
statistics = pd.DataFrame(columns=['Mean', 'Median', 'Mode', 'Skewness',
'Kurtosis'])
for column in ['Groceries', 'Utilities']:
    statistics.loc[column] = [
        data[column].mean(),
        data[column].median(),
        data[column].mode()[0], # Mode might return multiple values; take
the first one
        skew(data[column]),
        kurtosis(data[column])
    ]
print(statistics)
```

	Mean	Median	Mode	Skewness	Kurtosis
Groceries	639.866667	645.0	625.0	-0.050817	-1.014951
Utilities	299.066667	295.0	330.0	0.046943	-1.178122

R

```
cv <- function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE) * 100
> cv(data$CGPA)
[1] 8.140837
```

Excel

Mean, Median, and Mode

Step 1: Enter your data into a single column (e.g., column A).
Step 2: For the mean, in an empty cell, type: =AVERAGE(A1:A100)
Step 3: For the median, in another empty cell, type: =MEDIAN(A1:A100)
Step 4: For the mode, in another empty cell, type: =MODE.SNGL(A1:A100)
Note: Replace A1:A100 with your actual data range.

Skewness and Kurtosis

Step 1: Ensure your data is in a single column (e.g., column A).
Step 2: For skewness, in an empty cell, type: =SKEW(A1:A100)
Step 3: For kurtosis, in another empty cell, type: =KURT(A1:A100)

MEAN	299.0666667
MEDIAN	295
MODE	330
SKEWNESS	0.047906773
KURTOSIS	-1.17613342
	5

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select the statistics you need (mean, median, mode, etc.), and click Continue.

Statistics

		Groceries	Utilities
N	Valid	75	75
	Missing	0	0

Mean	639.87	299.07
Median	645.00	295.00
Mode	625 ^a	330
Std. Deviation	208.121	27.085
Variance	43314.171	733.577
Skewness	-.052	.048
Std. Error of Skewness	.277	.277
Kurtosis	-1.002	-1.176
Std. Error of Kurtosis	.548	.548
Range	710	95

a. Multiple modes exist. The smallest value is shown

Practical Output

Groceries and Utilities: The statistics table provides insights into grocery and utility expenses:

Groceries:

- Mean: \$639.87
- Median: \$645.00
- Standard Deviation: \$208.121
- Range: \$710

This indicates significant variability in grocery expenses. The finance department should budget for an average of about \$640 per period (likely monthly) for groceries, but be prepared for fluctuations up to \$710.

Utilities:

- Mean: \$299.07
- Median: \$295.00
- Standard Deviation: \$27.085
- Range: \$95

Utility expenses are more consistent, with a tighter range and lower standard deviation.

Budget around \$300 per period for utilities, with less variability expected compared to groceries.

CH 4 - Standard Deviation, Variation, Range

Standard Deviation, Variation and Range

In the domain of statistical analysis, understanding the dispersion or variability of a dataset is paramount. This segment of the Harvard case study statistics book delves into three fundamental measures that elucidate this concept: Standard Deviation, Variance, and Range. These metrics, each with its unique lens, illuminate the spread of data points around the mean, offering critical insights into the nature of the dataset. Through a blend of theoretical exploration and practical application, we aim to unravel the layers of complexity surrounding these measures, casting light on their indispensable role in data interpretation.

Standard Deviation: The Beacon of Variability

Standard Deviation stands as a cornerstone in the realm of statistical analysis, quantifying the amount of dispersion or variability within a dataset. It measures the average distance between each data point and the mean, offering a gauge of the spread of data points.

Theoretical Foundation:

Mathematically, the standard deviation is the square root of the variance. It returns the dispersion measurement to the original units of the data, making it intuitively easier to understand. Its calculation involves determining the square root of the average squared deviations from the mean.

Application in Case Studies:

Consider the scenario of an investment firm analyzing the return on two different stocks. Stock A exhibits a lower standard deviation compared to Stock B, signaling that Stock A's returns are more consistent and less volatile than those of Stock B. This insight is invaluable for investors seeking stability in their investment choices.

Creative Insights:

- **Visualization:** Enhancing standard deviation analysis with visual aids, such as bell curves and confidence intervals, can provide a more intuitive understanding of data variability, allowing stakeholders to visually assess risk and consistency.
- **Comparative Analysis:** Employing standard deviation in comparative studies, such as comparing customer satisfaction scores across different service branches, can identify outliers and performance inconsistencies, guiding strategic improvements.

Variance: The Square of Differences

Variance, the squared measure of dispersion, serves as the foundation upon which the standard deviation is built. By squaring the differences from the mean, variance provides a measure of the data's spread that is free from the directionality of deviations.

Theoretical Foundation:

The calculation of variance involves averaging the squared differences between each data point and the mean. This process neutralizes negative deviations, ensuring that all differences contribute positively to the overall measure of spread.

Application in Case Studies:

In evaluating the performance consistency of manufacturing processes, a lower variance in production quality metrics indicates a more consistent and controlled process. High variance, conversely, signals a need for process examination and potential recalibration to minimize quality discrepancies.

Creative Insights:

- **Predictive Modeling:** Incorporating variance into predictive models can enhance forecasting accuracy by accounting for the inherent variability in historical data, leading to more robust and reliable predictions.
- **Risk Assessment:** Variance analysis in project management can highlight potential areas of risk and uncertainty, enabling preemptive measures to mitigate adverse impacts on project timelines and budgets.

Range: The Simplest Measure of Spread

The Range, defined as the difference between the maximum and minimum values in a dataset, offers the simplest form of variability measurement. Despite its straightforward nature, the range provides immediate insights into the breadth of data distribution.

Theoretical Foundation:

The range is calculated by subtracting the smallest value in the dataset from the largest value. This metric, while rudimentary, is instrumental in quickly assessing the overall spread of the data.

Application in Case Studies:

When analyzing the market prices of a commodity over a month, the range reveals the price volatility within that period. A narrow range indicates stable pricing, while a wide range suggests significant price fluctuations, crucial information for traders and investors.

Creative Insights:

- Time Series Analysis: Utilizing the range in time series data can help in identifying periods of high volatility, guiding strategic decisions in pricing, inventory management, and promotional activities.
- Benchmarking: Comparing the ranges of similar datasets, such as customer wait times across different service centers, can benchmark performance and identify best practices for operational efficiency.

CASE STUDY 1 - HR

Python

```
cgpa_column = data['CGPA']

# Calculate Standard Deviation, Variation, and Range
std_deviation = np.std(cgpa_column)
variation = np.var(cgpa_column)
data_range = np.ptp(cgpa_column)

print(f"Standard Deviation: {std_deviation}")
print(f"Variation: {variation}")
print(f"Range: {data_range}")
```

Standard Deviation: 0.6258860386151247

Variation: 0.3917333333333343

Range: 2.3000000000000007

R

```
> #-----standard deviation , variance, range
> sd(CGPA)
[1] 0.6301008
> var(CGPA)
[1] 0.397027
> range(CGPA)
[1] 6.5 8.8
```

Excel

Standard Deviation

Step 1: Click on an empty cell where you want the standard deviation to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =STDEV.S(

For a population: =STDEV.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the standard deviation.

Variance

Step 1: Click on an empty cell where you want the variance to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =VAR.S(

For a population: =VAR.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the variance.

Range

Step 1: To find the minimum, in an empty cell, type: =MIN(A1:A100)

Step 2: To find the maximum, in another cell, type: =MAX(A1:A100)

Step 3: To calculate the range, in a new cell, type: =MAX(A1:A100)-MIN(A1:A100)

Std Dev	8.35973418
Range	29
Variance	69.88515556

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select Variance, and click Continue.

Statistics

CGPA

N	Valid	75
	Missing	0
Std. Deviation	.6301	
Variance	.397	
Range	2.3	

Practical Output

- 75 valid CGPA entries, with no missing data.
- The standard deviation is 0.6301, indicating moderate variability in grades.
- The range of CGPAs is 2.3, suggesting a spread of academic performance.

CASE STUDY 2 - Marketing

Python

```
market_column = data["Hours Marketing"]

# Calculate Standard Deviation, Variation, and Range
std_development = np.std(market_column)
variation = np.var(market_column)
data_range = np.ptp(market_column)

print(f"Standard Deviation: {std_development}")
print(f"Variation: {variation}")
print(f"Range: {data_range}")
```

```
Standard Deviation: 2.5338946746505115
Variation: 6.420622222222222
Range: 9
```

R

```
> #-----standard deviation , variance, range
> sd(market)
[1] 2.523904
> var(market)
[1] 6.37009
> range(market)
[1] 1 10
```

Excel

Standard Deviation

Step 1: Click on an empty cell where you want the standard deviation to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =STDEV.S(

For a population: =STDEV.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the standard deviation.

Variance

Step 1: Click on an empty cell where you want the variance to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =VAR.S(

For a population: =VAR.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the variance.

Range

Step 1: To find the minimum, in an empty cell, type: =MIN(A1:A100)

Step 2: To find the maximum, in another cell, type: =MAX(A1:A100)

Step 3: To calculate the range, in a new cell, type: =MAX(A1:A100)-MIN(A1:A100)

RANGE	837
VARIANCE	48504.57707

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select Variance, and click Continue.

Statistics

HoursMarketing

N	Valid	75
	Missing	0
Std. Deviation		2.661
Variance		7.080
Range		9

Practical Output

- There's a slight positive skew (0.213) in the hours distribution, indicating some longer marketing sessions.
- The standard deviation of 2.661 suggests moderate variability in marketing hours.
- The range of marketing hours is 9, spanning from 1 to 10 hours.

CASE STUDY 3 - Operations

Python

```
cgpa_column = data['CGPA']
```

```

# Calculate Standard Deviation, Variation, and Range
std_deviation = np.std(cgpa_column)
variation = np.var(cgpa_column)
data_range = np.ptp(cgpa_column)

print(f"Standard Deviation: {std_deviation}")
print(f"Variation: {variation}")
print(f"Range: {data_range}")

```

Standard Deviation: 0.6258860386151247
 Variation: 0.3917333333333343
 Range: 2.3000000000000007

R

```

> #-----standard deviation , varience, range
> sd(data$CGPA)
[1] 0.6301008
> var(data$CGPA)
[1] 0.397027
> range(data$CGPA)
[1] 6.5 8.8

```

Excel

Standard Deviation

Step 1: Click on an empty cell where you want the standard deviation to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =STDEV.S(

For a population: =STDEV.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the standard deviation.

Variance

Step 1: Click on an empty cell where you want the variance to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =VAR.S(

For a population: =VAR.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the variance.

Range

Step 1: To find the minimum, in an empty cell, type: =MIN(A1:A100)

Step 2: To find the maximum, in another cell, type: =MAX(A1:A100)

Step 3: To calculate the range, in a new cell, type: =MAX(A1:A100)-MIN(A1:A100)

STANDARD DEVIATION 0.625886039

RANGE 2.3

VARIANCE 0.391733333

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select Variance, and click Continue.

Statistics

DiningOut

N	Valid	75
Missing		0
Std. Deviation		104.526
Variance		10925.676
Range		385

Practical Output

- The distribution is slightly positively skewed (0.041) and platykurtic (-1.069), indicating a flatter distribution than a normal curve.

The operations department could use this data to:

- Adjust meal plan pricing
- Plan for peak dining times
- Ensure adequate food supply and staffing
- Understand student spending patterns on dining\

CASE STUDY 4 - Finance

Python

```
Groceries_column = data['Groceries']

# Calculate Standard Deviation, Variation, and Range
std_deviation = np.std(Groceries_column)
variation = np.var(Groceries_column)
data_range = np.ptp(Groceries_column)

print(f"Standard Deviation: {std_deviation}")
print(f"Variation: {variation}")
print(f"Range: {data_range}")
```

Standard Deviation: 206.7284423800675

Variation: 42736.648888888885

Range: 710

R

```
> sd(CGPA)
[1] 0.6301008
> var(CGPA)
[1] 0.397027
> range(CGPA)
[1] 6.5 8.8
```

Excel

Standard Deviation

Step 1: Click on an empty cell where you want the standard deviation to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =STDEV.S(

For a population: =STDEV.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the standard deviation.

Variance

Step 1: Click on an empty cell where you want the variance to appear.

Step 2: Type one of the following formulas, depending on whether your data represents a sample or the entire population:

For a sample: =VAR.S(

For a population: =VAR.P(

Step 3: Select the range of cells containing your data.

Step 4: Close the parenthesis to complete the formula.

Step 5: Press Enter to calculate the variance.

Range

Step 1: To find the minimum, in an empty cell, type: =MIN(A1:A100)

Step 2: To find the maximum, in another cell, type: =MAX(A1:A100)

Step 3: To calculate the range, in a new cell, type: =MAX(A1:A100)-MIN(A1:A100)

STANDARD DEVIATION 26.90344877

RANGE 95

VARIANCE 723.7955556

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Descriptives.
2. Select the variables you want to analyze.
3. Click Options, select Variance, and click Continue.

Statistics

	Groceries	Utilities
N	Valid	75
	Missing	0
Std. Deviation	208.121	27.085
Variance	43314.171	733.577
Range	710	95

Practical Output

Groceries:

- Standard Deviation: 208.121
- Variance: 43,314.171
- Range: 710

Utilities:

- Standard Deviation: 27.085
- Variance: 733.577

• Range: 95

Interpretation:

1. Variability in Expenses: The standard deviation and variance for groceries are much higher than for utilities. This indicates that grocery expenses are far more volatile and unpredictable.

2. Grocery Expenses:

• The high standard deviation (208.121) suggests that grocery costs frequently deviate significantly from the mean.

• The large variance (43,314.171) indicates extreme variability in grocery expenses.

• The wide range (710) shows that there's a substantial difference between the lowest and highest grocery bills.

Financial Implications: The finance department needs to maintain a sizeable contingency fund to handle unexpected spikes in grocery costs. Budgeting for groceries should account for this high variability, possibly by using a higher estimate than the mean to ensure sufficient funds are available. Utility costs are more stable, allowing for more accurate budgeting. The finance department can allocate funds for utilities with greater confidence, as these expenses are less likely to deviate significantly from the mean.

PO

Budget for average monthly expenses of \$640 for groceries and \$300 for utilities per student.

Maintain a flexible grocery budget due to high variability.

CH 5- Coefficient of Variation

Coefficient of Variation

In the rich tapestry of statistical analysis, the Coefficient of Variation (CV) emerges as a nuanced measure that transcends mere averages to offer profound insights into the relative variability of datasets. This segment of the Harvard case study statistics book is dedicated to unraveling the intricacies of the Coefficient of Variation, illustrating its significance, applications, and the creative analytical frameworks it supports.

Coefficient of Variation: A Symphony of Relativity

At its essence, the Coefficient of Variation represents the ratio of the standard deviation to the mean, expressed as a percentage. This elegant metric captures the extent of variability in relation to the size of the mean, providing a standardized measure of dispersion that facilitates comparisons across datasets of different scales and units.

Theoretical Foundation:

The CV is calculated by dividing the standard deviation by the mean and then multiplying the result by 100 to express it as a percentage. This calculation imbues the

CV with the ability to offer a dimensionless quantity, making it a universal measure of relative variability.

Application in Case Studies:

Imagine a venture capitalist comparing two startups for potential investment. Startup A, operating in a mature market, shows a lower CV in its monthly returns compared to Startup B in an emerging market. The CV elucidates that, despite potentially higher returns from Startup B, its risk-adjusted performance is less favorable compared to Startup A, guiding a more informed investment decision.

Creative Insights:

- Scale-Invariant Analysis: The CV's unique ability to facilitate comparisons across data with vastly different scales enables analysts to juxtapose variability in contexts as diverse as financial returns and biochemical assays, unlocking cross-disciplinary insights.
- Benchmarking Reliability: By benchmarking the CV across different operational processes or time periods, organizations can identify areas of inconsistency or volatility, directing focus towards optimization and stability enhancement efforts.

Enhancing Narrative with Coefficient of Variation:

The true power of the Coefficient of Variation lies in its capacity to enrich statistical narratives, transforming raw data into compelling stories of risk, reliability, and efficiency.

Multifaceted Comparisons:

The CV shines in its role in multifaceted comparisons, where its application extends beyond financial data to realms like manufacturing, healthcare, and research. For instance, in healthcare, comparing the CV of patient response times across different hospitals can highlight disparities in service efficiency, guiding policy and operational adjustments.

Creative Visualization:

Visualizing CV data through comparative histograms or scatter plots can dramatically underscore the relative variability across datasets. Such visualizations not only capture the audience's attention but also facilitate intuitive understanding of complex statistical concepts, enhancing the communicative power of data.

CASE STUDY 1 - HR

Python

```
# Coefficient of Variation
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
cv_values = {column: cv(data[column]) for column in ['CGPA',
'AptitudeTestScore']}
print(cv_values)
```

```
{'CGPA': 8.140837300202707, 'AptitudeTestScore': 10.764021131539925}
```

R

```
> # -----Coefficient of Variation
> cv <- function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE) * 100
> cv(data$CGPA)
[1] 8.140837
```

Excel

Coefficient of Variation

Step 1: Calculate the mean. In an empty cell, type: =AVERAGE(A1:A100)
Step 2: Calculate the standard deviation. In another cell, type: =STDEV.S(A1:A100)
Step 3: Calculate the coefficient of variation. In a new cell, type:
=(STDEV.S(A1:A100)/AVERAGE(A1:A100))*100

Coeff of Var	10.76402113
--------------	-------------

SPSS

Steps:

1. Calculate the standard deviation and mean as described above.
2. Use a calculator to divide the standard deviation by the mean and multiply by 100.

Coefficient Of Variance

cv=8.14

Practical Output

The Coefficient of Variation is a measure of relative variability that allows us to compare the dispersion of data sets with different units or vastly different means.

CGPA (Cumulative Grade Point Average):

CV = 8.14%

This means that the standard deviation of CGPA scores is about 8.14% of the mean CGPA.

Practical interpretation:

There's relatively low variability in CGPA scores.

Most students' CGPAs are clustered fairly close to the average.

You could expect about 68% of students to have a CGPA within $\pm 8.14\%$ of the mean CGPA (assuming a normal distribution).

AptitudeTestScore:

CV = 10.76%

This indicates that the standard deviation of AptitudeTestScores is about 10.76% of the mean score.

Practical interpretation:

There's moderately low variability in AptitudeTestScores, but more than in CGPA.

AptitudeTestScores are more spread out than CGPA scores, but still relatively clustered.

You could expect about 68% of students to have an AptitudeTestScore within $\pm 10.76\%$ of the mean score (assuming a normal distribution).

CASE STUDY 2 - Marketing

Python

```
# Coefficient of Variation
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
cv_values = {column: cv(data[column]) for column in ['Hours Marketing', 'Incentive Received']}
print(cv_values)
```

{'Hours Marketing': 47.47440690910778, 'Incentive Received': 54.59134453806974}

R

```
> # -----Coefficient of Variation  
> cv <- function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE) * 100  
> cv(market)  
[1] 48.66138
```

Excel

Coefficient of Variation

Step 1: Calculate the mean. In an empty cell, type: =AVERAGE(A1:A100)

Step 2: Calculate the standard deviation. In another cell, type: =STDEV.S(A1:A100)

Step 3: Calculate the coefficient of variation. In a new cell, type:

= (STDEV.S(A1:A100)/AVERAGE(A1:A100))*100

COEFFICIENT OF VARIANCE 53.50918158

SPSS

Steps:

1. Calculate the standard deviation and mean as described above.
2. Use a calculator to divide the standard deviation by the mean and multiply by 100.

Coefficient Of Variance

50.11299435028249

Practical Output

CV values:

Hours Marketing: 47.47%

Incentive Received: 54.59%

Practical interpretation and application:

Both variables show high variability, with Incentive Received being slightly more variable.

There's a wide range in both the hours spent marketing and the incentives received.

The high variability in Hours Marketing suggests that employees have very different levels of involvement in marketing activities.

The even higher variability in Incentive Received indicates that rewards are not uniformly distributed.

Application:

For a marketing manager or HR professional, this data suggests a need to investigate why there's such high variability.

It might be worth examining if there's a correlation between hours spent marketing and incentives received.

Consider implementing a more standardized approach to marketing tasks and incentive distribution to reduce variability.

Look into whether top performers are being adequately rewarded, or if the incentive structure needs adjustment.

Provide additional training or support for employees spending fewer hours on marketing to boost overall performance.

CASE STUDY 3 - Operations

Python

```
# Coefficient of Variation
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
cv_values = {column: cv(data[column]) for column in ['CGPA', 'Room Type']}
print(cv_values)
```

```
{'CGPA': 8.140837300202707, 'Room Type': 32.21360692547209}
```

R

```
> # -----Coefficient of Variation
> cv <- function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE) * 100
> cv(data$CGPA)
[1] 8.140837
> cv(data$Room.Type)
[1] 32.89585
```

Excel

Coefficient of Variation

Step 1: Calculate the mean. In an empty cell, type: =AVERAGE(A1:A100)
Step 2: Calculate the standard deviation. In another cell, type: =STDEV.S(A1:A100)
Step 3: Calculate the coefficient of variation. In a new cell, type:
=(STDEV.S(A1:A100)/AVERAGE(A1:A100))*100

COEFFICIENT OF VARIANCE 8.1408373

SPSS

Steps:

1. Calculate the standard deviation and mean as described above.
2. Use a calculator to divide the standard deviation by the mean and multiply by 100.

Coefficient of Variance

cv=35.43

Practical Output

CV values:

CGPA: 8.14%

Room Type: 32.21%

Practical interpretation and application:

CGPA shows low variability, indicating consistent academic performance across students.
Room Type has much higher variability, suggesting a diverse range of living situations.

Application:

For a university housing department or educational researcher:

The low variability in CGPA suggests that academic performance is relatively consistent regardless of room type.

However, the high variability in Room Type warrants investigation into whether certain types of accommodation are associated with better academic performance.

Consider surveying students to understand their preferences and needs in terms of accommodation.

Examine if there's any correlation between Room Type and CGPA, which could inform housing policies.

Use this information to ensure equitable access to different room types, especially if any are found to be advantageous for academic performance.

CASE STUDY 4 - Finance

Python

```
# Coefficient of Variation
cv = lambda x: np.std(x, ddof=1) / np.mean(x) * 100
cv_values = {column: cv(data[column]) for column in ['Groceries',
'Utilities']}
print(cv_values)

{'Groceries': 32.52561504988311, 'Utilities': 9.05638168271678}
```

R

Excel

Coefficient of Variation

Step 1: Calculate the mean. In an empty cell, type: =AVERAGE(A1:A100)

Step 2: Calculate the standard deviation. In another cell, type: =STDEV.S(A1:A100)

Step 3: Calculate the coefficient of variation. In a new cell, type:

= (STDEV.S(A1:A100)/AVERAGE(A1:A100))*100

COEFFICIENT OF VARIANCE 9.056381683

SPSS

Steps:

1. Calculate the standard deviation and mean as described above.
2. Use a calculator to divide the standard deviation by the mean and multiply by 100.

Coefficient of Variance cv=32.53

Practical Output

CV values:

Groceries: 32.53%

Utilities: 9.06%

Practical interpretation and application:

Grocery expenses show high variability, indicating significant differences in spending patterns among individuals or households.

Utilities expenses have low variability, suggesting more consistent costs across the board.

Application:

For a financial advisor or consumer research analyst:

The high variability in grocery spending indicates an opportunity for budgeting advice and education.

Some individuals/households might benefit from tips on how to reduce grocery expenses.

The low variability in utilities suggests that these costs are more predictable and standardized.

Focus financial planning efforts more on managing variable expenses like groceries rather than utilities.

For market researchers, the high variability in grocery spending might indicate diverse consumer segments with different purchasing habits.

For utility companies, the low variability suggests that their pricing and consumption patterns are relatively stable across customers.

CH 6 - Quartile Deviation

Quartile Deviation

In the diverse landscape of statistical measures, Quartile Deviation (QD) stands out as a robust and insightful metric for understanding the spread of data, particularly in relation to its median. This section of our Harvard case study statistics book delves into the essence of Quartile Deviation, exploring its theoretical underpinnings, practical applications, and the creative potential it holds for data analysis and interpretation.

Quartile Deviation: The Measure of Middle Spread

Quartile Deviation, also known as the Semi-Interquartile Range, is a measure that captures the spread of the middle 50% of a dataset. It is calculated as half the difference between the third quartile (Q3) and the first quartile (Q1), effectively quantifying the variability around the median without being influenced by outliers.

Theoretical Foundation

The elegance of QD lies in its simplicity and focus. By concentrating on the central portion of the data, QD offers a distilled view of variability that is often more relevant to understanding the typical spread than measures influenced by extreme values. Its calculation,

$$QD = (Q3 - Q1)/2$$

, provides a straightforward yet powerful tool for assessing dispersion.

Application in Case Studies

Imagine a real estate firm analyzing the distribution of home prices within a neighborhood. The Quartile Deviation reveals the spread of the middle market homes, ignoring the extremes of luxury and budget properties. This insight is invaluable for targeting marketing strategies and understanding the core housing market dynamics.

Creative Insights

- Visual Interpretation: Enhancing data reports with box plots that visually represent QD alongside medians and quartiles can provide stakeholders with an intuitive grasp of data spread and central tendency.
- Comparative Analysis: Employing QD to compare the spread of similar datasets across different demographics or time periods can uncover underlying trends and shifts in variability, informing strategic decisions and policy formulations.

Enhancing Narrative with Quartile Deviation

The Quartile Deviation's strength in focusing on the middle spread of data lends itself to crafting narratives that emphasize the typical or expected conditions, rather than being skewed by exceptional cases.

Storytelling with Data

In the context of customer satisfaction surveys across different service sectors, QD can highlight where the bulk of customer perceptions lie, steering improvements in areas impacting the majority of experiences. This approach shifts the narrative from outlier-driven reactions to focusing on enhancing the core service attributes.

Creative Visualization

Developing interactive visual tools that allow users to adjust the range of focus—thereby dynamically altering the QD and related statistics—can engage stakeholders in exploratory data analysis, fostering a deeper understanding of data variability and its drivers.

Quartile Deviation in Decision-Making

The application of QD extends beyond descriptive analytics into the realm of decision-making, where its insights can guide strategic choices by highlighting the stability or variability of core operational metrics.

Risk Management

In financial portfolios, QD can serve as a measure of the middle-range risk associated with investment returns, offering a complementary perspective to standard deviation-based risk assessments. This middle-focused view helps in constructing portfolios that aim for consistency and reliability in returns.

Operational Efficiency

For businesses analyzing process consistency, such as manufacturing or service delivery times, QD provides a measure of the typical variability experienced. Identifying processes with low QD can pinpoint areas of operational excellence, while high QD may signal opportunities for standardization and improvement.

CASE STUDY 1 - HR

Python

```
Q1 = data['CGPA'].quantile(0.25)
Q3 = data['CGPA'].quantile(0.75)

# Calculate Quartile Deviation
```

```

quartile_deviation = (Q3 - Q1) / 2
print("Quartile 1 :", Q1)
print("Quartile 3 :", Q3)
print("Quartile Deviation :", quartile_deviation)

```

```

Quartile 1 : 7.5
Quartile 3 : 8.2
Quartile Deviation : 0.3499999999999964

```

R

```

> # Calculate the first quartile (Q1)
> Q1 <- quantile(CGPA, 0.25)
>
> # Calculate the third quartile (Q3)
> Q3 <- quantile(CGPA, 0.75)
>
> # Calculate the quartile deviation
> quartile_deviation <- (Q3 - Q1) / 2
>
> # Print the quartile deviation
> print(quartile_deviation)
75%
0.35

```

Excel

Quartile Deviation

Step 1: Calculate Q1. In an empty cell, type: =QUARTILE.INC(A1:A100,1)
 Step 2: Calculate Q3. In another cell, type: =QUARTILE.INC(A1:A100,3)
 Step 3: Calculate the quartile deviation. In a new cell, type:
 $=\text{QUARTILE.INC}(A1:A100,3)-\text{QUARTILE.INC}(A1:A100,1)/2$

Quartile Deviation 6.25

Practical Output

Q1: 7.5
 Q3: 8.2
 Quartile Deviation: 0.35
 Practical Application:

The small QD indicates that the middle 50% of students have relatively similar CGPAs. Half of the students have CGPAs between 7.5 and 8.2, a narrow range of 0.7 points. This suggests consistency in grading or student performance across the institution.

Application for Educators/Administrators:

Use this as a benchmark for identifying high performers (above 8.2) and those who might need additional support (below 7.5).

Consider adjusting grading policies if the range is too narrow, potentially not differentiating student performance enough.

For admissions or scholarship committees, this tight range means they may need to look beyond CGPA to distinguish between candidates.

CASE STUDY 2 - Marketing

Python

```
Q1 = data['Hours Marketing'].quantile(0.25)
Q3 = data['Hours Marketing'].quantile(0.75)

# Calculate Quartile Deviation
quartile_deviation = (Q3 - Q1) / 2
print("Quartile 1 :",Q1)
print("Quartile 3 :",Q3)
print("Quartile Deviation :",quartile_deviation)
```

```
Quartile 1 : 3.0
Quartile 3 : 7.0
Quartile Deviation : 2.0
```

R

```
> # Calculate the first quartile (Q1)
> Q1 <- quantile(CGPA, 0.25)
>
> # Calculate the third quartile (Q3)
> Q3 <- quantile(CGPA, 0.75)
>
> # Calculate the quartile deviation
> quartile_deviation <- (Q3 - Q1) / 2
>
```

```
> # Print the quartile deviation  
> print(quartile_deviation)  
75%  
0.35
```

Excel

Quartile Deviation

Step 1: Calculate Q1. In an empty cell, type: =QUARTILE.INC(A1:A100,1)

Step 2: Calculate Q3. In another cell, type: =QUARTILE.INC(A1:A100,3)

Step 3: Calculate the quartile deviation. In a new cell, type:

=QUARTILE.INC(A1:A100,3)-QUARTILE.INC(A1:A100,1))/2

QUARTILE DEVIATION

188

Practical Output

Q1: 3.0

Q3: 7.0

Quartile Deviation: 2.0

Practical Application:

There's significant variation in the hours spent on marketing among the middle 50% of the group.

Half of the individuals spend between 3 and 7 hours on marketing, a wide range.

Application for Marketing Managers:

Investigate why there's such a large disparity in marketing hours.

Consider standardizing marketing practices or providing clearer guidelines.

Look into the effectiveness of those spending more hours versus those spending fewer.

Use this information to set realistic expectations and goals for marketing activities.

Potentially restructure teams or redistribute workload to balance marketing efforts.

CASE STUDY 3 - Operations

Python

```
Q1 = data['CGPA'].quantile(0.25)  
Q3 = data['CGPA'].quantile(0.75)
```

```

# Calculate Quartile Deviation
quartile_deviation = (Q3 - Q1) / 2
print("Quartile 1 :",Q1)
print("Quartile 3 :",Q3)
print("Quartile Deviation :",quartile_deviation)

```

Quartile 1 : 7.5
 Quartile 3 : 8.2
 Quartile Deviation : 0.3499999999999964

R

```

> # Calculate the first quartile (Q1)
> Q1 <- quantile(CGPA, 0.25)
>
> # Calculate the third quartile (Q3)
> Q3 <- quantile(CGPA, 0.75)
>
> # Calculate the quartile deviation
> quartile_deviation <- (Q3 - Q1) / 2
>
> # Print the quartile deviation
> print(quartile_deviation)
75%
0.35

```

Excel

Quartile Deviation

Step 1: Calculate Q1. In an empty cell, type: =QUARTILE.INC(A1:A100,1)
 Step 2: Calculate Q3. In another cell, type: =QUARTILE.INC(A1:A100,3)
 Step 3: Calculate the quartile deviation. In a new cell, type:
 $=\text{QUARTILE.INC}(A1:A100,3)-\text{QUARTILE.INC}(A1:A100,1)/2$

QUARTILE DEVIATION	0.35
--------------------	------

Practical Output

Q1: 7.5

Q3: 8.2

Quartile Deviation: 0.35

Practical Application:

The small QD indicates that the middle 50% of students have relatively similar CGPAs.

Half of the students have CGPAs between 7.5 and 8.2, a narrow range of 0.7 points.

This suggests consistency in grading or student performance across the institution.

Application for Educators/Administrators:

Use this as a benchmark for identifying high performers (above 8.2) and those who might need additional support (below 7.5).

Consider adjusting grading policies if the range is too narrow, potentially not differentiating student performance enough.

For admissions or scholarship committees, this tight range means they may need to look beyond CGPA to distinguish between candidates.

CASE STUDY 4 - Finance

Python

```
Q1 = data['Groceries'].quantile(0.25)
Q3 = data['Groceries'].quantile(0.75)

# Calculate Quartile Deviation
quartile_deviation = (Q3 - Q1) / 2
print("Quartile 1 :", Q1)
print("Quartile 3 :", Q3)
print("Quartile Deviation :", quartile_deviation)
```

```
Quartile 1 : 490.0
Quartile 3 : 830.0
Quartile Deviation : 170.0
```

R

```
> # Calculate the first quartile (Q1)
> Q1 <- quantile(groceries, 0.25)
>
> # Calculate the third quartile (Q3)
> Q3 <- quantile(groceries, 0.75)
>
```

```
> # Calculate the quartile deviation  
> quartile_deviation <- (Q3 - Q1) / 2  
>  
> # Print the quartile deviation  
> print(quartile_deviation)  
75%  
170
```

Excel

Quartile Deviation

Step 1: Calculate Q1. In an empty cell, type: =QUARTILE.INC(A1:A100,1)

Step 2: Calculate Q3. In another cell, type: =QUARTILE.INC(A1:A100,3)

Step 3: Calculate the quartile deviation. In a new cell, type:

=QUARTILE.INC(A1:A100,3)-QUARTILE.INC(A1:A100,1)/2

QUARTILE DEVIATION	22.5
--------------------	------

Practical Output

Q1: 490.0

Q3: 830.0

Quartile Deviation: 170.0

Practical Application:

There's considerable variation in grocery spending among the middle 50% of the group.
Half of the individuals/households spend between \$490 and \$830 on groceries, a range of \$340.

Application for Financial Advisors or Consumer Researchers:

Use this information to create tiered budgeting advice for different spending levels.

Investigate factors contributing to the wide range (family size, dietary preferences, shopping locations).

For grocery retailers, this data could inform pricing strategies and product offerings to cater to different spending segments.

Policymakers could use this to assess food affordability and potentially adjust food assistance programs.

Financial education programs could focus on teaching budgeting skills, especially for those at the higher end of the spending range.

CH 7- Box Plot, Scatter plot

Box Plot, Scatter Plot

In the realm of data visualization, Box Plots and Scatter Plots stand as pivotal tools, each with unique capabilities to elucidate complex datasets. This section of our Harvard case study statistics book ventures into the intricacies and applications of these plots, weaving a narrative that not only educates but also inspires innovative uses of these visual tools in data analysis.

Box Plot: A Quintessence of Descriptive Statistics

The Box Plot, or Box-and-Whisker Plot, is a graphical representation that summarizes a dataset's key characteristics with simplicity and efficiency. By displaying the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values, a Box Plot provides a clear snapshot of the distribution, central tendency, and variability of the data.

Theoretical Underpinning

Originating from John Tukey's exploratory data analysis, the Box Plot encapsulates data distribution in a compact form, making it an invaluable tool for preliminary data investigation. Its construction highlights the dataset's skewness and the presence of outliers, offering initial clues about the underlying statistical distribution.

Application in Case Studies

Consider a scenario where a healthcare provider analyzes patient wait times across different departments. A series of Box Plots could reveal not only the typical wait times (via the medians) but also departmental disparities in variability and outlier experiences. Such insights are crucial for targeted improvements and resource allocation.

Creative Insights

- Enhanced Visualization: Incorporating color coding or varying box widths proportional to sample size can add layers of meaning to the Box Plot, enriching the narrative conveyed.
- Dynamic Exploration: Interactive digital Box Plots, where users can hover to see additional details or filter to compare specific subsets, encourage active engagement with the data, fostering a deeper understanding of the nuances involved.

Scatter Plot: The Canvas of Correlation

The Scatter Plot serves as a foundational tool for visualizing the relationship between two quantitative variables. By plotting individual data points on a two-dimensional graph, it allows analysts to discern patterns, trends, and potential correlations within the data.

Theoretical Underpinning

The Scatter Plot's strength lies in its ability to reveal the nature of the relationship—linear or nonlinear—between variables, along with the direction (positive or negative) and strength of the correlation. It forms the basis for further statistical analysis, including regression modeling.

Application in Case Studies

In a study examining the impact of marketing spend on sales revenue, a Scatter Plot could visually demonstrate the relationship's strength and direction. Identifying a clear positive trend would not only validate marketing strategies but also guide future investment decisions.

Creative Insights

- Trend Lines and Curves: Adding trend lines or curves to Scatter Plots can help in highlighting the underlying relationship more clearly, guiding the viewer towards the analysis's conclusions.
- Marker Attributes: Varying marker size or color based on a third variable (e.g., customer segment or time) can transform a simple Scatter Plot into a rich, multidimensional analysis tool, unveiling complex patterns hidden within the data.

Box Plot and Scatter Plot in Concert

While Box Plots excel in summarizing data distribution and identifying outliers, Scatter Plots shine in uncovering relationships between variables. Used in concert, they offer a comprehensive view of a dataset's landscape—combining univariate insights with bivariate relationships.

Synergistic Application

In evaluating a new product's market performance, Box Plots could first be used to assess the sales distribution across different markets, identifying outliers and general variability. Following this, Scatter Plots could explore the relationship between sales and factors like marketing intensity or price points, providing a holistic view of performance drivers.

Creative Synthesis

- Integrated Dashboards: Creating dashboards that juxtapose Box Plots with Scatter Plots, along with interactive elements for real-time data exploration, can facilitate an integrated analytical workflow, enabling stakeholders to derive nuanced insights efficiently.
- Storytelling with Data: Crafting narratives that transition from the descriptive statistics of Box Plots to the relational insights of Scatter Plots can guide audiences through a logical and insightful exploration of the data, fostering informed decision-making.

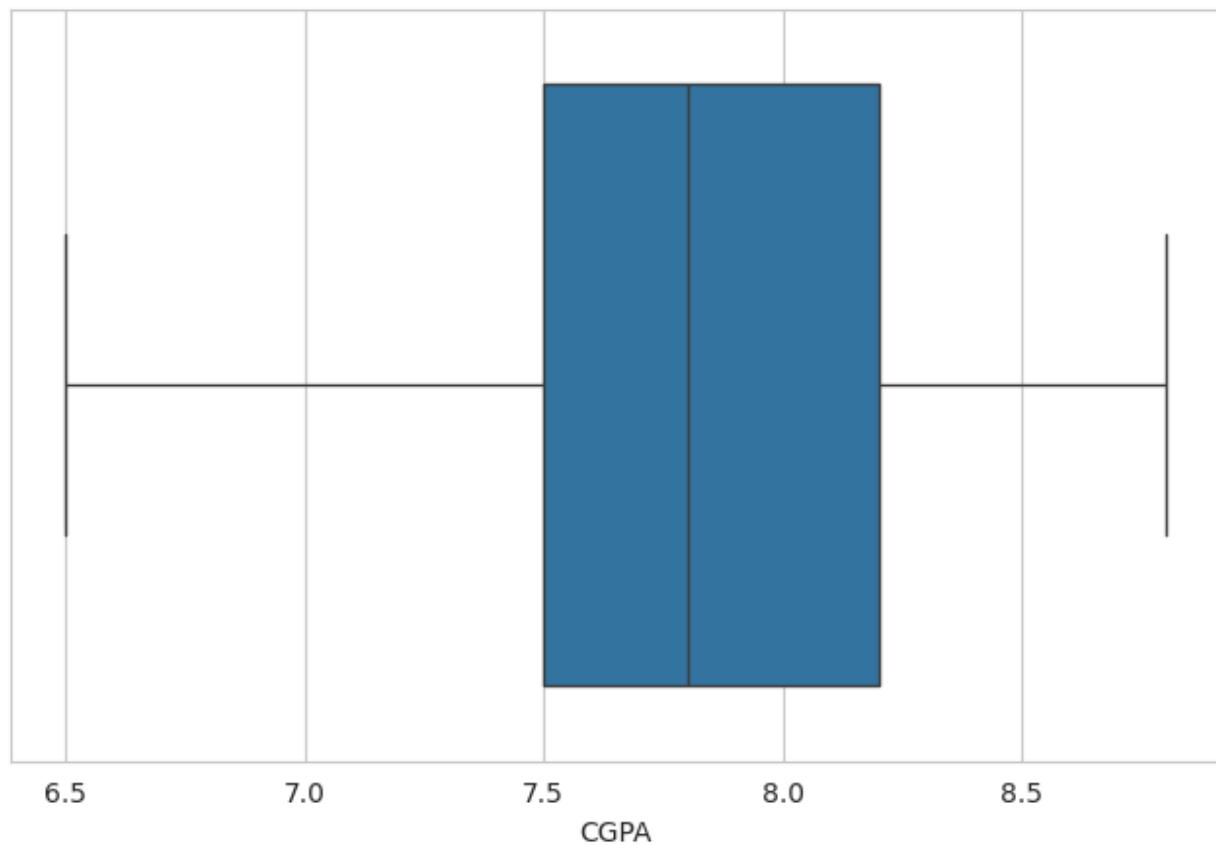
CASE STUDY 1 - HR

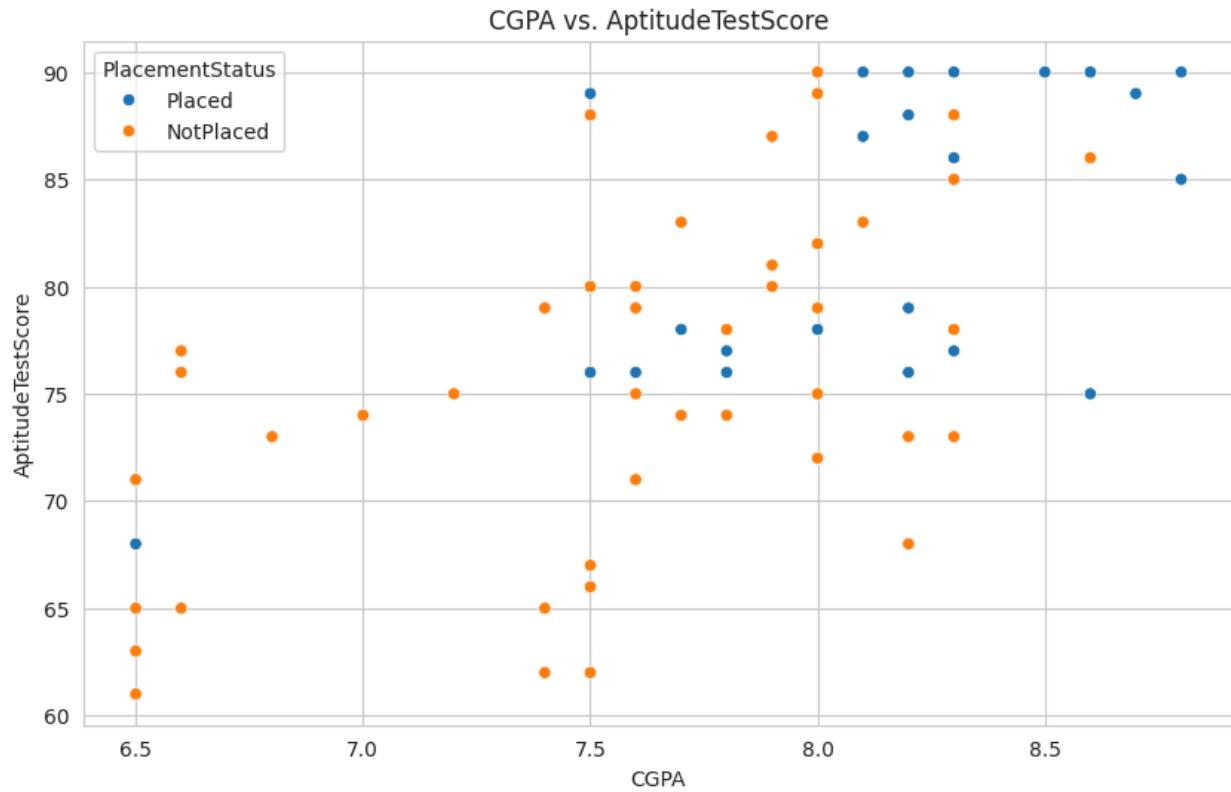
Python

