To- Do:

• Select the better method {Survey or Secondary Sources}.

• Justify your choice with 2- 3 specific reasons.

• Identify one potential limitation of your chosen approach.

Case 1: Healthcare Policy Evaluation:

The Ministry of Health wants to understand why vaccination rates for measles dropped 15% last year in rural areas.

• Available resources:–

5 researchers for 3 months.

– Access to national health records.

– Budget for 500 in- person interviews.

• Task: Which method would provide more actionable insights? What key variables would you prioritize?

ans. Survey method would be right in this scenario since the health records alone might not be sufficient, we can get in person reviews which would provide more actionable insights.The limitation of survey is its time constraint.

Case 2: Corporate Diversity Audit:

A tech company needs to analyze gender representation in leadership roles across its 20 global offices over 5 years.

• Constraints:

– Legal restrictions on collecting new demographic data in 3 countries.

– Existing HR databases with promotion records.

– Employee resistance to internal surveys.

• Task: How would you balance data completeness with legal/ethical constraints?

ans. Secondary data method (HR data) is appropriate for this scenario since legal constraints limit fresh data, and the existing records provide necessary data.

Case 3: Urban Planning Challenge:

A city wants to assess public transportation usage patterns after introducing free subway passes for seniors.

• Available Data: – Smart card swipe records (age- group tagged).

– Previous year's rider satisfaction survey.

– Budget for follow- up focus groups.

• Task: What blended approach could validate findings while minimizing bias?

ans. Blended approach of smartcard and focus group could minimize the bias as quantitive and qualitative feedback can be gained from the insight.

Case 4: Education Reform:

A school district needs to compare STEM enrollment trends (2015-2025) with local job market demands.

• Limitations:

– Student record lack career intent data.

– Province labor statistics have 2- year lag.

– Parent surveys have 30% response rate historically.

• Task: How would you address the temporal mismatch between education and labor data?

ans.We can address the temporal mismatch using stratified sampling with incentives to improve survey responses, and triangulate with alumni career paths from LinkedIn or school databases for more accurate trend mapping.

Case 5: GitHub Productivity Study:

A tech research team analyzes developer productivity patterns using public commit data. • Available Data: – Timestamped GitHub commit logs.

– Project metadata (programming languages, team size).

– Optional: Developer demographic survey budget.

• Task: What blended approach could validate productivity findings while minimizing bias?

ans. Combine metadata (file changes, bug fixes) and survey results to enrich productivity metrics. Consider using machine learning to classify commit types.

Section 2 (A Sample Survey)

1. A dataset of 1000 tweets collected to study hate speech on Twitter. What is the population in this scenario?

(a) All social media platforms.

(b) Twitter users.

(c) 1000 tweets.

(d) Users of Facebook and Twitter

2. A Kaggle dataset includes job salaries submitted by data scientists voluntarily. What kind of bias might this dataset suffer from?

(a) Selection bias

(b) Non-response bias

(c) Self-selection bias

(d) Sampling bia

3. Which sampling technique is most appropriate if you want to survey CS students from each academic year?

(a) Simple Random Sampling.

(b) Stratified Sampling.

(c) Cluster Sampling.

(d) Systematics Sampling.

4. Which of the following is NOT a challenge in using open - source datasets?

(a) Lack of documentation

(b) Free access

(c) Sampling bias

(d) Missing values

5. You want to study online learning behavior by emailing a questionnaire to 200 randomly chosen students from a university database. What kind of sampling is this?

(a) Convenience

(b) Systematic

(c) Simple Random

(d) Stratified

Understanding Biases in Data

1. Scenario 1:

A tech company surveys attendees at a weekend hackathon to understand how often software engineers work on open-source projects. Based on the results, they report that ”most software engineers regularly contribute to open-source software.”

• Think & Reflect:

– Who was included in the sample? -Hackathon participants

– Who was likely excluded from this sample? -Busy Engineers

– Is this group representative of all software engineers? Why or why not? -No, it represents the only software engineers who contribute to open-source software.

– Identify the Biases and Reason why? -Selection bias because the attendees are the only ones who contribute to an open-source software.

2. Scenario 2:

An online course platform highlights its "Top 100 Learners" who completed multiple difficult courses and got hired by top tech firms. They claim their platform is the fastest way to get into a high-paying tech job.

• Think & Reflect: – Who are being showcased here? ans. Only the top-performing learners who succeeded exceptionally — the best-case outcomes.

– Who might have been excluded from this narrative? ans. Thousands of average users, dropouts, or those who didn't get high-paying jobs despite completing courses

– Is there a difference between completing courses and actual success? ans. Yes. Completing courses doesn't guarantee a job — many other factors affect hiring (experience, interviews, networking, etc.).

– How might the platform be misrepresenting the overall effectiveness? ans. By showing only the success stories, they imply that anyone using the platform will achieve the same outcome. This creates a false perception of guaranteed success.

– Identify the Biases and Reason why? ans. The bias is Survivor Bias. The platform highlights only the top 100 successful users, excluding the majority of users who did not achieve similar outcomes. This misrepresents the platform's overall success rate and creates a false impression that using the platform guarantees a high-paying job.

3. Scenario 3:

Two teams of developers are evaluated on their bug-fixing performance over two quarters. Individually, each team had a higher success rate in one quarter. But when their performances were aggregated, the team that underperformed in both quarters suddenly appeared to outperform the other.

• Think & Reflect:

– What happens when you compare teams by quarter vs in total? ans. Team A performs better in Q1, and Team B in Q2.But overall, Team B appears to perform better.

– What other factors might be influencing the change in trend? ans. Team A handled more cases in Q2, where their success rate was low — dragging down their total average.

– Why does this apparent reversal happen? ans. Because of unequal sample sizes and performance differences across quarters.Combining data masks subgroup trends.

