

Global Superstore Data Wrangling Project

Overview

You will be working with three CSV files from the Global Superstore dataset (these are available as downloads from the course resources).

Your task is to interpret *high-level instructions* (like the kind you'd get from a typical manager, especially a non-technical one) – and translate those into explicit, well-crafted *prompts* asking Gemini for assistance in implementing solutions.

Important: Focus on clearly describing the relationships between datasets, along with what you're trying to achieve. Ask Gemini for recommendations and plans, not just code.

Part 1: Sales Performance by Manager Summary

Your goal for Part 1:

Create a summary breakout of sales performance by manager that matches the numbers in the screenshot below:

Person	Total_Sales	Total_Profit	Number_of_Orders
Anna Andreadi	725457.8245	108418.4489	1611
Cassandra Brandow	391721.9050	46749.4303	822
Chuck Magee	678781.2400	91522.7800	1401
Kelly Williams	501239.8908	39706.3625	1175

This data you'll need to create this summary comes from **both** of the following datasets:

Orders.csv:

	A	B	C	D	E	F	G	H	I	J	
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Sales	Quantity	Discount	Profit	Region	City
2		1 CA-2020-1	11/8/2020	11/11/2020	Second Class	261.96	2	0	41.9136	South	Hendersc
3		2 CA-2020-1	11/8/2020	11/11/2020	Second Class	731.94	3	0	219.582	South	Hendersc
4		3 CA-2020-1	6/12/2020	6/16/2020	Second Class	14.62	2	0	6.8714	West	Los Ange
5		4 US-2019-1	10/11/2019	10/18/2019	Standard Class	957.5775	5	0.45	-383.031	South	Fort Lauc
6		5 US-2019-1	10/11/2019	10/18/2019	Standard Class	22.368	2	0.2	2.5164	South	Fort Lauc
7		6 CA-2018-1	6/9/2018	6/14/2018	Standard Class	48.86	7	0	14.1694	West	Los Ange
8		7 CA-2018-1	6/9/2018	6/14/2018	Standard Class	7.28	4	0	1.9656	West	Los Ange
9		8 CA-2018-1	6/9/2018	6/14/2018	Standard Class	907.152	6	0.2	90.7152	West	Los Ange
10		9 CA-2018-1	6/9/2018	6/14/2018	Standard Class	18.504	3	0.2	5.7825	West	Los Ange
11		10 CA-2018-1	6/9/2018	6/14/2018	Standard Class	114.9	5	0	34.47	West	Los Ange
12		11 CA-2018-1	6/9/2018	6/14/2018	Standard Class	1706.184	9	0.2	85.3092	West	Los Ange
13		12 CA-2018-1	6/9/2018	6/14/2018	Standard Class	911.424	4	0.2	68.3568	West	Los Ange
14		13 CA-2021-1	4/15/2021	4/20/2021	Standard Class	15.552	3	0.2	5.4432	South	Concord
15		14 CA-2020-1	12/5/2020	12/10/2020	Standard Class	407.976	3	0.2	132.5922	West	Seattle
16		15 US-2019-1	11/22/2019	11/26/2019	Standard Class	68.81	5	0.8	-123.858	Central	Fort Wort
17		16 US-2019-1	11/22/2019	11/26/2019	Standard Class	2.544	3	0.8	-3.816	Central	Fort Wort
18		17 CA-2018-1	11/11/2018	11/18/2018	Standard Class	665.88	6	0	13.3176	Central	Madison

People.csv:

	A	B
1	Person	Region
2	Anna Andreadi	West
3	Chuck Magee	East
4	Kelly Williams	Central
5	Cassandra Brandow	South

What You Need to Figure Out:

- Look at both datasets - what column do they have in common?
- How can you use this to determine which sales belong to which person?
- What calculations are needed for the three metrics?

Metrics Definitions:

- Total Revenue = Sum of all sales (\$)
- Total Profit = Sum of all profit (\$)
- Order Count = Number of **unique** orders

Part 2: Handle Returned Orders

Your goal for Part 2:

It turns out that *some* of the orders in the Orders.csv dataset were ultimately returned by customers. Those returned orders can be found in the provided **Returns.csv** dataset:

	A	B
1	Returned	Order ID
2	Yes	CA-2018-100762
3	Yes	CA-2018-100762
4	Yes	CA-2018-100762
5	Yes	CA-2018-100762
6	Yes	CA-2018-100867

Your job will be to **exclude** these returned orders from the summary you created in Part 1.

The resulting summary should match the numbers in the screenshot below:

Person	Total_Sales	Total_Profit	Number_of_Orders
Anna Andreadi	617974.7670	88755.0711	1422
Cassandra Brandow	374412.8080	44530.8198	798
Chuck Magee	637076.0960	86537.9794	1357
Kelly Williams	487232.9114	43340.7899	1136

Things to Consider:

- The Returns dataset contains Order IDs that were returned
- These same Order IDs exist in your Orders dataset
- You want to EXCLUDE any order that appears in the Returns dataset from the summary results
- Not all orders were returned (most weren't!)

Part 3: Create a Professional Visualization

Your goal for Part 3:

Create a chart that **effectively visualizes** the data in your modified summary from Part 2.

Only include the “Total Profit” and “Order Count” metrics in this visualization, NOT “Total Revenue”.

(Hopefully) Helpful Hint:

Before getting it to actually *build* your visualization, try asking Gemini for recommendations on what chart type(s) would work best for this data. Then, carefully consider the pros and cons of all recommended options before proceeding.