```
# Box Plot - CGPA
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='CGPA')
plt.title('Box Plot of CGPA')
plt.xlabel('CGPA')
plt.show()

# Scatter Plot - CGPA vs. AptitudeTestScore
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='CGPA', y='AptitudeTestScore',
hue='PlacementStatus')
plt.title('CGPA vs. AptitudeTestScore')
plt.xlabel('CGPA')
plt.ylabel('AptitudeTestScore')
plt.show()
```

Box Plot of CGPA

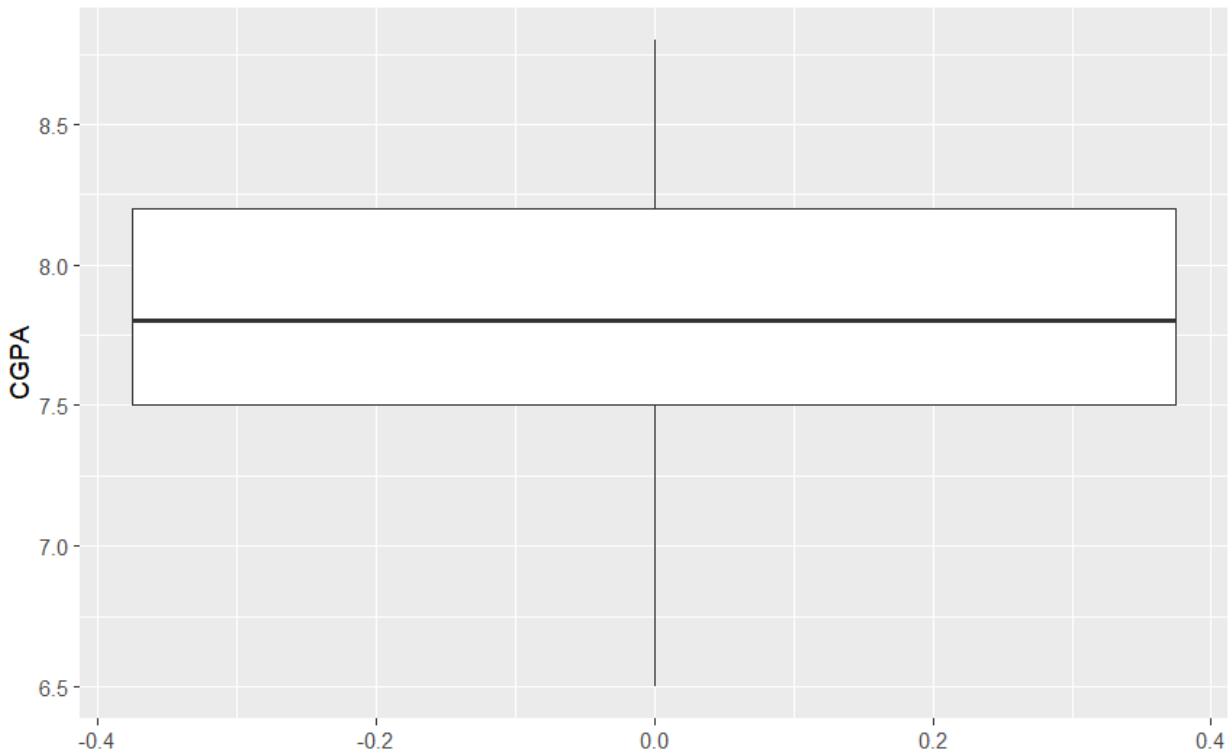




R

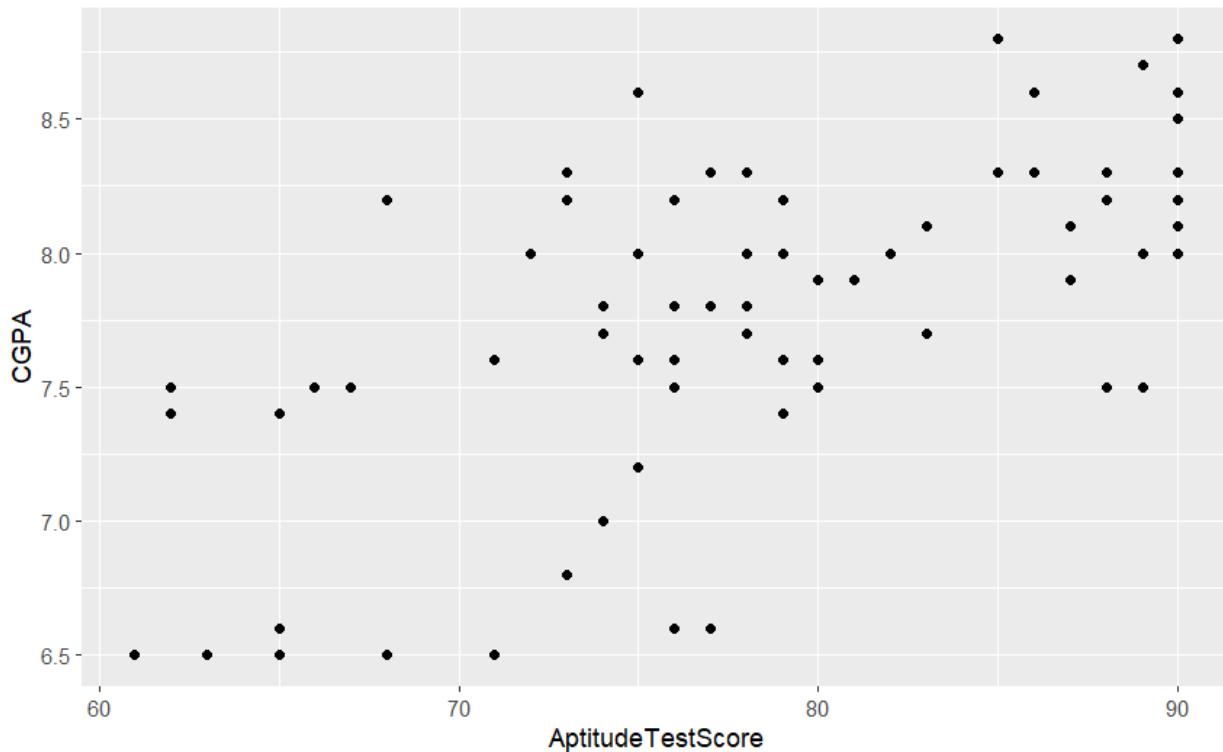
```
> # -----Box and Scatter Plot for CGPA
> ggplot(data, aes(y=CGPA)) + geom_boxplot() + labs(title="Box Plot of CGPA",
y="CGPA")
```

Box Plot of CGPA



```
> ggplot(data, aes(x=AptitudeTestScore, y=CGPA)) + geom_point() +
  labs(title="Scatter Plot of CGPA", y="CGPA")
```

Scatter Plot of CGPA



Excel

Box Plot

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Insert Statistic Chart" button in the Charts group.

Step 4: Select "Box and Whisker" from the dropdown menu.

Scatter Plot

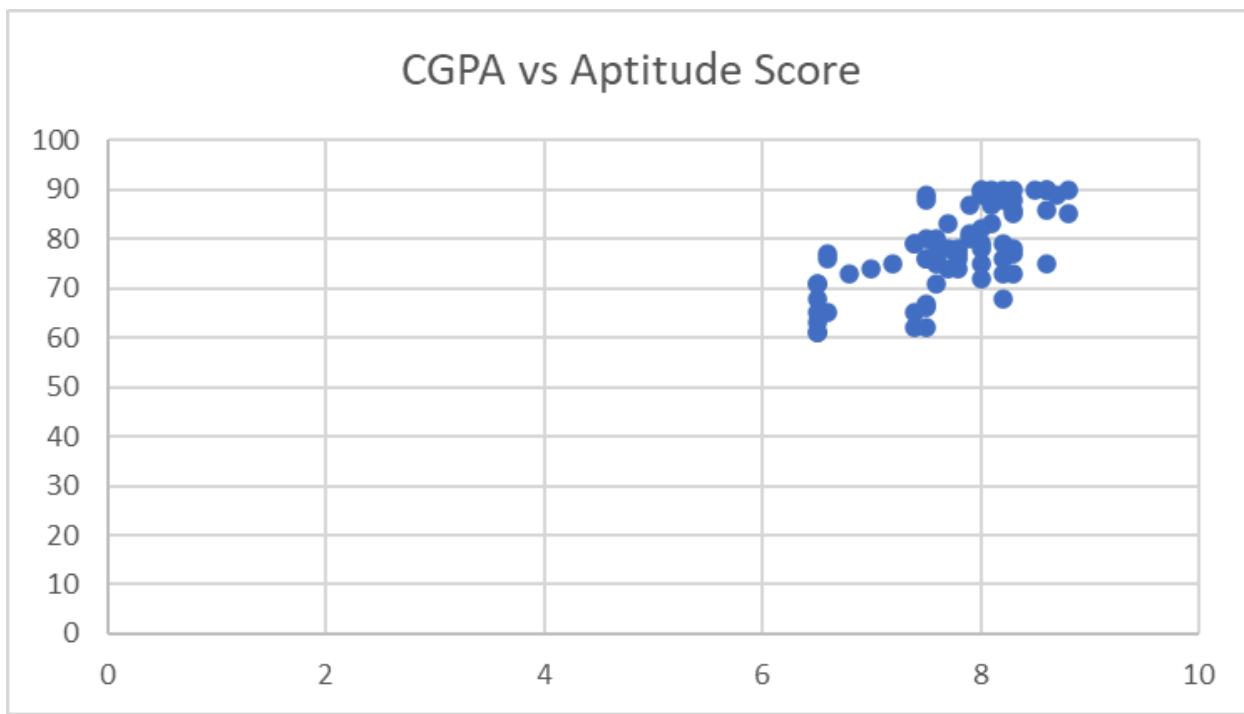
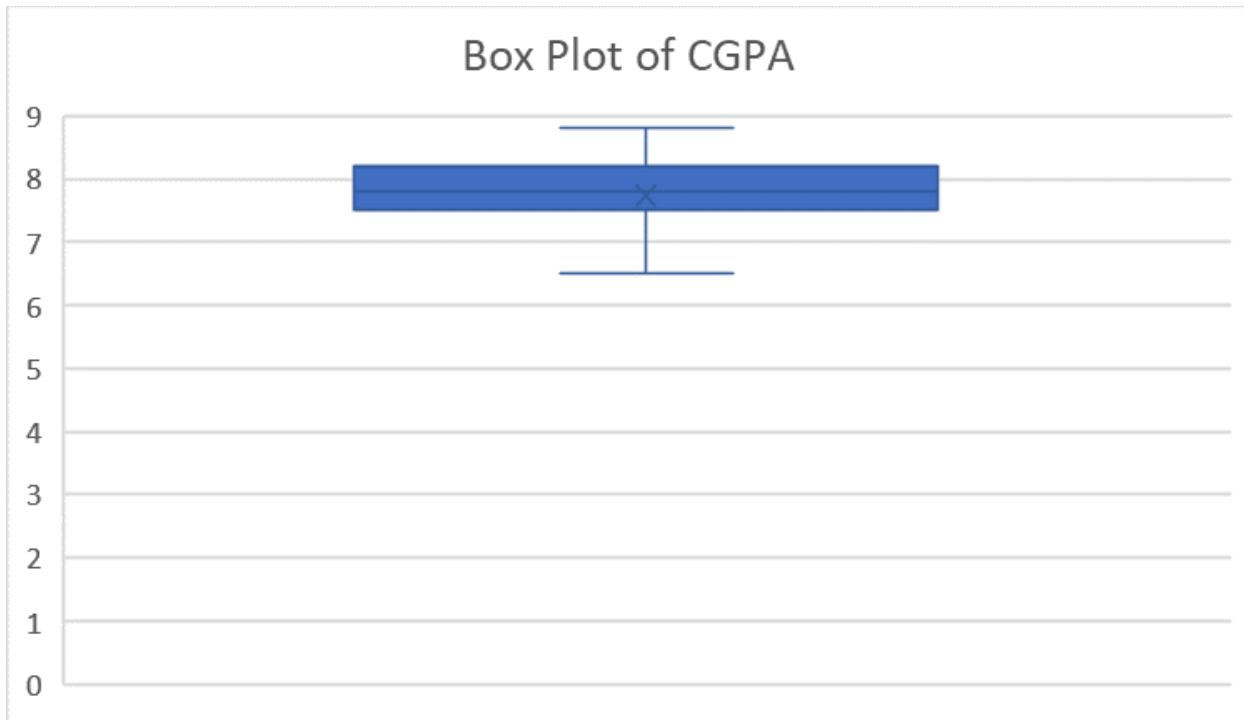
Step 1: Enter your x-axis data in one column and y-axis data in an adjacent column.

Step 2: Select both columns of data.

Step 3: Go to the Insert tab on the ribbon.

Step 4: Click on the "Scatter" button in the Charts group.

Step 5: Choose the desired scatter plot type from the dropdown menu.

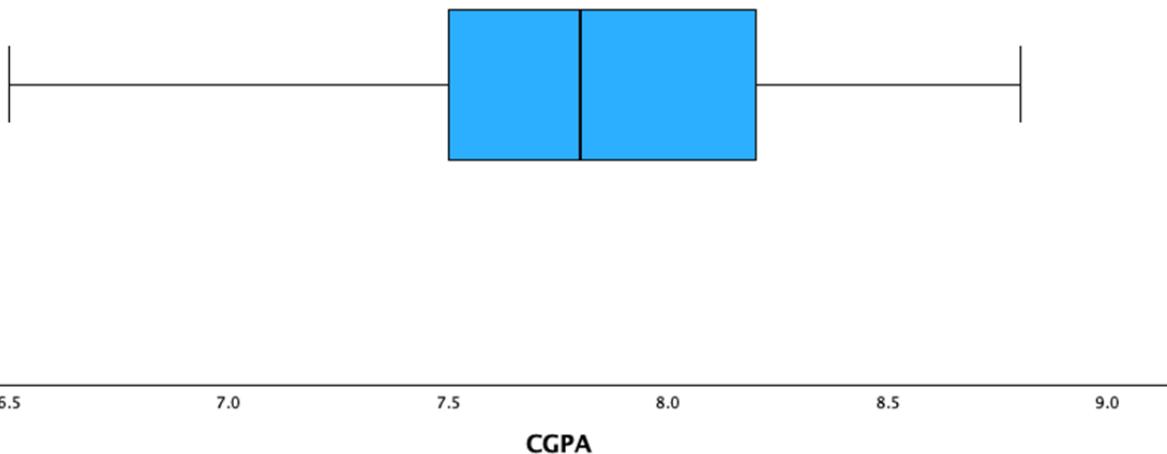


SPSS

Steps:

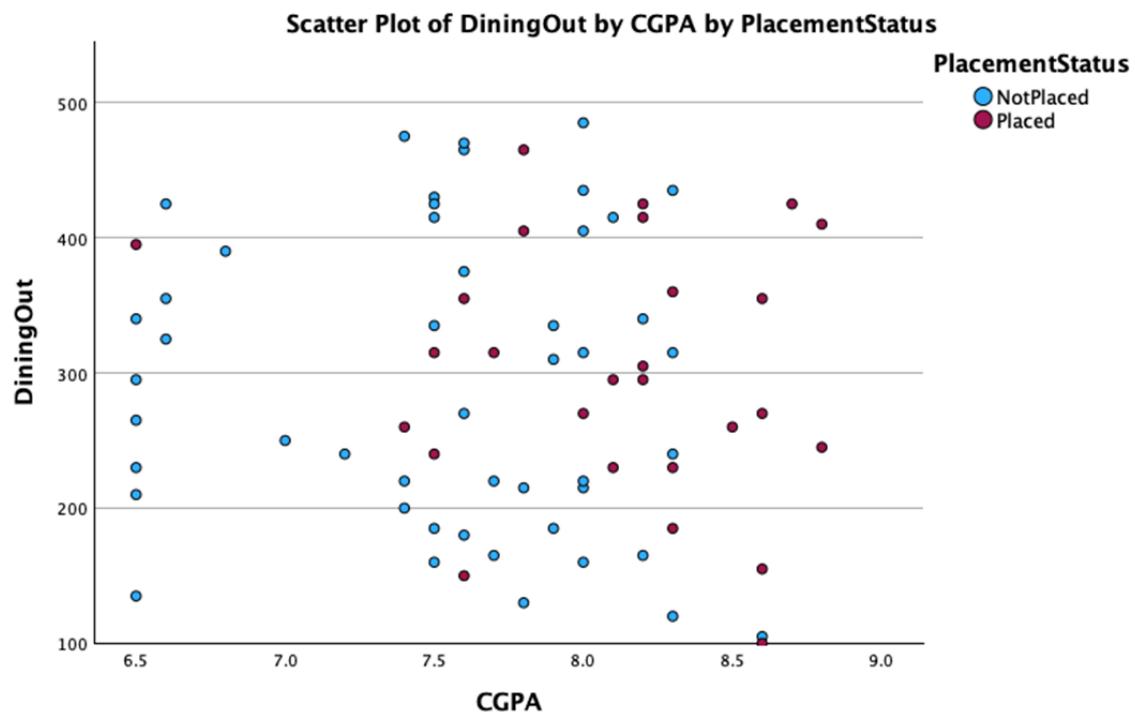
1. Go to Graphs > Chart Builder.
2. Select the Boxplot option.
3. Drag the variable to the y-axis and a categorical variable to the x-axis (if needed).

Simple Boxplot of CGPA



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Scatter/Dot option.
3. Drag the independent variable to the x-axis and the dependent variable to the y-axis.



Practical Output

The scatter plot shows a slight negative correlation between CGPA and dining out frequency. Students with higher CGPAs tend to dine out less often, though there's significant variation. The box plot indicates that most CGPAs fall between 7.5 and 8.0, with some outliers on both ends.

Practical applications:

Academic advising: Counselors could use this data to discuss time management and study habits with students who dine out frequently.

Campus dining services: Universities could use this information to improve on-campus dining options, potentially helping students maintain higher GPAs.

Student wellness programs: Develop programs that teach students how to balance academic performance with social activities and proper nutrition.

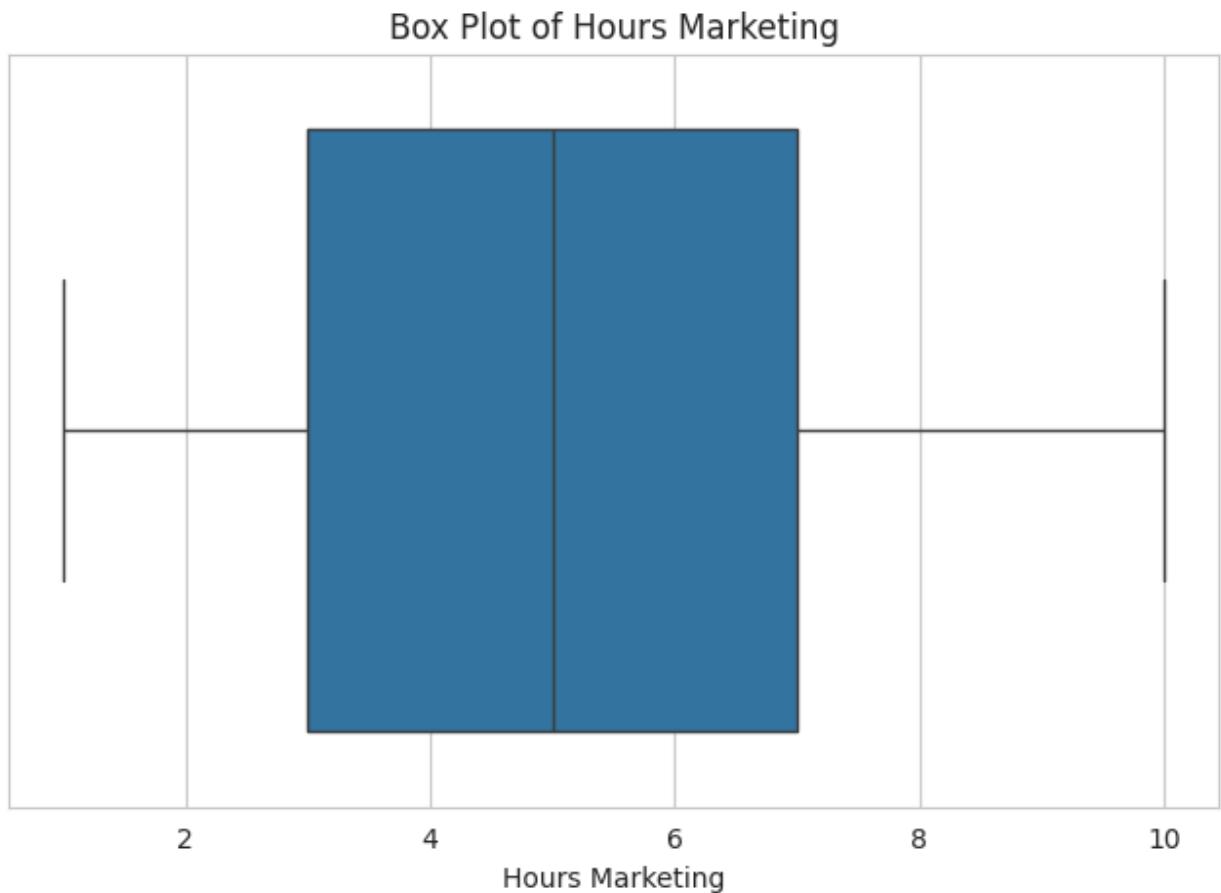
CASE STUDY 2 - Marketing

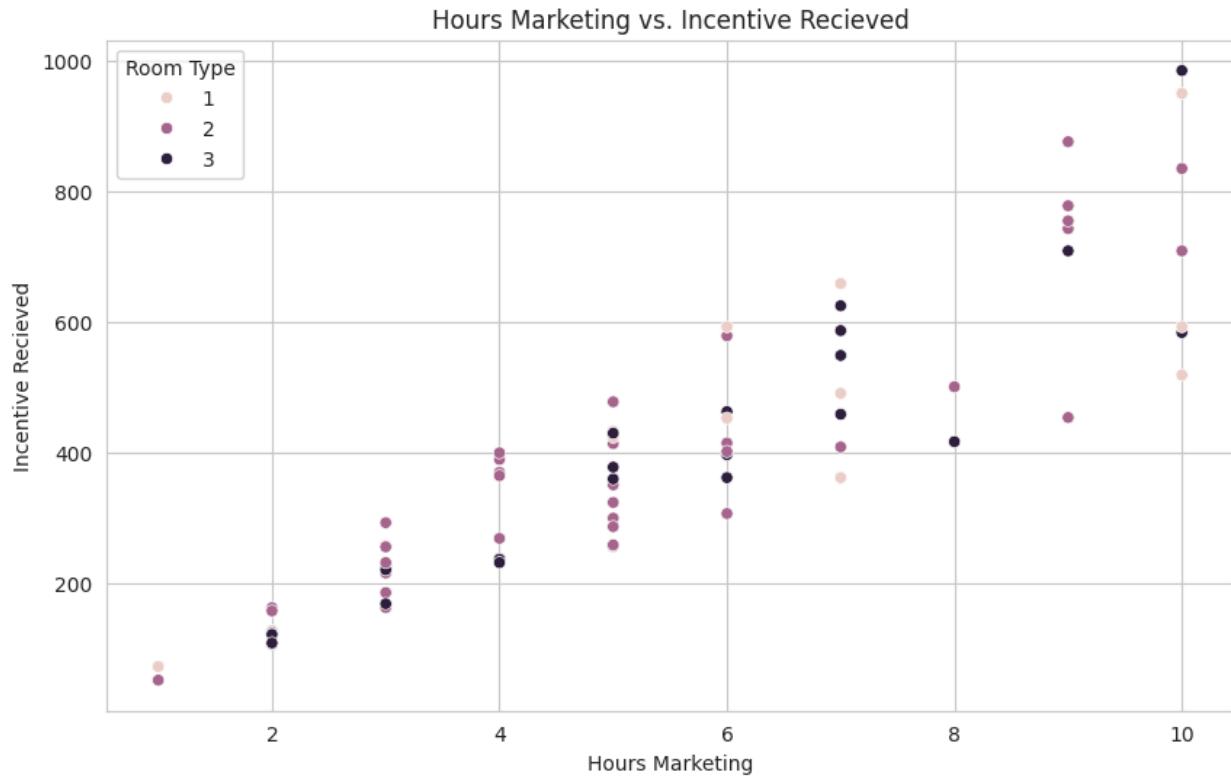
Python

```
# Box Plot - Hours Marketing
plt.figure(figsize=(8, 5))
```

```
sns.boxplot(data=data, x='Hours Marketing')
plt.title('Box Plot of Hours Marketing')
plt.xlabel('Hours Marketing')
plt.show()

# Scatter Plot - Hours Marketing vs. Incentive Received
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='Hours Marketing', y='Incentive Received', hue='Room Type')
plt.title('Hours Marketing vs. Incentive Recieved')
plt.xlabel('Hours Marketing')
plt.ylabel('Incentive Recieved')
plt.show()
```

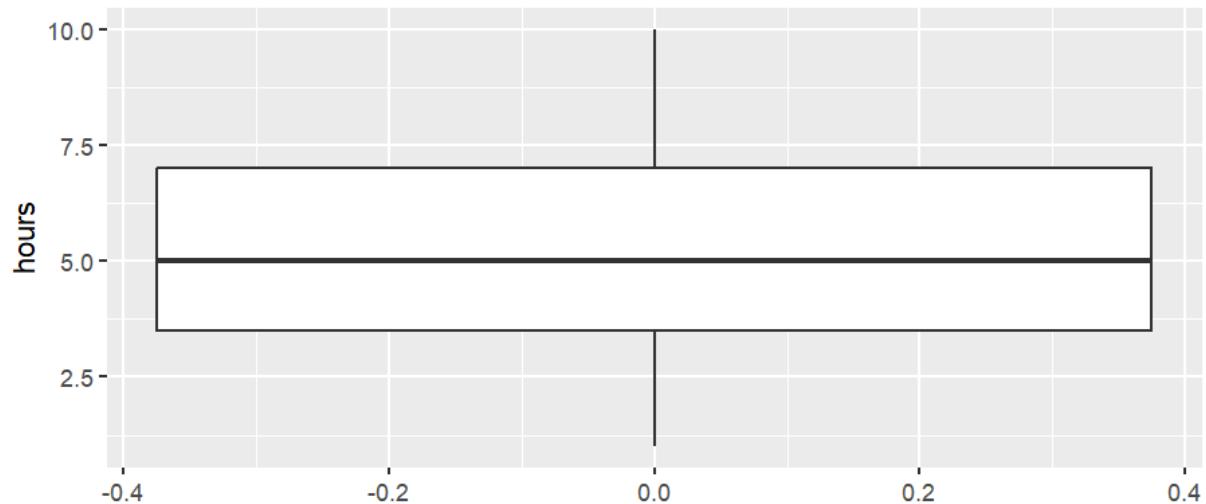




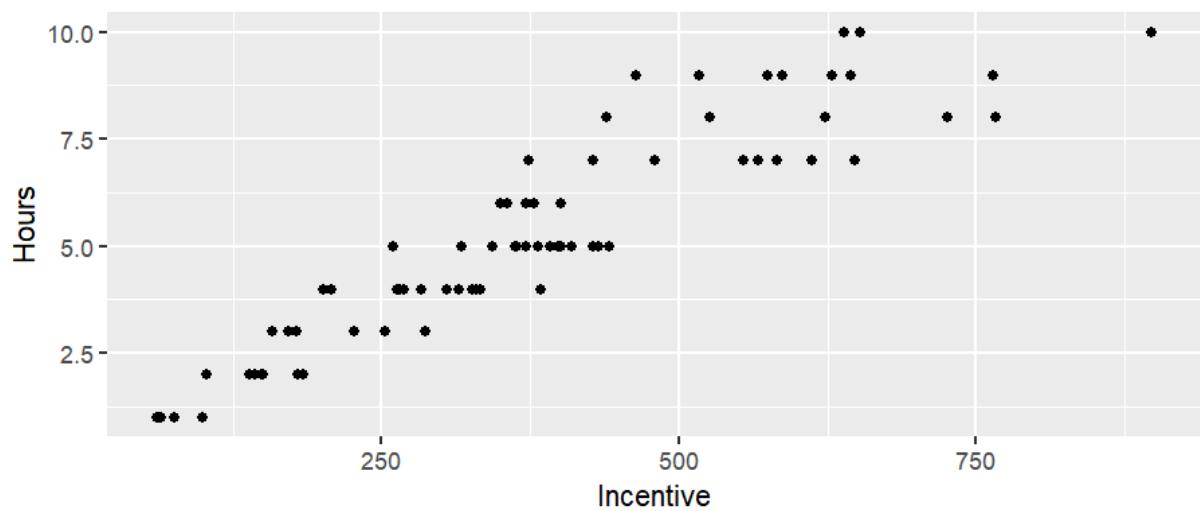
R