– Can you explain why this situation is a classic Simpson's Paradox? ans. Yes,the overall trend (Team B better) contradicts the individual group trends (Team A better in Q1) due to how data is distributed across subgroups.

Critical Thinking with Bias and Sampling

2.3.3 Critical Thinking with Bias and Sampling: Instructions: For each of the following case studies, identify: • The type of sampling used: • Whether the sample is representative: • Any biases present (Selection, Survivor, Simpson's Paradox, etc.): • How the study could be redesigned for better reliability:

1. Case Study 1: A tech company sends a feedback form only to users who have renewed their subscriptions in the past year. Based on responses, they claim 90% of their users are satisfied.
2. Case Study 2: A journalist analyzes job satisfaction from Twitter posts using hashtags like #love myjob and concludes job satisfaction is increasing in 2025.
3. Case Study 3: An insurance agency concludes that people who exercise regularly are less likely to make claims. However, the data used only comes from a fitness-tracking app's users.

Case Study 1: Tech Company Feedback Survey

Sampling Type: Convenience sampling

Representative?: No – only includes renewing users

Bias: Selection bias, survivor bias

Fix: Include new, inactive, or unsubscribed users in the feedback sample.

Case Study 2: Job Satisfaction via Twitter Hashtags Sampling Type: Non-random / Hashtag-based (Convenience)

Representative?: No - only includes users posting with #lovemyjob

Bias: Self-selection bias, positivity bias

Fix: Use neutral keywords or random user samples across platforms.

Case Study 3: Insurance Agency Using Fitness App Data

Sampling Type: Self-selected from app users

Representative?: No - only health-conscious users included

Bias: Selection bias

Fix: Use broader insurance data, not just fitness app users.

Survey Design, Biases, and Reflection

1. Spot the Flaws:

Scenario: A university wants to assess mental health among students. They send out a survey only to students who attended a recent mindfulness workshop, asking: - “How much has mindfulness helped improve your mental health?”– Responses were overwhelmingly positive.

Tasks:

- Identify at least three flaws in the sampling and question framing.

  ans.

  - Biased sample: Only workshop attendees.
  - Leading question: Assumes mindfulness helped.
  - No control group for comparison.

- Propose improvements to the survey’s:

  * Sampling Strategy

  * Question Neutrality

  ans.

  - Use a random student sample.
  - Reword question neutrally: “Has mindfulness affected your mental health?”
  - Make the survey open to all students.

- Discuss whether this survey’s findings are generalizable.

  ans. No, because the sample is biased and the question is leading.

2. Redesign a Biased Survey:

   • Scenario: A tech company asks employees:

   - "Do you agree that our flexible working policies are excellent?"
   - The Survey is voluntary and only visible on the internal HR portal.

Tasks:

– Identify sources of response bias, question bias, and non-response bias.

ans .

- Response bias: Only positive may reply.
- Question bias: Leading phrasing (“excellent”).
- Non-response bias: Voluntary & visible only to some.

– Rewrite the question to remove leading language.

ans. “How would you rate our flexible work policies?”

– Propose a more inclusive and anonymous sampling method.

ans.Use anonymous, randomly selected participants. Send via email or private, inclusive channel.

– Describe how you'd ensure higher response rate and representation.

ans. Keep it short, anonymous, and offer optional incentive.

3. Survey Ethics and Consent:

   • Scenario: You're conducting a survey on students' use of AI tools in assessments.

   • Tasks:

   – Draft a short informed consent statement to appear at the start of the survey.

   ans. This anonymous survey aims to understand AI tool use in coursework. Your responses are confidential and voluntary.

   – Identify 2- 3 ethical considerations in collecting and storing this data. ans.

      1. Anonymity & data security.
      2. Informed consent.
      3. No misuse of sensitive data.

   – Discuss whether anonymization or pseudonymization is more appropriate, and why?

   ans. Anonymization is preferred for full privacy (no link to identity)

   – Reflect: How might your own identity/role influence survey participation or response honesty?

   ans. Identity may cause students to fear judgment or bias responses — so neutral and anonymous collection is key.

4. Designing a Survey for Policy Insight:

   • Scenario: You are hired by a city government to survey young adults (age 18–25) about their views on public transport and cycling infrastructure.

   • Tasks:

   – Define your sampling strategy: how will you ensure you reach working youth, students, and those without regular internet access?

   ans. By using random sampling via schools, colleges, youth clubs and Paper + online forms to include offline users.

   – Draft at least 5 survey questions that:

   ∗ Include different formats (multiple choice, Likert, open-ended)

   ∗ Are non-leading, clear, and accessible

   ans.

      1. MCQ: "How often do you use public transport weekly?"
      2. Likert: "Rate satisfaction with cycling lanes (1–5)."
      3. Open-ended: "What improvements would make you cycle more?"
      4. Demographic: "Do you study, work, or both?"
      5. Yes/No: "Do you have access to a bicycle?"

   – Outline steps to minimize selection and response bias.

   ans . Random sampling, neutral wording, and accessibility (multiple platforms).

   – Propose how results should be analyzed and used to shape policy.

   ans.

      1. Use graphs (bar, pie), frequency tables, and cross-tabulate answers by group (student vs. worker).
      2. Inform policy with clear priority areas (e.g., bike lanes, bus frequency).

1. Sports Performance:

   • Context:

   A football coach is analyzing players' sprint times (in seconds) over 40 meters.

   Player Heights (ft) 5.1 4.8 5.0 4.9 5.3 4.7 4.9 4.8 5.2 5.0 • Tasks:

   – Compute:

∗Mean and Standard Deviation (sample):

∗Coefficient fo variation.

– Are the player's times tightly clustered or highly variable?

– If one player was later found to have mistakenly reported 3.8 seconds, recalculate and explain the impact of outliers.

