SI 201 Project 1 checkpoint

Jose Velazquez

Dataset: *Sample Superstore* (Kaggle) – CSV file

File path (planned): `data/sample_superstore.csv`

Columns I Will Use

- `Region` (text) – geographic region of the order
- `Category` (text) – high-level product category
- `Sub-Category` (text) – specific product subcategory
- `Sales` (numeric) – dollar amount of the sale
- `Profit` (numeric) – dollar profit from the sale
- `Order Date` (date) – used for month/year grouping if needed

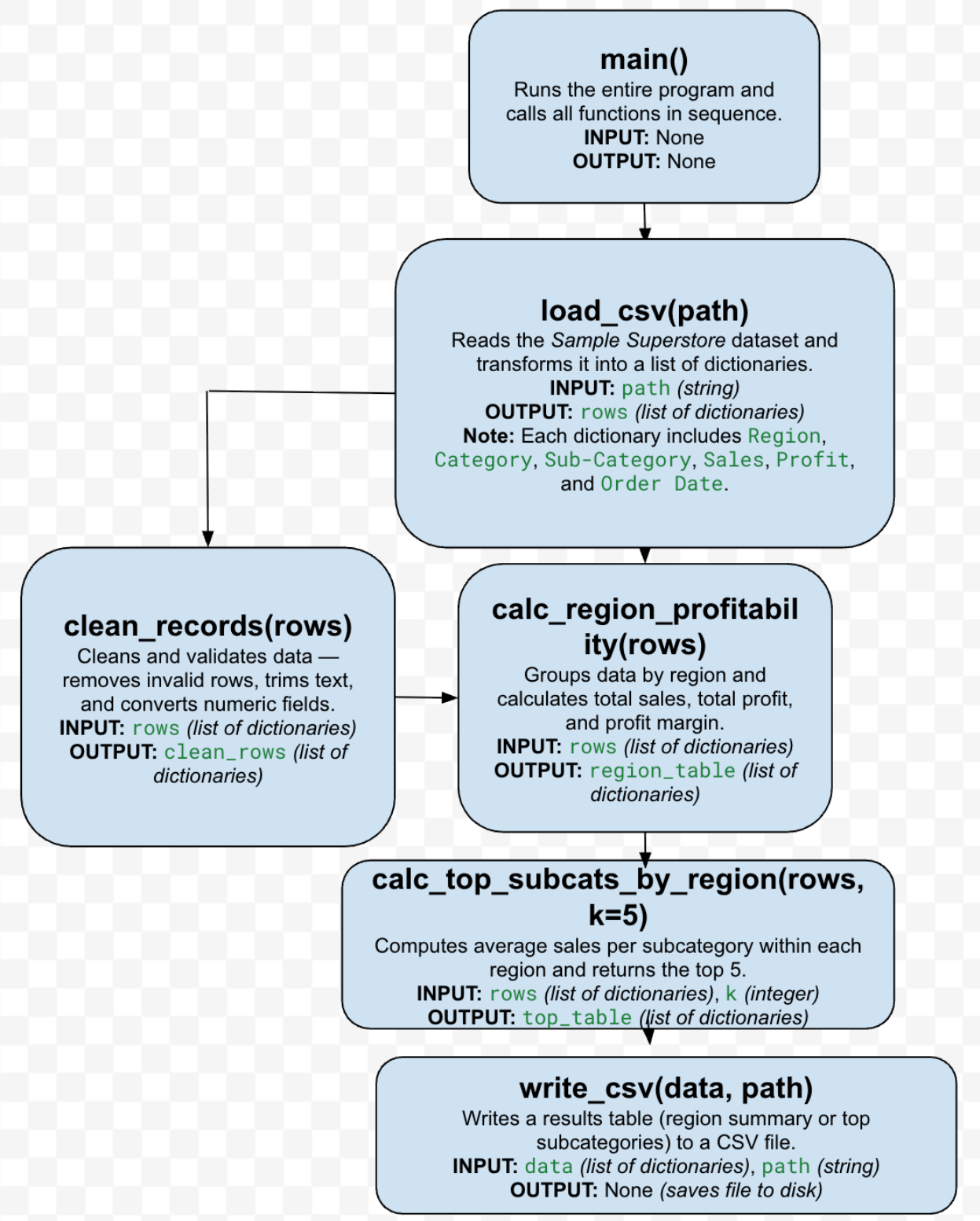
Calculations I Will Perform (solo = 2 total)

1. **Region Profitability Summary**
   - For each `Region`, compute:
     - `total_sales` (sum of `Sales`)
     - `total_profit` (sum of `Profit`)
     - `profit_margin` = `total_profit / total_sales`
   - Uses: `Region`, `Sales`, `Profit`
   - Output to: `results/region_profitability.csv`
2. **Top Subcategories by Average Order Value (AOV) in Each Region**
   - For each `Region`, compute the average `Sales` per `Sub-Category`.

- Return the **top 5 subcategories** per region by AOV (break ties alphabetically).
- Uses: `Region`, `Sub-Category`, `Sales`
- Output to: `results/top_subcats_by_region.csv`

Collaborators: No collaborators

Function Decomposition Diagram:


main()
Runs the entire program and calls all functions in sequence.
INPUT: None
OUTPUT: None
load_csv(path)
Reads the Sample Superstore dataset and transforms it into a list of dictionaries.
INPUT: path (string)
OUTPUT: rows (list of dictionaries)
Note: Each dictionary includes Region, Category, Sub-Category, Sales, Profit, and Order Date.
clean_records(rows)
Cleans and validates data — removes invalid rows, trims text, and converts numeric fields.
INPUT: rows (list of dictionaries)
OUTPUT: clean_rows (list of dictionaries)
calc_region_profitability(rows)
Groups data by region and calculates total sales, total profit, and profit margin.
INPUT: rows (list of dictionaries)
OUTPUT: region_table (list of dictionaries)
calc_top_subcats_by_region(rows, k=5)
Computes average sales per subcategory within each region and returns the top 5.
INPUT: rows (list of dictionaries), k (integer)
OUTPUT: top_table (list of dictionaries)
write_csv(data, path)
Writes a results table (region summary or top subcategories) to a CSV file.
INPUT: data (list of dictionaries), path (string)
OUTPUT: None (saves file to disk)