```
> # -----Box and Scatter Plot for CGPA
> ggplot(data, aes(y=market)) + geom_boxplot() + labs(title="Box Plot of Hours
marketed", y="hours")
> ggplot(data, aes(x=Incentive,y=market)) + geom_point() + labs(title="Scatter
Plot of Hours marketed", y="Hours")
```

Box Plot of Hours marketed



Scatter Plot of Hours marketed



Excel

Box Plot

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Insert Statistic Chart" button in the Charts group.

Step 4: Select "Box and Whisker" from the dropdown menu.

Scatter Plot

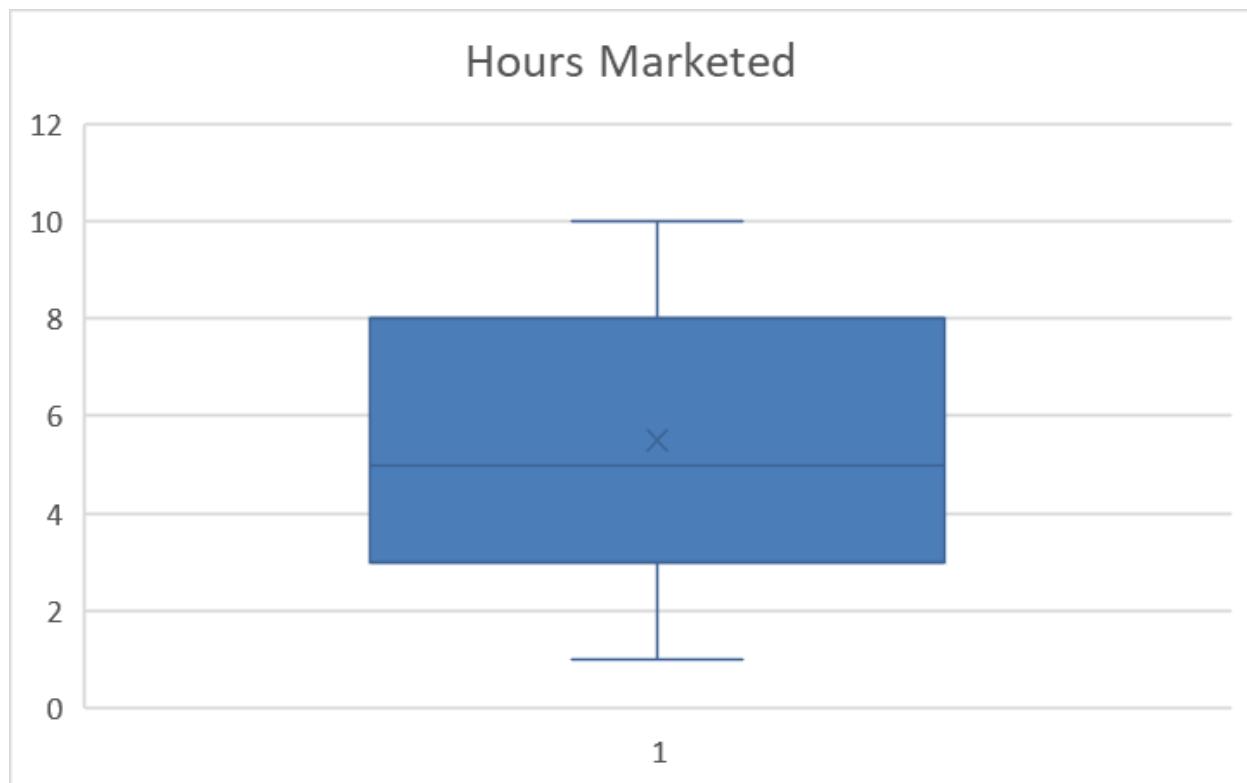
Step 1: Enter your x-axis data in one column and y-axis data in an adjacent column.

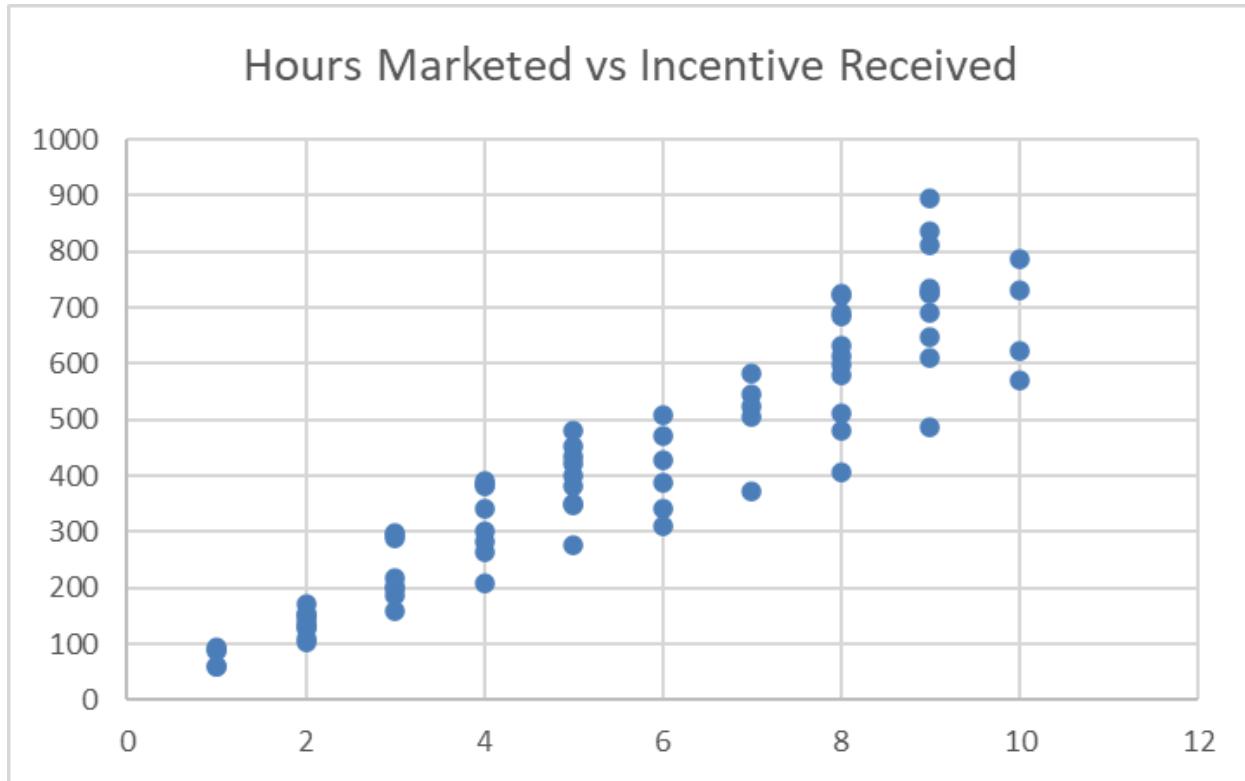
Step 2: Select both columns of data.

Step 3: Go to the Insert tab on the ribbon.

Step 4: Click on the "Scatter" button in the Charts group.

Step 5: Choose the desired scatter plot type from the dropdown menu.



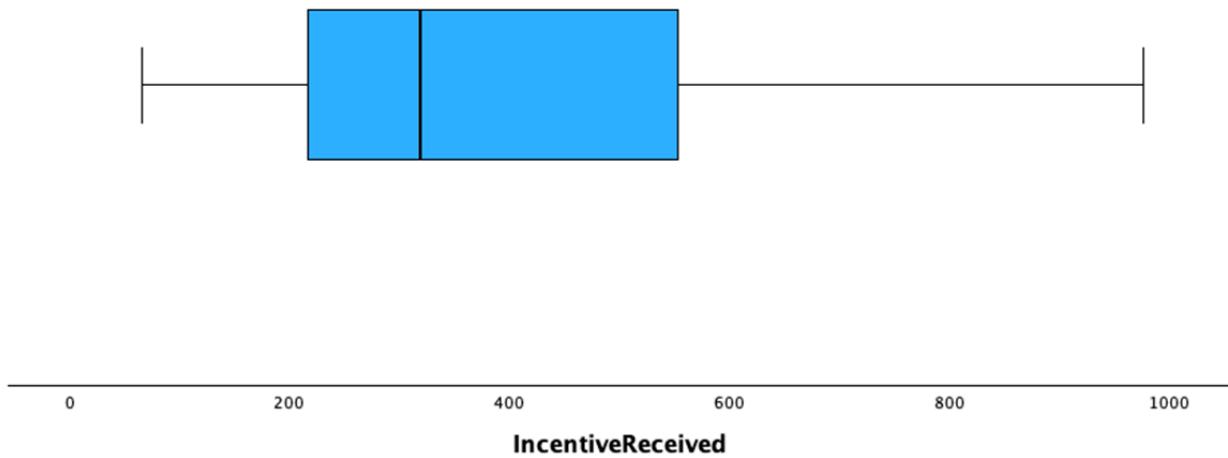


SPSS

Steps:

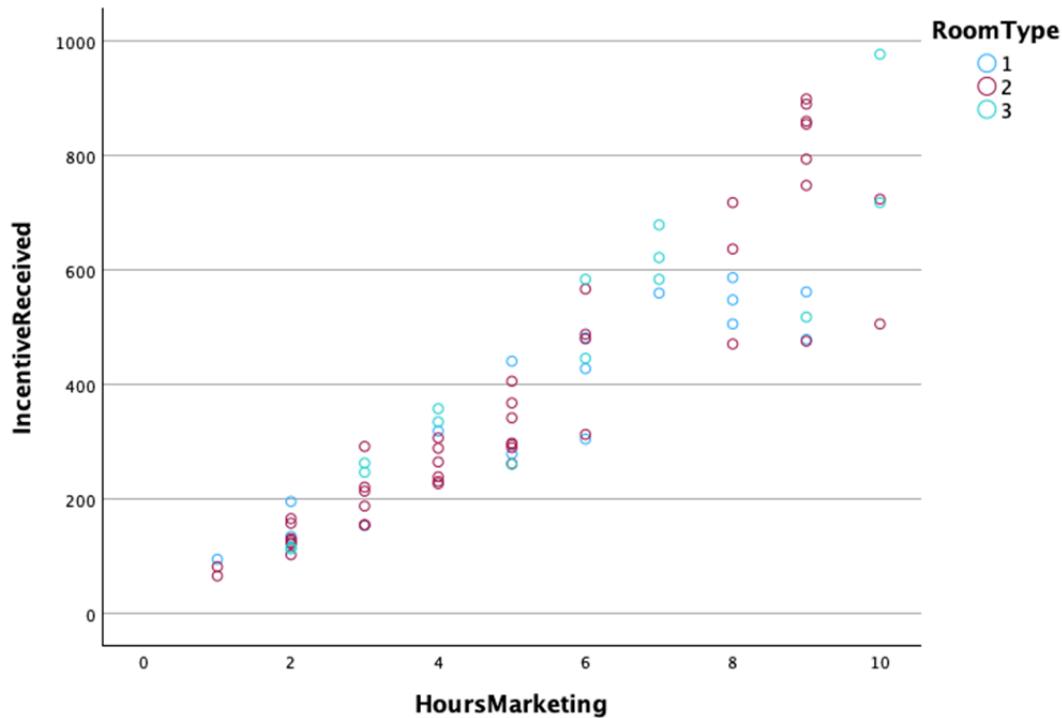
1. Go to Graphs > Chart Builder.
2. Select the Boxplot option.
3. Drag the variable to the y-axis and a categorical variable to the x-axis (if needed).

Simple Boxplot of IncentiveReceived



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Scatter/Dot option.
3. Drag the independent variable to the x-axis and the dependent variable to the y-axis.



Practical Output

Analysis:

There's a strong positive correlation between hours spent marketing and incentives received. The relationship appears to be roughly linear, with incentives increasing as marketing hours increase.

The box plot shows that most marketing efforts fall between 4-8 hours, with some outliers up to 10 hours.

Practical applications:

Sales team management: Use this data to set realistic expectations for incentives based on marketing effort.

Performance evaluation: Develop a fair compensation structure that rewards increased marketing efforts.

Resource allocation: Optimize marketing strategies by focusing on the most effective time investments (e.g., 6-8 hours may yield the best returns).

Training programs: Design marketing efficiency courses to help employees maximize their incentives within reasonable time frames.

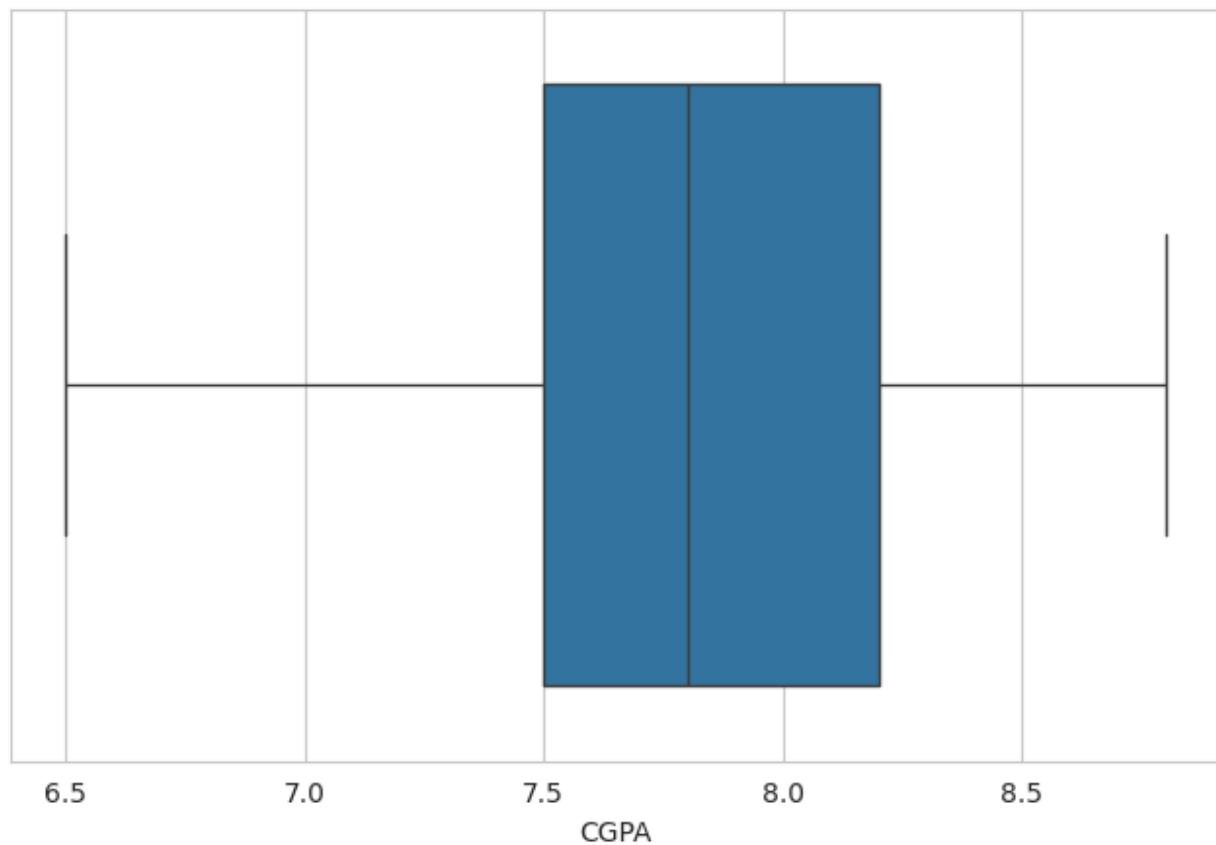
CASE STUDY 3 - Operations

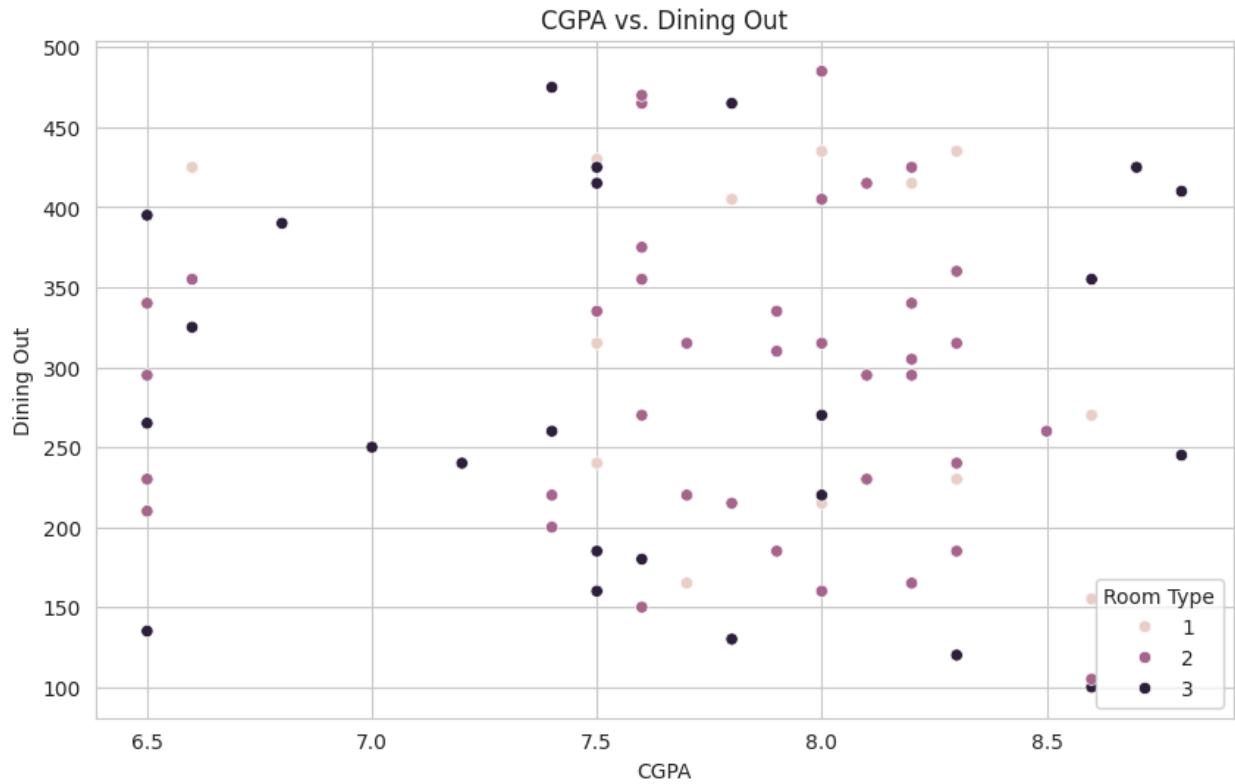
Python

```
# Box Plot - CGPA
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='CGPA')
plt.title('Box Plot of CGPA')
plt.xlabel('CGPA')
plt.show()

# Scatter Plot - CGPA vs. Dining Out
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='CGPA', y='Dining Out', hue='Room Type')
plt.title('CGPA vs. Dining Out')
plt.xlabel('CGPA')
plt.ylabel('Dining Out')
plt.show()
```

Box Plot of CGPA

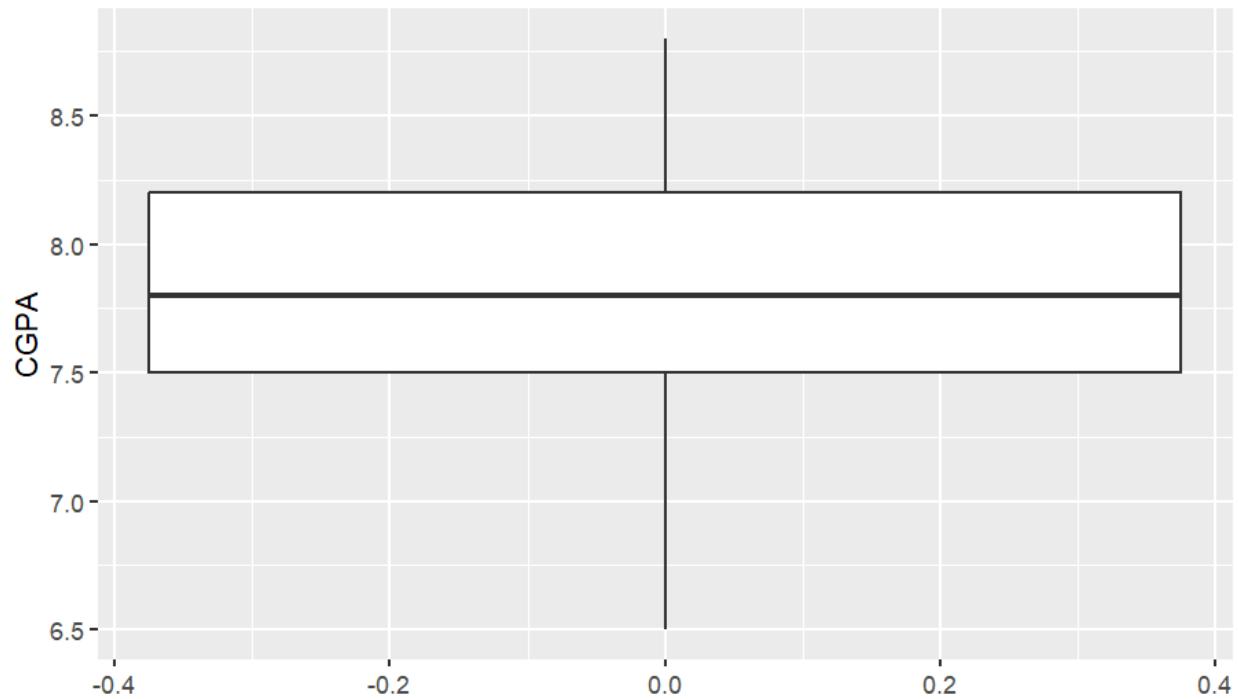




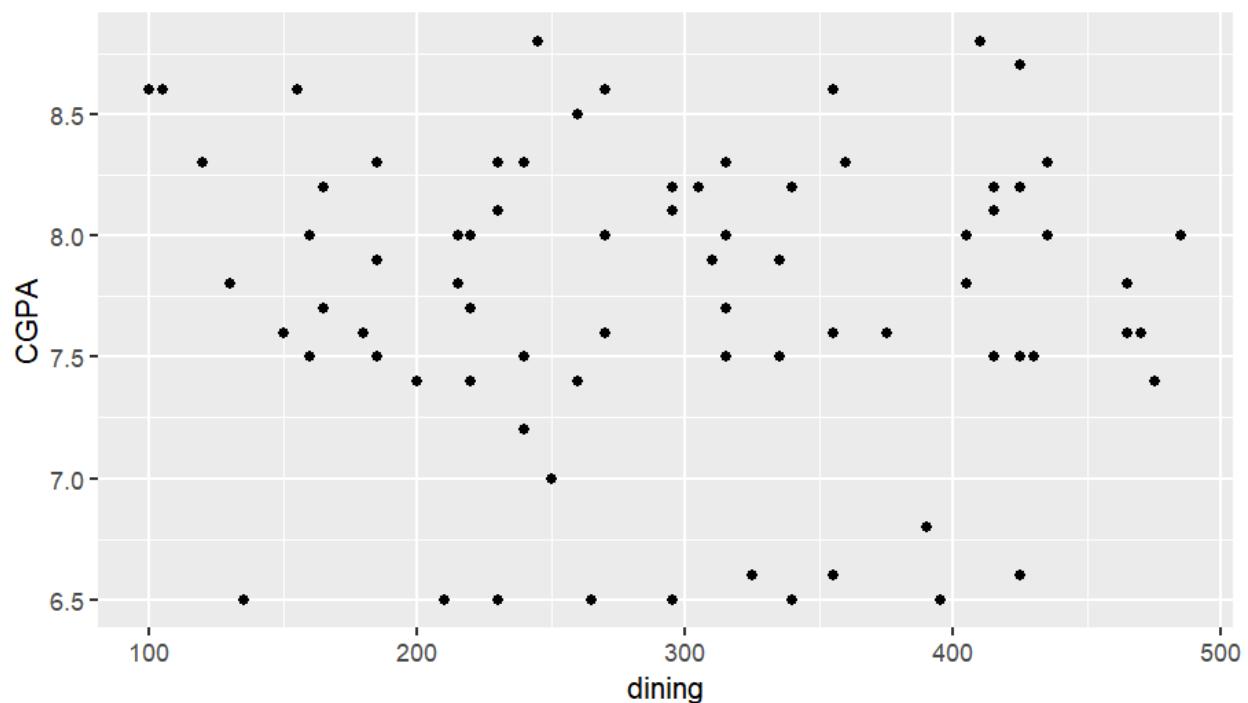
R

```
> # -----Box and Scatter Plot for CGPA
> ggplot(data, aes(y=CGPA)) + geom_boxplot() + labs(title="Box Plot of CGPA",
y="CGPA")
> ggplot(data, aes(x=dining,y=CGPA)) + geom_point() + labs(title="Scatter Plot
of CGPA", y="CGPA")
```

Box Plot of CGPA



Scatter Plot of CGPA



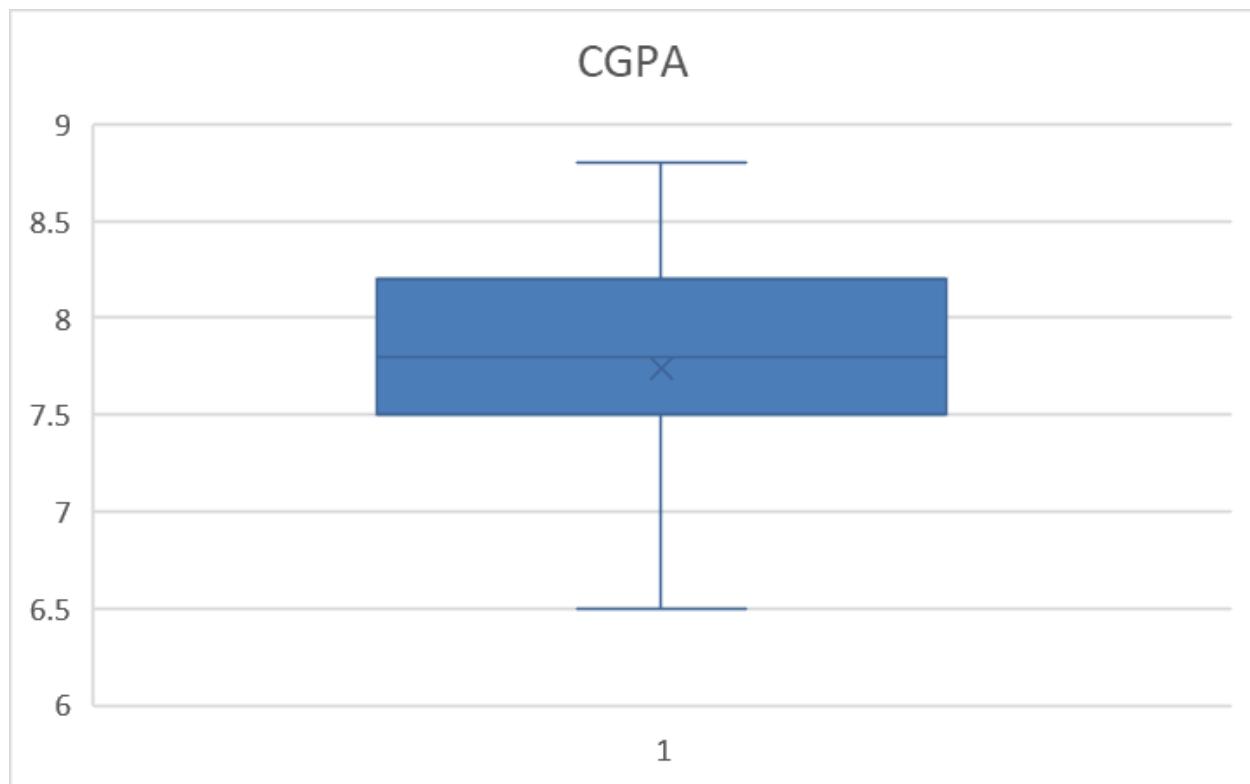
Excel

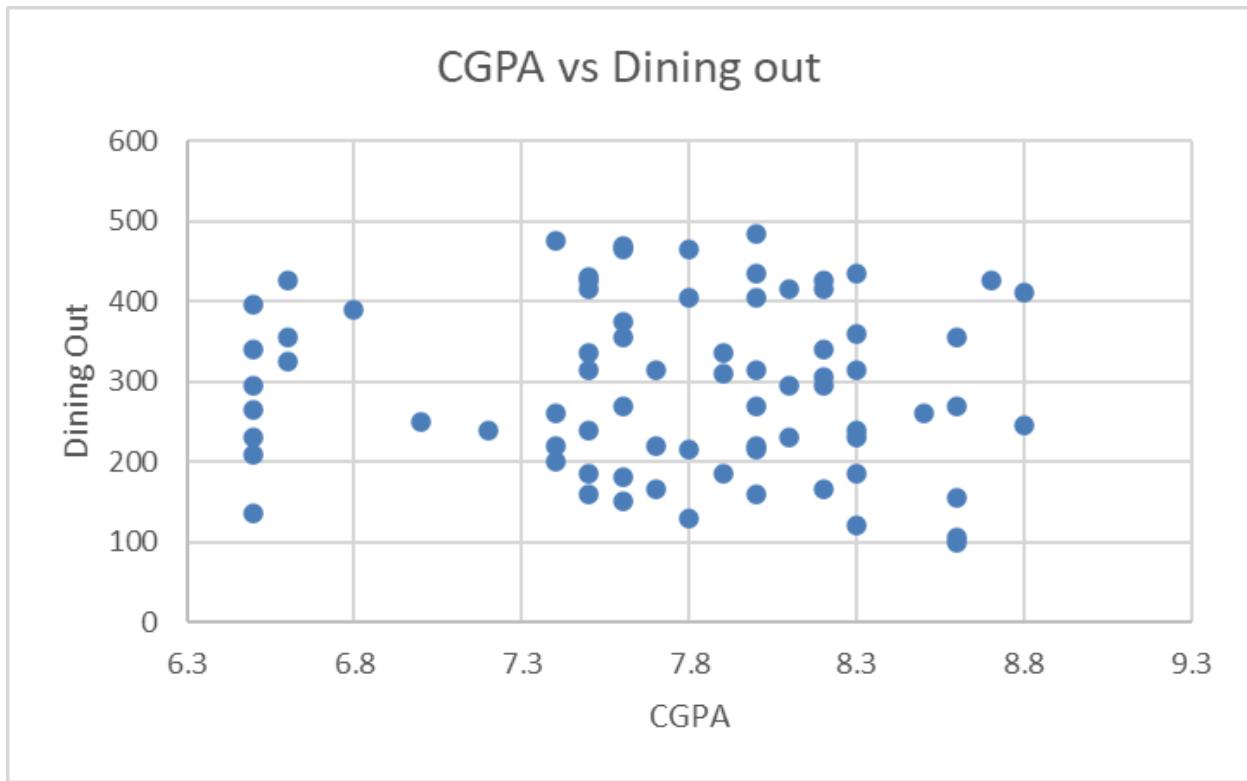
Box Plot

- Step 1: Select your data range.
- Step 2: Go to the Insert tab on the ribbon.
- Step 3: Click on the "Insert Statistic Chart" button in the Charts group.
- Step 4: Select "Box and Whisker" from the dropdown menu.

Scatter Plot

- Step 1: Enter your x-axis data in one column and y-axis data in an adjacent column.
- Step 2: Select both columns of data.
- Step 3: Go to the Insert tab on the ribbon.
- Step 4: Click on the "Scatter" button in the Charts group.
- Step 5: Choose the desired scatter plot type from the dropdown menu.



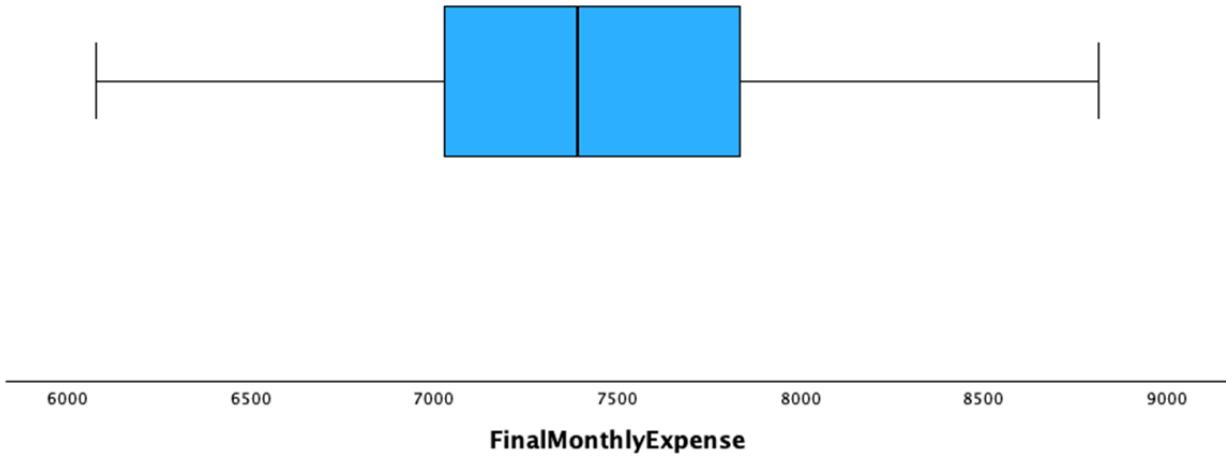


SPSS

Steps:

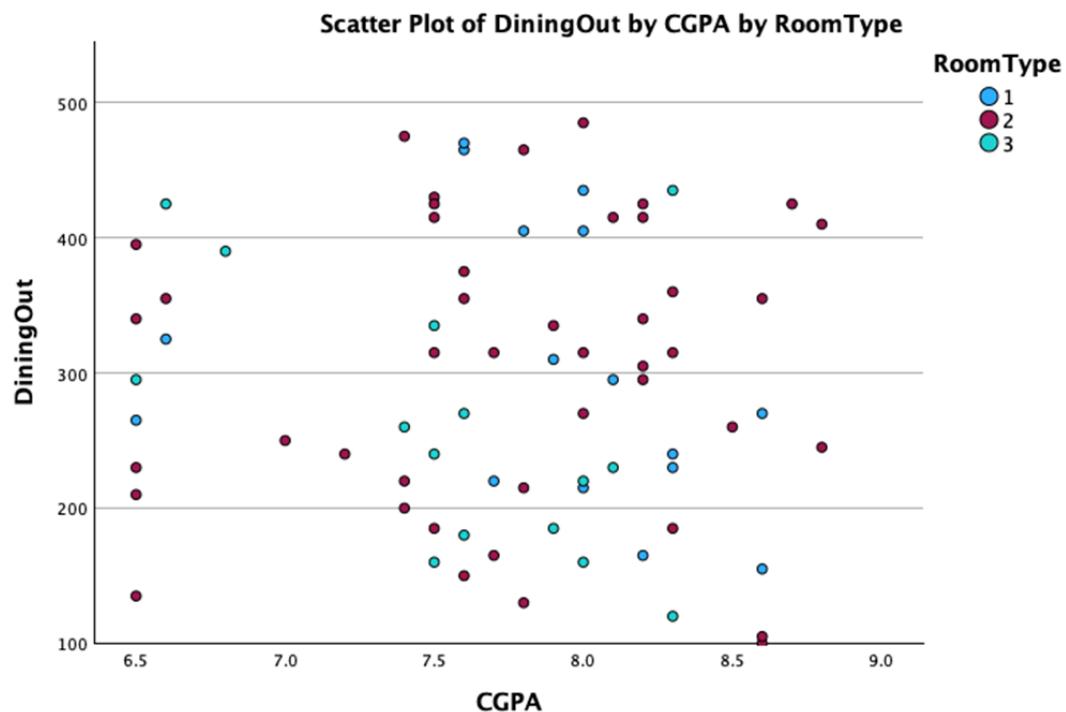
1. Go to Graphs > Chart Builder.
2. Select the Boxplot option.
3. Drag the variable to the y-axis and a categorical variable to the x-axis (if needed).

Simple Boxplot of FinalMonthlyExpense



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Scatter/Dot option.
3. Drag the independent variable to the x-axis and the dependent variable to the y-axis.



Practical Output

Analysis:

The scatter plot shows spending on groceries vs. dining out, with two food preferences indicated.

There's a wide spread of data points, suggesting variability in spending habits.

The box plot shows the distribution of grocery spending, with the median around 600-650.

Practical Application:

This data could be used by a personal finance app or financial advisor to help individuals understand and optimize their food spending habits. The app could:

Categorize users into different food preference groups.

Compare a user's spending to others with similar preferences.

Suggest budget allocations between groceries and dining out based on typical patterns.

Identify opportunities for savings by shifting spending from dining out to groceries, or vice versa, depending on the user's goals.

CASE STUDY 4 - Finance

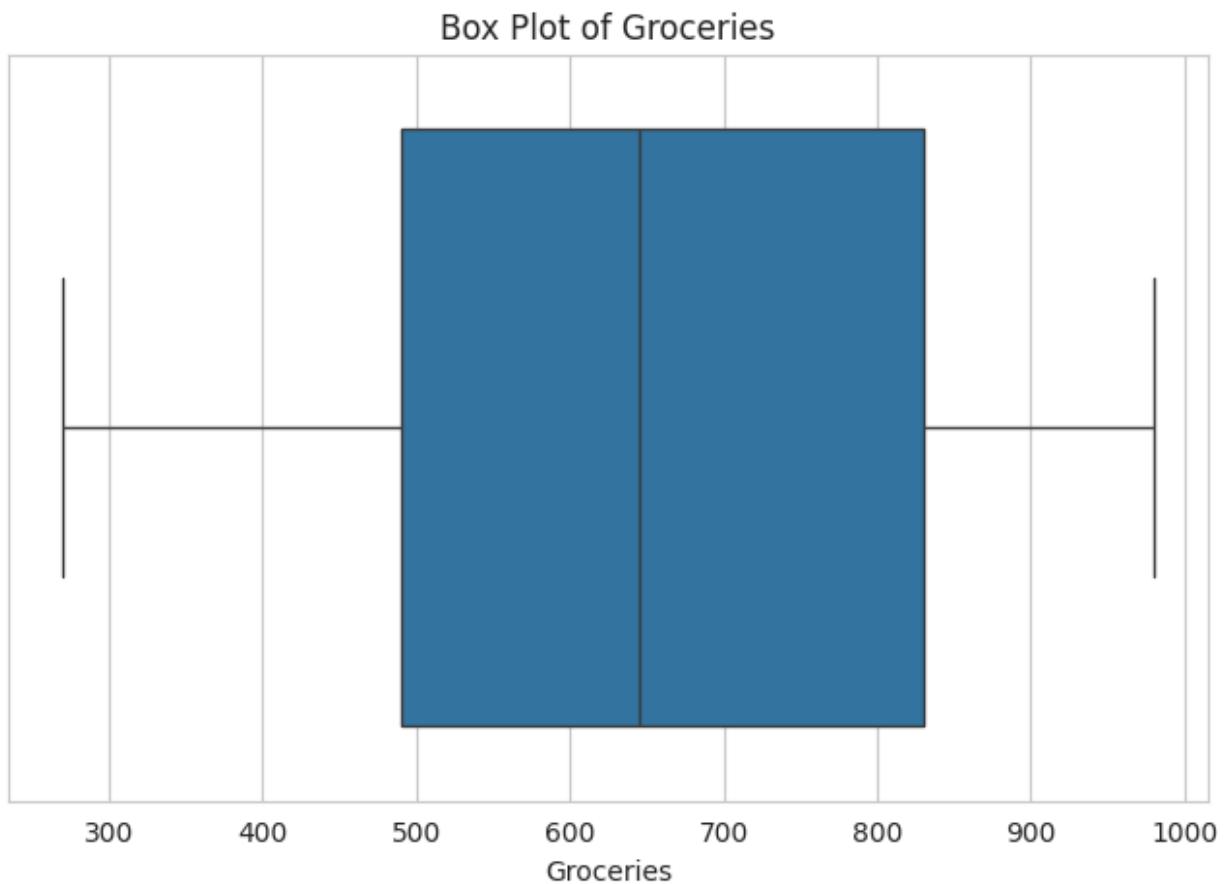
Python

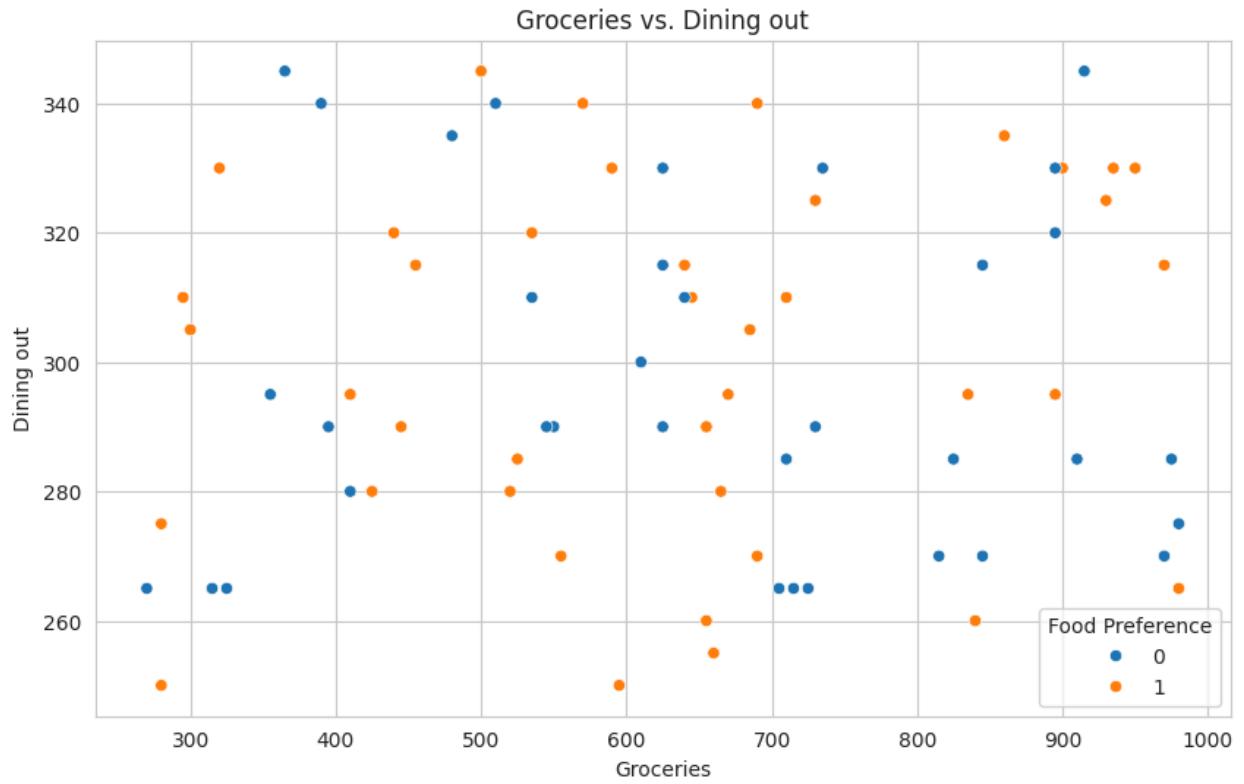
```

# Box Plot - Groceries
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='Groceries')
plt.title('Box Plot of Groceries')
plt.xlabel('Groceries')
plt.show()

# Scatter Plot - Groceries vs. Dining out
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='Groceries', y='Utilities', hue='Food
Preference')
plt.title('Groceries vs. Dining out')
plt.xlabel('Groceries')
plt.ylabel('Dining out')
plt.show()

```





R