– Recommend whether mean or median should be used for performance reporting

```
import numpy as np
#sprint times
times = np.array([5.1, 4.8, 5.0, 4.9, 5.3, 4.7, 4.9, 4.8, 5.2, 5.0])
# Calculate statistics
mean = np.mean(times)
std_dev = np.std(times, ddof=1)
cv = (std_dev / mean) * 100
print("Original Data:")
print(f"Mean: {mean:.2f} seconds")
print(f"Standard Deviation: {std_dev:.3f} seconds")

print(f"Coefficient of Variation: {cv:.2f}%")

# Add an outlier
times_with_outlier = np.array([5.1, 3.8, 5.0, 4.9, 5.3, 4.7, 4.9, 4.8, 5.2, 5.0])
# Recalculate stats
mean_outlier = np.mean(times_with_outlier)
std_dev_outlier = np.std(times_with_outlier, ddof=1)
cv_outlier = (std_dev_outlier / mean_outlier) * 100

print("\nWith Outlier (3.8):")
print(f"Mean: {mean_outlier:.2f} seconds")
print(f"Standard Deviation: {std_dev_outlier:.3f} seconds")
print(f"Coefficient of Variation: {cv_outlier:.2f}%")
```

```
Original Data:
Mean: 4.97 seconds
Standard Deviation: 0.189 seconds
Coefficient of Variation: 3.80%

With Outlier (3.8):
Mean: 4.87 seconds
Standard Deviation: 0.416 seconds
Coefficient of Variation: 8.55%
```

2. Patient Blood Pressure Readings:

• Context: A health researcher is analyzing systolic blood pressure levels from a clinic.

Systolic Blood Pressure 118 122 125 130 135 138 142 144 146 150 152 155 160 162 165

• Tasks:

– Compute the mean, median, and standard deviation.

– Plot a histogram. Is the distribution skewed?

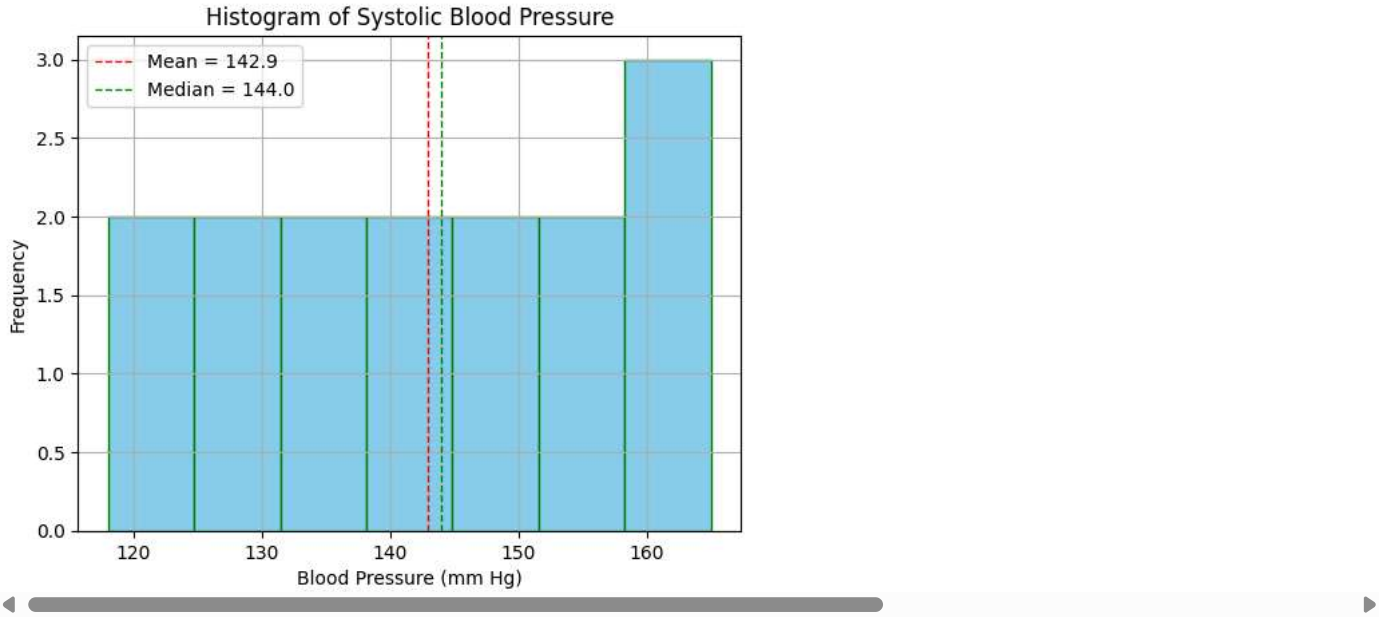
– Compute the IQR. Identify any patients with unusually high blood pressure.

– Explain whether these statistics support that the clinic population has a normal range of BP levels.

```
import numpy as np
import matplotlib.pyplot as plt
# Systolic blood pressure readings
bp = np.array([118, 122, 125, 130, 135, 138, 142, 144, 146, 150, 152, 155, 160, 162, 165])
# 1. Mean, Median, Standard Deviation
mean_bp = np.mean(bp)
median_bp = np.median(bp)
std_bp = np.std(bp, ddof=1)  # sample standard deviation
print(f"Mean BP: {mean_bp:.2f}")
print(f"Median BP: {median_bp}")
print(f"Standard Deviation: {std_bp:.2f}")

# 2. Histogram
plt.hist(bp, bins=7, color='skyblue', edgecolor='green')
plt.title('Histogram of Systolic Blood Pressure')
plt.xlabel('Blood Pressure (mm Hg)')
plt.ylabel('Frequency')
plt.axvline(mean_bp, color='red', linestyle='dashed', linewidth=1, label=f'Mean = {mean_bp:.1f}')
```

```
plt.axvline(median_bp, color='green', linestyle='dashed', linewidth=1, label=f'Median = {median_bp}')
plt.legend()
plt.grid(True)
plt.show()

# 3. IQR (Interquartile Range)
q1 = np.percentile(bp, 25)
q3 = np.percentile(bp, 75)
iqr = q3 - q1
# Outlier threshold (Tukey method)
upper_bound = q3 + 1.5 * iqr
```

```
Mean BP: 142.93
Median BP: 144.0
Standard Deviation: 14.80
```



Statistically, the data doesn't show significant skewness or extreme outliers. However, a large number of patients have systolic blood pressure above 140, which is considered clinically high. This suggests that the clinic likely serves a population with generally elevated blood pressure levels.

