Statistics and Data Analysis

Unit 01 – Lecture 04: In-class Activity (Cleaning + Summary + Plots)

Tofik Ali

School of Computer Science, UPES Dehradun

February 9, 2026

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Task Deliverables Solution Wrap-up

Agenda

What We Will Do Today

You will complete a small end-to-end preprocessing + EDA task:

- load a raw dataset

What We Will Do Today

You will complete a small end-to-end preprocessing + EDA task:

- load a raw dataset
- clean errors (missing, invalid, outliers)

What We Will Do Today

You will complete a small end-to-end preprocessing + EDA task:

- load a raw dataset
- clean errors (missing, invalid, outliers)
- create 2–3 engineered features

What We Will Do Today

You will complete a small end-to-end preprocessing + EDA task:

- load a raw dataset
- clean errors (missing, invalid, outliers)
- create 2–3 engineered features
- create summary tables

What We Will Do Today

You will complete a small end-to-end preprocessing + EDA task:

- load a raw dataset
- clean errors (missing, invalid, outliers)
- create 2–3 engineered features
- create summary tables
- create 3 plots and write short insights

Time Plan (55 minutes)

- 10 min: attendance + setup

Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)

Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)
- 10 min: discussion (compare approaches)

Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)
- 10 min: discussion (compare approaches)
- 5 min: wrap-up + exit question

Dataset

File: `data/campus_cafe_transactions.csv`

Columns:

- `date` (datetime), `category` (Snacks/Drinks/Stationery)
- `payment_mode` (Cash/UPI/Card)
- `units`, `unit_price`, `discount_pct`

Note: It intentionally includes missing values and invalid entries to clean.

Task 1: Load and Inspect (5 minutes)

- Print shape and first 5 rows

Task 1: Load and Inspect (5 minutes)

- Print shape and first 5 rows
- Check dtypes (are `units` numeric? is `date` datetime?)

Task 1: Load and Inspect (5 minutes)

- Print shape and first 5 rows
- Check dtypes (are `units` numeric? is `date` datetime?)
- Compute missingness % per column

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric
- Handle missing `discount_pct` (e.g., fill with 0)

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric
- Handle missing `discount_pct` (e.g., fill with 0)
- Handle missing `units` (drop or impute; justify)

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric
- Handle missing `discount_pct` (e.g., fill with 0)
- Handle missing `units` (drop or impute; justify)
- Handle invalid `units` (negative) **drop the row**

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric
- Handle missing `discount_pct` (e.g., fill with 0)
- Handle missing `units` (drop or impute; justify)
- Handle invalid `units` (negative) **drop the row**
- Handle discount outliers (cap to a reasonable range)

Task 2: Cleaning Rules (10 minutes)

- Convert `date` to datetime, numeric columns to numeric
- Handle missing `discount_pct` (e.g., fill with 0)
- Handle missing `units` (drop or impute; justify)
- Handle invalid `units` (negative) **drop the row**
- Handle discount outliers (cap to a reasonable range)
- Handle price outliers (replace with category median or cap; justify)

Task 3: Feature Engineering (5 minutes)

Create:

- $\texttt{gross\_amount} = \texttt{units} \times \texttt{unit\_price}$

Task 3: Feature Engineering (5 minutes)

Create:

- $\texttt{gross\_amount} = \texttt{units} \times \texttt{unit\_price}$
- $\texttt{net\_amount} = \texttt{gross\_amount} \times (1 - \texttt{discount\_pct}/100)$

Task 3: Feature Engineering (5 minutes)

Create:

- $\texttt{gross\_amount} = \texttt{units} \times \texttt{unit\_price}$
- $\texttt{net\_amount} = \texttt{gross\_amount} \times (1 - \texttt{discount\_pct}/100)$
- `is_weekend` from `date`

Task 4: Summary Tables (10 minutes)

Produce:

- total net revenue by `category`

Task 4: Summary Tables (10 minutes)

Produce:

- total net revenue by `category`
- total net revenue by `payment_mode`

Task 4: Summary Tables (10 minutes)

Produce:

- total net revenue by `category`
- total net revenue by `payment_mode`
- top 5 transactions by `net_amount`

Task 5: Plots (10 minutes)

Create and save:

- bar chart: net revenue by category

Task 5: Plots (10 minutes)

Create and save:

- bar chart: net revenue by category
- histogram: net amount per transaction

Task 5: Plots (10 minutes)

Create and save:

- bar chart: net revenue by category
- histogram: net amount per transaction
- line chart: daily net revenue

Final Deliverables (Submit/Show)

- Cleaned dataset saved as `data/campus_cafe_clean.csv`

Final Deliverables (Submit/Show)

- Cleaned dataset saved as `data/campus_cafe_clean.csv`
- Three plots saved in `images/`

Final Deliverables (Submit/Show)

- Cleaned dataset saved as `data/campus_cafe_clean.csv`
- Three plots saved in `images/`
- 3–5 short insights + 2 limitations/caveats

Solution Script (Python)

After attempting yourself, run:

```
python demo/activity_solution.py
```

Outputs:

- `data/campus_cafe_clean.csv`
- summary CSVs in `data/`
- plots in `images/`

Expected Key Results (Example)

Net revenue by category (after cleaning):

Category	Count	Net Revenue (INR)
Snacks	9	1129.25
Drinks	8	368.00
Stationery	7	283.90

Question: Why is Snacks revenue much higher?