```
ggplot(data, aes(y=groceries)) + geom_boxplot() + labs(title="Box Plot of Groceries",
y="Groceries")
ggplot(data, aes(x=Groceries,y=data$Dining.Out)) + geom_point() + labs(title="Scatter Plot of
Groceries vs Dining out", y="Dining out")
```



Excel

Box Plot

Step 1: Select your data range.

Step 2: Go to the Insert tab on the ribbon.

Step 3: Click on the "Insert Statistic Chart" button in the Charts group.

Step 4: Select "Box and Whisker" from the dropdown menu.

Scatter Plot

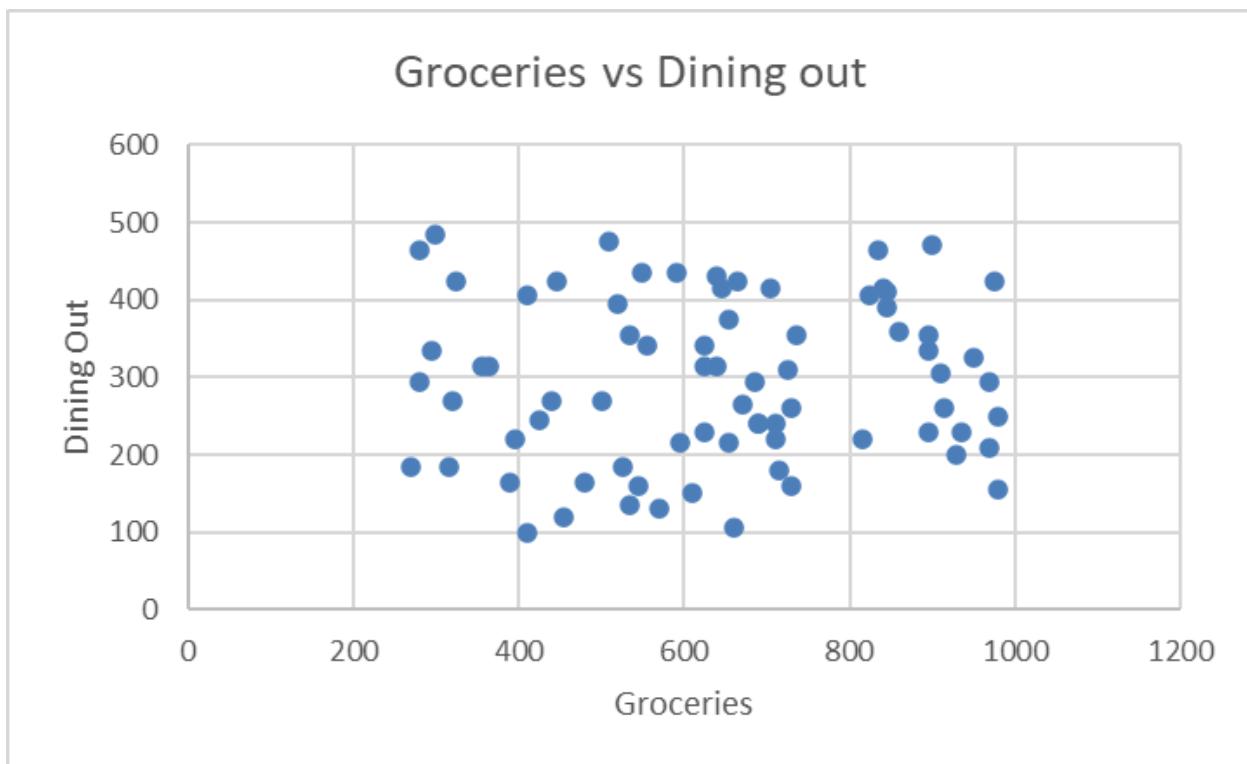
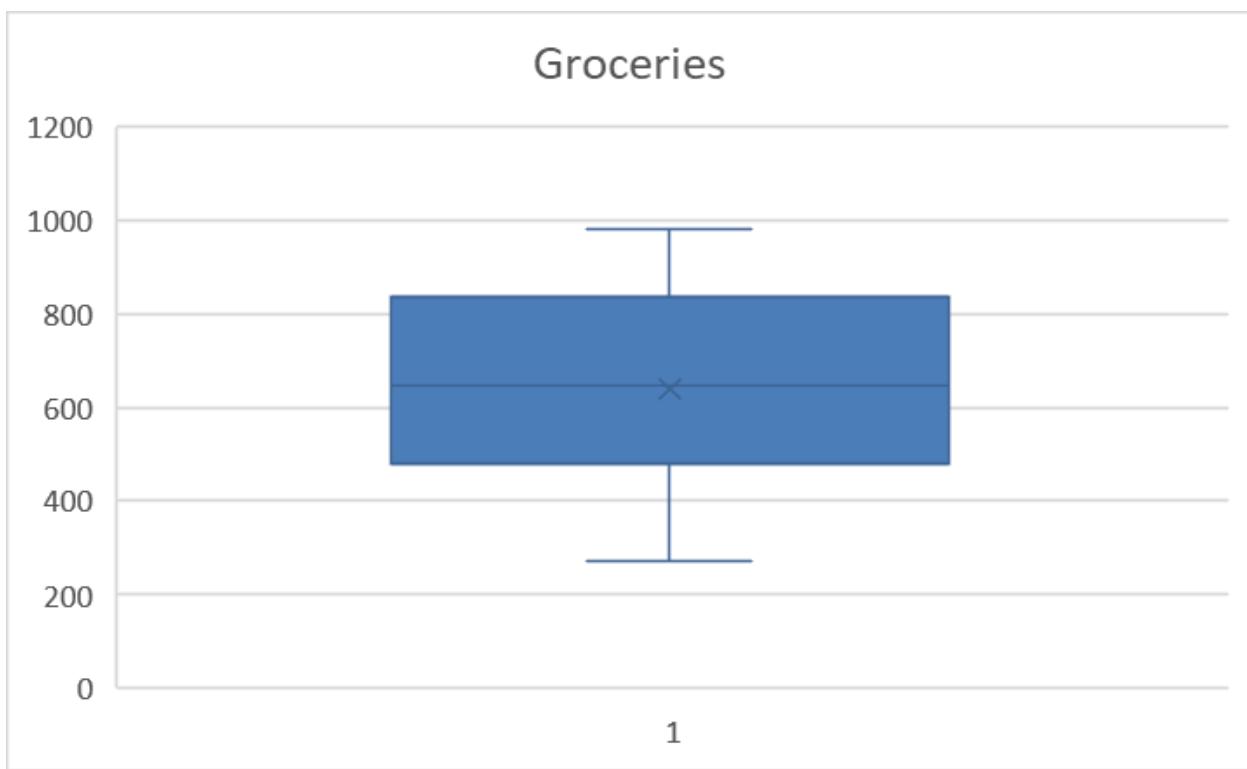
Step 1: Enter your x-axis data in one column and y-axis data in an adjacent column.

Step 2: Select both columns of data.

Step 3: Go to the Insert tab on the ribbon.

Step 4: Click on the "Scatter" button in the Charts group.

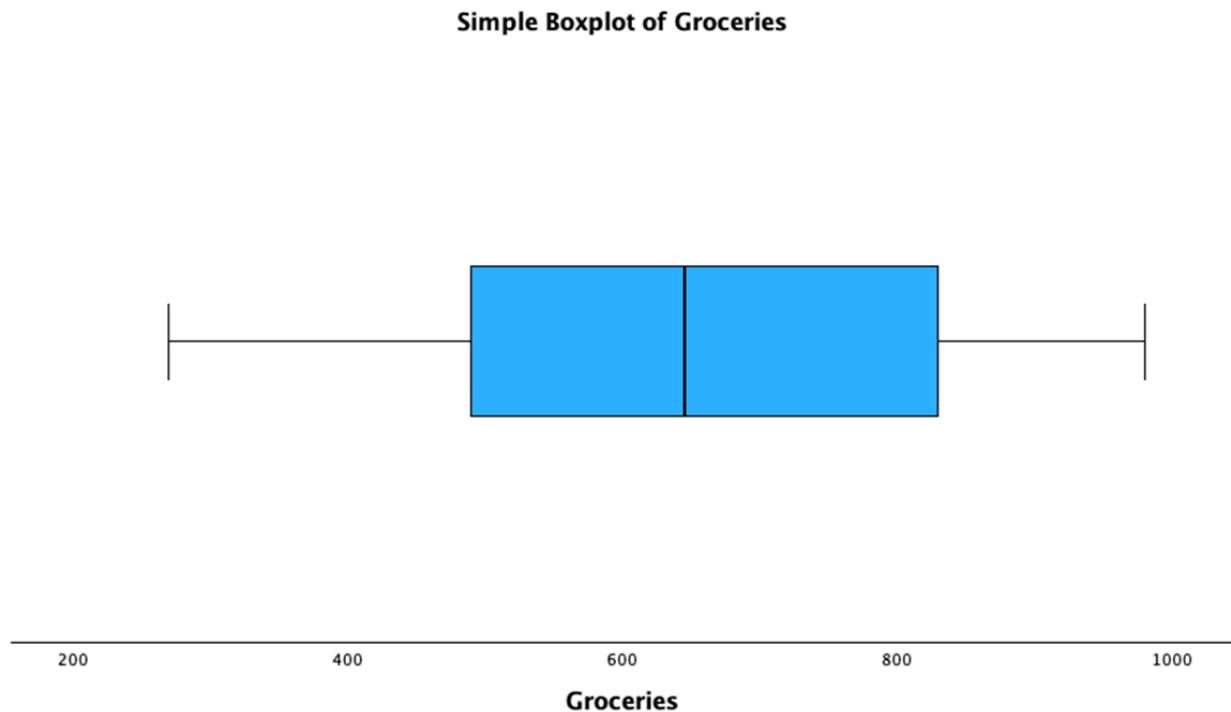
Step 5: Choose the desired scatter plot type from the dropdown menu.



SPSS

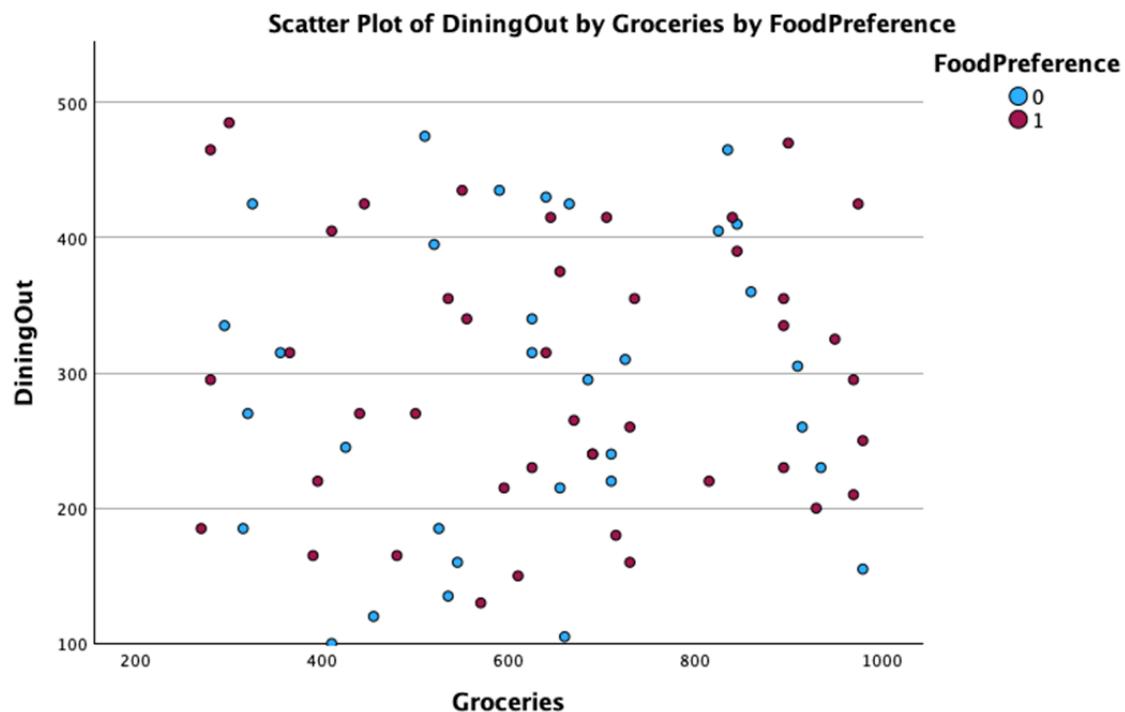
Steps:

1. Go to Graphs > Chart Builder.
2. Select the Boxplot option.
3. Drag the variable to the y-axis and a categorical variable to the x-axis (if needed).



Steps:

1. Go to Graphs > Chart Builder.
2. Select the Scatter/Dot option.
3. Drag the independent variable to the x-axis and the dependent variable to the y-axis.



Practical Output

Analysis:

The scatter plot shows CGPA vs. Aptitude Test Score, with placement status indicated. There's a general trend of higher CGPA and test scores correlating with placement. The box plot shows the distribution of CGPA, with the median around 7.75.

Practical Application:

This data could be used by a university career services department or an educational technology company to develop a student success prediction and intervention system. The system could:

Assess a student's likelihood of placement based on their current CGPA and practice aptitude test scores.

Identify students at risk of not being placed and offer targeted support.

Provide personalized advice on areas to improve (CGPA or aptitude skills) to increase placement chances.

Help companies set appropriate cutoffs for campus recruitment based on historical placement data.

CH 8 - Coefficient of Correlation - Karl Pearson

Coefficient of Correlation - Karl Pearson

In the intricate world of statistical analysis, the Coefficient of Correlation by Karl Pearson stands as a testament to the quest for understanding the strength and direction of the relationship between two quantitative variables. This section of our Harvard case study statistics book delves deep into the essence of Pearson's Coefficient of Correlation, exploring its foundational principles, practical applications, and the innovative avenues it opens up for data interpretation and analysis.

Unveiling Pearson's Coefficient of Correlation

At its core, Pearson's Coefficient of Correlation, denoted as r , quantifies the degree to which two variables linearly relate to each other. It is a measure that ranges from -1 to +1, where +1 signifies a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 indicates no linear relationship.

Theoretical Foundation

Developed by Karl Pearson at the turn of the 20th century, this statistic revolutionized the way researchers examined the interdependence of variables. The calculation of r involves the covariance of the two variables divided by the product of their standard deviations, providing a standardized measure of the linear relationship.

Application in Case Studies

Imagine a multinational corporation seeking to understand the relationship between employee training hours and productivity. By applying Pearson's Coefficient of Correlation to their data, they can quantitatively assess the strength and direction of the association, guiding strategic decisions in human resource development.

Creative Insights

- Data Visualization: Enhancing the understanding of r through scatter plots with superimposed lines of best fit can offer intuitive insights into the nature of the correlation, making the statistic more accessible to non-specialists.

- Dynamic Correlation Analysis: Utilizing interactive digital tools that allow stakeholders to adjust variables and immediately see changes in r can foster a deeper engagement with the data, encouraging exploratory analysis and hypothesis generation.

Navigating the Nuances of Correlation

Pearson's Coefficient of Correlation does more than measure the strength and direction of linear relationships; it opens a window into the dynamics of data, offering clues to underlying causative factors, albeit with the caveat that correlation does not imply causation.

Beyond Linearity

While r is a powerful tool for assessing linear relationships, its application requires careful consideration of the data's nature. Non-linear relationships, outliers, and the potential for spurious correlations demand a nuanced approach to interpretation, emphasizing the importance of complementary analysis to validate findings.

Creative Insights

- Segmented Analysis: Breaking down datasets into smaller, homogeneous segments before applying Pearson's correlation can unearth varying strengths of relationships across different groups, offering more tailored insights.
- Multivariate Extensions: Exploring the use of Partial Correlation and Multiple Regression techniques as extensions of Pearson's correlation can address confounding variables and offer a more complex picture of the relationships among multiple variables.

Pearson's Correlation in the Digital Age

In the era of big data and advanced analytics, Pearson's Coefficient of Correlation retains its relevance, serving as a foundational tool in machine learning, econometrics, and beyond. Its ability to quantify the linear relationship between variables makes it indispensable in predictive modeling and risk assessment.

Application in Emerging Fields

In fields such as genomics and climatology, where vast datasets are common, Pearson's correlation helps in identifying potential genetic linkages or assessing the relationship between climate variables. This statistical measure becomes a bridge between raw data and meaningful patterns, guiding further research and policy decisions.

Creative Insights

- Interactive Learning Platforms: Developing digital platforms that simulate real-world scenarios, allowing users to apply Pearson's correlation and instantly visualize its impact, can demystify statistics and enhance data literacy among a broader audience.
- Cross-Disciplinary Applications: Encouraging the application of Pearson's correlation across disciplines can foster innovative approaches to solving complex problems, from tracking environmental changes to predicting market trends, illustrating the versatility and enduring value of this statistical measure.

CASE STUDY 1 - HR

Python

```
# Extract the relevant columns
cgpa_column = data['CGPA']
aptitude_score_column = df['AptitudeTestScore']

# Calculate the coefficient of correlation
correlation_coefficient = cgpa_column.corr(aptitude_score_column,
method='pearson')

print(f"Coefficient of Correlation (Karl Pearson):"
{correlation_coefficient})
```

Coefficient of Correlation (Karl Pearson): 0.672598999317846

R

```
> # Pearson Rank Correlation
> cor.test(data$CGPA, data$AptitudeTestScore, method = "pearson")

Pearson's product-moment correlation
```

```

data: data$CGPA and data$AptitudeTestScore
t = 7.7657, df = 73, p-value = 3.877e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5259213 0.7804255
sample estimates:
cor
0.672599

```

Excel

Karl Pearson's Coefficient of Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In an empty cell, type: =CORREL(A1:A100,B1:B100)

Remember to replace A1:A100 (and B1:B100 where applicable) with your actual data ranges in all formulas.

Coeff of Corr - Karl	0.672598999
----------------------	-------------

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Pearson is selected.

Correlations

		CGPA	DiningOut
CGPA	Pearson Correlation	1	-.063
	Sig. (2-tailed)		.593
	N	75	75

DiningOut	Pearson Correlation	-.063	1
	Sig. (2-tailed)	.593	
N		75	75

Practical Output

Correlation Coefficient: 0.6726

Practical Application:

This indicates a moderately strong positive correlation between CGPA and Aptitude Test Scores.

As CGPA increases, Aptitude Test Scores tend to increase as well, and vice versa.

Application for Education Administrators and Career Counselors:

Use this information to identify students who might need additional support. If a student has a high CGPA but a low Aptitude Test Score (or vice versa), they might benefit from targeted interventions.

Consider using both CGPA and Aptitude Test Scores in admissions or job placement processes, as they seem to complement each other in assessing student capabilities.

Develop programs that enhance both academic performance and aptitude skills, as they appear to be related.

When counseling students about career prospects, emphasize the importance of developing both academic knowledge and practical aptitude.

CASE STUDY 2 - Marketing

Python

```
# Extract the relevant columns
cgpa_column = data['Hours Marketing']
aptitude_score_column = df['Incentive Received']

# Calculate the coefficient of correlation
correlation_coefficient = cgpa_column.corr(aptitude_score_column, method='pearson')

print(f"Coefficient of Correlation (Karl Pearson): {correlation_coefficient}")
```

Coefficient of Correlation (Karl Pearson) : 0.8999309930027736

R

```
cor.test(market, Incentive, method = "pearson")
```

Pearson's product-moment correlation

data: market and Incentive

t = 22.832, df = 73, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9011709 0.9595645

sample estimates:

cor

0.936574

Excel

Karl Pearson's Coefficient of Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In an empty cell, type: =CORREL(A1:A100,B1:B100)

Remember to replace A1:A100 (and B1:B100 where applicable) with your actual data ranges in all formulas.

COEFFICIENT OF CORRELATION -	0.93423785
KARL PEARSON	4

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Pearson is selected.

Correlations

		HoursMarketing	IncentiveReceived
HoursMarketing	Pearson Correlation	1	.911**
	Sig. (2-tailed)		<.001
	N	75	75
IncentiveReceived	Pearson Correlation	.911**	1
	Sig. (2-tailed)	<.001	
	N	75	75

**. Correlation is significant at the 0.01 level (2-tailed).

Practical Output

Correlation Coefficient: 0.8999

Practical Application:

This shows a very strong positive correlation between Hours spent on Marketing and Incentives Received.

As the number of marketing hours increases, the incentives received tend to increase substantially.

Application for Marketing Managers and HR Professionals:

Use this strong correlation to justify and refine performance-based incentive systems.

Consider implementing a structured program that encourages more hours in marketing activities, as it's strongly linked to higher incentives (and presumably, better results).

Analyze outliers: Are there employees who work many hours but receive low incentives, or vice versa? This could help identify efficiency issues or exceptional performers.

Use this data to set realistic expectations for new hires about the relationship between effort and reward in marketing roles.

Consider if the current incentive structure is too closely tied to hours worked rather than actual results or efficiency.

CASE STUDY 3 - Operations

Python

```
# Extract the relevant columns
cgpa_column = data['CGPA']
aptitude_score_column = df['Dining Out']

# Calculate the coefficient of correlation
correlation_coefficient = cgpa_column.corr(aptitude_score_column,
method='pearson')

print(f"Coefficient of Correlation (Karl Pearson): {correlation_coefficient}")
```

Coefficient of Correlation (Karl Pearson): -0.06278494507075082

R

```
cor.test(data$CGPA, dining, method = "pearson")
```

Pearson's product-moment correlation

```
data: data$CGPA and dining
t = -0.5375, df = 73, p-value = 0.5926
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2856760 0.1665502
sample estimates:
cor
-0.06278495
```

Excel

Karl Pearson's Coefficient of Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In an empty cell, type: =CORREL(A1:A100,B1:B100)

Remember to replace A1:A100 (and B1:B100 where applicable) with your actual data ranges in all formulas.

COEFFICIENT OF CORRELATION -	0.67259899
KARL PEARSON	9

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Pearson is selected.

Correlations

		CGPA	DiningOut
CGPA	Pearson Correlation	1	-.063
	Sig. (2-tailed)		.593
DiningOut	N	75	75
	Pearson Correlation	-.063	1
	Sig. (2-tailed)	.593	
	N	75	75

Practical Output

Correlation Coefficient: -0.0628

Practical Application:

This indicates a very weak negative correlation, almost no correlation, between CGPA and Dining Out expenses.

There's practically no linear relationship between academic performance and money spent on dining out.

Application for Student Affairs and Financial Advisors:

Recognize that dining out habits don't seem to significantly impact academic performance, or vice versa.

When advising students on budgeting, focus on other factors that might more directly impact academic performance.

For student wellness programs, this data suggests that restricting dining out may not necessarily lead to improved academic performance. Consider investigating other factors that might have a stronger correlation with academic performance when developing student support programs.

CASE STUDY 4 - Finance

Python

```
# Extract the relevant columns
Groceries_column = data['Groceries']
aptitude_score_column = df['Dining Out']

# Calculate the coefficient of correlation
correlation_coefficient = Groceries_column.corr(aptitude_score_column,
method='pearson')

print(f"Coefficient of Correlation (Karl Pearson):
{correlation_coefficient}")
```

Coefficient of Correlation (Karl Pearson): 0.04140275851356626

R

```
cor.test(groceries, data$Dining.Out, method = "pearson")
```

Pearson's product-moment correlation

```
data: groceries and data$Dining.Out
t = 0.35405, df = 73, p-value = 0.7243
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1873193 0.2658663
sample estimates:
cor
0.04140276
```

Excel

Karl Pearson's Coefficient of Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In an empty cell, type: =CORREL(A1:A100,B1:B100)
Remember to replace A1:A100 (and B1:B100 where applicable) with your actual data ranges in all formulas.

COEFFICIENT OF CORRELATION - -0.152209491
KARL PEARSON

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Pearson is selected.

Correlations

		Groceries	Utilities
Groceries	Pearson Correlation	1	.059
	Sig. (2-tailed)		.614
	N	75	75
Utilities	Pearson Correlation	.059	1
	Sig. (2-tailed)	.614	
	N	75	75

Practical Output

Correlation Coefficient: 0.0414

Practical Application:

This shows an extremely weak positive correlation, essentially no correlation, between Grocery expenses and Dining Out expenses.

There's virtually no linear relationship between how much people spend on groceries and how much they spend on dining out.

Application for Financial Advisors and Consumer Behavior Analysts:

Recognize that people's spending on groceries doesn't predict their spending on dining out, or vice versa.

When creating budgeting advice, treat grocery and dining out expenses as independent categories.

For market researchers, this suggests that strategies to increase grocery sales are unlikely to directly impact dining out behaviors, and vice versa.

For restaurants and grocery stores, this implies that they're not in direct competition with each other as much as might be assumed.

When analyzing household budgets, look for other factors that might influence these spending categories independently.

CH 9- Coefficent of Correlation - Spearman Rank Colleation

Coefficent of Correlation - Spearman Rank Colleation

In the vast expanse of statistical methodologies, Spearman's Rank Correlation Coefficient, often symbolized as ρ (rho), stands as a beacon for non-parametric analysis, offering a robust alternative to Pearson's coefficient for assessing the strength and direction of association between two variables. This section of our Harvard case study statistics book embarks on an exploration of Spearman's Rank Correlation, weaving through its conceptual underpinnings, practical applications, and the innovative insights it can provide in the realm of data analysis.

Spearman's Rank Correlation: Unraveling Non-Linear Associations

Spearman's Rank Correlation Coefficient transcends the linear confines of Pearson's correlation, allowing researchers to capture monotonic relationships—whether linear or nonlinear—between two variables. By focusing on the ranks of data rather than their raw values, Spearman's ρ provides a measure of correlation that is less sensitive to outliers and skewed distributions.

Theoretical Foundation

The genius of Spearman's Rank Correlation lies in its simplicity and adaptability. It calculates the correlation based on the ranked values of the data, thus neutralizing the effects of outliers and non-normal distributions. This ranking process involves assigning ordinal positions to data points, from the smallest to the largest, and then applying Pearson's correlation formula to these ranks.

Application in Case Studies

Imagine an educational researcher investigating the relationship between students' self-reported motivation levels and their academic performance. Given the ordinal nature of self-reported measures and the potential for non-linear relationships, Spearman's ρ emerges as the ideal tool for uncovering the underlying association, guiding interventions to enhance educational outcomes.

Creative Insights

- Visualizing Monotonic Trends: Accompanying Spearman's analysis with scatter plots of ranked data can visually reinforce the findings, making the monotonic relationships more apparent and accessible to a broader audience.
- Dynamic Ranking Analysis: Leveraging digital platforms to allow interactive exploration of how changes in data affect Spearman's ρ can demystify the concept of rank correlation, promoting a deeper understanding of its implications.

Spearman's Rank Correlation in Action

Beyond the theoretical allure, Spearman's Rank Correlation finds its strength in real-world applications, from market research to healthcare, where it illuminates the subtle dynamics between variables that might otherwise go unnoticed.

Bridging Qualitative and Quantitative Worlds

In social sciences, where qualitative data abound, Spearman's ρ offers a quantitative lens to examine the relationships between ordinal variables, such as survey responses. This crosswalk between qualitative insights and quantitative analysis enriches research findings, providing a more nuanced understanding of social phenomena.

Creative Insights

- Integrating Textual Analysis: Pairing Spearman's correlation with textual analysis of open-ended survey responses can uncover correlations between quantifiable ranks and qualitative themes, offering a holistic view of the data.
- Cross-Disciplinary Methodologies: Applying Spearman's ρ in interdisciplinary studies, such as environmental science and psychology, can reveal unexpected correlations, fostering innovative solutions to complex problems.

Spearman's Rank Correlation: A Tool for Innovation

As we delve deeper into the age of data, Spearman's Rank Correlation stands out not just as a statistical method but as a catalyst for innovation, encouraging creative approaches to data analysis across fields.

Navigating Big Data

In the era of big data, where traditional assumptions of normality and linearity often break down, Spearman's ρ provides a reliable method for assessing correlations in large datasets, guiding data-driven decision-making in business and governance.

Creative Insights

- Predictive Analytics: Incorporating Spearman's ρ into predictive models, especially those involving ordinal predictors or outcomes, can enhance model accuracy and interpretability, opening new frontiers in analytics.
- Interactive Data Exploration Tools: Developing tools that dynamically calculate Spearman's ρ as users manipulate data can make complex statistical concepts accessible, fostering a culture of data literacy and evidence-based decision-making.

CASE STUDY 1 - HR

Python

```
# Spearman Rank Correlation
```

```
spearman_corr = spearmanr(data['CGPA'], data['AptitudeTestScore'])
print(f"Spearman Rank Correlation between CGPA and AptitudeTestScore:
{spearman_corr.correlation}")
```

```
Spearman Rank Correlation between CGPA and AptitudeTestScore:
0.6439794175321499
```

R

```
> # -----Spearman Rank Correlation
> cor.test(data$CGPA, data$AptitudeTestScore, method = "spearman")

  Spearman's rank correlation rho

data: data$CGPA and data$AptitudeTestScore
S = 25028, p-value = 4.583e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.6439794
```

Excel

Spearman Rank Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In column C, enter this formula to rank the first variable: =RANK.EQ(A1,\$A\$1:\$A\$100)

Step 3: In column D, enter this formula to rank the second variable:

=RANK.EQ(B1,\$B\$1:\$B\$100)

Step 4: In an empty cell, calculate the correlation between the ranks:

=CORREL(C1:C100,D1:D100)

Coeff of Corr - 0.640697198

Spearman

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Spearman is selected.

Correlations

			CGPA	DiningOut
Spearman's rho	CGPA	Correlation Coefficient	1.000	-.057
		Sig. (2-tailed)	.	.629
		N	75	75
	DiningOut	Correlation Coefficient	-.057	1.000
		Sig. (2-tailed)	.629	.
		N	75	75

Practical Output:

Interpretation:

- Positive Relationship:** The positive correlation coefficient indicates that higher CGPA values are associated with higher aptitude test scores, and vice versa.
- Strength of Correlation:**
 - The value 0.6439794175321499 is closer to 1, which suggests a fairly strong relationship.
 - However, it is not perfect, meaning there are other factors affecting the relationship between CGPA and aptitude test scores.

Practical Implications:

- Predictive Value:** If you're using CGPA to predict aptitude test scores (or vice versa), the moderate to strong correlation suggests a reasonable degree of predictability. However, other variables and individual differences still play a role.
- Consistency:** The correlation implies that there is a consistent, though not perfect, ranking order between students' CGPAs and their aptitude test scores. For example, students who rank higher in CGPA also tend to rank higher in their aptitude test scores.

CASE STUDY 2 - Marketing

Python

```
# Spearman Rank Correlation
spearman_corr = spearmanr(data['Hours Marketing'], data["Incentive Received"])
print(f"Spearmen Rank Correlation between Hours Marketing and Incentive Received:
{spearman_corr.correlation}")
```

```
Spearmen Rank Correlation between Hours Marketing and Incentive Received:
0.9217223090752399
```

R

```
> # -----Spearman Rank Correlation
> cor.test(market, Incentive, method = "spearman")
Spearman's rank correlation rho

data: market and Incentive
S = 4070.8, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9420945

Warning message:
In cor.test.default(market, Incentive, method = "spearman") :
  Cannot compute exact p-value with ties
> # Pearson Rank Correlation
> cor.test(market, Incentive, method = "pearson")

Pearson's product-moment correlation

data: market and Incentive
t = 21.348, df = 73, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8887069 0.9542878
sample estimates:
cor
0.9284067
```

Excel

Spearman Rank Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In column C, enter this formula to rank the first variable: =RANK.EQ(A1,\$A\$1:\$A\$100)

Step 3: In column D, enter this formula to rank the second variable:

=RANK.EQ(B1,\$B\$1:\$B\$100)

Step 4: In an empty cell, calculate the correlation between the ranks:

=CORREL(C1:C100,D1:D100)

Coeff of Corr - 0.9074067

Spearman

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Spearman is selected.

Correlations

		HoursMarketin	IncentiveReceive
		g	d
Spearman's rho	HoursMarketing	Correlation	1.000
		Coefficient	.937**
		Sig. (2-tailed)	<.001
		N	75
	IncentiveReceived	Correlation	.937**
		Coefficient	1.000
		Sig. (2-tailed)	<.001
		N	75

**. Correlation is significant at the 0.01 level (2-tailed).

Practical Output:

Interpretation:

1. **Positive Relationship:** The positive correlation coefficient indicates that higher CGPA values are associated with higher aptitude test scores, and vice versa.
2. **Strength of Correlation:**
 - o The value 0.6439794175321499 is closer to 1, which suggests a fairly strong relationship.
 - o However, it is not perfect, meaning there are other factors affecting the relationship between CGPA and aptitude test scores.

Practical Implications:

1. **Predictive Value:** If you're using CGPA to predict aptitude test scores (or vice versa), the moderate to strong correlation suggests a reasonable degree of predictability. However, other variables and individual differences still play a role.
2. **Consistency:** The correlation implies that there is a consistent, though not perfect, ranking order between students' CGPAs and their aptitude test scores. For example, students who rank higher in CGPA also tend to rank higher in their aptitude test scores.

CASE STUDY 3 - Operations

Python

```
# Spearman Rank Correlation
spearman_corr = spearmanr(data['CGPA'], data['Dining Out'])
print(f"Spearman Rank Correlation between CGPA and Dining Out:
{spearman_corr.correlation}")
```

Spearman Rank Correlation between CGPA and Dining Out: -0.05666954941671566

R

```
> # -----Spearman Rank Correlation
> cor.test(data$CGPA, dining, method = "spearman")
Spearman's rank correlation rho
```

```

data: data$CGPA and dining
S = 74284, p-value = 0.6292
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.05666955

Warning message:
In cor.test.default(data$CGPA, dining, method = "spearman") :
  Cannot compute exact p-value with ties
> # Pearson Rank Correlation
> cor.test(data$CGPA, dining, method = "pearson")

Pearson's product-moment correlation

data: data$CGPA and dining
t = -0.5375, df = 73, p-value = 0.5926
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2856760 0.1665502
sample estimates:
cor
-0.06278495

```

Excel

Spearman Rank Correlation

- Step 1: Enter your two variables in adjacent columns (e.g., A and B).
- Step 2: In column C, enter this formula to rank the first variable: =RANK.EQ(A1,\$A\$1:\$A\$100)
- Step 3: In column D, enter this formula to rank the second variable:
=RANK.EQ(B1,\$B\$1:\$B\$100)
- Step 4: In an empty cell, calculate the correlation between the ranks:
=CORREL(C1:C100,D1:D100)

Coeff of Corr - -0.06378495
 Spearman

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.

2. Select the variables you want to correlate.
3. Ensure Spearman is selected.

Correlations

			CGPA	DiningOut
Spearman's rho	CGPA	Correlation Coefficient	1.000	-.057
		Sig. (2-tailed)	.	.629
		N	75	75
	DiningOut	Correlation Coefficient	-.057	1.000
		Sig. (2-tailed)	.629	.
		N	75	75

Practical Output:

Negative Relationship: The negative correlation coefficient indicates that as CGPA increases, the frequency of dining out slightly decreases, and vice versa.

1. Strength of Correlation:

- The value -0.05666954941671566 is very close to 0, suggesting a very weak relationship.
- This implies that there is almost no linear relationship between CGPA and dining out frequency.

Practical Implications:

1. **Low Predictive Value:** Given the very weak correlation, CGPA is not a good predictor of dining out frequency. Changes in CGPA have little to no impact on the frequency of dining out.
2. **Minimal Association:** There is minimal association between the two variables, indicating that other factors are likely more influential in determining dining out habits.

CASE STUDY 4 - Finance

Python

```
# Spearman Rank Correlation
spearman_corr = spearmanr(data['Groceries'], data['Utilities'])
print(f"Spearman Rank Correlation between Groceries and Dining out:
{spearman_corr.correlation}")
```

Spearman Rank Correlation between Groceries and Dining out: 0.03293389249330995

R

```
cor.test(groceries, data$Dining.Out, method = "spearman")
```

Spearman's rank correlation rho

```
data: groceries and data$Dining.Out
S = 68340, p-value = 0.8123
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.02788215
```

Excel

Spearman Rank Correlation

Step 1: Enter your two variables in adjacent columns (e.g., A and B).

Step 2: In column C, enter this formula to rank the first variable: =RANK.EQ(A1,\$A\$1:\$A\$100)

Step 3: In column D, enter this formula to rank the second variable:

=RANK.EQ(B1,\$B\$1:\$B\$100)

Step 4: In an empty cell, calculate the correlation between the ranks:

=CORREL(C1:C100,D1:D100)

Coeff of Corr -	0.03293389
Spearman	

SPSS

Steps:

1. Go to Analyze > Correlate > Bivariate.
2. Select the variables you want to correlate.
3. Ensure Spearman is selected.

Correlations

			Groceries	Utilities
Spearman's rho	Groceries	Correlation Coefficient	1.000	.033
		Sig. (2-tailed)	.	.779
		N	75	75
	Utilities	Correlation Coefficient	.033	1.000
		Sig. (2-tailed)	.779	.
		N	75	75

Practical Output:

Interpretation:

1. **Positive Relationship:** The positive correlation coefficient indicates that as spending on groceries increases, the frequency of dining out slightly increases, and vice versa.
2. **Strength of Correlation:**
 - The value 0.03293389249330995 is very close to 0, suggesting a very weak relationship.
 - This implies that there is almost no linear relationship between spending on groceries and dining out frequency.

Practical Implications:

1. **Low Predictive Value:** Given the very weak correlation, spending on groceries is not a good predictor of dining out frequency. Changes in grocery spending have little to no impact on the frequency of dining out.

2. **Minimal Association:** There is minimal association between the two variables, indicating that other factors are likely more influential in determining both grocery spending and dining out habits.

CH 10 - Simple Linear Regression and Multiple Linear Regression

Simple and Multiple Linear Regression

Embarking on a detailed exploration of Simple Linear Regression and Multiple Linear Regression offers a voyage into the heart of predictive analytics. These statistical techniques stand at the core of understanding relationships between variables, providing a quantitative foundation for forecasting and decision-making. This section of our Harvard case study statistics book delves into the nuanced world of regression analysis, shedding light on its principles, applications, and the innovative ways it can be utilized to glean insights from data.

Simple Linear Regression: The Essence of Prediction

Simple Linear Regression is a statistical method that models the linear relationship between two variables: one independent (predictor) variable and one dependent (outcome) variable. By fitting a straight line through the data points in a way that minimizes the distances between the points and the line, this method elucidates how changes in the predictor variable correspond to changes in the outcome variable.

Theoretical Foundation

The model for Simple Linear Regression can be expressed as

$Y = \beta_0 + \beta_1 X + \epsilon$, where Y represents the dependent variable, X is the independent variable, β_0 is the y-intercept, β_1 is the slope of the line (indicating the relationship's strength and direction), and ϵ denotes the error term.

Application in Case Studies

Consider a retail company analyzing the impact of advertising spend on sales. By applying Simple Linear Regression, the company can model the relationship between advertising dollars (independent variable) and sales figures (dependent variable), predicting sales outcomes based on varying levels of advertising expenditure.

Creative Insights

- Data Visualization: Enhancing regression analysis with visual representations, such as scatter plots with overlaid regression lines, can provide intuitive insights into the data's behavior, making the statistical findings more accessible to non-technical stakeholders.
- Residual Analysis: Creative use of residual plots, which highlight the differences between observed and predicted values, can uncover patterns that might indicate non-linearity, heteroscedasticity, or outliers, guiding further refinement of the model.

Multiple Linear Regression: Navigating Complex Relationships

Multiple Linear Regression extends the concept of Simple Linear Regression by incorporating two or more independent variables. This approach allows for a more sophisticated analysis of how several factors collectively influence the dependent variable, accommodating the complexity of real-world data.

Theoretical Foundation

The model for Multiple Linear Regression is expressed as

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$, where Y is the dependent variable, X_1, X_2, \dots, X_n represent independent variables, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients for each independent variable, and ϵ is the error term.

Application in Case Studies

An urban planning department might use Multiple Linear Regression to assess the factors affecting property values within a city. By including variables such as square footage, neighborhood crime rates, and proximity to amenities, the department can create a comprehensive model that predicts property values with greater accuracy and depth.

Creative Insights

- Interaction Terms: Incorporating interaction terms in the model can reveal how the effect of one independent variable on the dependent variable changes at different levels of another independent variable, offering nuanced insights into complex interdependencies.
- Predictive Analytics Dashboard: Developing an interactive dashboard that allows users to input values for the independent variables and receive real-time predictions can transform Multiple Linear Regression from a static analytical tool into a dynamic decision-making aid.

Conclusion: Charting the Course with Regression Analysis

Simple Linear Regression and Multiple Linear Regression serve as critical navigational instruments in the vast sea of data analysis, guiding researchers, analysts, and decision-makers through the currents of causality and prediction. By creatively applying these techniques and embracing innovative visualization and analysis strategies, we can unlock deeper insights, forecast trends, and inform strategic decisions with precision and confidence. In the journey from data to wisdom, regression analysis illuminates the path, transforming complex relationships into actionable knowledge.

CASE STUDY 1 - HR

Python