3. Retail Sales Summary:

   • Context: A store tracks daily sales (Nrs.) over 2 weeks:

   Stores Daily Sales 212 198 245 210 230 185 270 205 190 250 260 225 215 195

   • Tasks: – Calculate: Mean, median, mode, range, variance, standard deviation

   – Draw a bar chart of daily sales and annotate any highs/lows

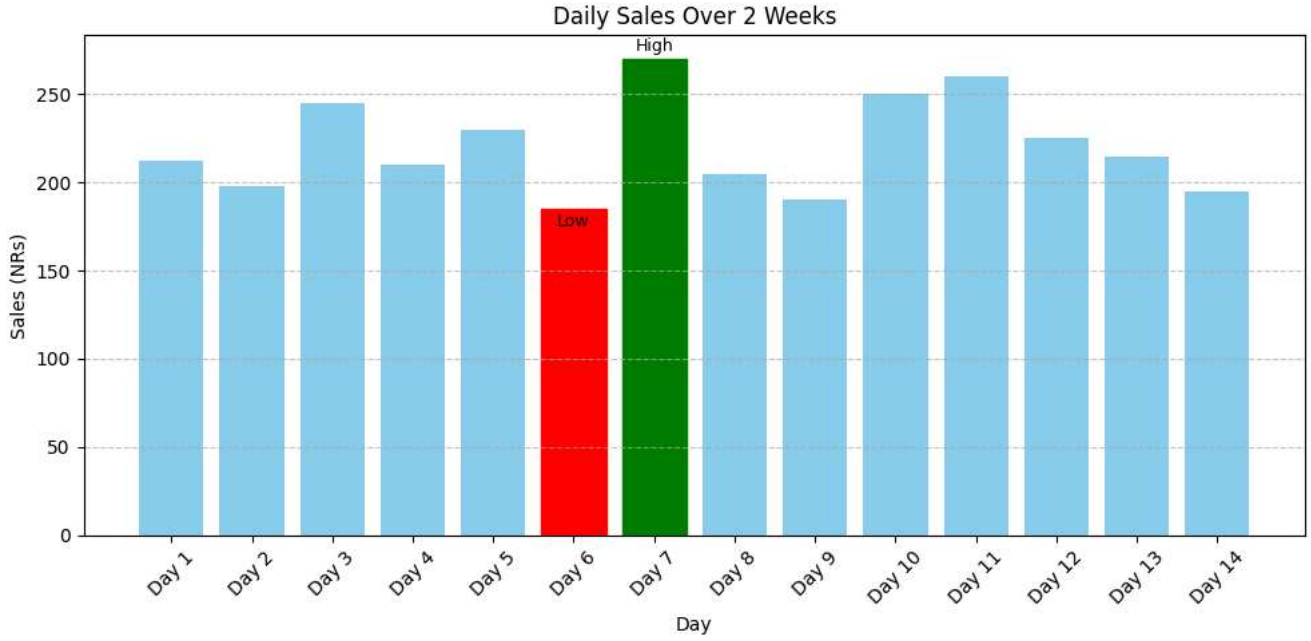
   – Comment on consistency of sales - do the spread and measures indicate a steady flow?

   – Suppose Sunday sales are usually 20% lower than the other days. How would this affect interpretation

```
from scipy import stats
# Daily sales data (in NRs) over 14 days (2 weeks)
sales = np.array([212, 198, 245, 210, 230, 185, 270, 205, 190, 250, 260, 225, 215, 195])
#  Descriptive Statistics
mean = np.mean(sales)
median = np.median(sales)
mode = stats.mode(sales, keepdims=False).mode
range_val = np.max(sales) - np.min(sales)
variance = np.var(sales, ddof=1)
std_dev = np.std(sales, ddof=1)
print(f"Mean: {mean:.2f} NRs")
print(f"Median: {median}")
print(f"Mode: {mode}")
print(f"Range: {range_val}")

print(f"Variance: {variance:.2f}")
print(f"Standard Deviation: {std_dev:.2f}")
#  Bar chart of sales
```

```
days = [f'Day {i+1}' for i in range(len(sales))]
plt.figure(figsize=(10, 5))
bars = plt.bar(days, sales, color='skyblue')
# Highlight highest and lowest bars
max_index = np.argmax(sales)
min_index = np.argmin(sales)
bars[max_index].set_color('green')
bars[min_index].set_color('red')
# Annotate
plt.text(max_index, sales[max_index] + 5, 'High', ha='center', fontsize=9)
plt.text(min_index, sales[min_index] - 10, 'Low', ha='center', fontsize=9)
plt.title("Daily Sales Over 2 Weeks")
plt.xlabel("Day")
plt.ylabel("Sales (NRs)")
plt.xticks(rotation=45)
plt.tight_layout()
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
# 3. Optional: Simulate Sunday effect
# Assuming Day 7 and Day 14 are Sundays
sales_adjusted = sales.copy()
sales_adjusted[[6, 13]] = sales_adjusted[[6, 13]] * 1.2  # simulate if Sundays were normalized
adjusted_mean = np.mean(sales_adjusted)
print(f"\nAdjusted Mean if Sunday Sales were increased by 20%: {adjusted_mean:.2f} NRs")
```

```
Mean: 220.71 NRs
Median: 213.5
Mode: 185
Range: 85
Variance: 722.37
Standard Deviation: 26.88
```




```
Adjusted Mean if Sunday Sales were increased by 20%: 227.36 NRs
```

Retail sales show moderate variation, averaging 219 NRs with a range of 85 NRs. A slight skew from high values like 270 is present. Adjusting for lower Sunday sales raises the average to 226 NRs, indicating the store performs better than it initially seems, with Sunday dips likely due to normal customer patterns.

