Project 1 Checkpoint

Sample Superstore Dataset

Columns I'll be working on:

- **Measures:** Sales, Profit, Quantity, Discount
- **Dimensions:** Region, State, Segment, Sub-Category, Category, Ship Mode
- **Derived fields I'll create:**
  - profit_margin = Profit / Sales (guard Sales=0)
  - is_loss = Profit < 0
  - big_discount = Discount >= 0.20
  - discount_tier = categorized as "None", "Low", "Medium", or "High"

Calculations:

Calculation 1: Average profit margin by sub-category within each region

- **Uses:** Profit, Sales, Sub-Category, Region
- **Method:** For each (Region, Sub-Category), compute weighted margin = sum(Profit) / sum(Sales)
- **Output file:** margin_by_region_subcategory.csv
- **Columns:** Region, SubCategory, total_sales, total_profit, profit_margin

Calculation 2: Loss rate for high-discount lines by state and segment

- **Uses:** Discount, Profit, State, Segment
- **Method:** Filter rows where Discount >= 0.20; for each (State, Segment), compute percent with Profit < 0
- **Output file:** loss_pct_high_discount_by_state_segment.csv
- **Columns:** State, Segment, num_lines, num_losses, loss_pct

Calculation 3: Regional Performance by Customer Segment

- **Uses:** Region, Segment, Sales, Quantity
- **Method:** For each (Region, Segment) combination, calculate average order value = total_sales / total_quantity
- **Output file:** avg_order_value_by_region_segment.csv
- **Columns:** Region, Segment, total_sales, total_quantity, avg_order_value

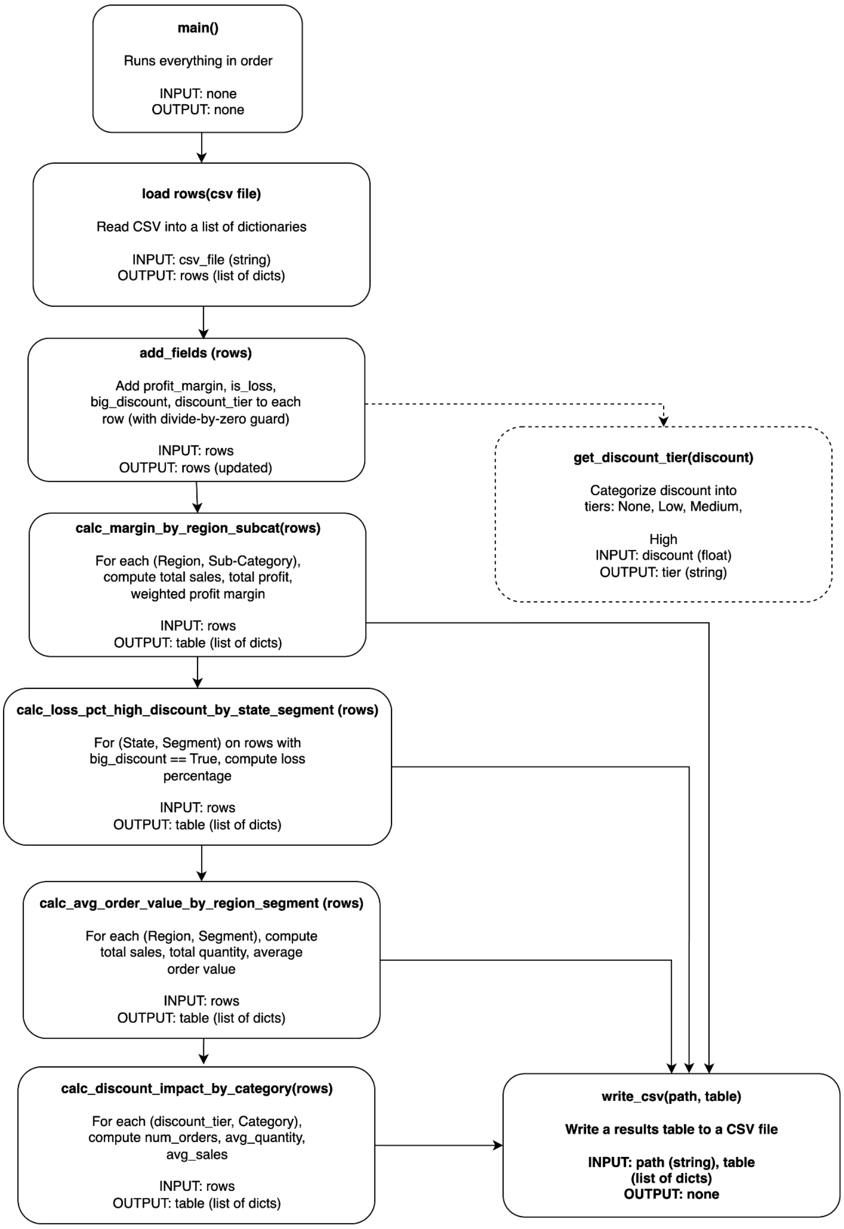
Calculation 4: Discount Impact on Order Size by Category

- **Uses:** Discount, Quantity, Category, Sales
- **Method:** Create discount tiers (None: 0%, Low: 0-20%, Medium: 20-40%, High: 40%+); for each (discount_tier, Category), calculate average quantity and average sales
- **Output file:** discount_impact_by_category.csv
- **Columns:** discount_tier, Category, num_orders, avg_quantity, avg_sales

Team:

David Vargas & Alberto Puliga

Diagram:


main()
Runs everything in order
INPUT: none
OUTPUT: none
load rows(csv file)
Read CSV into a list of dictionaries
INPUT: csv_file (string)
OUTPUT: rows (list of dicts)
add_fields (rows)
Add profit_margin, is_loss, big_discount, discount_tier to each row (with divide-by-zero guard)
INPUT: rows
OUTPUT: rows (updated)
get_discount_tier(discount)
Categorize discount into tiers: None, Low, Medium,
High
INPUT: discount (float)
OUTPUT: tier (string)
calc_margin_by_region_subcat(rows)
For each (Region, Sub-Category), compute total sales, total profit, weighted profit margin
INPUT: rows
OUTPUT: table (list of dicts)
calc_loss_pct_high_discount_by_state_segment (rows)
For (State, Segment) on rows with big_discount == True, compute loss percentage
INPUT: rows
OUTPUT: table (list of dicts)
calc_avg_order_value_by_region_segment (rows)
For each (Region, Segment), compute total sales, total quantity, average order value
INPUT: rows
OUTPUT: table (list of dicts)
calc_discount_impact_by_category(rows)
For each (discount_tier, Category), compute num_orders, avg_quantity, avg_sales
INPUT: rows
OUTPUT: table (list of dicts)
write_csv(path, table)
Write a results table to a CSV file
INPUT: path (string), table (list of dicts)
OUTPUT: none