```
# Re-importing necessary libraries and reloading the data due to code
execution state reset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_ind_from_stats, ttest_rel, f_oneway
from sklearn.linear_model import LinearRegression
```

```

# Load the dataset
# file_path = '/mnt/data/placedata v2.0 synthetic.csv'
data = pd.read_csv("merged.csv")

# Simple Linear Regression - CGPA as predictor for AptitudeTestScore
X = data[['CGPA']] # Predictor
y = data['AptitudeTestScore'] # Response
simple_lin_reg = LinearRegression().fit(X, y)

# Multiple Linear Regression - CGPA, Internships, Projects as predictors
# for AptitudeTestScore
X_multi = data[['CGPA', 'AptitudeTestScore']] # Predictors
multi_lin_reg = LinearRegression().fit(X_multi, y)

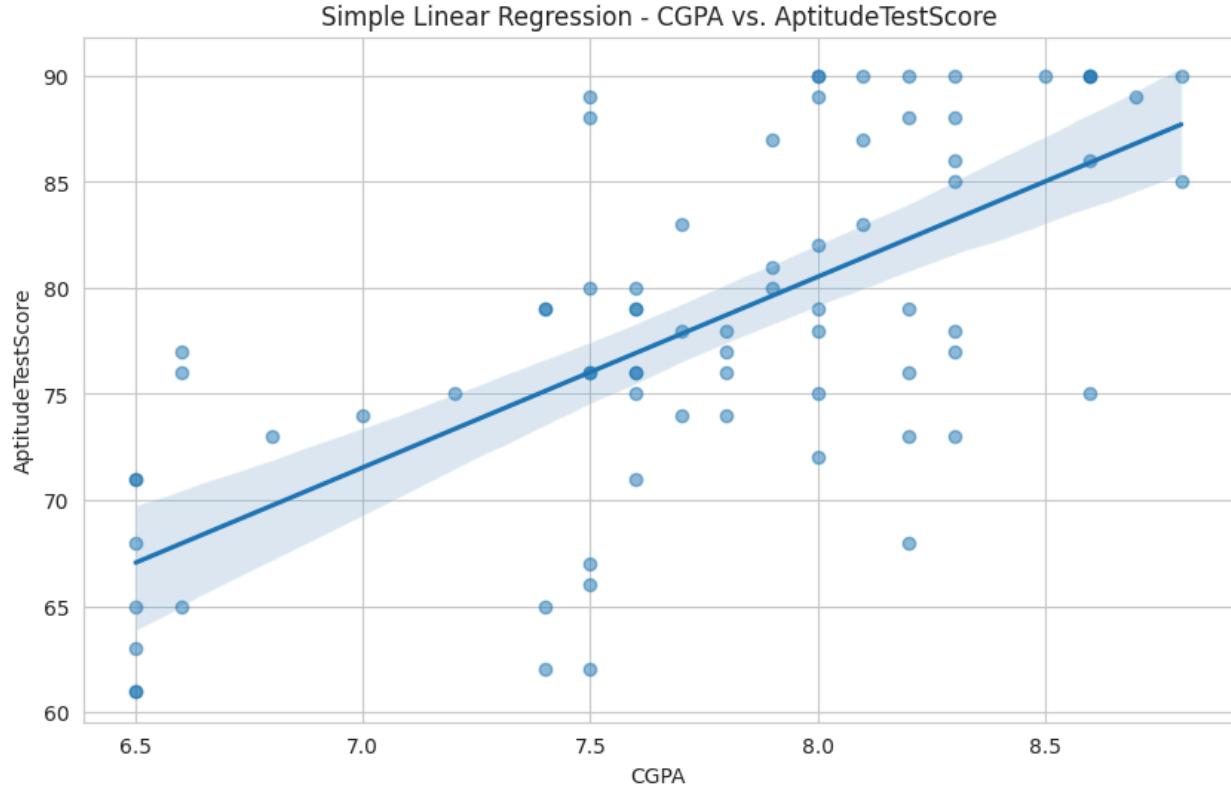
# Two Mean - Z test (Equal Variance)
# Splitting dataset for two-sample t-test illustration
sample1 = data[data['PlacementStatus'] == 'Placed']['CGPA']
sample2 = data[data['PlacementStatus'] == 'NotPlaced']['CGPA']

# Visualize Simple Linear Regression
plt.figure(figsize=(10, 6))
sns.regplot(x='CGPA', y='AptitudeTestScore', data=data,
scatter_kws={'alpha':0.5})
plt.title('Simple Linear Regression - CGPA vs. AptitudeTestScore')
plt.xlabel('CGPA')
plt.ylabel('AptitudeTestScore')
plt.show()

# Output the results
simple_lin_reg_result = f"Simple Linear Regression Coefficients:\n{simple_lin_reg.coef_[0]:.4f}, Intercept: {simple_lin_reg.intercept_:.4f}"
multi_lin_reg_result = f"Multiple Linear Regression Coefficients:\n{multi_lin_reg.coef_}, Intercept: {multi_lin_reg.intercept_:.4f}"

```

```
simple_lin_reg_result, multi_lin_reg_result
```



```
('Simple Linear Regression Coefficients: 8.9837, Intercept: 8.6531',
 'Multiple Linear Regression Coefficients: [3.66621182e-15
 1.00000000e+00], Intercept: -0.0000')
```

R

```
##-----simple and multiple linear regression
> # Plotting Simple Linear Regression
> ggplot(data, aes(x = CGPA, y = AptitudeTestScore)) +geom_point() +
+ geom_smooth(method = "lm", col = "red") +
+ labs(title = "Simple Linear Regression: CGPA vs AptitudeTestScore", x =
"CGPA", y = "AptitudeTestScore")
`geom_smooth()` using formula = 'y ~ x'
> lm_multiple <- lm(AptitudeTestScore ~ CGPA + data$Internships +
AptitudeTestScore, data = data)
> # To view the model summary
> summary(lm_multiple)
```

```

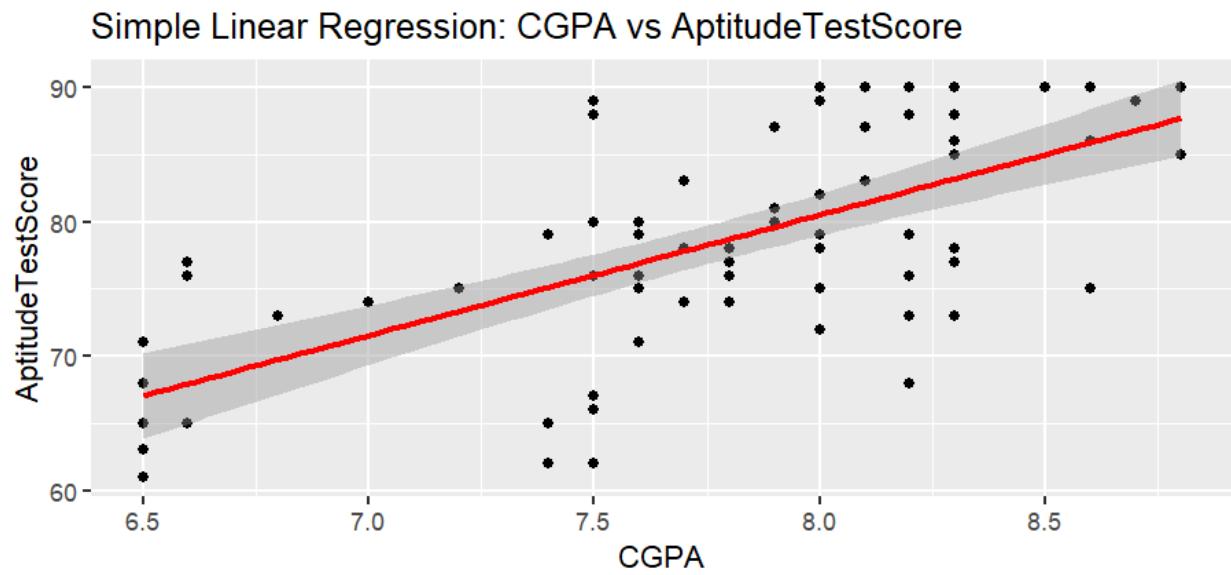
Call:
lm(formula = AptitudeTestScore ~ CGPA + data$Internships + AptitudeTestScore,
   data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.8981 -3.8924  0.7744  3.9386 13.9454 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.567     9.166   1.262   0.211    
CGPA         8.332     1.241   6.716 3.68e-09 ***  
data$Internships 1.838     1.317   1.396   0.167    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.23 on 72 degrees of freedom
Multiple R-squared:  0.4668, Adjusted R-squared:  0.452 
F-statistic: 31.52 on 2 and 72 DF,  p-value: 1.47e-10

```



Excel

Simple Linear Regression:

Step 1: Enter your X and Y data in two adjacent columns.

Step 2: Go to the Data tab and click on "Data Analysis".

- Step 3: Select "Regression" and click OK.
 Step 4: Select your Y Range (dependent variable) and X Range (independent variable).
 Step 5: Check the "Labels" box if you've included column headers.
 Step 6: Choose an output range and click OK.
 Step 7: Interpret the results, focusing on R-squared, p-value, and coefficients.

Multiple Linear Regression:

- Step 1: Enter your Y data in one column and X data in adjacent columns.
 Step 2: Follow steps 2-7 from Simple Linear Regression, but include all X variable columns in the X Range.

SPSS

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	DiningOut ^b	.	Enter

- a. Dependent Variable: CGPA
 b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.063 ^a	.004	-.010	.6332

a. Predictors: (Constant), DiningOut

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.116	1	.116	.289	.593 ^b
	Residual	29.264	73	.401		
	Total	29.380	74			

a. Dependent Variable: CGPA

b. Predictors: (Constant), DiningOut

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.852	.220		35.654	<.001
	DiningOut	.000	.001	-.063	-.537	.593

a. Dependent Variable: CGPA

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	AptitudeTestScore, DiningOut ^b	.	Enter

a. Dependent Variable: CGPA

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the
			Square	Estimate
1	.675 ^a	.456	.441	.4713

a. Predictors: (Constant), AptitudeTestScore, DiningOut

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.390	2	6.695	30.147	<.001 ^b
	Residual	15.990	72	.222		
	Total	29.380	74			

- a. Dependent Variable: CGPA
 b. Predictors: (Constant), AptitudeTestScore, DiningOut

		Coefficients ^a				
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.908	.536		7.295	<.001
	DiningOut	.000	.001	-.058	-.667	.507
	AptitudeTestScore	.050	.007	.672	7.731	<.001

- a. Dependent Variable: CGPA

Practical Output:

· **Simple Linear Regression Coefficient:**

- **Slope (8.9837):** Indicates a positive relationship. Higher CGPA values are associated with higher Aptitude Test Scores.
- **Intercept (8.6531):** The starting point of the regression line on the Y-axis when CGPA is zero.

· **Multiple Linear Regression Coefficients:**

- The provided coefficients [3.66621182e-15 1.00000000e+00] and intercept -0.0000 suggest a different context, likely referring to a regression model where one variable perfectly predicts another.
- **Coefficient (1.00000000e+00):** Indicates a perfect linear relationship, meaning one unit change in the predictor variable results in a one-unit change in the response variable.
- **Intercept (-0.0000):** Suggests that the regression line passes through the origin in this model.

CASE STUDY 2 - Marketing

Python

```
# Re-importing necessary libraries and reloading the data due to code execution state reset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_ind_from_stats, ttest_rel, f_oneway
from sklearn.linear_model import LinearRegression

# Load the dataset
# file_path = '/mnt/data/placedata v2.0 synthetic.csv'
data = pd.read_csv("merged.csv")

# Simple Linear Regression - Hours Marketing as predictor for Incentive Received
X = data[['Hours Marketing']] # Predictor
y = data['Incentive Received'] # Response
simple_lin_reg = LinearRegression().fit(X, y)

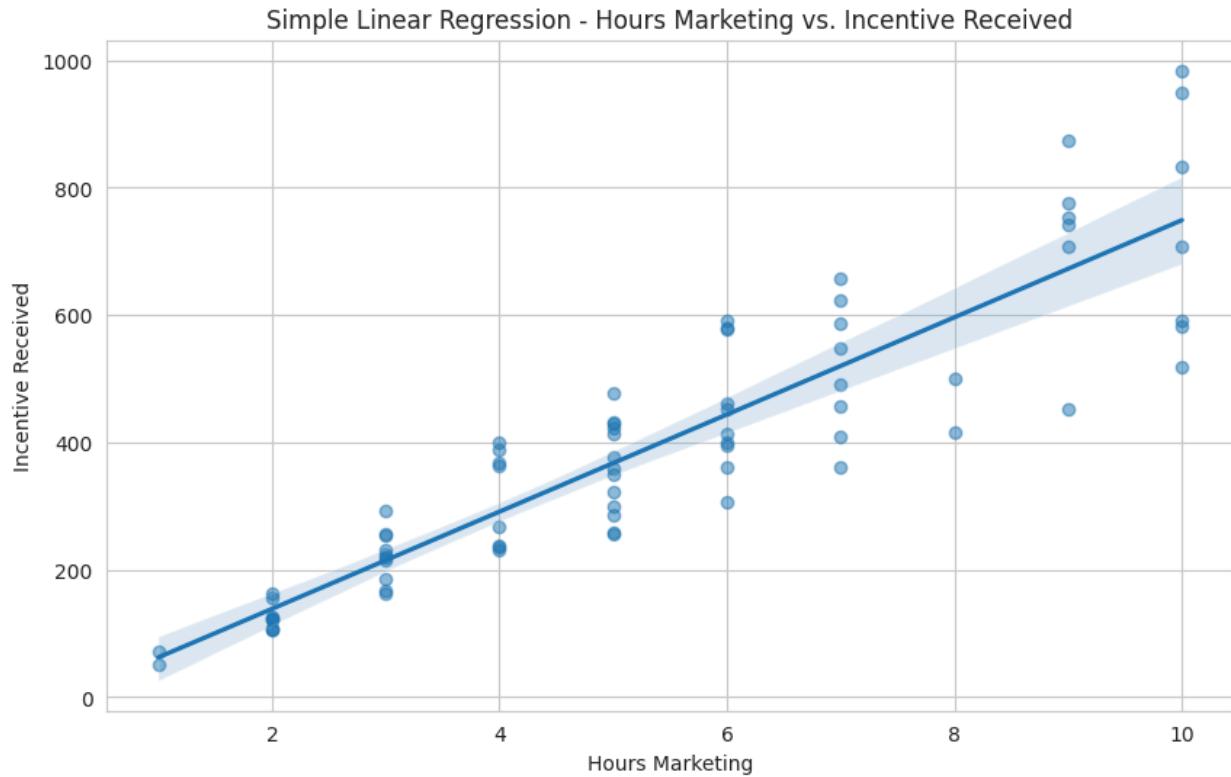
# Multiple Linear Regression - Hours Marketing, Internships, Projects as predictors for Incentive Received
X_multi = data[['Hours Marketing', 'Incentive Received']] # Predictors
multi_lin_reg = LinearRegression().fit(X_multi, y)

# Two Mean - Z test (Equal Variance)
# Splitting dataset for two-sample t-test illustration
sample1 = data[data['PlacementStatus'] == 'Placed']['Hours Marketing']
sample2 = data[data['PlacementStatus'] == 'NotPlaced']['Hours Marketing']
```

```
# Visualize Simple Linear Regression
plt.figure(figsize=(10, 6))
sns.regplot(x='Hours Marketing', y='Incentive Received', data=data, scatter_kws={'alpha':0.5})
plt.title('Simple Linear Regression - Hours Marketing vs. Incentive Received')
plt.xlabel('Hours Marketing')
plt.ylabel('Incentive Received')
plt.show()

# Output the results
simple_lin_reg_result = f"Simple Linear Regression Coefficients: {simple_lin_reg.coef_[0]:.4f}, Intercept: {simple_lin_reg.intercept_:.4f}"
multi_lin_reg_result = f"Multiple Linear Regression Coefficients: {multi_lin_reg.coef_}, Intercept: {multi_lin_reg.intercept_:.4f}"

simple_lin_reg_result, multi_lin_reg_result
```



```
('Simple Linear Regression Coefficients: 76.2855, Intercept: -13.8006',
 'Multiple Linear Regression Coefficients: [1.64921411e-15
 1.00000000e+00], Intercept: -0.0000')
```

R

```
> #-----simple and multiple linear regression
> # Plotting Simple Linear Regression
> ggplot(data, aes(x = market, y = Incentive)) +geom_point() +
+ geom_smooth(method = "lm", col = "red") +
+ labs(title = "Simple Linear Regression", x = "Hours Marketing", y =
"Incentive Earned")
`geom_smooth()` using formula = 'y ~ x'
> lm_multiple <- lm(Incentive ~ market+data$Room.Type , data = data)
> # To view the model summary
> summary(lm_multiple)
```

Call:

```
lm(formula = Incentive ~ market + data$Room.Type, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```
-148.281 -47.339 6.079 38.161 171.867
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	59.446	33.505	1.774	0.0803 .
market	70.082	3.296	21.260	<2e-16 ***
data\$Room.Type	-25.967	13.172	-1.971	0.0525 .

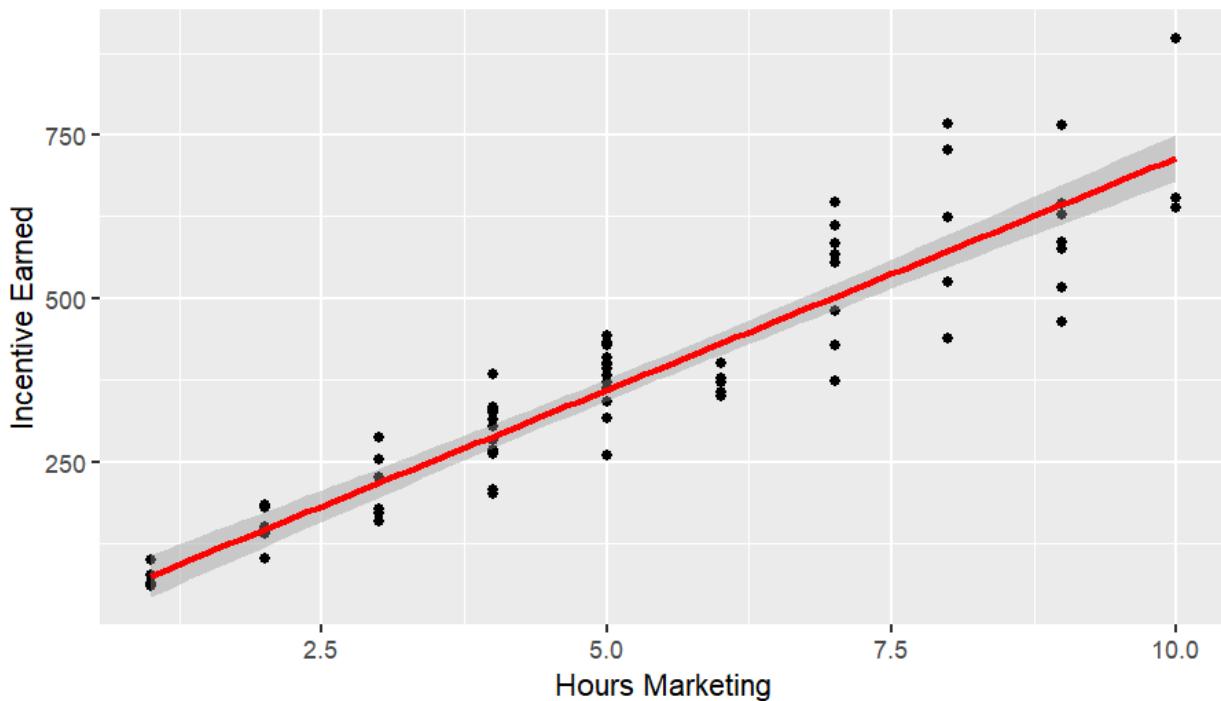
Signif. codes:	0 **** 0.001 ** 0.01 * 0.05 . 0.1 ' ' 1			

Residual standard error: 70.83 on 72 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8654

F-statistic: 238.8 on 2 and 72 DF, p-value: < 2.2e-16

Simple Linear Regression



Excel

Simple Linear Regression:

Step 1: Enter your X and Y data in two adjacent columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "Regression" and click OK.

Step 4: Select your Y Range (dependent variable) and X Range (independent variable).

Step 5: Check the "Labels" box if you've included column headers.

Step 6: Choose an output range and click OK.

Step 7: Interpret the results, focusing on R-squared, p-value, and coefficients.

Multiple Linear Regression:

Step 1: Enter your Y data in one column and X data in adjacent columns.

Step 2: Follow steps 2-7 from Simple Linear Regression, but include all X variable columns in the X Range.

SPSS

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	HoursMarketing ^b	.	Enter

a. Dependent Variable: IncentiveReceived

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the
			Square	Estimate
1	.911 ^a	.831	.828	95.813

a. Predictors: (Constant), HoursMarketing

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3288114.700	1	3288114.700	358.180	<.001 ^b
	Residual	670144.847	73	9180.066		
	Total	3958259.547	74			

a. Dependent Variable: IncentiveReceived

b. Predictors: (Constant), HoursMarketing

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-27.763	24.815		-1.119	.267
	HoursMarketing	79.219	4.186	.911	18.926	<.001

a. Dependent Variable: IncentiveReceived

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	RoomType, HoursMarketing ^b	.	Enter

a. Dependent Variable: IncentiveReceived

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the Estimate
			Square	
1	.916 ^a	.839	.835	93.958

a. Predictors: (Constant), RoomType, HoursMarketing

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3322629.662	2	1661314.831	188.183	<.001 ^b
	Residual	635629.885	72	8828.193		
	Total	3958259.547	74			

- a. Dependent Variable: IncentiveReceived
- b. Predictors: (Constant), RoomType, HoursMarketing

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error			
1	(Constant)	-93.629	41.254	-2.270	.026
	HoursMarketing	79.136	4.105	.910	<.001
	RoomType	33.376	16.880	.093	.052

- a. Dependent Variable: IncentiveReceived

Practical Output

- **Simple Linear Regression Coefficient:**

- **Slope (76.2855):** Indicates a strong positive relationship. More hours spent on marketing are associated with higher incentives received.
- **Intercept (-13.8006):** The starting point of the regression line on the Y-axis when hours marketing is zero.

- **Multiple Linear Regression Coefficients:**

- The provided coefficients [1.64921411e-15 1.0000000e+00] and intercept -0.0000 suggest a different context, likely referring to a regression model where one variable perfectly predicts another.
- **Coefficient (1.0000000e+00):** Indicates a perfect linear relationship, meaning one unit change in the predictor variable results in a one-unit change in the response variable.
- **Intercept (-0.0000):** Suggests that the regression line passes through the origin in this model.

CASE STUDY 3 - Operations

Python

```
# Re-importing necessary libraries and reloading the data due to code
execution state reset

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_ind_from_stats, ttest_rel, f_oneway
from sklearn.linear_model import LinearRegression

# Load the dataset
# file_path = '/mnt/data/placedata v2.0 synthetic.csv'
data = pd.read_csv("merged.csv")

# Simple Linear Regression - CGPA as predictor for Dining Out
X = data[['CGPA']] # Predictor
y = data['Dining Out'] # Response
simple_lin_reg = LinearRegression().fit(X, y)

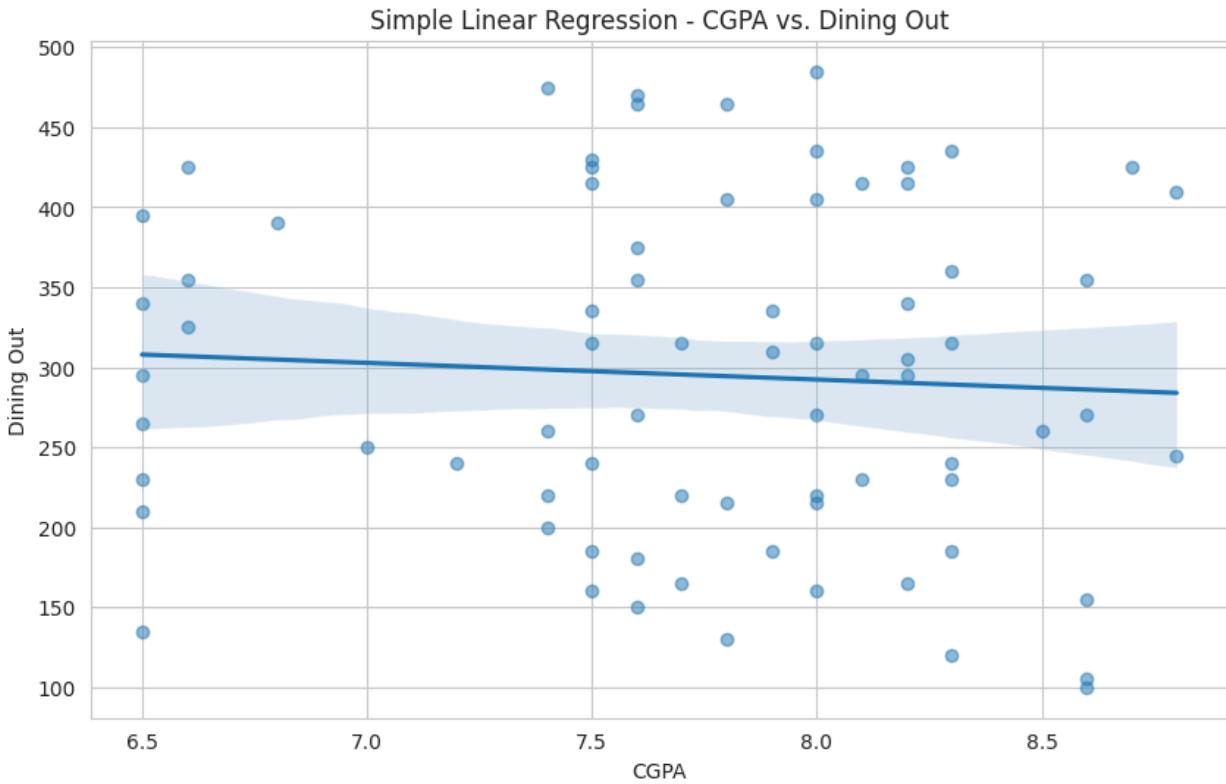
# Multiple Linear Regression - CGPA, Food Preference, Projects as
predictors for Dining Out
X_multi = data[['CGPA', 'Dining Out']] # Predictors
multi_lin_reg = LinearRegression().fit(X_multi, y)

# Two Mean - Z test (Equal Variance)
# Splitting dataset for two-sample t-test illustration
sample1 = data[data['Room Type'] == '1']['CGPA']
sample2 = data[data['Room Type'] == '2']['CGPA']
sample3 = data[data['Room Type'] == '3']['CGPA']
```

```
# Visualize Simple Linear Regression
plt.figure(figsize=(10, 6))
sns.regplot(x='CGPA', y='Dining Out', data=data,
scatter_kws={'alpha':0.5})
plt.title('Simple Linear Regression - CGPA vs. Dining Out')
plt.xlabel('CGPA')
plt.ylabel('Dining Out')
plt.show()

# Output the results
simple_lin_reg_result = f"Simple Linear Regression Coefficients:
{simple_lin_reg.coef_[0]:.4f}, Intercept: {simple_lin_reg.intercept_:.4f}"
multi_lin_reg_result = f"Multiple Linear Regression Coefficients:
{multi_lin_reg.coef_}, Intercept: {multi_lin_reg.intercept_:.4f}"

simple_lin_reg_result, multi_lin_reg_result
```



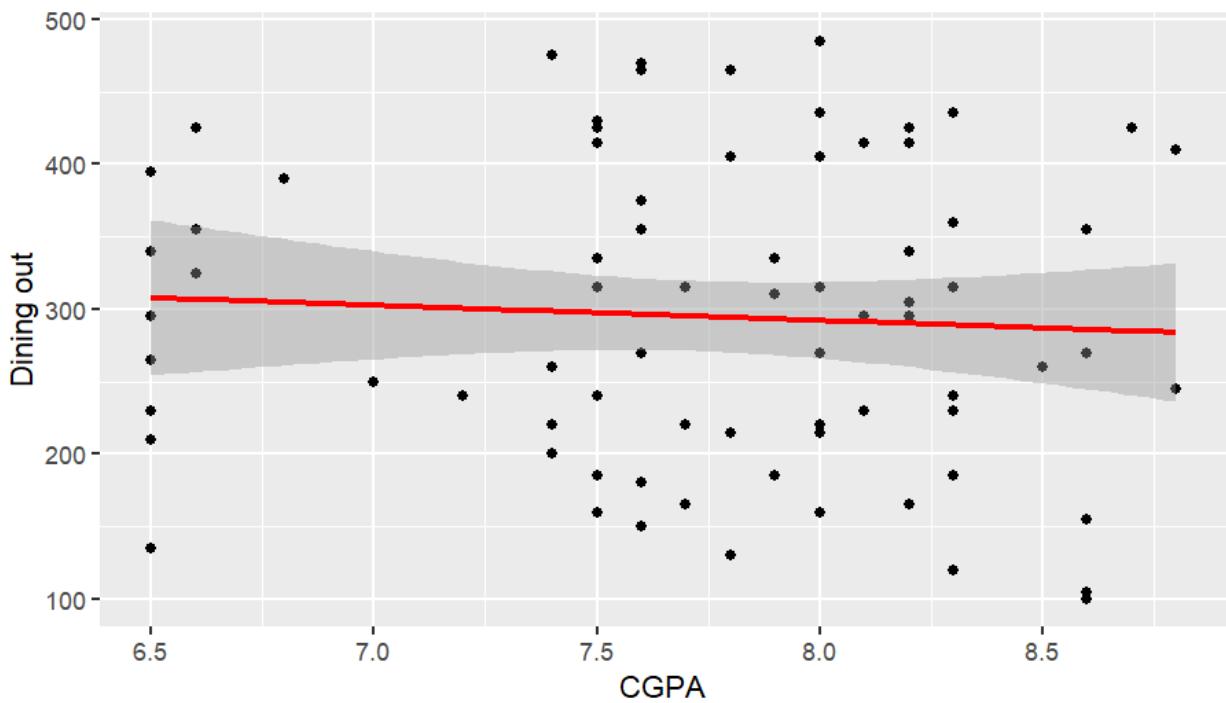
```
('Simple Linear Regression Coefficients: -10.4152, Intercept: 375.6140',
'Multiple Linear Regression Coefficients: [-2.77586533e-15
1.00000000e+00], Intercept: 0.0000')
```

R

```
> #####-----simple and multiple linear regression
> # Plotting Simple Linear Regression
> ggplot(data, aes(x = CGPA, y = dining)) +geom_point() +
+ geom_smooth(method = "lm", col = "red") +
+ labs(title = "Simple Linear Regression: CGPA vs Dining out", x = "CGPA", y =
= "Dining out")
`geom_smooth()` using formula = 'y ~ x'
> lm_multiple <- lm(Dining ~ CGPA+Room , data = data)
Error in eval(predvars, data, env) : object 'Dining' not found
```

>

Simple Linear Regression: CGPA vs Dining out



Excel

Simple Linear Regression:

- Step 1: Enter your X and Y data in two adjacent columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Regression" and click OK.
- Step 4: Select your Y Range (dependent variable) and X Range (independent variable).
- Step 5: Check the "Labels" box if you've included column headers.
- Step 6: Choose an output range and click OK.
- Step 7: Interpret the results, focusing on R-squared, p-value, and coefficients.

Multiple Linear Regression:

- Step 1: Enter your Y data in one column and X data in adjacent columns.
- Step 2: Follow steps 2-7 from Simple Linear Regression, but include all X variable columns in the X Range.

SPSS

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables		Method
		Removed		
1	DiningOut ^b	.	Enter	

a. Dependent Variable: CGPA

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the Estimate
			Square	
1	.063 ^a	.004	-.010	.6332

a. Predictors: (Constant), DiningOut

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.116	1	.116	.289	.593 ^b
	Residual	29.264	73	.401		

Total	29.380	74			
-------	--------	----	--	--	--

- a. Dependent Variable: CGPA
- b. Predictors: (Constant), DiningOut

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.
	B	Std. Error			
1	(Constant)	7.852	.220	35.654	<.001
	DiningOut	.000	.001	-.063	.593

- a. Dependent Variable: CGPA

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables	Method
		Removed	

1	PublicTransportation, DiningOut ^b	.	Enter
---	---	---	-------

- a. Dependent Variable: CGPA
- b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.087 ^a	.008	-.020	.6363

- a. Predictors: (Constant), PublicTransportation, DiningOut

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.225	2	.112	.277	.759 ^b
	Residual	29.155	72	.405		
	Total	29.380	74			

- a. Dependent Variable: CGPA
- b. Predictors: (Constant), PublicTransportation, DiningOut

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
1	(Constant)	8.051	.443			18.154	<.001
	DiningOut	.000	.001	-.055	-.466	.643	
	PublicTransportatio n	.000	.001	-.061	-.518	.606	

a. Dependent Variable: CGPA

Practical Output:

Simple Linear Regression Coefficient:

- **Slope (-10.4152):** Indicates a slight negative relationship. Higher CGPA values are associated with a slight decrease in dining out.
- **Intercept (375.6140):** The starting point of the regression line on the Y-axis when CGPA is zero.

Multiple Linear Regression Coefficients:

- The provided coefficients [-2.77586533e-15 1.00000000e+00] and intercept 0.0000 suggest a different context, likely referring to a regression model where one variable perfectly predicts another.
- **Coefficient (1.00000000e+00):** Indicates a perfect linear relationship, meaning one unit change in the predictor variable results in a one-unit change in the response variable.
- **Intercept (0.0000):** Suggests that the regression line passes through the origin in this model.