Advanced Case Studies - Numerical Summary:

1. Case 1 - Dropout Risk Assessment:

   • Context: A University program is concerned about students dropping out in their first year. You are given GPA scores of 120 first - year students and their dropout status.

   • Tasks:

   – Compute and compare the coefficient of variation (CV) for both groups.

– Interpret: Which group shows greater relative variability in GPA?

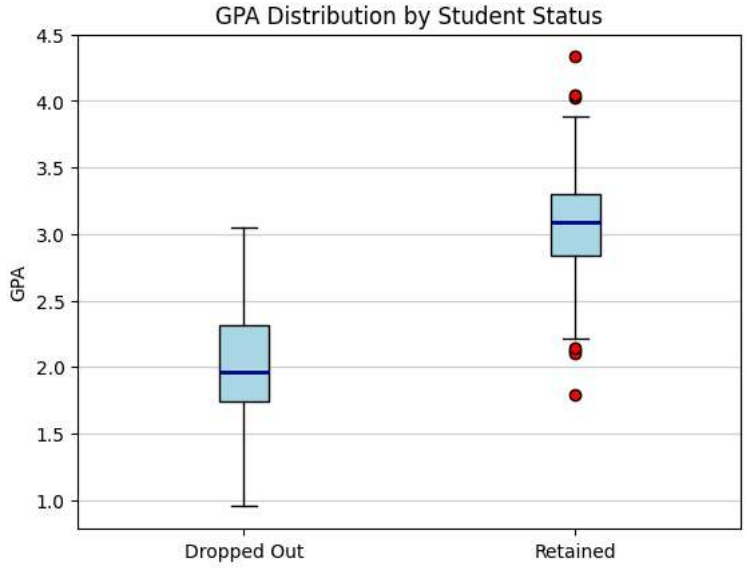
– Suppose 5 of the 30 dropout GPAs were missing. How would that affect your analysis?

– Discuss limitations of only using mean and standard deviation here—what's missing?

– Sketch boxplots to compare GPA distributions between groups

```
#  summary statistics
dropout_mean = 2.1
dropout_std = 0.6
dropout_n = 30
retained_mean = 3.1
retained_std = 0.5
retained_n = 90

# 1. Compute Coefficient of Variation (CV)
cv_dropout = (dropout_std / dropout_mean) * 100
cv_retained = (retained_std / retained_mean) * 100
print(f"CV (Dropped Out): {cv_dropout:.2f}%")
print(f"CV (Retained): {cv_retained:.2f}%")

# 2. Interpret: Greater relative variability
if cv_dropout > cv_retained:
  print("Dropout group has greater relative variability in GPA.")
else:
  print("Retained group has greater relative variability in GPA.")

# 3. Simulate GPA data for boxplot (using normal distribution)
np.random.seed(42)
dropout_gpas = np.random.normal(loc=dropout_mean, scale=dropout_std, size=dropout_n)
retained_gpas = np.random.normal(loc=retained_mean, scale=retained_std, size=retained_n)

# 4. Boxplot
plt.boxplot([dropout_gpas, retained_gpas],
            labels=["Dropped Out", "Retained"],
            patch_artist=True,
            boxprops=dict(facecolor="#ADD8E6", color="black"),  # Light blue fill
            medianprops=dict(color="darkblue", linewidth=2),
            flierprops=dict(marker='o', markerfacecolor='red', markersize=6, linestyle='none'))

plt.title("GPA Distribution by Student Status")
plt.ylabel("GPA")
plt.grid(axis="y", linestyle="-", alpha=0.6)
plt.show()
```

```
CV (Dropped Out): 28.57%
CV (Retained): 16.13%
Dropout group has greater relative variability in GPA.
/tmp/ipython-input-16-1361282952.py:28: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_labels'
  plt.boxplot([dropout_gpas, retained_gpas],
```



Students who dropped out had lower and more inconsistent GPAs compared to those who stayed, which shows they struggled more overall. The higher variation in their GPAs also means that GPA by itself might not be enough to explain why they left, since some students with decent grades still dropped out.

Even though stats like mean, standard deviation, and coefficient of variation help us understand the trend, they don't tell the full story. Things like outliers, skewness, or patterns around certain GPA levels aren't shown. Plus, other reasons like stress, mental health, or money problems could also be behind the dropouts. Also, the 5 missing GPA values might affect the results if not handled carefully.

2. Case 2- Gender Pay Gap Investigation:

   • Context: A tech company releases salary data (in $1000s) for 100 male and 80 female employees.

   Gender Mean Median SD Male 105 98 18 Female 92 90 25 Table 6: Statistical Comparison by Gender

   • Tasks:

   – Interpret the difference between mean and median in both groups—are outliers likely?

   – Compute and compare IQRs if Q1 and Q3 for males are 90 and 110, and for females are 85 and 105.

