Individual Assignment – 3: Descriptive statistics

Due on Saturday, September 27, 2025, at 8 pm.

Read first:

- **Points:** 70
  - ✓ Python file: 30 points
  - ✓ Word file: 40 points
- **Submission files:**
  - ✓ Deliverables (submit only these 2 items):
    1. Python file name: Name "IA_3_IST615_Fall25_FirstName_LastName.ipynb".
       - Write your code on a Jupyter Notebook after you click Jupyter Lab on Anaconda Navigator.
    2. Report: Name "IA_3_IST615_Fall25_FirstName_LastName.docx".
       - Not a PDF file
  - ✓ If you submit a different file than the required file format, you get ZERO.
  - ✓ There are NO formatting requirements for the Word file (e.g., Times New Roman, double-space).
  - ✓ Your Word submission will be evaluated for originality by SafeAssign. You can review the SafeAssign report after the submission.
- Read the syllabus for the rules about late submissions.

Use of LLMs

- Use an LLM for creating and verifying code. Make sure that you prompt properly and check the code on a Jupyter notebook. If you receive an error in the output, you should continue with prompts to understand what this error is and how it can be corrected. You must have a generated output in a Jupyter notebook.
- Add an "LLM Use" section at the end of the Word document (Report).
  - ✓ Some examples of "What you asked (prompts)".
  - ✓ What you liked and disliked about the responses you received (provide several examples).
- *Not including a "LLM Use" section will result in a ZERO for your grade.*

Assignment Objectives:
1. Compute and interpret descriptive statistics.
2. Create and interpret box-and-whisker plots and histograms.

This assignment is based on the dataset you generated in your Individual Assignment – 2. Download the CSV file "smaller_online_retail_clean.csv" first, and have it in your working directory for Jupyter Notebook. Then, start working in your .ipynb file.

CSV file includes:

- 5,398 rows
- **Numeric variables**: Quantity, UnitPrice, SalesAmount (float)
- **Identifier**: CustomerID (treat as ID; don't analyze)
- **Categorical variables**: Country, InvoiceMonth, IsReturn (bool), plus InvoiceNo, StockCode, Description, InvoiceDate (text)

Start with the Jupyter file attached to the assignment on Blackboard.

Rename IA_3_IST615_Fall25.ipynb by adding your first and last names following Fall25.

This renamed file should match the required filename in the Deliverables section.

PART 1: DESCRIPTIVE STATISTICS (15 points)

1. **Load and inspect**
   - Import the packages and load the file.
     - ✓ *Available in the .ipynb file.*
   - Read the CSV into a DataFrame.
   - Print the shape and the first few rows.
   - Treat CustomerID as an ID. Exclude it from stats/plots.
   - Treat InvoiceNo and StockCode as identifiers too. Do not analyze or plot them.
2. **Overall statistics table**
   - Columns to analyze (numeric): Quantity, UnitPrice, SalesAmount.
   - Create a table with one row per numeric column and these fields:
     - ✓ count (non-missing)
     - ✓ mean and median
     - ✓ min, max, range (max − min)
     - ✓ variance (sample) and std dev (sample) (use ddof=1)
       - *ddof = "Delta Degrees Of Freedom."*
         It's the small adjustment used in formulas for variance and standard deviation.
     - ✓ MAD (from mean) (mean absolute deviation from the mean)
     - ✓ Q1 (25th percentile), Q3 (75th), IQR (Q3 − Q1)
     - ✓ skewness and kurtosis (excess)
   - *We are not collecting the mode for these variables because they are continuous-like. Use median/IQR and mean/std to summarize center and spread.*
   - Python tips:
     - ✓ Use Series/column methods like: .count(), .mean(), .median(), .min(), .max(), .var(ddof=1), .std(ddof=1), .quantile(0.25) and .quantile(0.75), .skew(), .kurt().
     - ✓ For MAD (from mean): compute the column's mean, subtract it from the column, take absolute values, then average them.
     - ✓ Collect the values into a list of dicts, then convert to a DataFrame.
     - ✓ Before saving, create the folder with os.makedirs("ba_outputs", exist_ok=True).
     - ✓ Save to ba_outputs/descriptive_stats_report.csv.
3. **Grouped statistics (choose one grouping)**
   - Pick exactly one grouping column:
     - ✓ InvoiceMonth (seasonality) or
     - ✓ Country (regional differences) or
     - ✓ IsReturn (returns vs. non-returns).