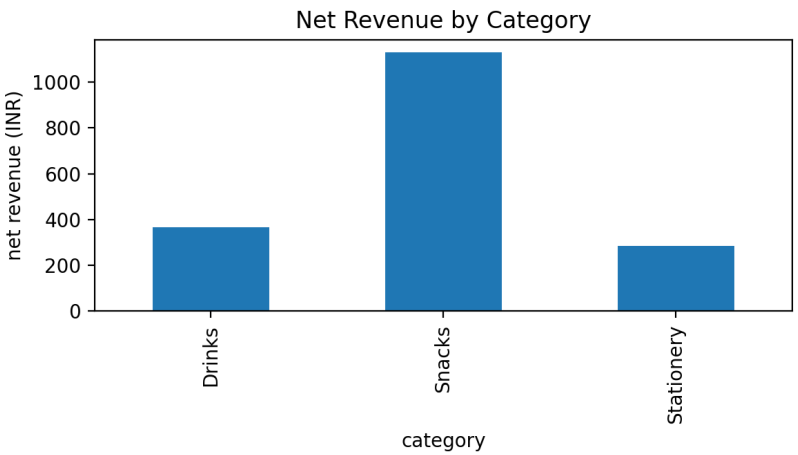
Expected Key Results (Payment Mode)

Net revenue by payment mode (after cleaning):

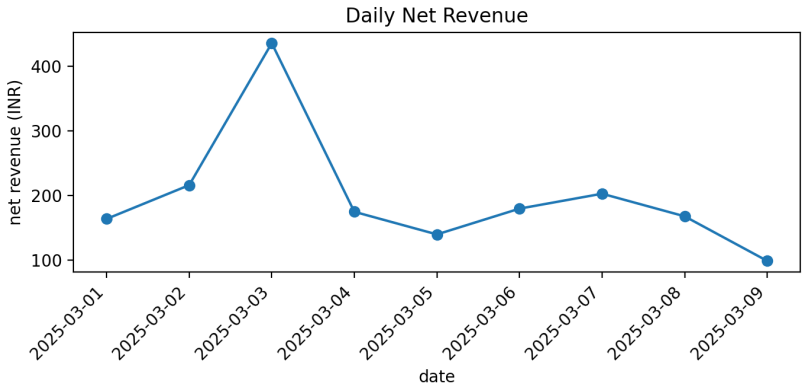
Mode	Count	Net Revenue (INR)
UPI	12	1124.15
Cash	6	304.00
Card	6	353.00

Example Plots

Revenue by Category



Daily Revenue



Write Insights + Caveats

Examples of insights:

- Snacks contribute the largest share of revenue in this sample week.

Examples of caveats:

Write Insights + Caveats

Examples of insights:

- Snacks contribute the largest share of revenue in this sample week.
- UPI is the most common payment mode and also highest revenue.

Examples of caveats:

Write Insights + Caveats

Examples of insights:

- Snacks contribute the largest share of revenue in this sample week.
- UPI is the most common payment mode and also highest revenue.
- A few large transactions dominate total revenue (check top-5 table).

Examples of caveats:

Write Insights + Caveats

Examples of insights:

- Snacks contribute the largest share of revenue in this sample week.
- UPI is the most common payment mode and also highest revenue.
- A few large transactions dominate total revenue (check top-5 table).

Examples of caveats:

- Small dataset (only a few days) $\Rightarrow$ not representative.

Write Insights + Caveats

Examples of insights:

- Snacks contribute the largest share of revenue in this sample week.
- UPI is the most common payment mode and also highest revenue.
- A few large transactions dominate total revenue (check top-5 table).

Examples of caveats:

- Small dataset (only a few days) $\Rightarrow$ not representative.
- Cleaning choices (caps/median replacement) can change results.

Wrap-up

- A good summary combines: cleaning + engineered features + tables + plots

Exit question: What is one cleaning rule you applied and why?

Wrap-up

- A good summary combines: cleaning + engineered features + tables + plots
- Document your rules so results are reproducible

Exit question: What is one cleaning rule you applied and why?

Wrap-up

- A good summary combines: cleaning + engineered features + tables + plots
- Document your rules so results are reproducible
- Always communicate limitations honestly

Exit question: What is one cleaning rule you applied and why?