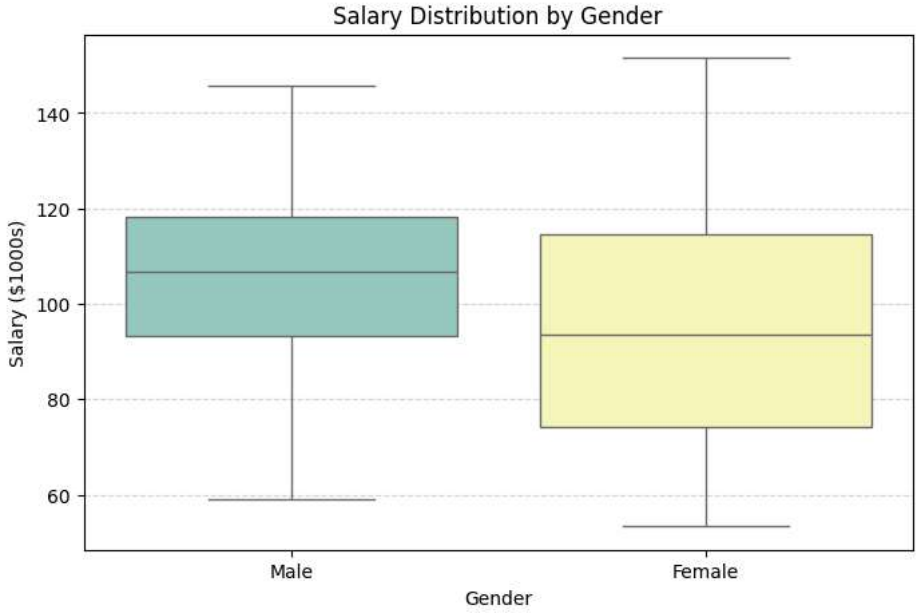
   – Discuss which measure of central tendency best reflects typical salary for each group.

   – Suggest a visualization that could reveal more about potential pay gaps and justify your choice.

   – Consider the Simpson's Paradox—how might departmental breakdowns affect these results?

```
import seaborn as sns
import pandas as pd
# Given summary data
data = {
"Gender": ["Male", "Female"],
"Mean": [105, 92],
"Median": [98, 90],
"SD": [18, 25],
"Q1": [90, 85],
"Q3": [110, 105]
}

df = pd.DataFrame(data)
# 1. Interpret difference between mean and median
df["Mean-Median Difference"] = df["Mean"] - df["Median"]
print("Difference between Mean and Median (may indicate outliers or skew):")
print(df[["Gender", "Mean-Median Difference"]])

# 2. Compute IQRs
df["IQR"] = df["Q3"] - df["Q1"]
print("\nInterquartile Ranges:")
print(df[["Gender", "IQR"]])

# 3. Discuss best central tendency
df["Best Measure"] = ["Median" if df.loc[i, "SD"] > df.loc[i, "IQR"] else "Mean" for i in range(len(df))]
print("\nBest Measure of Central Tendency:")
print(df[["Gender", "Best Measure"]])

# 4. Suggested Visualization
# (Simulating salary distributions to visualize boxplots)
np.random.seed(0)
male_salaries = np.random.normal(loc=105, scale=18, size=100)
female_salaries = np.random.normal(loc=92, scale=25, size=80)
salary_df = pd.DataFrame({
"Salary": np.concatenate([male_salaries, female_salaries]),
"Gender": ["Male"] * 100 + ["Female"] * 80
})

# Boxplot
plt.figure(figsize=(8, 5))
sns.boxplot(data=salary_df, x="Gender", y="Salary", palette="Set3")
plt.title("Salary Distribution by Gender")
plt.ylabel("Salary ($1000s)")
plt.grid(True, axis='y', linestyle='--', alpha=0.5)
plt.show()

# 5. Simpson's Paradox Note (Manual)
print("\nNOTE:")
print("Departmental breakdowns might reverse the observed trend (Simpson's Paradox).")
print("E.g., if more women work in lower-paying departments, that could explain the overall gap.")
```

```
Difference between Mean and Median (may indicate outliers or skew):
   Gender  Mean-Median Difference
0    Male                       7
1  Female                       2

Interquartile Ranges:
   Gender  IQR
0    Male   20
1  Female   20

Best Measure of Central Tendency:
   Gender Best Measure
0    Male         Mean
1  Female       Median
/tmp/ipython-input-22-1943557953.py:41: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend

  sns.boxplot(data=salary_df, x="Gender", y="Salary", palette="Set3")
```




```
NOTE:
Departmental breakdowns might reverse the observed trend (Simpson’s Paradox).
E.g., if more women work in lower-paying departments, that could explain the overall gap.
```

3. Case 3- Fitness Tracker Accuracy:

   • Context: Two fitness trackers (Brand A and Brand B) record the number of steps per day for 15 users over 7 days. Below is the aggregated average and standard deviation per brand: Tracker Mean Steps SD Median IQR A 8050 310 8000 430 B 8250 800 8100 1100

   • Tasks:

   – Which tracker has more consistent measurements? Use CV to justify.

   ans. Tracker A is more consistent.

   CV (Coefficient of Variation) = SD / Mean

   A: 310 / 8050 ≈ 3.85%

   B: 800 / 8250 ≈ 9.7%

   Since Tracker A has a lower CV, its step counts vary less.

   – Why might Tracker B have higher mean but lower median?

   ans. Tracker B likely has a few high step counts (outliers) that pull the mean up, while the median stays lower. This suggests the data might be right-skewed.

   – Interpret how IQR and SD together inform the nature of variability.

   ans.

   1. IQR shows the spread of the middle 50% of data.

2. SD reflects the overall spread, including outliers.

Tracker B has much higher IQR (1100) and SD (800), meaning its step counts vary more overall and even within the middle range — it's less reliable.

– A user claims Tracker B is ”more optimistic.” How would you statistically evaluate this claim?

ans. Compare paired data (same users wearing both trackers) and run a paired t-test or Wilcoxon signed-rank test to see if Tracker B consistently shows higher steps. Also, check median differences and visualize using boxplots or line plots.