You're asked to group **by one column** for three reasons:

- One grouping keeps the analysis focused. You can clearly compare centers (median/mean), spreads (IQR/std), and shapes (skew/outliers) within that dimension without mixing messages.
- Multiple groupings explode the number of tables and plots. One grouping ensures you finish with depth, not just breadth.
- Each grouping answers a different question. Choose the one that best matches the business pattern you want to explore:

  - ✓ **InvoiceMonth** → *Seasonality*. "Do orders or prices shift across months?"
  - ✓ **Country** → *Regional differences*. "Do markets differ in level/variability?"
  - ✓ **IsReturn** → *Returns behavior*. "Do orders with returns look different (value, prices, items)?"

- For that one grouping:
  - ✓ Compute, by group, for each numeric column (Quantity, UnitPrice, SalesAmount):
    - count, mean, median, std, Q1 (0.25), Q3 (0.75), IQR (= Q3 − Q1).
  - ✓ Export exactly one CSV named: ba_outputs/grouped_stats_by_<your_group>.csv
    - Examples: grouped_stats_by_InvoiceMonth.csv, grouped_stats_by_Country.csv.

PART 2: VISUALIZATIONS (15 points)

1. Box-and-whisker plots (with mean marker)

- You will create one PNG per numeric column: Quantity, UnitPrice, SalesAmount.
- Create an output folder: ba_outputs/boxplots/.
- For each numeric column:
  - ✓ Drop missing values.
  - ✓ Make a vertical box plot (one box).
    - In Matplotlib ≥3.9, set the x-tick label using tick_labels=[<column name>].
  - ✓ Compute the mean and overlay it as a single marker (e.g., a diamond) at x=1, y=mean.
    - Add a small legend entry: "Mean".
    - Verify the legend appears in each saved image.
  - ✓ Title: Box-and-Whisker Plot: <column>.
  - ✓ Save as ba_outputs/boxplots/<column>_box.png.

2. Histograms (with mean line)

- You will create one PNG per numeric column.
- Create an output folder: ba_outputs/histograms/.
- For each numeric column:

- ✓ Drop missing values.
- ✓ Plot a histogram (Use the default bins).
- ✓ Compute the mean and draw a vertical dashed line at x = mean; add a legend "Mean".
- ✓ Title: Histogram: <column>; label axes.
- ✓ Save as ba_outputs/histograms/<column>_hist.png.

3. Zoomed box plots & histograms (within Tukey fences)

- Purpose: Improve readability when extreme outliers compress the plots. This does not change any descriptive statistics; it's for visualization only.
- You will create one zoomed box plot and one zoomed histogram per numeric column: Quantity, UnitPrice, SalesAmount.
- Create output folders: ba_outputs/boxplots_zoom/ and ba_outputs/histograms_zoom/.
- For each numeric column:
  - ✓ Drop missing values.
  - ✓ Compute Tukey fences using the full data: Lower = Q1 − 1.5×IQR, Upper = Q3 + 1.5×IQR (where IQR = Q3 − Q1).
  - ✓ Filter to values within the fences (visualization only).
  - ✓ Box plot (vertical, one box) for the filtered data.
    - In Matplotlib ≥3.9, set the x-tick label to <column> (within fences).
    - Compute the mean of the filtered values and overlay it as a single marker (e.g., diamond) at x=1, y=mean.
    - Add a legend entry: "Mean" and ensure it appears.
    - Title: Box-and-Whisker (Zoomed): <column>.
    - Save as: ba_outputs/boxplots_zoom/<column>_box_zoom.png.
  - ✓ Histogram for the filtered data (default bins).
    - Draw a vertical dashed line at x = mean; add legend "Mean".
    - Title: Histogram (Zoomed within fences): <column>; label axes.
    - Save as: ba_outputs/histograms_zoom/<column>_hist_zoom.png.