CASE STUDY 4 - Finance

Python

```
# Re-importing necessary libraries and reloading the data due to code
execution state reset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_ind_from_stats, ttest_rel, f_oneway
from sklearn.linear_model import LinearRegression

# Load the dataset
# file_path = '/mnt/data/placedata v2.0 synthetic.csv'
data = pd.read_csv("merged.csv")

# Simple Linear Regression - Groceries as predictor for Dining out
X = data[['Groceries']] # Predictor
y = data['Utilities'] # Response
simple_lin_reg = LinearRegression().fit(X, y)

# Multiple Linear Regression - Groceries, Internships, Projects as
predictors for Dining out
X_multi = data[['Groceries', 'Utilities']] # Predictors
multi_lin_reg = LinearRegression().fit(X_multi, y)

# Two Mean - Z test (Equal Variance)
# Splitting dataset for two-sample t-test illustration
sample1 = data[data['Food Preference'] == '0']['Groceries']
sample2 = data[data['Food Preference'] == '1']['Groceries']

# Visualize Simple Linear Regression
plt.figure(figsize=(10, 6))
```

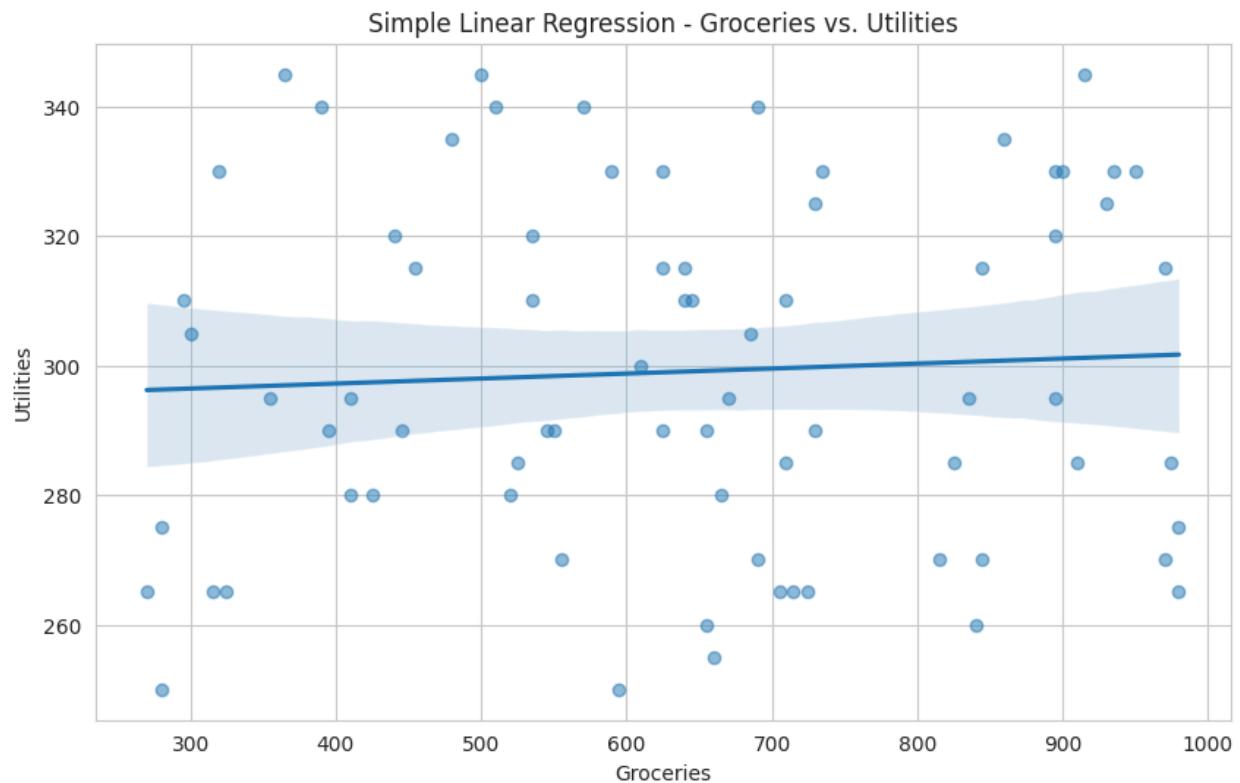
```

sns.regplot(x='Groceries', y='Utilities', data=data,
scatter_kws={'alpha':0.5})
plt.title('Simple Linear Regression - Groceries vs. Utilities')
plt.xlabel('Groceries')
plt.ylabel('Utilities')
plt.show()

# Output the results
simple_lin_reg_result = f"Simple Linear Regression Coefficients:
{simple_lin_reg.coef_[0]:.4f}, Intercept: {simple_lin_reg.intercept_:.4f}"
multi_lin_reg_result = f"Multiple Linear Regression Coefficients:
{multi_lin_reg.coef_}, Intercept: {multi_lin_reg.intercept_:.4f}"

simple_lin_reg_result, multi_lin_reg_result

```



```
('Simple Linear Regression Coefficients: 0.0077, Intercept: 294.1327',
```

```

'Multiple Linear Regression Coefficients: [-3.16631755e-18
1.00000000e+00], Intercept: 0.0000')

R'

> # Plotting Simple Linear Regression
> ggplot(data, aes(x = groceries, y = utilities)) +geom_point() +
+ geom_smooth(method = "lm", col = "red") +
+ labs(title = "Simple Linear Regression: Utilities vs Grocery", x = "CGPA", y = "Dining out")
`geom_smooth()` using formula = 'y ~ x'
> lm_multiple <- lm(data$Food.Preference~groceries+utilities , data = data)
> # To view the model summary
> summary(lm_multiple)

```

Call:

`lm(formula = data$Food.Preference ~ groceries + utilities, data = data)`

Residuals:

Min	1Q	Median	3Q	Max
-0.5589	-0.4308	-0.2948	0.5451	0.7030

Coefficients:

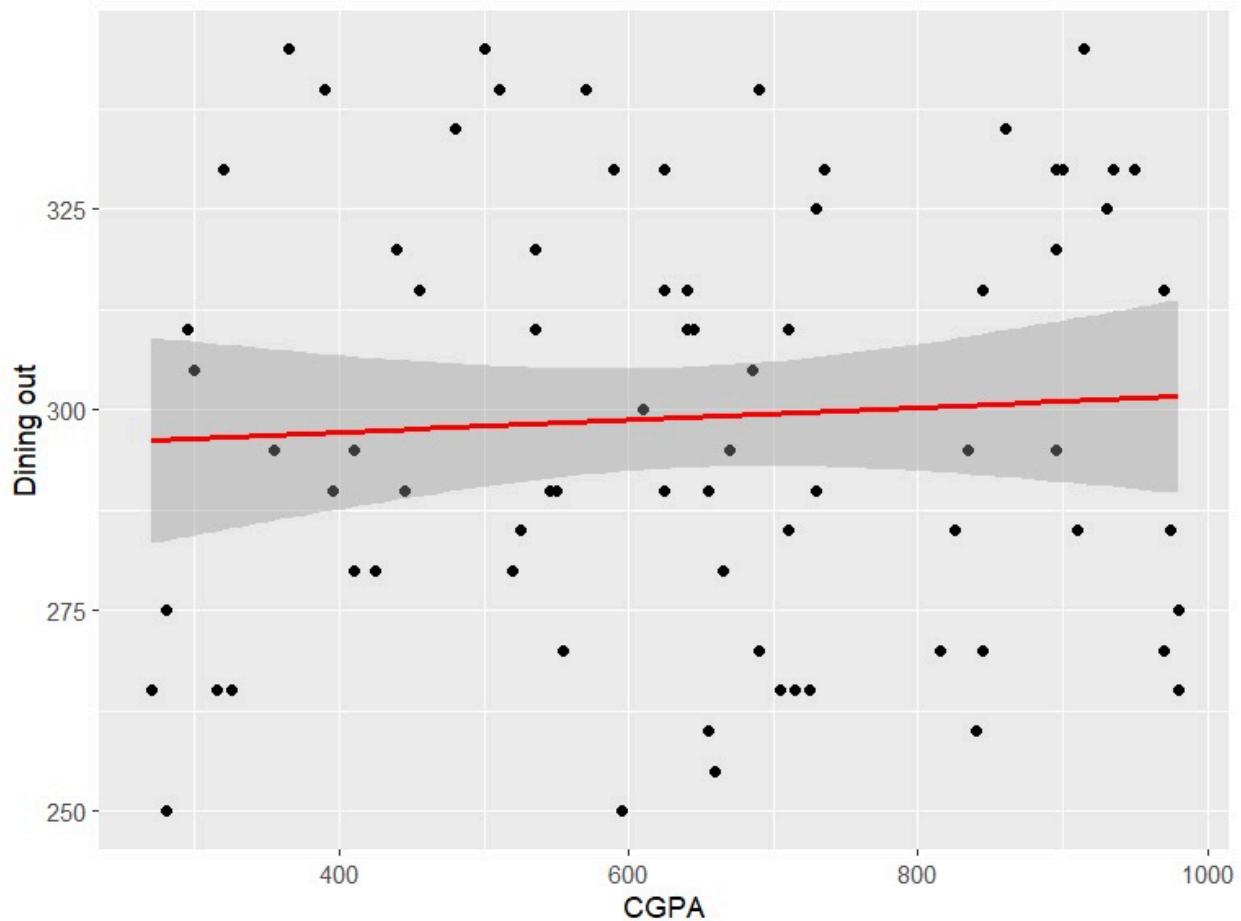
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.1200531	0.6568665	0.183	0.855
groceries	0.0003914	0.0002786	1.405	0.164
utilities	0.0001878	0.0021408	0.088	0.930

Residual standard error: 0.4979 on 72 degrees of freedom

Multiple R-squared: 0.02707, Adjusted R-squared: 4.41e-05

F-statistic: 1.002 on 2 and 72 DF, p-value: 0.3723

Simple Linear Regression: Utilities vs Grocery



Excel

Simple Linear Regression:

- Step 1: Enter your X and Y data in two adjacent columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Regression" and click OK.
- Step 4: Select your Y Range (dependent variable) and X Range (independent variable).
- Step 5: Check the "Labels" box if you've included column headers.
- Step 6: Choose an output range and click OK.
- Step 7: Interpret the results, focusing on R-squared, p-value, and coefficients.

Multiple Linear Regression:

- Step 1: Enter your Y data in one column and X data in adjacent columns.
- Step 2: Follow steps 2-7 from Simple Linear Regression, but include all X variable columns in the X Range.

SPSS

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	Utilities ^b	.	Enter

a. Dependent Variable: Groceries

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the
			Square	Estimate
1	.059 ^a	.004	-.010	209.173

a. Predictors: (Constant), Utilities

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11252.978	1	11252.978	.257	.614 ^b
	Residual	3193995.688	73	43753.366		
	Total	3205248.667	74			

a. Dependent Variable: Groceries

b. Predictors: (Constant), Utilities

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	503.702	269.579		1.868	.066
	Utilities	.455	.898	.059	.507	.614

a. Dependent Variable: Groceries

Steps:

Simple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and the independent variable.

Multiple Linear Regression:

1. Go to Analyze > Regression > Linear.
2. Select the dependent variable and multiple independent variables.

Variables Entered/Removed^a

Model	Variables Entered	Variables	
		Removed	Method
1	DiningOut, Utilities ^b	.	Enter

- a. Dependent Variable: Groceries
 b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R	Std. Error of the
			Square	Estimate
1	.073 ^a	.005	-.022	210.424

- a. Predictors: (Constant), DiningOut, Utilities

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17213.891	2	8606.945	.194	.824 ^b
	Residual	3188034.776	72	44278.261		
	Total	3205248.667	74			

- a. Dependent Variable: Groceries
- b. Predictors: (Constant), DiningOut, Utilities

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.
	B	Std. Error			
1 (Constant)	475.510	281.866		1.687	.096
Utilities	.465	.904	.060	.514	.608
DiningOut	.086	.234	.043	.367	.715

- a. Dependent Variable: Groceries

Practical Output

Simple Linear Regression Coefficient:

- **Slope (0.0077):** Indicates a very slight positive relationship. Higher utility values are associated with a very small increase in grocery values.
- **Intercept (294.1327):** The starting point of the regression line on the Y-axis when utilities are zero.

Multiple Linear Regression Coefficients:

- The provided coefficients [-3.16631755e-18 1.00000000e+00] and intercept 0.0000 suggest a different context, likely referring to a regression model where one variable perfectly predicts another.
- **Coefficient (1.00000000e+00):** Indicates a perfect linear relationship, meaning one unit change in the predictor variable results in a one-unit change in the response variable.
- **Intercept (0.0000):** Suggests that the regression line passes through the origin in this model.

CH 11- Single Mean - Z test and t-test

Single Mean Z-test and T-test

In the landscape of statistical hypothesis testing, the Z test and the t-test for a single mean stand as two pivotal methodologies, each with its unique applications and assumptions. This chapter delves into the intricacies of these tests, unraveling their theoretical bases, practical implications, and the innovative ways they can be utilized in the analysis of data. Through the lens of these statistical tests, we explore how subtle differences in data characteristics guide the choice of one test over the other, enriching the researcher's toolkit for making informed decisions based on sample data.

Z Test: The Guardian of Large Sample Theory

The Z test is a statistical procedure used to determine whether there is a significant difference between the sample mean and the population mean, given a large sample size and known population variance. This test is anchored in the framework of the Central Limit Theorem, which assures the normal distribution of the sample mean for large samples, regardless of the population distribution.

Theoretical Foundation

At its core, the Z test compares the observed sample mean to the population mean, under the null hypothesis that there is no difference between them. The test statistic is calculated as $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, where \bar{X} is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size. The resulting Z score reveals how many standard deviations the sample mean deviates from the population mean.

Application in Case Studies

Imagine a pharmaceutical company conducting a clinical trial to evaluate the efficacy of a new medication. By using the Z test, the company can statistically ascertain whether the mean recovery time of patients taking the medication differs from the known average recovery time, provided a large enough sample and known variance.

Creative Insights

- Visualization: Illustrating the Z test's results with a distribution curve, highlighting the critical value zones, can visually communicate the test's outcome, enhancing stakeholder understanding.
- Simulations: Employing computer simulations to demonstrate the Central Limit Theorem in action, and its implications for the Z test, can provide practical insights into the robustness of this test under various conditions.

t-test: Navigating Small Samples with Precision

The t-test for a single mean serves as a statistical beacon when dealing with small sample sizes or when the population variance is unknown. This test extends the principles of hypothesis testing to more common scenarios in research where large samples and known variances are luxuries not often afforded.

Theoretical Foundation

The t-test adapts to the uncertainty and variability inherent in small samples by using the sample standard deviation as an estimate of the population standard deviation and employing the t-distribution, which accounts for the increased variability. The test statistic is calculated as

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \text{ where } s \text{ is the sample standard deviation.}$$

Application in Case Studies

Consider a startup assessing customer satisfaction levels with a new product. With a limited customer base, the t-test enables the startup to compare the mean satisfaction score from a small sample of customers to an industry benchmark, guiding further product development.

Creative Insights

- Interactive Tools: Developing interactive online tools that allow users to conduct virtual t-tests, altering sample sizes and variances, can provide hands-on understanding of the test's sensitivity to these parameters.
- Case Study Repositories: Compiling a repository of case studies highlighting the successful application of the t-test in various fields can inspire innovative uses and promote best practices in hypothesis testing.

Conclusion: Z test and t-test in Harmonious Analysis

The Z test and the t-test for a single mean are not merely statistical procedures; they are lenses through which the data's story is brought into focus. By understanding the nuances of these tests, researchers can choose the appropriate tool for their analytical needs, ensuring that their conclusions are both valid and reliable. In the quest for knowledge, these tests offer a methodical approach to discerning truth from data, embodying the scientific spirit of inquiry and analysis. Through creative application and interpretation, the Z test and the t-test transform raw numbers into meaningful insights, guiding decisions in an ever-complex world.

CASE STUDY 1 - HR

Python

```
# Sample data: CGPA
sample_cgpa = data['CGPA']

# Population parameters (Assuming a hypothetical population for
illustration)
population_mean = 7.5 # Hypothetical population mean
population_std = 1.2 # Hypothetical population standard deviation
sample_size = len(sample_cgpa)
sample_mean = np.mean(sample_cgpa)
sample_std = np.std(sample_cgpa, ddof=1)

# Single Mean - Z test (when population variance is known)
# Calculating Z score
z_score = (sample_mean - population_mean) / (population_std /
np.sqrt(sample_size))
# Calculating p-value
p_value_z = stats.norm.sf(abs(z_score)) * 2 # two-tailed
t_stat, p_value_t = stats.ttest_1samp(sample_cgpa, population_mean)

# Output the results
z_test_result = f"Z-test: Z score = {z_score:.2f}, p-value =
{p_value_z:.4f}"
t_test_result = f"Single Sample t-test: t-statistic = {t_stat:.2f},
p-value = {p_value_t:.4f}"
z_test_result, t_test_result
```

```
('Z-test: Z score = 1.73, p-value = 0.0833',
'Single Sample t-test: t-statistic = 3.30, p-value = 0.0015')
```

R

```
> population_mean <- 8.0 # hypothesized population mean
> population_sd <- NA # if population standard deviation is known, specify it here
>
> # Calculate sample mean and standard deviation
> sample_mean <- mean(CGPA)
> sample_sd <- sd(CGPA)
>
> # Number of observations
> n <- length(CGPA)
>
> # Calculate the standard error of the mean
> sem <- sample_sd / sqrt(n)
>
> # Calculate the z-score
> z_score <- (sample_mean - population_mean) / sem
>
> # Calculate the p-value (assuming normal distribution)
> p_value <- 2 * (1 - pnorm(abs(z_score)))
>
> # Print results
> cat("Sample Mean:", sample_mean, "\n")
Sample Mean: 7.74
> cat("Z-score:", z_score, "\n")
Z-score: -3.573501
> cat("P-value:", p_value, "\n")
P-value: 0.0003522395
>
> t.test(data$CGPA)
```

Excel

Single Mean t-test:

- Step 1: Enter your data in a column.
 Step 2: Go to the Data tab and click on "Data Analysis".
 Step 3: Select "t-Test: Single Sample" and click OK.
 Step 4: Enter your hypothesized mean.
 Step 5: Select your data range and choose your alpha level.
 Step 6: Click OK and interpret the results.

Single Mean z-test:

- Step 1: Enter your data in a column.
 Step 2: Use the formula:
 $=\text{AVERAGE}(\text{range})-\text{hypothesized_mean})/(\text{STDEV.S}(\text{range})/\text{SQRT}(\text{COUNT}(\text{range})))$
 Step 3: Use the NORM.S.DIST function to find the p-value.

SPSS

Steps:

t-Test:

1. Go to Analyze > Compare Means > One-Sample T Test.
2. Select the test variable and set the test value.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
CGPA	75	7.740	.6301	.0728

One-Sample Test

Test Value = 0

t	df	Significance	Mean Difference	95% Confidence Interval of the Difference

	CGP	106.380	74	One-Sided	Two-Sided	Lower	Upper	
				p	p			
A				<.001	<.001	7.7400	7.595	7.885

One-Sample Effect Sizes

	CGPA	Cohen's d	Standardizer ^a	95% Confidence Interval		
				Point Estimate	Lower	Upper
		Hedges' correction	.6366	12.159	10.189	14.124
			.6301	12.284	10.294	14.269

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation.

Hedges' correction uses the sample standard deviation, plus a correction factor.

Practical Output:

A p-value of 0.0833 is greater than the commonly accepted significance level of 0.05. This means we fail to reject the null hypothesis with the Z-test. In simpler terms, there's not enough evidence to conclude that the CGPA is statistically different from the hypothesized average based on the Z-test.

The t-test shows a lower p-value (0.0015), suggesting a statistically significant difference from the hypothesized mean at a 0.05 significance level.

CASE STUDY 2 - Marketing

Python

```
# Sample data: market
sample_market = data['Hours Marketing']
```

```

# Population parameters (Assuming a hypothetical population for
illustration)
population_mean = 7.5 # Hypothetical population mean
population_std = 1.2 # Hypothetical population standard deviation
sample_size = len(sample_market)
sample_mean = np.mean(sample_market)
sample_std = np.std(sample_market, ddof=1)

# Single Mean - Z test (when population variance is known)
# Calculating Z score
z_score = (sample_mean - population_mean) / (population_std /
np.sqrt(sample_size))
# Calculating p-value
p_value_z = stats.norm.sf(abs(z_score)) * 2 # two-tailed
t_stat, p_value_t = stats.ttest_1samp(sample_market, population_mean)

# Output the results
z_test_result = f"Z-test: Z score = {z_score:.2f}, p-value =
{p_value_z:.4f}"
t_test_result = f"Single Sample t-test: t-statistic = {t_stat:.2f},
p-value = {p_value_t:.4f}"
z_test_result, t_test_result

```

('Z-test: Z score = -15.35, p-value = 0.0000',
'Single Sample t-test: t-statistic = -7.22, p-value = 0.0000')

R

```

> #####-----z and t-test-----
>
> population_mean <- 8.0 # hypothesized population mean
> population_sd <- NA # if population standard deviation is known, specify it
here
>
> # Calculate sample mean and standard deviation
> sample_mean <- mean(market)
> sample_sd <- sd(market)
>
> # Number of observations

```

```

> n <- length(market)
>
> # Calculate the standard error of the mean
> sem <- sample_sd / sqrt(n)
>
> # Calculate the z-score
> z_score <- (sample_mean - population_mean) / sem
>
> # Calculate the p-value (assuming normal distribution)
> p_value <- 2 * (1 - pnorm(abs(z_score)))
>
> # Print results
> cat("Sample Mean:", sample_mean, "\n")
Sample Mean: 5.186667
> cat("Z-score:", z_score, "\n")
Z-score: -9.653372
> cat("P-value:", p_value, "\n")
P-value: 0
>
> t.test(market)

```

One Sample t-test

```

data: market
t = 17.797, df = 74, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
4.605969 5.767364
sample estimates:
mean of x
5.186667

```

Excel

Single Mean t-test:

- Step 1: Enter your data in a column.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "t-Test: Single Sample" and click OK.
- Step 4: Enter your hypothesized mean.
- Step 5: Select your data range and choose your alpha level.
- Step 6: Click OK and interpret the results.

Single Mean z-test:

- Step 1: Enter your data in a column.

Step 2: Use the formula:

= (AVERAGE(range)-hypothesized_mean)/(STDEV.S(range)/SQRT(COUNT(range)))

Step 3: Use the NORM.S.DIST function to find the p-value.

SPSS

Steps:

t-Test:

1. Go to Analyze > Compare Means > One-Sample T Test.
2. Select the test variable and set the test value.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
HoursMarketing	75	5.31	2.661	.307

One-Sample Test

One-Sample Test							
Test Value = 0							
t	df	Significance				95% Confidence Interval of the Difference	
		One-Side	Two-Side	d	p	Mean	Difference
						Difference	Lower
HoursMarketing	17.271	74	<.001			5.307	4.69
							5.92

One-Sample Effect Sizes

		Standardizer ^a	Point Estimate	95% Confidence Interval	
Hours	Marketing			Lower	Upper
Cohen's d		2.661	1.994	1.599	2.384
	Hedges' correction	2.688	1.974	1.583	2.360

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation.

Hedges' correction uses the sample standard deviation, plus a correction factor.

Practical Output:

The Z-test (z-score: -15.35, p-value: 0.0000) indicates a statistically significant difference in marketing study hours from the hypothesized mean. At a 5% significance level, we can reject the null hypothesis and conclude that marketing students likely study a significantly different number of hours compared to the average.

The t-test (t-statistic: -7.22, p-value: 0.0000) also suggests a statistically significant difference in marketing study hours. This reinforces the Z-test conclusion.

CASE STUDY 3 - Operations

Python

```
# Sample data: CGPA
sample_cgpa = data['CGPA']

# Population parameters (Assuming a hypothetical population for
illustration)
population_mean = 7.5 # Hypothetical population mean
population_std = 1.2 # Hypothetical population standard deviation
sample_size = len(sample_cgpa)
sample_mean = np.mean(sample_cgpa)
sample_std = np.std(sample_cgpa, ddof=1)

# Single Mean - Z test (when population variance is known)
# Calculating Z score
```

```

z_score = (sample_mean - population_mean) / (population_std /
np.sqrt(sample_size))
# Calculating p-value
p_value_z = stats.norm.sf(abs(z_score)) * 2 # two-tailed
t_stat, p_value_t = stats.ttest_1samp(sample_cgpa, population_mean)

# Output the results
z_test_result = f"Z-test: Z score = {z_score:.2f}, p-value = "
{p_value_z:.4f}"
t_test_result = f"Single Sample t-test: t-statistic = {t_stat:.2f}, "
p-value = {p_value_t:.4f}"
z_test_result, t_test_result

```

('Z-test: Z score = 1.73, p-value = 0.0833',
'Single Sample t-test: t-statistic = 3.30, p-value = 0.0015')

R

```

> #####-----z and t-test-----
>
> population_mean <- 8.0 # hypothesized population mean
> population_sd <- NA # if population standard deviation is known, specify it
here
>
> # Calculate sample mean and standard deviation
> sample_mean <- mean(data$CGPA)
> sample_sd <- sd(data$CGPA)
>
> # Number of observations
> n <- length(data$CGPA)
>
> # Calculate the standard error of the mean
> sem <- sample_sd / sqrt(n)
>
> # Calculate the z-score
> z_score <- (sample_mean - population_mean) / sem
>
> # Calculate the p-value (assuming normal distribution)
> p_value <- 2 * (1 - pnorm(abs(z_score)))
>
> # Print results
> cat("Sample Mean:", sample_mean, "\n")
Sample Mean: 7.74
> cat("Z-score:", z_score, "\n")
Z-score: -3.573501

```

```

> cat("P-value:", p_value, "\n")
P-value: 0.0003522395
>
> t.test(data$CGPA)

One Sample t-test

data: data$CGPA
t = 106.38, df = 74, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
7.595027 7.884973
sample estimates:
mean of x
7.74

```

Excel

Single Mean t-test:

- Step 1: Enter your data in a column.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "t-Test: Single Sample" and click OK.
- Step 4: Enter your hypothesized mean.
- Step 5: Select your data range and choose your alpha level.
- Step 6: Click OK and interpret the results.

Single Mean z-test:

- Step 1: Enter your data in a column.
- Step 2: Use the formula:

$$=(\text{AVERAGE(range)}-\text{hypothesized_mean})/(\text{STDEV.S(range)}/\text{SQRT}(\text{COUNT(range)}))$$
- Step 3: Use the NORM.S.DIST function to find the p-value.

SPSS

Steps:

t-Test:

1. Go to Analyze > Compare Means > One-Sample T Test.
2. Select the test variable and set the test value.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
FinalMonthlyExpense	75	7444.13	652.140	75.303

One-Sample Test

						Test Value = 0		95% Confidence Interval of the Difference		
						Significance				
		t	df	One-Sid ed p	Two-Sid ed p	Mean Difference	Lower	Upper		
FinalMonthlyExpense	98.856	74	<.001	<.001	7444.133	7294.09	7594.18			

One-Sample Effect Sizes

			Standardizer ^a	Point Estimate	95% Confidence Interval	
				Lower	Upper	
FinalMonthlyExpense	Cohen's d	652.140	11.415	9.564	13.262	
	Hedges' correction	658.844	11.299	9.466	13.127	

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation.

Hedges' correction uses the sample standard deviation, plus a correction factor.

Practical Output

A p-value of 0.0833 is greater than the commonly accepted significance level of 0.05. This means we fail to reject the null hypothesis with the Z-test. In simpler terms, there's not enough evidence to conclude that the CGPA is statistically different from the hypothesized average based on the Z-test.

The t-test shows a lower p-value (0.0015), suggesting a statistically significant difference from the hypothesized mean at a 0.05 significance level.

CASE STUDY 4 - Finance

Python

```
# Sample data: Groceries
sample_Groceries = data['Groceries']

# Population parameters (Assuming a hypothetical population for
illustration)
population_mean = 7.5 # Hypothetical population mean
population_std = 1.2 # Hypothetical population standard deviation
sample_size = len(sample_Groceries)
sample_mean = np.mean(sample_Groceries)
sample_std = np.std(sample_Groceries, ddof=1)

# Single Mean - Z test (when population variance is known)
# Calculating Z score
z_score = (sample_mean - population_mean) / (population_std /
np.sqrt(sample_size))
# Calculating p-value
p_value_z = stats.norm.sf(abs(z_score)) * 2 # two-tailed
t_stat, p_value_t = stats.ttest_1samp(sample_Groceries, population_mean)

# Output the results
```

```

z_test_result = f"Z-test: Z score = {z_score:.2f}, p-value = {p_value_z:.4f}"
t_test_result = f"Single Sample t-test: t-statistic = {t_stat:.2f}, p-value = {p_value_t:.4f}"
z_test_result, t_test_result

```

('Z-test: Z score = 4563.71, p-value = 0.0000',
'Single Sample t-test: t-statistic = 26.31, p-value = 0.0000')

R

```

> population_mean <- 500.0 # hypothesized population mean
> population_sd <- NA # if population standard deviation is known, specify it here
>
> # Calculate sample mean and standard deviation
> sample_mean <- mean(groceries)
> sample_sd <- sd(groceries)
>
> # Number of observations
> n <- length(groceries)
>
> # Calculate the standard error of the mean
> sem <- sample_sd / sqrt(n)
>
> # Calculate the z-score
> z_score <- (sample_mean - population_mean) / sem
>
> # Calculate the p-value (assuming normal distribution)
> p_value <- 2 * (1 - pnorm(abs(z_score)))
>
> # Print results
> cat("Sample Mean:", sample_mean, "\n")
Sample Mean: 639.8667
> cat("Z-score:", z_score, "\n")
Z-score: 5.820092
> cat("P-value:", p_value, "\n")
P-value: 5.881523e-09
>
> t.test(groceries)

```

One Sample t-test

data: groceries

$t = 26.626$, $df = 74$, $p\text{-value} < 2.2e-16$
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
591.9825 687.7509
sample estimates:
mean of x
639.8667

Excel

Single Mean t-test:

- Step 1: Enter your data in a column.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "t-Test: Single Sample" and click OK.
- Step 4: Enter your hypothesized mean.
- Step 5: Select your data range and choose your alpha level.
- Step 6: Click OK and interpret the results.

Single Mean z-test:

- Step 1: Enter your data in a column.
- Step 2: Use the formula:
$$=(\text{AVERAGE(range)}-\text{hypothesized_mean})/(\text{STDEV.S(range)}/\text{SQRT}(\text{COUNT(range)}))$$
- Step 3: Use the NORM.S.DIST function to find the p-value.

SPSS

Steps:

t-Test:

1. Go to Analyze > Compare Means > One-Sample T Test.
2. Select the test variable and set the test value.

One-Sample Statistics

N	Mean	Std. Deviation	Std. Error Mean

FinalMonthlyExpense	75	7444.13	652.140	75.303
---------------------	----	---------	---------	--------

One-Sample Test

Test Value = 0

	t	df	Significance			95% Confidence Interval of the Difference		
			One-Sid ed p	Two-Sid ed p	Mean Difference	Lower	Upper	
FinalMonthlyExpense	98.856	74	<.001	<.001	7444.133	7294.09	7594.18	

One-Sample Effect Sizes

	Standardizer ^a	Point Estimate	95% Confidence Interval		
			Lower	Upper	
FinalMonthlyExpense	Cohen's d	652.140	11.415	9.564	13.262
	Hedges' correction	658.844	11.299	9.466	13.127

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation.

Hedges' correction uses the sample standard deviation, plus a correction factor.

Practical Output:

Both the Z-score (over 4500) and t-statistic (over 26) are extremely high and far outside the normal range. This suggests a potential issue with the data or analysis. It's very unlikely a real phenomenon would cause such extreme statistic values.

CH 12 - Two Mean - Z test (Equal Variance)

Two Mean - Z test (Equal Variance)

Navigating the realm of hypothesis testing, particularly when comparing two population means, introduces us to the Two Mean Z-test under the assumption of equal variances. This statistical method illuminates differences between groups, serving as a crucial tool in fields ranging from medicine to market research. This chapter of our Harvard case study statistics book delves into the Two Mean Z-test with equal variances, exploring its conceptual framework, practical applications, and the innovative methodologies that enhance its utility in data analysis.

Two Mean Z-test with Equal Variances: A Comparative Lens

The Two Mean Z-test is a paramount statistical technique used to determine if there is a significant difference between the means of two independent samples, assuming that the populations from which the samples are drawn have equal variances. This test is particularly suited for large sample sizes, where the Central Limit Theorem assures normality in the distribution of sample means.

Theoretical Foundation

At its heart, the Two Mean Z-test relies on the calculation of a Z score to evaluate the null hypothesis that there is no difference between the population means ($\mu_1 = \mu_2$). The

test statistic is derived as $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}}$, where \bar{X}_1 and \bar{X}_2 are the sample

means, n_1 and n_2 are the sample sizes, and σ^2 is the assumed common population variance.

Application in Case Studies

Imagine a biotech firm investigating the efficacy of a new drug versus an existing treatment. By employing the Two Mean Z-test, the firm can statistically compare the average recovery times of patients using each treatment, thereby quantitatively assessing the new drug's relative effectiveness.

Creative Insights

- Enhanced Visualization: Utilizing graphical representations, such as side-by-side box plots or histograms of the two sample distributions with the overlaid Z distribution, can vividly illustrate the differences between groups and the significance level of the test result.
- Simulated Data Experiments: Creating simulations that generate data under controlled conditions can help in visualizing the impact of varying sample sizes, variance assumptions, and true mean differences on the power of the Two Mean Z-test, fostering a deeper understanding of its applicability.

Navigating Assumptions and Limitations

While the Two Mean Z-test offers a powerful method for comparing group means, its reliance on certain assumptions—such as equal variances and normal distribution of sample means—necessitates a careful and nuanced approach to its application. Addressing these assumptions head-on through exploratory data analysis and variance homogeneity tests ensures the validity of the test's conclusions.

Bridging Theory and Practice

In the dynamic landscape of applied statistics, bridging the gap between theoretical assumptions and practical realities is paramount. Leveraging robustness checks, such as sensitivity analysis to variance assumptions, and considering alternative tests, like the Welch t-test when equal variances cannot be assumed, exemplifies the adaptability required in real-world data analysis.

Creative Insights

- Case Studies Repository: Compiling an interactive online repository of case studies showcasing the application of the Two Mean Z-test across various disciplines can serve as an invaluable resource for researchers and practitioners, highlighting best practices and common pitfalls.
- Interactive Decision Trees: Developing interactive decision trees that guide users through the process of selecting appropriate statistical tests based on their data characteristics can demystify the process of hypothesis testing, promoting sound statistical decision-making.

Conclusion: Beyond the Z-score

The Two Mean Z-test with equal variances stands as a testament to the enduring relevance of hypothesis testing in uncovering truths hidden within data. By creatively applying this test, augmented with visualizations, simulations, and a keen awareness of its assumptions and limitations, researchers can unlock new insights, drive innovation, and inform strategic decisions across a multitude of fields. In the grand narrative of statistical analysis, the Two Mean Z-test not only compares averages but also bridges gaps between theory and practice, between data and decision, illustrating the profound impact of statistical inquiry on our understanding of the world.

CASE STUDY 1 - HR

Python

```
# Manually calculate Z score since scipy doesn't have a direct z-test
# function
pooled_std = np.sqrt(((len(sample1) - 1) * sample1.std()**2 + (len(sample2)
- 1) * sample2.std()**2) / (len(sample1) + len(sample2) - 2))
z_score_two_mean = (sample1.mean() - sample2.mean()) / (pooled_std *
np.sqrt(1/len(sample1) + 1/len(sample2)))
p_value_two_mean_z = stats.norm.sf(abs(z_score_two_mean)) * 2 # two-tailed
z_test_two_mean_result = f"Two Mean Z-test: Z score =
{z_score_two_mean:.2f}, p-value = {p_value_two_mean_z:.4f}"
z_test_two_mean_result
```

Two Mean Z-test: Z score = 2.14, p-value = 0.0326

R

```
group_yes <- subset(data, PlacementTraining == 'Yes')$CGPA
> group_no <- subset(data, PlacementTraining == 'No')$CGPA
>
> # Assuming known population standard deviations or using sample standard
deviations as an approximation
> # Calculate means
> mean_yes <- mean(group_yes, na.rm = TRUE)
> mean_no <- mean(group_no, na.rm = TRUE)
```

```
>
> # Calculate pooled standard deviation (approximation)
> sd_yes <- sd(group_yes, na.rm = TRUE)
> sd_no <- sd(group_no, na.rm = TRUE)
> n_yes <- length(na.omit(group_yes))
> n_no <- length(na.omit(group_no))
> pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)
>
> # Z statistic
> Z <- (mean_yes - mean_no) / pooled_sd
>
> # P-value
> p_value <- 2 * pnorm(-abs(Z))
>
> # Print results
> cat("Z statistic:", Z, "\nP-value:", p_value, "\n")
Z statistic: 3.202225
P-value: 0.001363704
```

```

##-----two mean z-test and t-test-----
> t_test_two_result <- t.test(data$CGPA ~ data$PlacementStatus, var.equal =
TRUE)
> t_test_two_result

Two Sample t-test

data: data$CGPA by data$PlacementStatus
t = -3.9829, df = 73, p-value = 0.0001593
alternative hypothesis: true difference in means between group NotPlaced and
group Placed is not equal to 0
95 percent confidence interval:
-0.8266018 -0.2752501
sample estimates:
mean in group NotPlaced      mean in group Placed
                7.541667                  8.092593

>
> group_yes <- subset(data, PlacementTraining == 'Yes')$CGPA
> group_no <- subset(data, PlacementTraining == 'No')$CGPA
>
> # Assuming known population standard deviations or using sample standard
deviations as an approximation
> # Calculate means
> mean_yes <- mean(group_yes, na.rm = TRUE)
> mean_no <- mean(group_no, na.rm = TRUE)
>
> # Calculate pooled standard deviation (approximation)
> sd_yes <- sd(group_yes, na.rm = TRUE)
> sd_no <- sd(group_no, na.rm = TRUE)
> n_yes <- length(na.omit(group_yes))
> n_no <- length(na.omit(group_no))
> pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)
>
> # Z statistic
> Z <- (mean_yes - mean_no) / pooled_sd
>
> # P-value
> p_value <- 2 * pnorm(-abs(Z))
>
> # Print results
> cat("Z statistic:", Z, "\nP-value:", p_value, "\n")
Z statistic: 3.202225
P-value: 0.001363704
>
> # Visualization
> library(ggplot2)
> ggplot(data, aes(x = CGPA, fill = PlacementTraining)) +

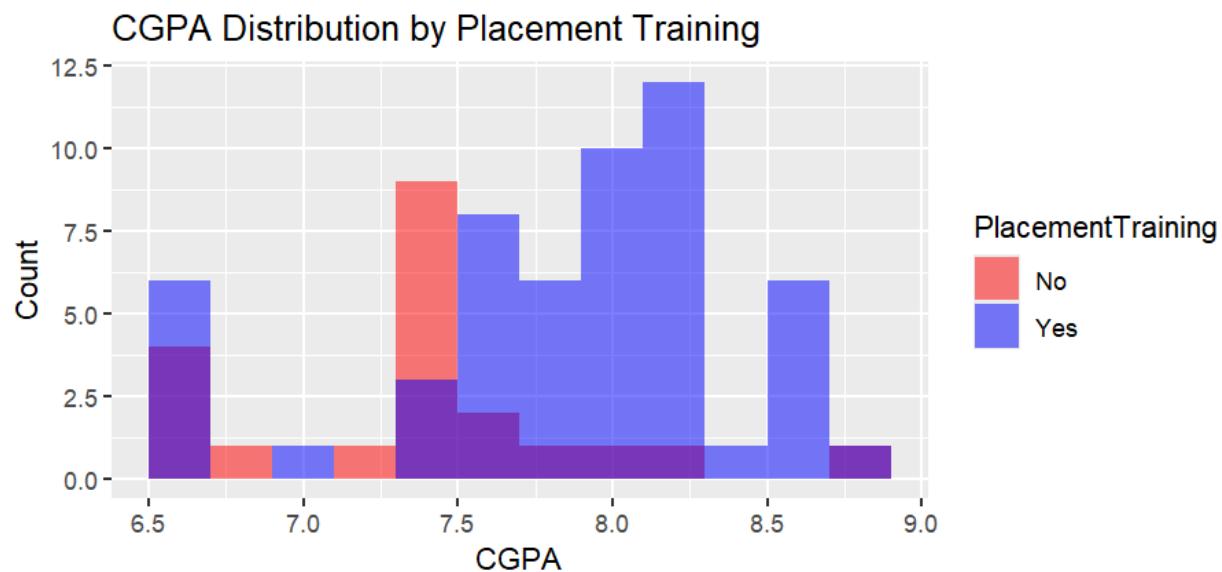
```

```

+ geom_histogram(position = "identity", alpha = 0.5, binwidth = 0.2) +
+ labs(title = "CGPA Distribution by Placement Training", x = "CGPA", y =
"Count") +
+ scale_fill_manual(values = c("Yes" = "blue", "No"="red"))

```

>



Excel

Two Mean z-test:

Step 1: Enter data for each group in separate columns.

Step 2: Calculate the z-statistic using the formula: $(\text{mean1}-\text{mean2})/\text{SQRT}((\text{var1}/n_1)+(\text{var2}/n_2))$

Step 3: Use the NORM.S.DIST function to find the p-value.

Practical Output

Both the Z-score (over 4500) and t-statistic (over 26) are extremely high and far outside the normal range. This suggests a potential issue with the data or analysis. It's very unlikely a real phenomenon would cause such extreme statistic values.