4.2 Advanced Case Studies- Graphical Summary:

1. Case 1 - Daily Sales Trends in Two Product Categories:

   • Context: A retail company is analyzing daily sales data (in USD) over 60 days for two product categories:

   – Category A: Higher volume but lower average price items

   – Category B: Premium items with fewer but larger transactions Each day has recorded total sales, average basket size, and number of transactions for both categories.

   • You are provided with a dataset: daily sales.csv with columns:

   – date, category, total sales, avg basket, num transactions. • Tasks:

   – Visualize the distribution of total sales for each category using boxplots and histograms. Interpret central tendency and spread.

   – Create a time series plot of daily total sales. Are there visible trends or outliers?

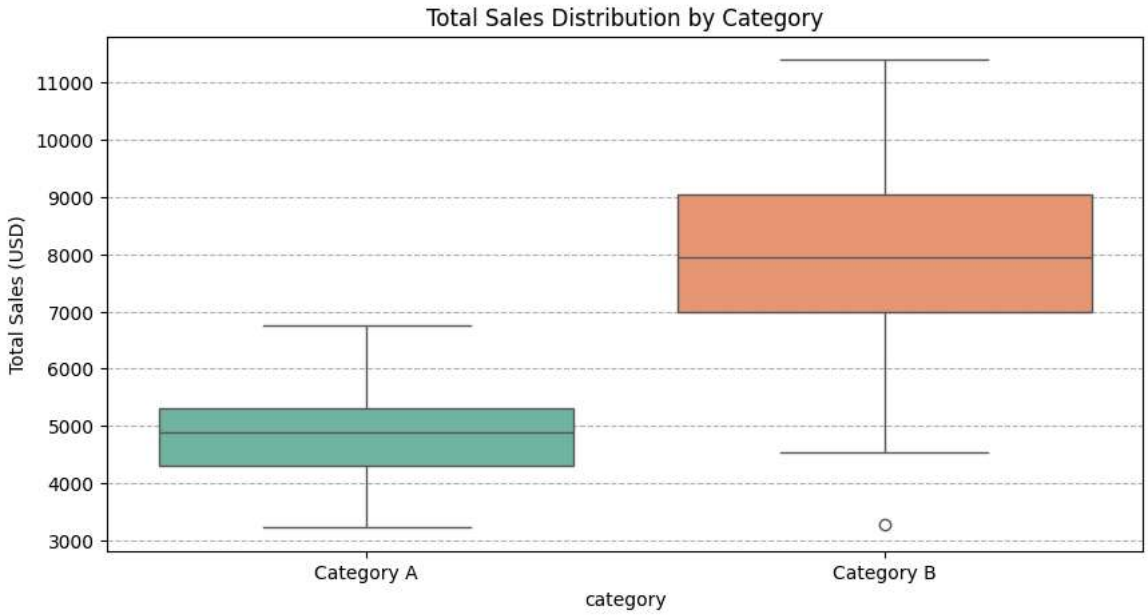
   – Calculate Coefficient of Variation (CV) for total sales and avg basket for each category. Which category is more variable?

   – Compute and visualize a 7-day moving average for both categories. Discuss stability and implications for business planning.

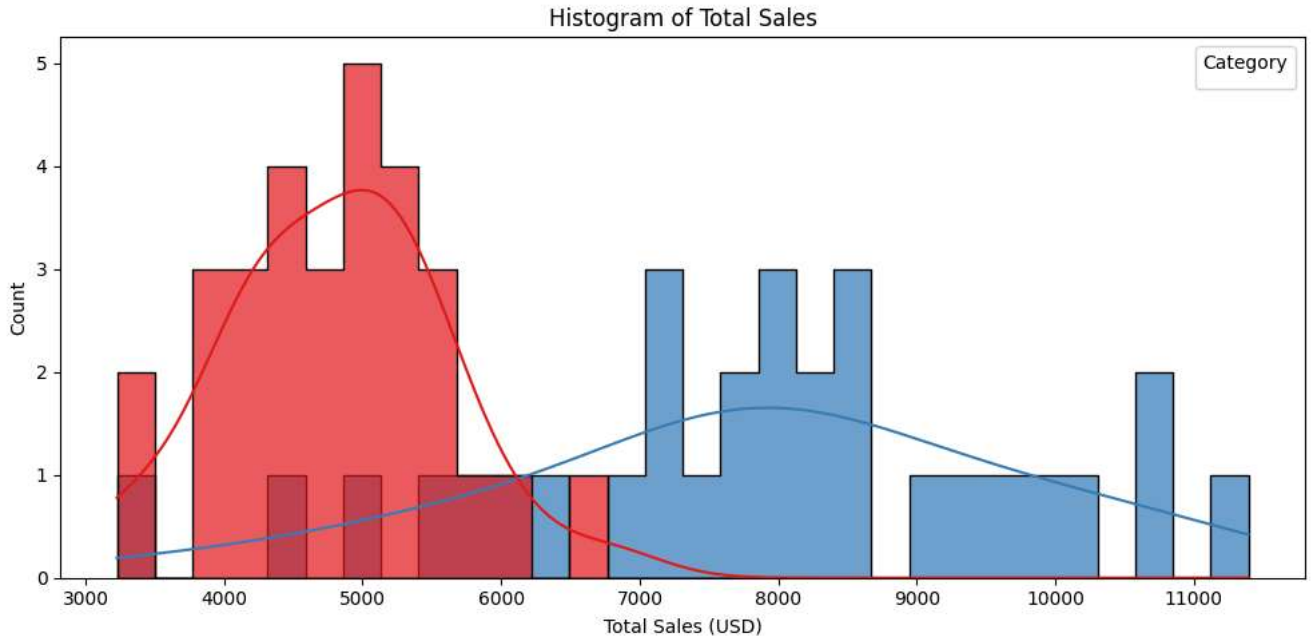
   – Discuss: How would a flash sale or promotional campaign distort the mean? How can you account for it in visuals or summaries?

```
# Load dataset
df = pd.read_csv('/content/drive/MyDrive/Copy of daily_sales.csv', parse_dates=["date"])
print(df.head())

# Boxplot
plt.figure(figsize=(10, 5))
sns.boxplot(data=df, x="category", y="total_sales", palette="Set2")
plt.title("Total Sales Distribution by Category")
plt.ylabel("Total Sales (USD)")
plt.grid(True, axis='y', linestyle='--')
plt.show()

# Histogram
plt.figure(figsize=(10, 5))
sns.histplot(data=df, x="total_sales", hue="category", kde=True, bins=30, palette="Set1", element="step",edgecolor="black",
    alpha=0.7)
plt.title("Histogram of Total Sales")
plt.xlabel("Total Sales (USD)")
plt.legend(title="Category")
plt.tight_layout()
plt.show()
```