IMPORTANT NOTE:

In your report, reference both the original and zoomed plots. State that zoomed plots are for clarity and that all reported statistics (mean, median, IQR, etc.) are based on the full dataset.

PART 3: REPORT (40 points)

- Use your numbers (from Part 1) and your plots (from Part 2).

1. Variable write-ups

Do this for Quantity, UnitPrice, and SalesAmount.

For **each** variable, include the four sections below. Refer to the exact values from your table and the saved figures.

A. Center & Skew

- Report median and mean (use the mean marker on the box plot and the mean line on the histogram).
- State whether the mean is above or below the median and what that implies about skew (right-skew if mean > median; left-skew if mean < median).
- Conclude which measure you will use for a "typical" value (median & IQR if skewed/outliers; mean & std if roughly symmetric).

B. Spread & Outliers (from the box plot)

- Report IQR (= Q3 − Q1) as the spread of the middle 50%, and say if it's small/large relative to the variable's scale.
- Provide the numeric fences: Lower = Q1 − 1.5×IQR, Upper = Q3 + 1.5×IQR.
- State where outliers appear (mainly high side, low side, or both) and whether they seem numerous or rare.

C. Shape (from the histogram)

- **Peaks (visual "modes")**:
  - ✓ Say unimodal or multimodal. If multimodal, offer a plausible reason (e.g., price tiers, promotions, regions).
    (You do not compute the statistical mode—describe the peaks you see.)
- **Skew vs mean line**:
  - ✓ Note whether the mean line sits toward the longer tail (right/left) and confirm it matches your box-plot skew.
- **Gaps/spikes**:
  - ✓ Mention any gaps (possible subgroups) or sharp spikes (rounding/bucketed prices).
    - Gaps = empty or very low bars between clusters of bars.
      - What it can mean: potential subgroups (e.g., low-price items vs. premium items), data rules (e.g., minimum order quantities), or seasonal availability.
    - Spikes = narrow, high bars at specific values.
      - What it can mean: rounding or pricing buckets (e.g., UnitPrice at $5, $10, $20), pack sizes (e.g., Quantity near multiples like 6, 12), or promotional price points.

D. Business takeaway (1 sentence)

- Example: *"UnitPrice is right-skewed with a few very high-priced items, so the median better reflects typical pricing."*

2. Grouped analysis (the one grouping you chose in Part 1.3)

A. Justification (1–2 sentences):

- ✓ Why you chose InvoiceMonth *(seasonality)* or Country *(regional differences)* or IsReturn *(returns vs. non-returns)*.

B. Insights (2–3 bullets):

- ✓ Compare groups using your grouped table. Reference at least one specific statistic per bullet (e.g., medians, IQRs, or outlier counts).
  - Example: *"November's median SalesAmount is ~40% higher than April's and has a wider IQR, indicating peak-season uplift and variability."*
  - Example: *"Return invoices show higher IQR in Quantity and more high-side outliers."*

C. Small-group note (if applicable):

- ✓ If any group had < 20 rows, say whether you excluded it or combined into 'Other'.
- ✓ If no group has < 20 rows, you may omit this item.

3. One-paragraph summary

In 4–6 sentences, synthesize your findings across the three variables. Answer:

1. Which variable is most skewed and why?
2. Where do outliers matter most for business decisions?
3. Which summary statistic (median/IQR vs mean/std) should be reported for each variable going forward?

4. LLM Use

The instructions are above on the first page.