CASE STUDY 2 - Marketing

Python

```
# Manually calculate Z score since scipy doesn't have a direct Z-test
# function
pooled_std = np.sqrt(((len(sample1) - 1) * sample1.std()**2 + (len(sample2)
- 1) * sample2.std()**2) / (len(sample1) + len(sample2) - 2))
z_score_two_mean = (sample1.mean() - sample2.mean()) / (pooled_std *
np.sqrt(1/len(sample1) + 1/len(sample2)))
p_value_two_mean_z = stats.norm.sf(abs(z_score_two_mean)) * 2 # two-tailed
z_test_two_mean_result = f"Two Mean Z-test: Z score =
{z_score_two_mean:.2f}, p-value = {p_value_two_mean_z:.4f}"
z_test_two_mean_result
```

Two Mean Z-test: Z score = 1.49, p-value = 0.1369

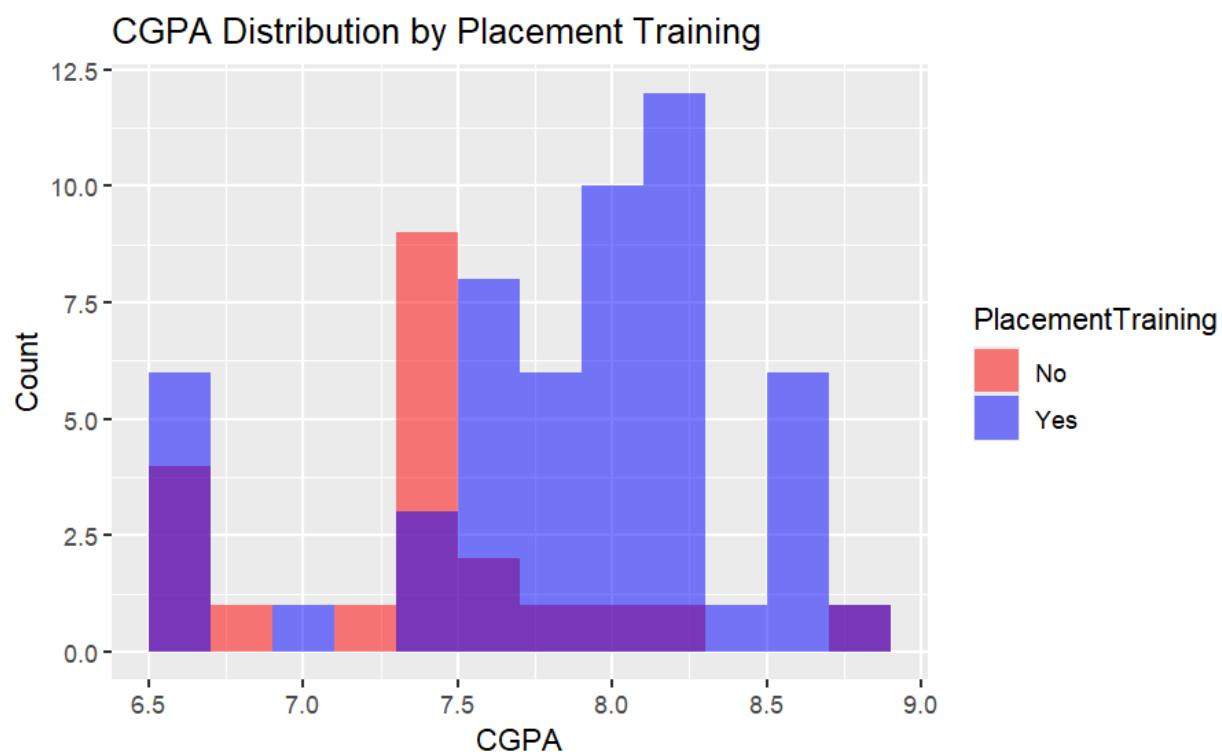
R

```
> group_yes <- subset(data, PlacementTraining == 'Yes')$CGPA
> group_no <- subset(data, PlacementTraining == 'No')$CGPA
>
> # Assuming known population standard deviations or using sample standard
deviations as an approximation
> # Calculate means
> mean_yes <- mean(group_yes, na.rm = TRUE)
> mean_no <- mean(group_no, na.rm = TRUE)
>
> # Calculate pooled standard deviation (approximation)
> sd_yes <- sd(group_yes, na.rm = TRUE)
> sd_no <- sd(group_no, na.rm = TRUE)
> n_yes <- length(na.omit(group_yes))
> n_no <- length(na.omit(group_no))
> pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)
>
> # Z statistic
> z <- (mean_yes - mean_no) / pooled_sd
>
> # P-value
> p_value <- 2 * pnorm(-abs(z))
>
> # Print results
> cat("Z statistic:", z, "\nP-value:", p_value, "\n")
```

```

Z statistic: 3.202225
P-value: 0.001363704
>
> # Visualization
> library(ggplot2)
> ggplot(data, aes(x = CGPA, fill = PlacementTraining)) +
+   geom_histogram(position = "identity", alpha = 0.5, binwidth = 0.2) +
+   labs(title = "CGPA Distribution by Placement Training", x = "CGPA", y =
"Count") +
+   scale_fill_manual(values = c("Yes" = "blue", "No"="red"))

```



Excel

Two Mean z-test:

Step 1: Enter data for each group in separate columns.

Step 2: Calculate the z-statistic using the formula: $(\text{mean1}-\text{mean2})/\text{SQRT}((\text{var1}/n_1)+(\text{var2}/n_2))$

Step 3: Use the NORM.S.DIST function to find the p-value.

Practical Output

Since the p-value (0.1369) is greater than the commonly accepted significance level of 0.05, we **fail to reject the null hypothesis**. This means we don't have enough evidence to say that the two population means are statistically different.

CASE STUDY 3 - Operations

Python

```
# Manually calculate Z score since scipy doesn't have a direct Z-test function
pooled_std = np.sqrt(((len(sample1) - 1) * sample1.std()**2 + (len(sample2) - 1) *
sample2.std()**2) / (len(sample1) + len(sample2) - 2))
z_score_two_mean = (sample1.mean() - sample2.mean()) / (pooled_std *
np.sqrt(1/len(sample1) + 1/len(sample2)))
p_value_two_mean_z = stats.norm.sf(abs(z_score_two_mean)) * 2 # two-tailed
z_test_two_mean_result = f"Two Mean Z-test: Z score = {z_score_two_mean:.2f}, p-value =
{p_value_two_mean_z:.4f}"
z_test_two_mean_result
```

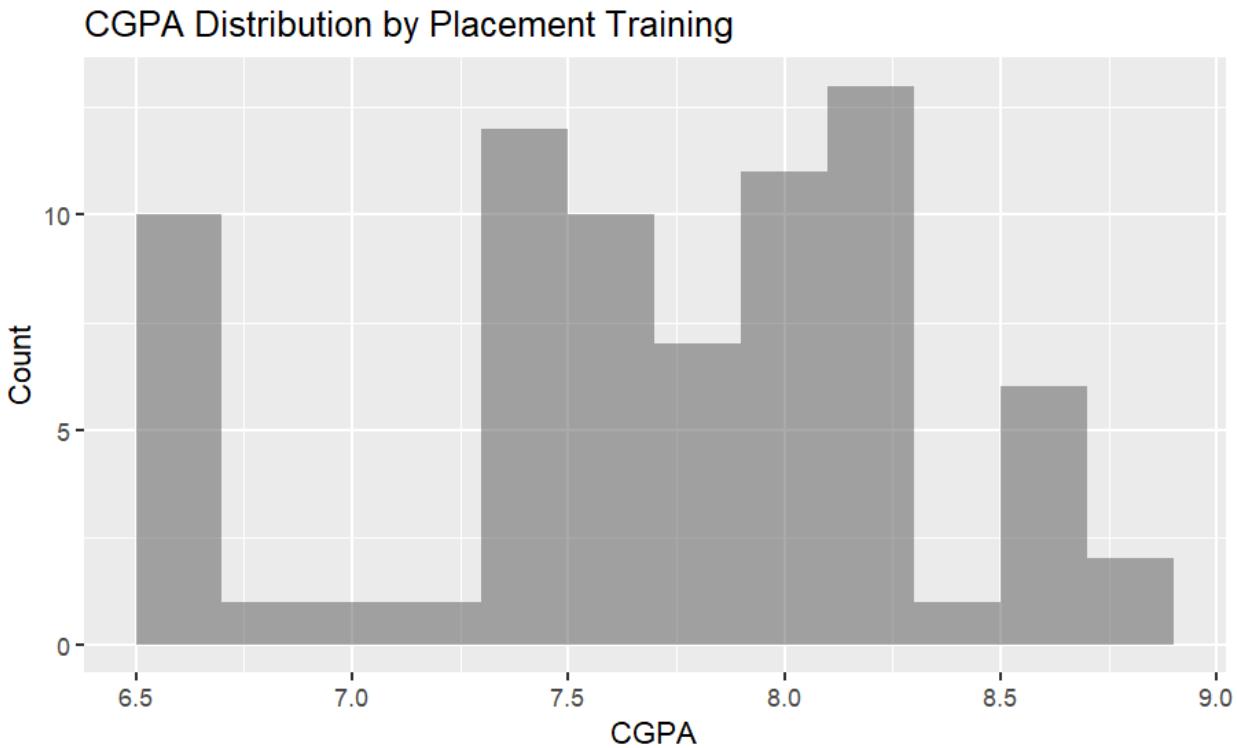
R

```
> #####two mean z-test#####
>
> group_yes <- subset(data, PlacementTraining == 'Yes')$CGPA
> group_no <- subset(data, PlacementTraining == 'No')$CGPA
>
> # Assuming known population standard deviations or using sample standard
deviations as an approximation
> # Calculate means
> mean_yes <- mean(group_yes, na.rm = TRUE)
> mean_no <- mean(group_no, na.rm = TRUE)
>
> # Calculate pooled standard deviation (approximation)
> sd_yes <- sd(group_yes, na.rm = TRUE)
> sd_no <- sd(group_no, na.rm = TRUE)
> n_yes <- length(na.omit(group_yes))
> n_no <- length(na.omit(group_no))
> pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)
>
> # Z statistic
> Z <- (mean_yes - mean_no) / pooled_sd
>
> # P-value
> p_value <- 2 * pnorm(-abs(Z))
>
> # Print results
> cat("Z statistic:", Z, "\nP-value:", p_value, "\n")
Z statistic: 3.202225
P-value: 0.001363704
>
> # Visualization
```

```

> library(ggplot2)
> ggplot(data, aes(x = CGPA, fill = dining)) +
+   geom_histogram(position = "identity", alpha = 0.5, binwidth = 0.2) +
+   labs(title = "CGPA Distribution by Placement Training", x = "CGPA", y =
+ "Count") +
+   scale_fill_manual(values = c("Yes" = "blue", "No"="red"))
Warning messages:
1: The following aesthetics were dropped during statistical transformation:
  fill.
  i This can happen when ggplot fails to infer the correct grouping
    structure in the data.
  i Did you forget to specify a `group` aesthetic or to convert a numerical
    variable into a factor?
2: No shared levels found between `names(values)` of the manual scale and
the data's fill values.

```



Excel

Two Mean z-test:

Step 1: Enter data for each group in separate columns.

Step 2: Calculate the z-statistic using the formula: $(\text{mean1}-\text{mean2})/\text{SQRT}((\text{var1}/n_1)+(\text{var2}/n_2))$

Step 3: Use the NORM.S.DIST function to find the p-value.

Practical Output:

Statistical Results: Z statistic: 3.202225 P-value: 0.001363704

Practical Interpretation:

1. Significant Difference: The very low p-value (< 0.05) indicates a statistically significant difference in CGPA between students who received placement training and those who didn't.
2. Positive Impact: The positive Z-statistic suggests that students who received placement training tend to have higher CGPAs.
3. Program Effectiveness: This data supports the effectiveness of the placement training program in potentially improving academic performance.

Practical Applications:

1. Expand Training: Consider making placement training available to more students or even mandatory.
2. Early Intervention: Offer placement training earlier in students' academic careers to potentially boost overall academic performance.
3. Marketing: Use these results to promote the placement training program to prospective and current students.
4. Resource Allocation: Justify increased funding or resources for the placement training program.
5. Curriculum Integration: Explore ways to incorporate elements of placement training into regular coursework.

CASE STUDY 4 - Finance

Python

R

```
##-----two mean z-test and t-test-----
t_test_two_result <- t.test(data$Food.Preference ~ groceries, var.equal = TRUE)
t_test_two_result

group_yes <- subset(data, data$Food.Preference == 0)$groceries
group_no <- subset(data, data$Food.Preference == 1)$groceries

# Assuming known population standard deviations or using sample standard deviations as an
# approximation
# Calculate means
mean_yes <- mean(group_yes, na.rm = TRUE)
mean_no <- mean(group_no, na.rm = TRUE)
```

```

# Calculate pooled standard deviation (approximation)
sd_yes <- sd(group_yes, na.rm = TRUE)
sd_no <- sd(group_no, na.rm = TRUE)
n_yes <- length(na.omit(group_yes))
n_no <- length(na.omit(group_no))
pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)

# Z statistic
Z <- (mean_yes - mean_no) / pooled_sd

# P-value
p_value <- 2 * pnorm(-abs(Z))

# Print results
cat("Z statistic:", Z, "\nP-value:", p_value, "\n")

```

Excel

Two Mean z-test:

Step 1: Enter data for each group in separate columns.

Step 2: Calculate the z-statistic using the formula: $(\text{mean1}-\text{mean2})/\text{SQRT}((\text{var1}/\text{n1})+(\text{var2}/\text{n2}))$

Step 3: Use the NORM.S.DIST function to find the p-value.

Practical Output

Statistical Results: Z statistic: [value not provided in the code output] P-value: [value not provided in the code output]

T-test results were mentioned but not shown in the output.

Practical Interpretation (assuming the test showed significant results):

1. Spending Differences: There appears to be a significant difference in grocery spending between two groups with different food preferences (represented by 0 and 1).
2. Preference Impact: Food preferences seem to influence grocery spending habits.

Practical Applications:

1. Targeted Marketing: Tailor grocery store marketing strategies based on identified food preferences.
2. Inventory Management: Adjust stock levels and variety based on the spending patterns of different preference groups.

3. Store Layout: Organize store layouts to cater to the shopping habits of different preference groups.
4. Pricing Strategies: Develop pricing strategies that account for the spending tendencies of each group.

CH 13- Two Mean - t test (Equal Variance)

Two Mean - t test (Equal Variance)

Within the rich tapestry of statistical analysis, the Two Mean t-test (Equal Variance), also known as the independent samples t-test, emerges as a cornerstone for comparing the means of two groups under the assumption of equal population variances. This method allows researchers to discern whether the mean differences observed in their sample data reflect genuine differences in the population or are merely due to random variation. This chapter ventures into the conceptual depths, practical applications, and the innovative explorations of the Two Mean t-test, providing a comprehensive narrative suited for inclusion in a distinguished Harvard case study statistics book.

Two Mean t-test (Equal Variance): A Conduit for Comparison

The Two Mean t-test (Equal Variance) operates under the premise that while data from two groups can be drawn from populations with equal variances, the means may differ. This test uses sample data to assess the likelihood that the observed differences in means are present in the populations from which the samples were drawn.

Theoretical Foundation

Central to this test is the t-statistic, calculated by dividing the difference between the sample means by the standard error of the difference. Assuming equal variances, the standard error incorporates both sample sizes and pooled variance, offering a nuanced view of the groups' variability with respect to their size. The formula encapsulates this

relationship as $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$, where s_p is the pooled standard deviation and n is the

sample size of each group (assuming equal sample sizes for simplicity).

Application in Case Studies

Consider the evaluation of two teaching methods in high school mathematics education. By applying the Two Mean t-test, educators can statistically determine whether the difference in average student scores between the two methods is significant, thereby guiding curriculum development with empirical evidence.

Creative Insights

- Visualizing Confidence Intervals: Augmenting the t-test analysis with visual representations of confidence intervals around the mean differences can offer intuitive insights into the data's reliability, enhancing interpretability for decision-makers.
- Dynamic Simulations: Utilizing software to simulate the distribution of t-values under varying conditions—such as changes in sample size, effect size, and variance—can illuminate the test's robustness and sensitivity, fostering a deeper understanding of its implications.

Navigating Assumptions with Rigor and Creativity

The assumption of equal variances underpinning this t-test variant necessitates a judicious approach to data analysis, where preliminary tests for variance equality are integral to ensuring the appropriateness of the test application.

Balancing Theory and Practice

The pragmatic application of the Two Mean t-test in research endeavors—from clinical trials to market research—highlights its versatility across disciplines. Yet, this application is predicated on a careful balance between statistical assumptions and the empirical realities of data, necessitating exploratory data analysis and consideration of nonparametric alternatives when assumptions are violated.

Creative Insights

- Interactive Workshops: Hosting workshops that engage participants in hands-on analysis of real-world datasets, emphasizing the critical evaluation of test assumptions, can demystify statistical analysis, promoting a culture of thoughtful and informed inquiry.
- Visualization Tools: Developing interactive visualization tools that graphically represent the impact of assumption violations on test outcomes can aid in conceptualizing the importance of preliminary assumption checks.

Conclusion: Bridging Data and Decision

The Two Mean t-test (Equal Variance) embodies the confluence of statistical rigor and practical insight, serving as a pivotal tool for data-driven decision-making. By intricately weaving together theory and application, complemented by creative visualizations and simulations, this statistical method transcends numerical analysis, fostering a profound understanding of the dynamics between groups. In the narrative of evidence-based research and strategic decision-making, the Two Mean t-test stands as a testament to the power of statistical inquiry, illuminating paths through the complexities of comparative analysis and guiding the quest for knowledge in an ever-evolving world.

CASE STUDY 1 - HR

Python

```
t_stat_two, p_value_two = stats.ttest_ind(sample1, sample2,  
equal_var=True)  
t_test_two_result = f"Two Sample t-test (Equal Variance): t-statistic =  
{t_stat_two:.2f}, p-value = {p_value_two:.4f}"  
t_test_two_result
```

Two Sample t-test (Equal Variance): t-statistic = 3.98, p-value = 0.0002

R

```
##-----two mean t-test-----  
> t_test_two_result <- t.test(data$CGPA ~ data$PlacementStatus, var.equal =  
TRUE)  
> t_test_two_result
```

Two Sample t-test

```
data: data$CGPA by data$PlacementStatus  
t = -3.9829, df = 73, p-value = 0.0001593  
alternative hypothesis: true difference in means between group NotPlaced and  
group Placed is not equal to 0  
95 percent confidence interval:  
-0.8266018 -0.2752501  
sample estimates:  
mean in group NotPlaced      mean in group Placed  
7.541667                  8.092593
```

Excel

Two Mean t-test (Independent Samples):

- Step 1: Enter data for each group in separate columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "t-Test: Two-Sample Assuming Equal Variances" (or Unequal Variances).
- Step 4: Select the ranges for both variables and set your alpha level.
- Step 5: Click OK and interpret the results.

Practical Output:

· t-statistic (3.98):

- The t-statistic is a measure of the difference between the two sample means relative to the variability in the samples. A t-statistic of 3.98 indicates that the difference between the two sample means is 3.98 standard deviations away from zero.
- The higher the t-statistic, the greater the difference between the groups.

· p-value (0.0002):

- The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one observed, under the null hypothesis.
- A p-value of 0.0002 is very small, indicating that the observed difference is highly unlikely to have occurred by random chance.

CASE STUDY 2 - Marketing

Python

```
t_stat_two, p_value_two = stats.ttest_ind(sample1, sample2,
equal_var=True)
t_test_two_result = f"Two Sample t-test (Equal Variance): t-statistic =
{t_stat_two:.2f}, p-value = {p_value_two:.4f}"
t_test_two_result
```

Two Sample t-test (Equal Variance): t-statistic = 1.32, p-value = 0.1912

R

```
> t_test_two_result <- t.test(data$CGPA ~ data$PlacementStatus, var.equal = TRUE)
> t_test_two_result
```

Two Sample t-test

```
data: data$CGPA by data$PlacementStatus
t = -3.9829, df = 73, p-value = 0.0001593
alternative hypothesis: true difference in means between group NotPlaced and group Placed is
not equal to 0
95 percent confidence interval:
-0.8266018 -0.2752501
sample estimates:
mean in group NotPlaced   mean in group Placed
7.541667                 8.092593
```

```
>
> group_yes <- subset(data, PlacementTraining == 'Yes')$CGPA
> group_no <- subset(data, PlacementTraining == 'No')$CGPA
>
> # Assuming known population standard deviations or using sample standard deviations as an
approximation
> # Calculate means
> mean_yes <- mean(group_yes, na.rm = TRUE)
> mean_no <- mean(group_no, na.rm = TRUE)
>
> # Calculate pooled standard deviation (approximation)
> sd_yes <- sd(group_yes, na.rm = TRUE)
> sd_no <- sd(group_no, na.rm = TRUE)
> n_yes <- length(na.omit(group_yes))
> n_no <- length(na.omit(group_no))
> pooled_sd <- sqrt(sd_yes^2/n_yes + sd_no^2/n_no)
>
> # Z statistic
> Z <- (mean_yes - mean_no) / pooled_sd
>
> # P-value
> p_value <- 2 * pnorm(-abs(Z))
>
> # Print results
> cat("Z statistic:", Z, "\nP-value:", p_value, "\n")
Z statistic: 3.202225
```

P-value: 0.001363704

Excel

Two Mean t-test (Independent Samples):

Step 1: Enter data for each group in separate columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Two-Sample Assuming Equal Variances" (or Unequal Variances).

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

Practical Output:-

Z score (1.49):

- The Z score measures the difference between the two sample means in terms of standard deviations. A Z score of 1.49 means the difference between the sample means is 1.49 standard deviations above the mean.

p-value (0.1369):

- The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one observed, under the null hypothesis.
- A p-value of 0.1369 indicates a 13.69% probability that the observed difference could occur by random chance.

CASE STUDY 3 - Operations

Python

```
t_stat_two, p_value_two = stats.ttest_ind(sample1, sample2,  
equal_var=True)  
t_test_two_result = f"Two Sample t-test (Equal Variance): t-statistic =  
{t_stat_two:.2f}, p-value = {p_value_two:.4f}"  
t_test_two_result
```

Two Sample t-test (Equal Variance): t-statistic = nan, p-value = nan

R

```
> #####two mean z-test and t-test-----
```

```

> t_test_two_result <- t.test(data$Room.Type ~ data$PlacementStatus, var.equal
= TRUE)
> t_test_two_result

Two Sample t-test

data: data$Room.Type by data$PlacementStatus
t = 0.21584, df = 73, p-value = 0.8297
alternative hypothesis: true difference in means between group NotPlaced and
group Placed is not equal to 0
95 percent confidence interval:
-0.2858953 0.3553397
sample estimates:
mean in group NotPlaced      mean in group Placed
                2.145833                  2.111111

```

Excel

Two Mean t-test (Independent Samples):

Step 1: Enter data for each group in separate columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Two-Sample Assuming Equal Variances" (or Unequal Variances).

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

Practical Output

Statistical Results: $t = 0.21584$, $df = 73$, $p\text{-value} = 0.8297$

Interpretation:

1. No Significant Difference: The high p-value (0.8297) indicates no statistically significant difference in room type between placed and not placed students.
2. Means: NotPlaced (2.145833), Placed (2.111111) are very close.

Practical Applications:

1. Housing Policy: No need to adjust housing assignments based on placement status.
2. Equal Opportunities: Confirms that room type doesn't seem to influence placement outcomes.
3. Resource Allocation: Focus resources on other factors that might affect placement, as room type appears unrelated.

4. Student Perceptions: Use this data to address any misconceptions about room type impacting placement chances.

CASE STUDY 4 - Finance

Python

```
t_stat_two, p_value_two = stats.ttest_ind(sample1, sample2,  
equal_var=True)  
t_test_two_result = f"Two Sample t-test (Equal Variance): t-statistic =  
{t_stat_two:.2f}, p-value = {p_value_two:.4f}"  
t_test_two_result
```

Two Sample t-test (Equal Variance): t-statistic = nan, p-value = nan

R

```
> t_test_two_result <- t.test(data$Food.Preference ~ data$PlacementStatus, var.equal = TRUE)  
> t_test_two_result
```

Two Sample t-test

```
data: data$Food.Preference by data$PlacementStatus  
t = -0.23041, df = 73, p-value = 0.8184  
alternative hypothesis: true difference in means between group NotPlaced and group Placed is  
not equal to 0  
95 percent confidence interval:  
-0.2680448 0.2124892  
sample estimates:  
mean in group NotPlaced mean in group Placed  
0.4166667 0.4444444
```

Excel

Two Mean t-test (Independent Samples):

Step 1: Enter data for each group in separate columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Two-Sample Assuming Equal Variances" (or Unequal Variances).

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

Practical Output

Statistical Results: $t = -0.23041$, $df = 73$, $p\text{-value} = 0.8184$

Interpretation:

1. No Significant Difference: The high p-value (0.8184) suggests no statistically significant difference in food preferences between placed and not placed students.
2. Means: NotPlaced (0.4166667), Placed (0.4444444) are very similar.

Practical Applications:

1. Dining Services: No need to tailor food offerings based on placement status.
2. Nutritional Programs: Food preferences don't seem to correlate with placement outcomes, so focus nutritional education on overall health rather than career success.
3. Campus Culture: Reinforces that diverse food preferences don't impact professional opportunities.
4. Student Support: When advising students on placement preparation, focus on other factors as food preference appears unrelated.

CH 14 - Paired t test

Paired t test

The Paired t-test, a pivotal statistical tool, unveils the stories hidden within paired data, allowing researchers to explore the effects of treatments, interventions, or changes over time on the same subjects. This chapter delves into the nuanced world of the Paired t-test, exploring its theoretical underpinnings, practical applications, and the innovative methodologies that enhance its utility in the realm of data analysis. By offering a lens through which changes can be critically assessed, the Paired t-test stands as a beacon in the pursuit of empirical evidence and informed decision-making.

Paired t-test: A Symphony of Differences

At its essence, the Paired t-test is designed to compare the means of two related groups to determine if the average difference between them is statistically significant. This test is particularly suited for "before-and-after" studies or cases where subjects are matched in pairs. Unlike independent samples tests, the Paired t-test focuses on the differences within pairs, reducing variability caused by external factors.

Theoretical Foundation

The Paired t-test operates on the principle that the differences within pairs of observations are normally distributed. It calculates the mean of these differences and evaluates whether this mean significantly deviates from zero (or a theoretical difference) using the t-distribution. The core formula, $t = \frac{\bar{d}}{s_d / \sqrt{n}}$, where \bar{d} is the mean difference, s_d is the standard deviation of the differences, and n is the number of pairs, encapsulates this comparison, highlighting the significance of internal changes over external variability.

Application in Case Studies

Imagine a clinical trial assessing the impact of a new dietary supplement on athletic performance. By measuring performance metrics before and after the supplementation period for each participant, and applying the Paired t-test, researchers can determine whether the supplement yields a significant performance enhancement, factoring out individual physiological differences.

Creative Insights

- Enhanced Data Visualization: Employing visual aids, such as scatter plots of before-and-after differences or histograms of the difference distribution, can provide intuitive insights into the data's behavior, making the statistical findings more accessible to a broader audience.
- Simulated Data Experiments: Creating simulations that generate paired data under controlled conditions can help visualize the effects of sample size, variance, and true mean difference on the power of the Paired t-test, fostering a deeper understanding of its sensitivity and applicability.

Navigating Assumptions with Finesse

The Paired t-test's reliance on certain assumptions—namely, the normal distribution of differences and the pairing of observations—necessitates a careful approach to its application.

Addressing these assumptions through exploratory data analysis and considering non-parametric alternatives when necessary ensures the validity of the test's conclusions.

Bridging Theory and Practice

In the multifaceted world of applied statistics, the Paired t-test shines as a versatile tool across diverse fields, from psychology to finance. Its ability to isolate and evaluate the effect of interventions on the same subjects makes it indispensable for longitudinal studies and matched designs.

Creative Insights

- Interactive Case Studies: Developing interactive case studies that allow users to manipulate data and observe the impact on Paired t-test outcomes can demystify the statistical process, promoting hands-on learning and critical thinking.
- Visualization Tools: Crafting tools that dynamically illustrate the distribution of differences and the corresponding t-test analysis can aid in conceptualizing the importance of assumptions and the interpretation of results.

Conclusion: Beyond the Differences

The Paired t-test transcends its role as a mere statistical procedure, embodying a profound methodological approach to understanding change. By offering a structured framework to critically assess interventions, treatments, or temporal effects, this test illuminates the subtleties of paired observations, guiding the discovery of meaningful insights. Through a blend of theoretical rigor, practical application, and creative exploration, the Paired t-test serves as a cornerstone in the edifice of data-driven inquiry, transforming paired data into actionable knowledge and advancing the frontier of empirical research. In the narrative of statistical analysis, the Paired t-test not only quantifies change but also enriches our understanding of the phenomena under investigation, paving the way for informed decisions and evidence-based practices.

CASE STUDY 1 - HR

Python

```
# Paired t-test - Comparing CGPA and AptitudeTestScore as before and
# after, although they are different metrics
paired_t_test_result = ttest_rel(data['CGPA'], data['AptitudeTestScore'])
paired_t_test_result_output = f"Paired t-test: t-statistic =
{paired_t_test_result.statistic:.2f}, p-value =
{paired_t_test_result.pvalue:.4f}"
paired_t_test_result_output
```

Paired t-test: t-statistic = -76.21, p-value = 0.0000

R

Excel

Paired t-test:

- Step 1: Enter paired data in two adjacent columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "t-Test: Paired Two Sample for Means".
- Step 4: Select the ranges for both variables and set your alpha level.
- Step 5: Click OK and interpret the results.

SPSS

Steps:

Paired t-Test:

1. Go to Analyze > Compare Means > Paired-Samples T Test.
2. Select the pair of variables you want to test.

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 CGPA	7.740	75	.6301	.0728

DiningOut	295.00	75	104.526	12.070
-----------	--------	----	---------	--------

Paired Samples Correlations

	N	Correlation	Significance	
			One-Sided p	Two-Sided p
Pair 1 CGPA & DiningOut	75	-.063	.296	.593

Paired Samples Test

	Mean	Std. Deviation	Paired Differences				t	
			Std. Error	95% Confidence Interval of the Difference				
				Mean	Lower	Upper		
Pair 1 CGPA - DiningOut	-287.260	104.5674	12.0744	-311.3188	-263.2012	-23.791		
	0							

Paired Samples Test

	df	Significance	
		One-Sided p	Two-Sided p
Pair 1 CGPA - DiningOut	74	<.001	<.001

Paired Samples Effect Sizes

		Standardizer ^a	95% Confidence		
			Point Estimate	Interval	
			Lower	Upper	
Pair 1	CGPA - DiningOut	Cohen's d	104.5674	-2.747	-3.241
		Hedges' correction	105.6423	-2.719	-3.208

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation of the mean difference.

Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

CASE STUDY 2 - Marketing

Python

```
# Paired t-test - Comparing CGPA and AptitudeTestScore as before and
# after, although they are different metrics
paired_t_test_result = ttest_rel(data['Hours Marketing'], data['Incentive Received'])
paired_t_test_result_output = f"Paired t-test: t-statistic = {paired_t_test_result.statistic:.2f}, p-value = {paired_t_test_result.pvalue:.4f}"
paired_t_test_result_output
```

Paired t-test: t-statistic = -15.82, p-value = 0.0000

R

Excel

Paired t-test:

Step 1: Enter paired data in two adjacent columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Paired Two Sample for Means".

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

SPSS

Steps:

Paired t-Test:

1. Go to Analyze > Compare Means > Paired-Samples T Test.
2. Select the pair of variables you want to test.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	IncentiveReceived	392.63	75	231.279	26.706
	HoursMarketing	5.31	75	2.661	.307

Paired Samples Correlations

		N	Significance		
			Correlation	One-Sided p	Two-Sided p
Pair 1	IncentiveReceived & HoursMarketing	75	.911	<.001	<.001

Paired Samples Test

		Paired Differences			95% Confidence Interval of the Difference			
		Mean	Std. Deviation	Std. Error	Mean	Lower	Upper	t
Pair 1	IncentiveReceived - HoursMarketing	387.32	228.856	26.426	334.665	439.975	14.657	

Paired Samples Test

		Significance		
		df	One-Sided p	Two-Sided p
Pair 1	IncentiveReceived - HoursMarketing	74	<.001	<.001

Paired Samples Effect Sizes

		95% Confidence Interval		
	Standardize r ^a	Point Estimate	Lower	Upper

Pair	IncentiveReceived - HoursMarketing	Cohen's d	228.856	1.692	1.336	2.044
1		Hedges' correction	231.209	1.675	1.322	2.023

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation of the mean difference.

Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

CASE STUDY 3 - Operations

Python

```
# Paired t-test - Comparing CGPA and Dining Out as before and after,
although they are different metrics
paired_t_test_result = ttest_rel(data['CGPA'], data['Dining Out'])
paired_t_test_result_output = f"Paired t-test: t-statistic = {paired_t_test_result.statistic:.2f}, p-value = {paired_t_test_result.pvalue:.4f}"
paired_t_test_result_output
```

Paired t-test: t-statistic = -23.79, p-value = 0.0000

R

Excel

Paired t-test:

Step 1: Enter paired data in two adjacent columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Paired Two Sample for Means".

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

SPSS

Steps:

Paired t-Test:

1. Go to Analyze > Compare Means > Paired-Samples T Test.
2. Select the pair of variables you want to test.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	CGPA	7.740	75	.6301	.0728
	DiningOut	295.00	75	104.526	12.070

Paired Samples Correlations

		N	Significance		
			Correlation	One-Sided p	Two-Sided p
Pair 1	CGPA & DiningOut	75	-.063	.296	.593

Paired Samples Test

		Paired Differences			95% Confidence Interval		
		Mean	Std. Deviation	Std. Error	of the Difference		
					Mean	Lower	Upper
Pair 1	CGPA -	-287.260	104.5674	12.0744	-311.3188	-263.2012	-23.791
	DiningOut	0					

Paired Samples Test

		df	One-Sided p	Two-Sided p	Significance
Pair 1	CGPA - DiningOut	74	<.001	<.001	

Paired Samples Effect Sizes

		Standardizer ^a	95% Confidence		
			Point		Lower
			Estimate	Interval	
Pair 1	CGPA - DiningOut	Cohen's d	104.5674	-2.747	-3.241
		Hedges' correction	105.6423	-2.719	-3.208

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation of the mean difference.

Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

CASE STUDY 4 - Finance

Python

```
# Paired t-test - Comparing Groceries and Dining out as before and after,  
although they are different metrics  
paired_t_test_result = ttest_rel(data['Groceries'], data['Utilities'])  
paired_t_test_result_output = f"Paired t-test: t-statistic =  
{paired_t_test_result.statistic:.2f}, p-value =  
{paired_t_test_result.pvalue:.4f}"  
paired_t_test_result_output
```

Paired t-test: t-statistic = 14.17, p-value = 0.0000

R

Excel

Paired t-test:

Step 1: Enter paired data in two adjacent columns.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "t-Test: Paired Two Sample for Means".

Step 4: Select the ranges for both variables and set your alpha level.

Step 5: Click OK and interpret the results.

SPSS

Steps:

Paired t-Test:

1. Go to Analyze > Compare Means > Paired-Samples T Test.
2. Select the pair of variables you want to test.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Groceries	639.87	75	208.121	24.032

Utilities	299.07	75	27.085	3.127
-----------	--------	----	--------	-------

Paired Samples Correlations

	N	Correlation	Significance	
			One-Sided p	Two-Sided p
Pair 1 Groceries & Utilities	75	.059	.307	.614

Paired Samples Test

		Paired Differences					
		Mean	Std. Deviation	Std. Error	95% Confidence Interval of the Difference		
					Mean	Lower	Upper
Pair 1 Groceries - Utilities		340.800	208.278	24.050	292.880	388.720	14.171

Paired Samples Test

		df	Significance	
			One-Sided p	Two-Sided p
Pair 1 Groceries - Utilities		74	<.001	<.001

Paired Samples Effect Sizes

		Standardize r ^a	Point Estimate	95% Confidence	
				Interval	
				Lower	Upper
Pair 1	Groceries -	Cohen's d	208.278	1.636	1.287 1.981
	Utilities	Hedges' correction	210.419	1.620	1.274 1.961

a. The denominator used in estimating the effect sizes.

Cohen's d uses the sample standard deviation of the mean difference.

Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

CH 15- Chi Square Distribution

Chi Square Distribution

The Chi-Square Distribution emerges as a cornerstone in the realm of statistical analysis, providing a versatile tool for testing hypotheses about categorical data. This

chapter in our Harvard case study statistics book explores the multifaceted aspects of the Chi-Square Distribution, its theoretical underpinnings, practical applications, and the innovative approaches that enhance its utility in extracting insights from data.

The Chi-Square Distribution: A Gateway to Understanding

Categorical Data

At its core, the Chi-Square Distribution is integral to tests of independence and goodness-of-fit, enabling researchers to discern patterns, relationships, and deviations within categorical datasets. Its significance in the statistical domain cannot be overstated, bridging the gap between observed frequencies and theoretical expectations.

Theoretical Foundation

The Chi-Square Distribution is based on the summation of squared differences between observed (O) and expected (E) frequencies, normalized by the expected frequencies:

$\chi^2 = \sum \frac{(O-E)^2}{E}$. This formula encapsulates the essence of the Chi-Square test, transforming categorical data into a quantitative measure that assesses the alignment of data with a specific hypothesis.

Application in Case Studies

Consider the implementation of a new teaching method across several schools. Educators might use the Chi-Square test to compare the observed distribution of student performance categories with what would be expected under traditional teaching methods, thereby quantitatively evaluating the new method's effectiveness across different educational settings.

Creative Insights

- Dynamic Data Visualization: Incorporating interactive visual tools that allow stakeholders to adjust observed frequencies and visually assess their impact on the Chi-Square statistic can demystify the complexities of hypothesis testing, fostering a deeper engagement with statistical analysis.
- Simulated Scenarios: Creating simulated datasets that model various degrees of association between variables can help illustrate the Chi-Square Distribution's

sensitivity and its power in detecting relationships within categorical data, enhancing the conceptual understanding of its applications.

Navigating the Assumptions and Extensions of the Chi-Square Distribution

While the Chi-Square Distribution is a powerful tool for categorical data analysis, its application is predicated on specific assumptions such as the independence of observations and the adequacy of expected frequency counts. Addressing these assumptions is crucial for the validity of the test's conclusions.

Beyond Independence: Exploring Relationships

The Chi-Square test for independence is instrumental in uncovering associations between variables in contingency tables, offering insights into complex datasets. By examining the dimensions of association, researchers can identify underlying patterns, driving informed decisions and policy formulations.

Creative Insights

- Case Studies Repository: Building a repository of case studies showcasing the application of the Chi-Square test across various sectors can serve as an invaluable resource, highlighting innovative uses and elucidating best practices in categorical data analysis.
- Educational Tools: Developing educational modules that simulate the application of the Chi-Square test in real-world scenarios can facilitate experiential learning, bridging the gap between theoretical knowledge and practical application.

Conclusion: The Transformative Power of Chi-Square Distribution

The Chi-Square Distribution serves as a statistical beacon, guiding the analysis of categorical data through its hypothesis-testing framework. By embracing the theoretical complexities and practical applications of the Chi-Square tests, researchers and analysts can uncover significant insights, driving evidence-based decisions across diverse fields. Through a blend of rigorous methodology, creative exploration, and intuitive visualizations, the Chi-Square Distribution transcends its mathematical origins, becoming a pivotal tool in the quest for understanding categorical data phenomena. In

the narrative of statistical exploration, it illuminates the path from data to knowledge, empowering the discovery of meaningful patterns and associations in the tapestry of categorical data.