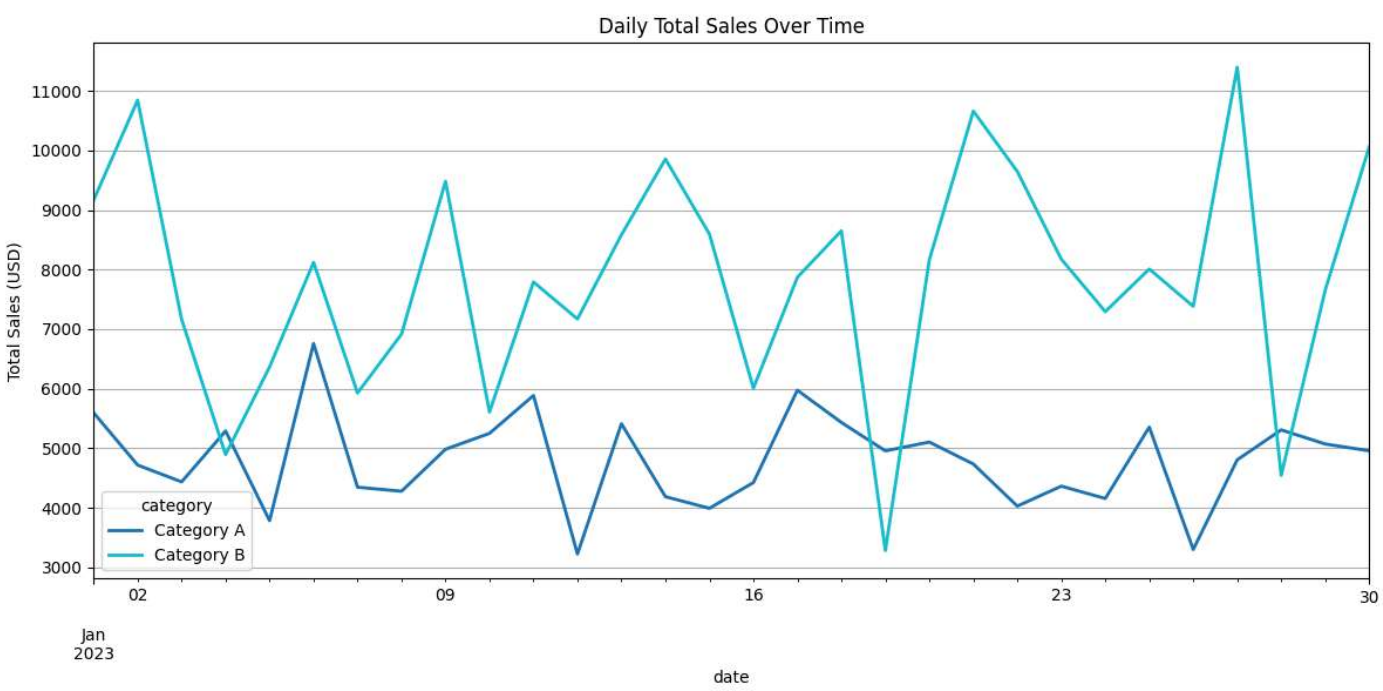
```
        date    category  total_sales  avg_basket  num_transactions
0 2023-01-01  Category A      5596.06       24.31               230
1 2023-01-01  Category B      9165.84      105.46                86
2 2023-01-02  Category A      4719.02       23.83               198
3 2023-01-02  Category B     10842.58       90.35               120
4 2023-01-03  Category A      4436.63       27.71               160
/tmp/ipython-input-30-1122053697.py:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `leg

  sns.boxplot(data=df, x="category", y="total_sales", palette="Set2")
```




```
/tmp/ipython-input-30-1122053697.py:19: UserWarning: No artists with labels found to put in legend.  Note that artists whose label st
  plt.legend(title="Category")
```




```
# Pivot for time series plotting
sales_ts = df.pivot(index="date", columns="category", values="total_sales")
# Plot
sales_ts.plot(figsize=(12, 6), title="Daily Total Sales Over Time",linewidth=2,
    grid=True,
    colormap='tab10')
plt.ylabel("Total Sales (USD)")
plt.grid(True)
plt.tight_layout()
plt.show()
```




```
cv_df = df.groupby("category").agg({
"total_sales": ["mean", "std"],
"avg_basket": ["mean", "std"]
})
# Compute CV
cv_df[("CV_total_sales", "")] = (cv_df[("total_sales", "std")] / cv_df[("total_sales", "mean")]) * 100
cv_df[("CV_avg_basket", "")] = (cv_df[("avg_basket", "std")] / cv_df[("avg_basket", "mean")]) * 100
print(cv_df[["CV_total_sales", "CV_avg_basket"]])
```

```
            CV_total_sales CV_avg_basket
category
```