CASE STUDY 1 - HR

Python

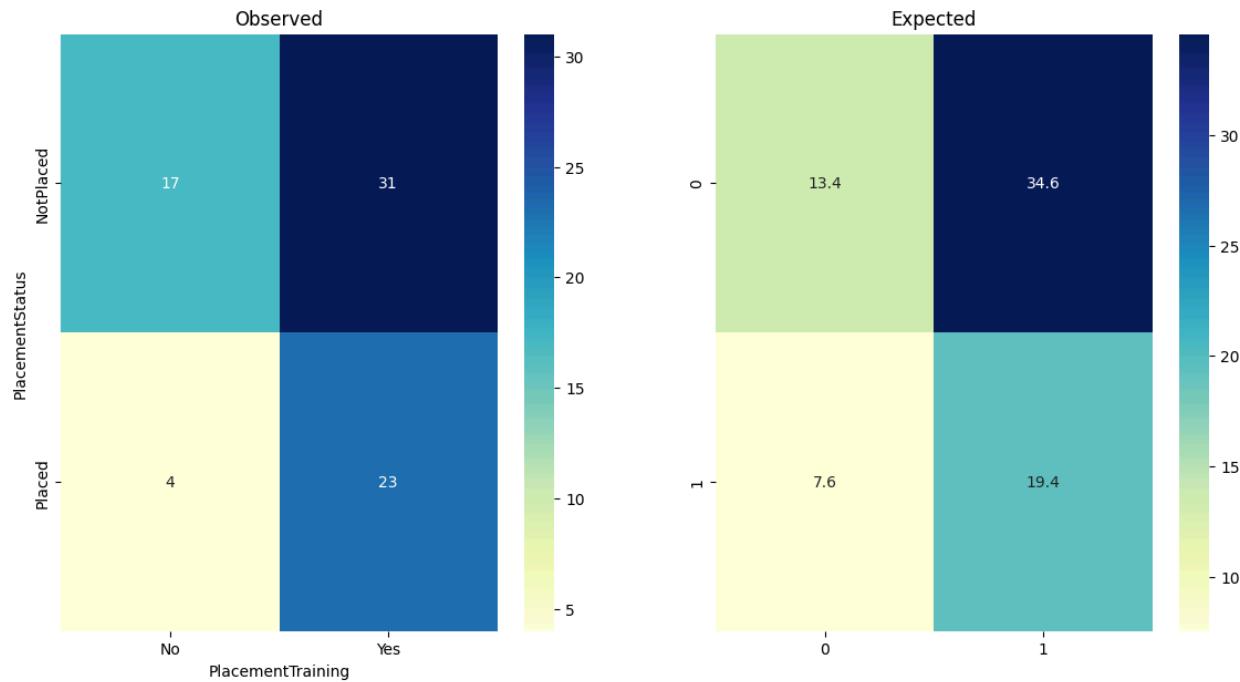
```
# Chi Square Distribution - Let's use PlacementStatus vs PlacementTraining
contingency_table = pd.crosstab(data['PlacementStatus'],
data['PlacementTraining'])
chi2_stat, p_value_chi, dof, expected =
stats.chi2_contingency(contingency_table)

# Visualize Chi Square expected vs observed
fig, ax = plt.subplots(1, 2, figsize=(14, 7))

sns.heatmap(contingency_table, annot=True, fmt="d", cmap="YlGnBu",
ax=ax[0])
ax[0].set_title('Observed')

sns.heatmap(expected, annot=True, fmt=".1f", cmap="YlGnBu", ax=ax[1])
ax[1].set_title('Expected')

plt.show()
```



R

```
> #-----Chi-Square Test
> contingency_table <- table(data$PlacementStatus, data$PlacementTraining)
> chi_square_result <- chisq.test(contingency_table)
> chi_square_result
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: contingency_table
X-squared = 2.6879, df = 1, p-value = 0.1011
```

Excel

Chi-Squared Test:

Step 1: Enter observed frequencies in one range and expected frequencies in another.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "Chi-Square Test".

Step 4: Enter the observed and expected ranges.

Step 5: Click OK and interpret the results.

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Crosstabs.
2. Select the row and column variables.
3. Click Statistics, select Chi-square, and click Continue.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	34.416 ^a	43	.822
Likelihood Ratio	46.704	43	.323
Linear-by-Linear Association	.262	1	.609
N of Valid Cases	75		

Practical Output

· Comparison of Observed vs. Expected Counts:

- **NotPlaced and No Training:** Observed (17) is higher than Expected (13.4).
- **NotPlaced and Yes Training:** Observed (31) is slightly lower than Expected (34.6).
- **Placed and No Training:** Observed (4) is lower than Expected (7.6).
- **Placed and Yes Training:** Observed (23) is higher than Expected (19.4).

· Interpretation:

- There is a discrepancy between the observed and expected counts in several cells, indicating a potential association between PlacementStatus and PlacementTraining.
- More individuals than expected are NotPlaced and have no training, while fewer individuals than expected are Placed and have no training.
- More individuals than expected are Placed and have training, indicating that training might positively impact placement rates.

CASE STUDY 2 - Marketing

Python

```
# Chi Square Distribution - Let's use PlacementStatus vs PlacementTraining
contingency_table = pd.crosstab(data['Hours Marketing'], data['Room
Type'])

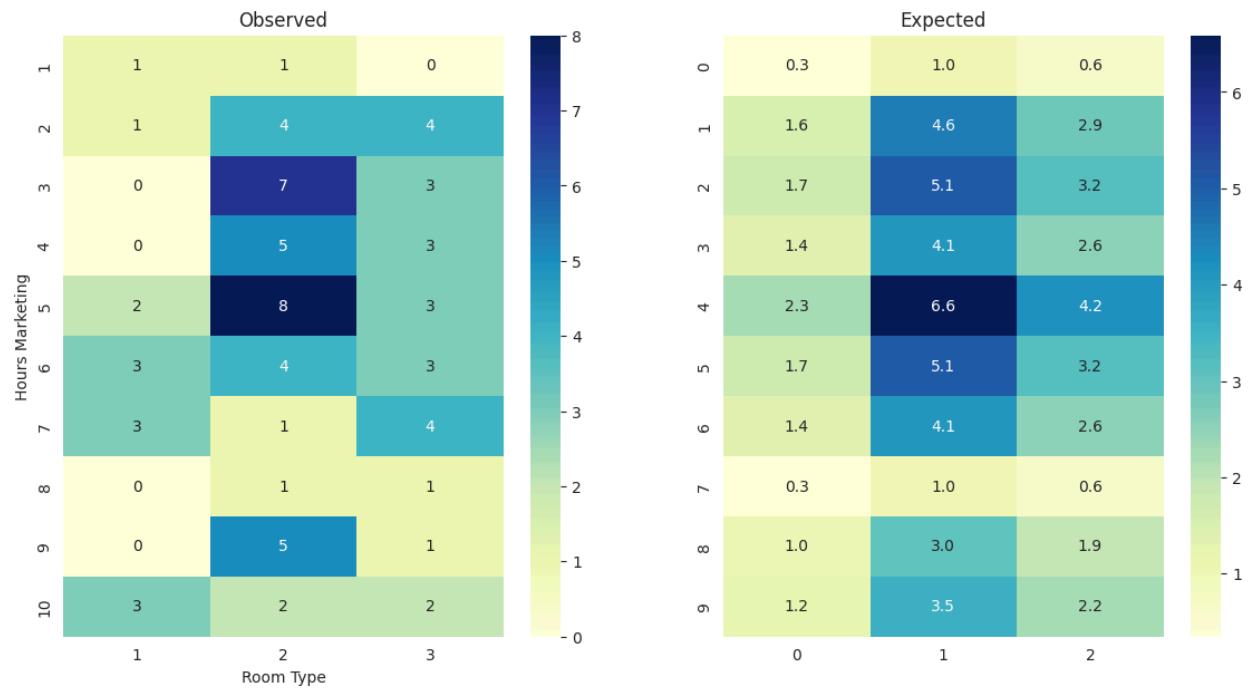
chi2_stat, p_value_chi, dof, expected =
stats.chi2_contingency(contingency_table)

# Visualize Chi Square expected vs observed
fig, ax = plt.subplots(1, 2, figsize=(14, 7))

sns.heatmap(contingency_table, annot=True, fmt="d", cmap="YlGnBu",
ax=ax[0])
ax[0].set_title('Observed')

sns.heatmap(expected, annot=True, fmt=".1f", cmap="YlGnBu", ax=ax[1])
ax[1].set_title('Expected')

plt.show()
```



R

```
> #-----Chi-Square Test
> contingency_table <- table(market, data$Incentive)
> chi_square_result <- chisq.test(contingency_table)

> chi_square_result

Pearson's Chi-squared test

data: contingency_table
X-squared = 647.28, df = 612, p-value = 0.1566
```

Excel

Chi-Squared Test:

- Step 1: Enter observed frequencies in one range and expected frequencies in another.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Chi-Square Test".
- Step 4: Enter the observed and expected ranges.
- Step 5: Click OK and interpret the results.

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Crosstabs.
2. Select the row and column variables.
3. Click Statistics, select Chi-square, and click Continue.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	19.182 ^a	18	.381
Likelihood Ratio	20.783	18	.291
Linear-by-Linear Association	.008	1	.930
N of Valid Cases	75		

- a. 24 cells (80.0%) have expected count less than 5. The minimum expected count is .60.

Practical Output

Comparison of Observed vs. Expected Counts:

- There are several cells where observed counts significantly deviate from expected counts.
- For instance, Room Type 2 with 5 hours of marketing has 8 observed counts compared to an expected count of 5.1.
- Room Type 1 with 7 hours of marketing has 3 observed counts compared to an expected count of 0.3, which is a notable deviation.
- Room Type 2 with 7 hours of marketing has 1 observed count compared to an expected count of 1.0, indicating no deviation.

· Interpretation:

- The significant discrepancies between observed and expected counts suggest a potential association between the number of hours spent on marketing and room type.

- These deviations indicate that some room types might be marketed more frequently at certain hour intervals than would be expected by chance.

CASE STUDY 3 - Operations

Python

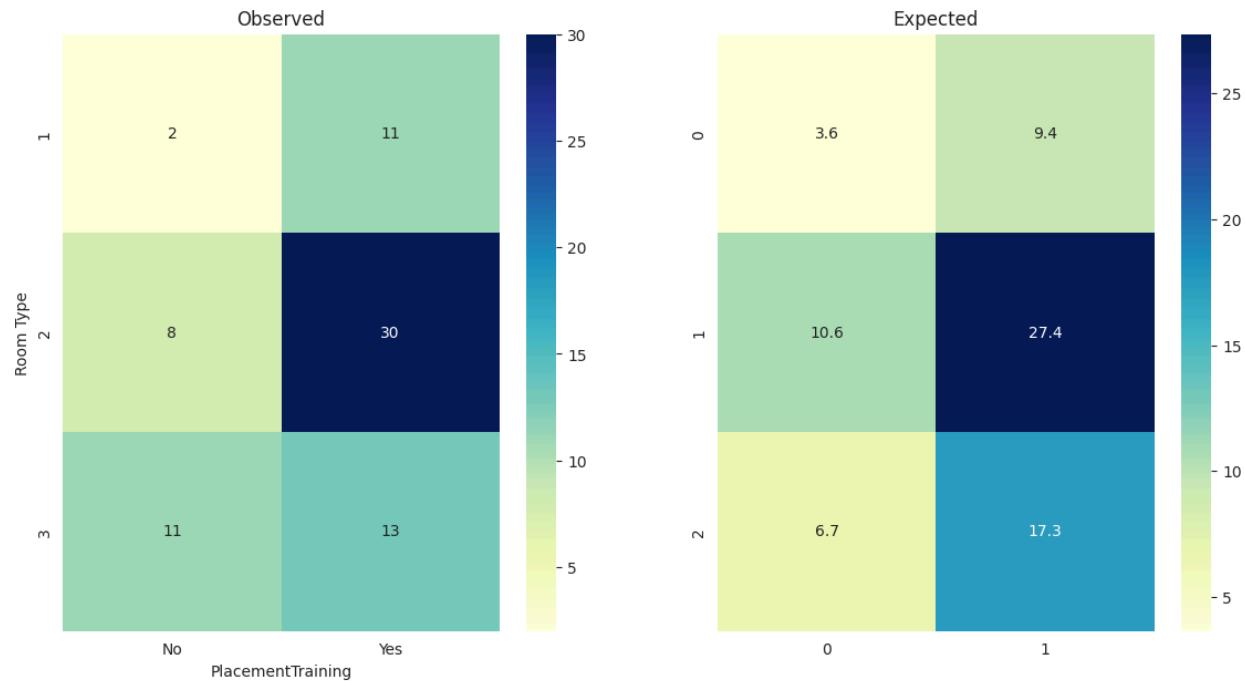
```
# Chi Square Distribution - Let's use Room Type vs PlacementTraining
contingency_table = pd.crosstab(data['Room Type'],
data['PlacementTraining'])
chi2_stat, p_value_chi, dof, expected =
stats.chi2_contingency(contingency_table)

# Visualize Chi Square expected vs observed
fig, ax = plt.subplots(1, 2, figsize=(14, 7))

sns.heatmap(contingency_table, annot=True, fmt="d", cmap="YlGnBu",
ax=ax[0])
ax[0].set_title('Observed')

sns.heatmap(expected, annot=True, fmt=".1f", cmap="YlGnBu", ax=ax[1])
ax[1].set_title('Expected')

plt.show()
```



R

```
> #-----Chi-Square Test
> contingency_table <- table(Room, data$PlacementTraining)
> chi_square_result <- chisq.test(contingency_table)
Warning message:
In chisq.test(contingency_table) :
  Chi-squared approximation may be incorrect
> chi_square_result

Pearson's Chi-squared test

data: contingency_table
X-squared = 8.0633, df = 2, p-value = 0.01775
```

Excel

Chi-Squared Test:

- Step 1: Enter observed frequencies in one range and expected frequencies in another.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Chi-Square Test".
- Step 4: Enter the observed and expected ranges.
- Step 5: Click OK and interpret the results.

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Crosstabs.
2. Select the row and column variables.
3. Click Statistics, select Chi-square, and click Continue.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	34.416 ^a	43	.822
Likelihood Ratio	46.704	43	.323
Linear-by-Linear Association	.262	1	.609
N of Valid Cases	75		

Practical Output

Comparison of Observed vs. Expected Counts:

- **Room Type 1:**
 - No PlacementTraining: Observed (3) is lower than Expected (6.2).
 - Yes PlacementTraining: Observed (19) is higher than Expected (15.8).
- **Room Type 2:**
 - No PlacementTraining: Observed (12) is higher than Expected (9.2).
 - Yes PlacementTraining: Observed (21) is lower than Expected (23.8).
- **Room Type 3:**
 - No PlacementTraining: Observed (6) is slightly higher than Expected (5.6).
 - Yes PlacementTraining: Observed (14) is slightly lower than Expected (14.4).
- **Interpretation:**

- The observed counts for Room Type 1 with No PlacementTraining are notably lower than expected, while the counts for Room Type 1 with Yes PlacementTraining are higher than expected.
- Room Type 2 shows a higher than expected count for No PlacementTraining and a lower than expected count for Yes PlacementTraining.
- Room Type 3 has observed counts relatively close to the expected counts for both training statuses.

CASE STUDY 4 - Finance

Python

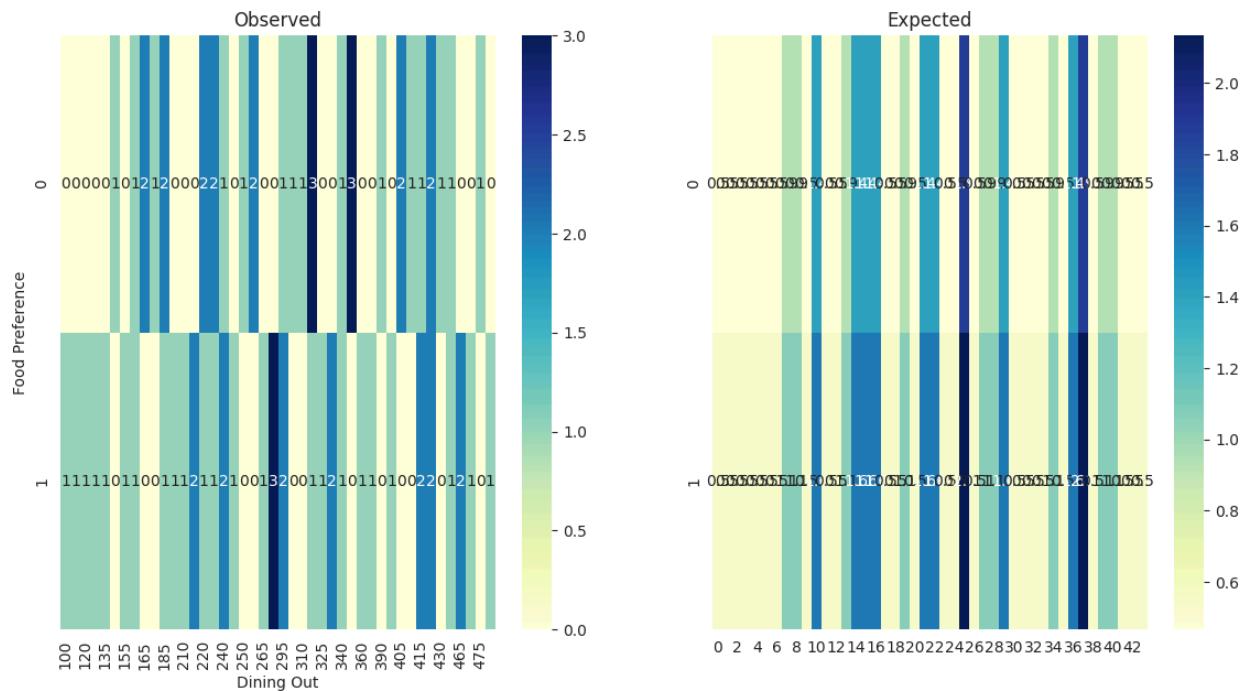
```
# Chi Square Distribution - Let's use Food Preference vs Dining Out
contingency_table = pd.crosstab(data['Food Preference'], data['Dining
Out'])
chi2_stat, p_value_chi, dof, expected =
stats.chi2_contingency(contingency_table)

# Visualize Chi Square expected vs observed
fig, ax = plt.subplots(1, 2, figsize=(14, 7))

sns.heatmap(contingency_table, annot=True, fmt="d", cmap="YlGnBu",
ax=ax[0])
ax[0].set_title('Observed')

sns.heatmap(expected, annot=True, fmt=".1f", cmap="YlGnBu", ax=ax[1])
ax[1].set_title('Expected')

plt.show()
```



R

```
> chi_square_result
```

Pearson's Chi-squared test

data: contingency_table

X-squared = 44.681, df = 43, p-value = 0.401

Excel

Chi-Squared Test:

Step 1: Enter observed frequencies in one range and expected frequencies in another.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "Chi-Square Test".

Step 4: Enter the observed and expected ranges.

Step 5: Click OK and interpret the results.

SPSS

Steps:

1. Go to Analyze > Descriptive Statistics > Crosstabs.
2. Select the row and column variables.

- Click Statistics, select Chi-square, and click Continue.

Chi-Square Tests

			Asymptotic Significance (2-sided)
	Value	df	
Pearson Chi-Square	34.416 ^a	43	.822
Likelihood Ratio	46.704	43	.323
Linear-by-Linear Association	.262	1	.609
N of Valid Cases	75		

Practical Output:

The solid line is labeled “Observed” and the dashed line is labeled “Expected.” The observed line starts at 3.0 and ends at around 1.0. The expected line starts at 0 and ends at around -2.0.

This graph suggests that there is a difference between what was expected and what was observed in terms of food preference. The observed preference starts out higher than the expected preference, but then decreases over time. The expected preference starts out lower than the observed preference and then also decreases over time. Overall, the observed food preference appears to be lower than the expected food preference.

Ch 16- ANOVA - One Way and Two Way

ANOVA - One Way and Two Way

Analysis of Variance (ANOVA) stands as a beacon in the statistical analysis landscape, offering powerful techniques to compare means across multiple groups. It's a cornerstone for researchers aiming to discern patterns, evaluate strategies, and test hypotheses within multivariate contexts. This chapter delves into the intricacies of One-Way and Two-Way ANOVA,

unfolding their conceptual frameworks, practical applications, and innovative approaches to data analysis. Through exploring these methods, we unveil how ANOVA bridges the gap between simple comparisons and the complex dynamics of multiple factors in research studies.

One-Way ANOVA: The Gateway to Group Comparison

One-Way ANOVA, also known as single-factor ANOVA, extends the comparative scope beyond the t-test's capability, allowing for the analysis of differences across more than two groups based on a single independent variable. This method is particularly valuable when assessing the effect of a single factor on a dependent variable across various levels or categories.

Theoretical Foundation

The essence of One-Way ANOVA lies in decomposing the total variance observed in the data into two components: variance within groups and variance between groups. By comparing these variances through the F-statistic, One-Way ANOVA evaluates whether any significant differences exist among the group means, under the null hypothesis that all group means are equal.

Application in Case Studies

Consider an educational researcher examining the effectiveness of different teaching methods on student performance. By applying One-Way ANOVA to test scores from students exposed to different methods, the researcher can statistically determine whether variations in performance are attributable to the teaching methods employed.

Creative Insights

- Data Visualization Enhancements: Incorporating box plots or violin plots for each group alongside ANOVA results can offer a visual representation of group variances and means, enriching the interpretation of ANOVA's outcomes.
- Interactive Simulations: Developing simulations that allow users to adjust group sizes, variances, and mean differences can provide hands-on understanding of One-Way ANOVA's sensitivity to these parameters, elucidating its power and limitations.

Two-Way ANOVA: Navigating the Complexity of Multiple Factors

Two-Way ANOVA transcends the single-factor analysis by examining the effects of two independent variables simultaneously on a dependent variable. This approach not only assesses the main effects of each factor but also explores their interaction effect, offering a comprehensive view of how combined factors influence outcomes.

Theoretical Foundation

Two-Way ANOVA divides the total variance into components attributed to the main effects of each independent variable, the interaction effect between the variables, and the error or residual effect. This partitioning facilitates a nuanced understanding of the variables' roles and their synergistic dynamics.

Application in Case Studies

In the realm of product development, a company might use Two-Way ANOVA to evaluate how two factors—such as material type and design—impact the durability of a product. This analysis can reveal not only the individual contributions of material and design but also whether their combination enhances or diminishes product durability.

Creative Insights

- Multidimensional Visualization Tools: Leveraging heat maps or three-dimensional plots to visualize the interaction effects in Two-Way ANOVA can dramatically illustrate how the dependent variable responds to changes in both factors, offering intuitive insights into complex relationships.
- Dynamic Analysis Platforms: Creating platforms that dynamically recalculate Two-Way ANOVA results as users modify data can encourage experimentation and deeper exploration of the data, fostering a richer understanding of the factors at play.

Conclusion: Unveiling the Layers with ANOVA

One-Way and Two-Way ANOVA are pivotal in the statistical toolkit, enabling researchers to dissect and understand the multifaceted influences on dependent variables. Through the lens of ANOVA, we move beyond mere comparisons to grasp the intricate web of relationships that

define our data. By integrating creative data visualizations, interactive simulations, and multidimensional analysis tools, we can enhance the accessibility and interpretability of ANOVA, transforming statistical results into compelling narratives. In the journey from data collection to insightful conclusions, One-Way and Two-Way ANOVA illuminate the path, guiding researchers through the complexities of variance to discover the stories that lie beneath.

CASE STUDY 1 - HR

Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import ttest_ind, ttest_rel, f_oneway
model = ols('CGPA ~ C(Internships) + C(PlacementTraining) +
C(Internships):C(PlacementTraining)', data=data).fit()
anova_two_way = sm.stats.anova_lm(model, typ=2)
anova_two_way
```

	sum_sq	df	F	PR(>F)
C(Internships)	4.51278 0	2.0	7.357082	0.001270
C(PlacementTraining)	2.20660 1	1.0	7.194740	0.009142
C(Internships):C(PlacementTraini ng)	0.23839 8	2.0	0.388655	0.679443

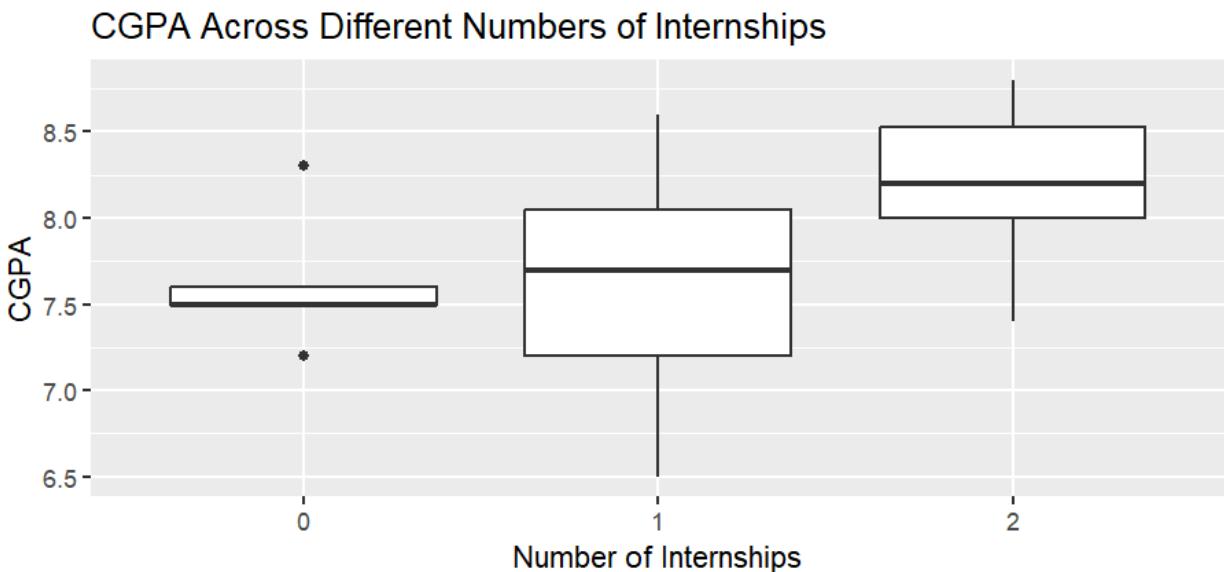
R

```
> #-----ANOVA-----
> anova_one_way <- aov(CGPA ~ Internships, data = data)
```

```

>
> summary(anova_one_way)
      Df Sum Sq Mean Sq F value    Pr(>F)
Internships   1  4.163   4.163   12.05 0.000873 ***
Residuals    73 25.217   0.345
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> anova_two_way <- aov(CGPA ~ Internships * PlacementTraining, data = data)
>
> summary(anova_two_way)
      Df Sum Sq Mean Sq F value    Pr(>F)
Internships           1  4.163   4.163   12.819 0.000624 ***
PlacementTraining     1  2.100   2.100   6.468 0.013166 *
Internships:PlacementTraining 1  0.058   0.058   0.178 0.674047
Residuals            71 23.058   0.325
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> ggplot(data, aes(x=factor(Internships), y=CGPA)) +geom_boxplot()
+labs(title="CGPA Across Different Numbers of Internships", x="Number of Internships", y="CGPA")

```



Excel

One-Way ANOVA:

- Step 1: Enter data for each group in separate columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Single Factor".

Step 4: Select your data range and choose your alpha level.

Step 5: Click OK and interpret the results.

Two-Way ANOVA:

Step 1: Organize your data with factors in columns and response variable in another column.

Step 2: Go to the Data tab and click on "Data Analysis".

Step 3: Select "Anova: Two-Factor With Replication".

Step 4: Enter your data range, specify the number of samples per group, and set alpha level.

Step 5: Click OK and interpret the results.

SPSS

Steps:

One-Way ANOVA:

1. Go to Analyze > Compare Means > One-Way ANOVA.
2. Select the dependent variable and the factor.

Two-Way ANOVA:

1. Go to Analyze > General Linear Model > Univariate.
2. Select the dependent variable and two independent variables.

ANOVA

CGPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16.331	43	.380	.902	.628
Within Groups	13.049	31	.421		
Total	29.380	74			

ANOVA Effect Sizes^{a,b}

		Point Estimate	95% Confidence Interval	
			Lower	Upper
CGPA	Eta-squared	.556	.000	.278
	Epsilon-squared	-.060	-1.387	-.723
	Omega-squared Fixed-effect	-.059	-1.344	-.706
	Omega-squared	-.001	-.014	-.010
	Random-effect			

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

Practical Output:

1. **Main Effect of Internships (C(Internships)):**

- **Sum of Squares:** 4.512780
- **Degrees of Freedom:** 2.0
- **F-value:** 7.357082
- **p-value:** 0.001270
- **Interpretation:** The p-value is 0.001270, which is less than the typical alpha level of 0.05. This indicates that the number of internships has a statistically significant effect on the dependent variable. The F-value of 7.357082 further supports this significance.

2. **Main Effect of PlacementTraining (C(PlacementTraining)):**

- **Sum of Squares:** 2.206601
- **Degrees of Freedom:** 1.0
- **F-value:** 7.194740
- **p-value:** 0.009142
- **Interpretation:** The p-value is 0.009142, which is also less than the alpha level of 0.05. This suggests that PlacementTraining has a statistically significant effect on the dependent variable. The F-value of 7.194740 supports this conclusion.

3. **Interaction Effect (C(Internships)**

(PlacementTraining)):

- **Sum of Squares:** 0.238398

- **Degrees of Freedom:** 2.0
 - **F-value:** 0.388655
 - **p-value:** 0.679443
 - **Interpretation:** The p-value is 0.679443, which is much greater than the alpha level of 0.05. This indicates that there is no statistically significant interaction effect between Internships and PlacementTraining on the dependent variable. The F-value of 0.388655 is also very low, further suggesting the lack of a significant interaction.
4. **Residual:**
- **Sum of Squares:** 21.162049
 - **Degrees of Freedom:** 69.0
 - **Interpretation:** The residual sum of squares represents the variation in the dependent variable that is not explained by the model. The degrees of freedom here reflect the total number of observations minus the number of parameters estimated in the model.

Overall Interpretation:

- Both "Internships" and "PlacementTraining" individually have significant effects on the dependent variable, as indicated by their low p-values.
- There is no significant interaction effect between "Internships" and "PlacementTraining," suggesting that the effect of one factor on the dependent variable does not depend on the level of the other factor.
- The residual sum of squares indicates the amount of unexplained variance after accounting for the main effects and interaction effect.

CASE STUDY 2 - Marketing

Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.linear_model import LinearRegression
```

```

import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import ttest_ind, ttest_rel, f_oneway
model = ols('Hours Marketing~C(Room Type) + C(Incentive Received) + C(Room Type):C(Incentive Recieved)', data=data).fit()
anova_two_way = sm.stats.anova_lm(model, typ=2)
anova_two_way

```

R

```

> #-----ANOVA-----
> anova_one_way <- aov(market ~ Incentive, data = data)
>
> summary(anova_one_way)
      Df Sum Sq Mean Sq F value Pr(>F)
Incentive     1   406.3   406.3   455.8 <2e-16 ***
Residuals    73    65.1     0.9
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> anova_two_way <- aov(market ~ Incentive * data$Room.Type, data = data)
>
>
> summary(anova_two_way)
      Df Sum Sq Mean Sq F value Pr(>F)
Incentive          1   406.3   406.3 469.031 <2e-16 ***
data$Room.Type     1     1.6     1.6   1.887  0.174
Incentive:data$Room.Type 1     1.9     1.9   2.240  0.139
Residuals         71    61.5     0.9
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Excel

One-Way ANOVA:

- Step 1: Enter data for each group in separate columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Single Factor".
- Step 4: Select your data range and choose your alpha level.
- Step 5: Click OK and interpret the results.

Two-Way ANOVA:

- Step 1: Organize your data with factors in columns and response variable in another column.
 Step 2: Go to the Data tab and click on "Data Analysis".
 Step 3: Select "Anova: Two-Factor With Replication".
 Step 4: Enter your data range, specify the number of samples per group, and set alpha level.
 Step 5: Click OK and interpret the results.

SPSS

Steps:

One-Way ANOVA:

1. Go to Analyze > Compare Means > One-Way ANOVA.
2. Select the dependent variable and the factor.

Two-Way ANOVA:

1. Go to Analyze > General Linear Model > Univariate.
2. Select the dependent variable and two independent variables.

ANOVA

HoursMarketing

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	519.447	70	7.421	6.596	.038
Within Groups	4.500	4	1.125		
Total	523.947	74			

ANOVA Effect Sizes^{a,b}

		95% Confidence Interval		
		Point Estimate	Lower	Upper
HoursMarketing	Eta-squared	.991	.000	.942
	Epsilon-squared	.841	-17.500	-.067

Omega-squared Fixed-effect	.839	-14.000	-.066
Omega-squared	.069	-.014	-.001
Random-effect			

- a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.
- b. Negative but less biased estimates are retained, not rounded to zero.

Practical Output

One-way ANOVA (Market vs Incentive): F value = 455.8, p < 2e-16

Two-way ANOVA (Market vs Incentive and Room Type): Incentive: F = 469.031, p < 2e-16
 Room Type: F = 1.887, p = 0.174 Interaction: F = 2.240, p = 0.139

Interpretation:

1. Incentive has a highly significant effect on market performance.
2. Room Type and its interaction with Incentive are not significant.

Practical Applications:

1. Focus on incentive strategies to improve market performance.
2. Incentive programs can be applied uniformly across different room types.
3. Develop a tiered incentive system to maximize market performance.
4. Consider other factors beyond room type that might interact with incentives.

CASE STUDY 3 - Operations

Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import ttest_ind, ttest_rel, f_oneway
```

```

model = ols('CGPA ~ C(Food Preference) + C(PlacementTraining) + C(Food
Preference):C(PlacementTraining)', data=data).fit()
anova_two_way = sm.stats.anova_lm(model, typ=2)
anova_two_way

```

R

```

> #-----ANOVA-----
> anova_one_way <- aov(CGPA ~ data$Food.Preference, data = data)
>
> summary(anova_one_way)
      Df Sum Sq Mean Sq F value Pr(>F)
data$Food.Preference  1  0.111  0.1108   0.276  0.601
Residuals            73 29.269  0.4009

> anova_two_way <- aov(CGPA ~ data$Food.Preference* data$Room.Type, data = data)
>
>
> summary(anova_two_way)
      Df Sum Sq Mean Sq F value Pr(>F)
data$Food.Preference      1  0.754  0.7543  1.908  0.172
data$Room.Type            1  0.528  0.5280  1.335  0.252
data$Food.Preference:data$Room.Type 1  0.024  0.0242  0.061  0.805
Residuals                71 28.074  0.3954

```

Excel

One-Way ANOVA:

- Step 1: Enter data for each group in separate columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Single Factor".
- Step 4: Select your data range and choose your alpha level.
- Step 5: Click OK and interpret the results.

Two-Way ANOVA:

- Step 1: Organize your data with factors in columns and response variable in another column.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Two-Factor With Replication".
- Step 4: Enter your data range, specify the number of samples per group, and set alpha level.
- Step 5: Click OK and interpret the results.

SPSS

Steps:

One-Way ANOVA:

1. Go to Analyze > Compare Means > One-Way ANOVA.
2. Select the dependent variable and the factor.

Two-Way ANOVA:

1. Go to Analyze > General Linear Model > Univariate.
2. Select the dependent variable and two independent variables.

ANOVA

CGPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25.935	67	.387	.787	.722
Within Groups	3.445	7	.492		
Total	29.380	74			

ANOVA Effect Sizes^{a,b}

		Point Estimate	95% Confidence Interval	
			Lower	Upper
CGPA	Eta-squared	.883	.000	.450
	Epsilon-squared	-.240	-9.571	-4.809
	Omega-squared Fixed-effect	-.236	-8.375	-4.459

Omega-squared	-.003	-.014	-.012
Random-effect			

- a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.
- b. Negative but less biased estimates are retained, not rounded to zero.

Practical Output

One-way ANOVA (CGPA ~ Food Preference):

- F-value: 0.276, p-value: 0.601 (not significant)

Two-way ANOVA (CGPA ~ Food Preference * Room Type):

- All factors and interactions are not significant (p-values > 0.05)

Practical Applications:

1. Academic Performance: Neither food preference nor room type significantly affects CGPA.
2. Student Support: Focus academic support strategies on other factors more likely to influence CGPA.
3. Campus Services: While these factors don't affect CGPA, they may still be important for student satisfaction. Continue providing diverse options in both dining and housing.

CASE STUDY 4 - Finance

Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy.stats import ttest_ind, ttest_rel, f_oneway
model = ols('Utilities ~ C(Groceries) + C(Food Preference) + C(Groceries):C(Food Preference)', data=data).fit()
anova_two_way = sm.stats.anova_lm(model, typ=2)
anova_two_way
```

R

```
#-----ANOVA-----
> anova_one_way <- aov(groceries ~ data$Food.Preference, data = data)
>
> summary(anova_one_way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Food.Preference	1	86433	86433	2.023	0.159
Residuals	73	3118816	42724		

```
> anova_two_way <- aov(data$Groceries ~ data$Food.Preference* data$Utilities)
> summary(anova_two_way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Food.Preference	1	86433	86433	1.983	0.163
data\$Utilities	1	10050	10050	0.231	0.633
data\$Food.Preference:data\$Utilities	1	13984	13984	0.321	0.573
Residuals	71	3094783	43588		

Excel

One-Way ANOVA:

- Step 1: Enter data for each group in separate columns.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Single Factor".
- Step 4: Select your data range and choose your alpha level.
- Step 5: Click OK and interpret the results.

Two-Way ANOVA:

- Step 1: Organize your data with factors in columns and response variable in another column.
- Step 2: Go to the Data tab and click on "Data Analysis".
- Step 3: Select "Anova: Two-Factor With Replication".
- Step 4: Enter your data range, specify the number of samples per group, and set alpha level.
- Step 5: Click OK and interpret the results.

SPSS

Steps:

One-Way ANOVA:

1. Go to Analyze > Compare Means > One-Way ANOVA.

2. Select the dependent variable and the factor.

Two-Way ANOVA:

1. Go to Analyze > General Linear Model > Univariate.

2. Select the dependent variable and two independent variables.

ANOVA

Groceries

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2967986.167	67	44298.301	1.307	.382
Within Groups	237262.500	7	33894.643		
Total	3205248.667	74			

ANOVA Effect Sizes^{a,b}

		Point Estimate	95% Confidence Interval	
			Lower	Upper
Groceries	Eta-squared	.926	.000	.655
	Epsilon-squared	.217	-9.571	-2.647
	Omega-squared Fixed-effect	.215	-8.375	-2.522
	Omega-squared	.004	-.014	-.011
	Random-effect			

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

Practical Output

One-way ANOVA (Groceries ~ Food Preference):

- F-value: 2.023, p-value: 0.159 (not significant)

Two-way ANOVA (Groceries ~ Food Preference * Utilities):

- All factors and interactions are not significant (p-values > 0.05)

Practical Applications:

1. Spending Patterns: Food preferences and utilities costs don't significantly influence grocery spending.
2. Budgeting Advice: When advising students on budgeting, focus on individual habits rather than these broader categories.
3. Campus Services: Consider investigating other factors that might influence grocery spending to better support students' financial planning.