A black screen with white text

AI-generated content may be incorrect.

**Part 2: Building Your Data Lake (45 minutes)**

📸 **Screenshot Required #3:** Your Cloud Storage bucket with uploaded data

A screenshot of a computer

AI-generated content may be incorrect.

📸 **Screenshot Required #4:** BigQuery dataset with loaded Superstore table

A screenshot of a computer

AI-generated content may be incorrect.

**Task 2.2: Verify Your Data Lake**

After completing the Scribd tutorial, verify your setup:

-- Run this in BigQuery to verify data loaded correctly

SELECT

COUNT(\*) as total\_rows,

COUNT(DISTINCT Customer\_ID) as unique\_customers,

MIN(Order\_Date) as earliest\_order,

MAX(Order\_Date) as latest\_order

FROM `your-project.your\_dataset.superstore\_sales`;

Expected results:

* Total rows: ~9,994
* Date range: 2020-2023

A screenshot of a computer

AI-generated content may be incorrect.

**Part 3: AI-Assisted Data Exploration (45 minutes)**

**Task 3.1: Initial Data Understanding with Gemini**

Create a new Colab notebook named Lab1\_AI\_Analysis.ipynb and complete these explorations:



**Exploration 1: Understanding the Business Context**

A screenshot of a computer

AI-generated content may be incorrect.

“### 1. What are our most and least profitable products and product categories?

\*\*Why it matters:\*\* This is the most fundamental question for any retail business. Not all sales are good sales. A product might sell in high volume but have a razor-thin or even negative profit margin, especially after discounts. Answering this question allows the business to:

\* \*\*Optimize Inventory:\*\* Double down on stocking profitable "winner" products.

\* \*\*Adjust Strategy:\*\* Re-evaluate pricing, reduce discounts, or find a cheaper supplier for low-profit items.

\* \*\*Make Pruning Decisions:\*\* Discontinue products that consistently lose money and drain resources.

\* \*\*Focus Marketing:\*\* Allocate marketing budget to promote the most profitable items.

\*\*Columns to Analyze:\*\*

\* `Profit` (the primary metric)

\* `Sales` (to understand the relationship between revenue and profitability)

\* `Category` & `Sub-Category` (to group the products)

\* `Discount` (often the key driver of low profitability)

\* `Quantity` (to see if high-volume items are profitable)

### 2. Who are our most valuable customers and what are their characteristics?

\*\*Why it matters:\*\* The Pareto Principle often applies to retail: roughly 80% of your profit comes from 20% of your customers. Identifying this core group is crucial for sustainable growth. This analysis helps you to:

\* \*\*Improve Customer Retention:\*\* Create loyalty programs and targeted offers to keep these high-value customers happy.

\* \*\*Personalize Marketing:\*\* Understand what these customers buy and tailor marketing campaigns to their preferences.

\* \*\*Find New "Lookalike" Customers:\*\* Analyze the segment, region, and product preferences of your best customers to find and attract similar prospects.

\* \*\*Increase Lifetime Value:\*\* Nurture these relationships to ensure they continue to purchase over the long term.

\*\*Columns to Analyze:\*\*

\* `Customer ID` (to group transactions by individual)

\* `Sales` & `Profit` (to calculate the total monetary value of each customer)

\* `Order Date` (to determine purchase frequency and recency)

\* `Segment` (to see if your best customers fall into a specific segment like 'Corporate' or 'Consumer')

\* `Region` & `State` (to identify geographic concentrations of top customers)

### 3. What is the impact of discounts on sales volume and profitability?

\*\*Why it matters:\*\* Discounts are a powerful but dangerous tool. They can drive revenue and clear out old inventory, but they can also destroy profit margins and devalue your brand. You need to know if your discount strategy is working. This investigation will reveal:

\* \*\*The "Sweet Spot":\*\* The discount level that maximizes sales or quantity without sacrificing too much profit.

\* \*\*Unprofitable Promotions:\*\* Scenarios where high discounts led to a net loss on sales.

\* \*\*Product/Category Sensitivity:\*\* Which products see a huge sales lift from a small discount, and which don't. This helps in creating smarter, targeted promotions.

\*\*Columns to Analyze:\*\*

\* `Discount` (the independent variable you are testing)

\* `Profit` (the key performance indicator for success)

\* `Sales` (to measure the top-line impact)

\* `Quantity` (to measure the impact on sales volume)

\* You should also group this analysis by `Category`, `Sub-Category`, and `Region` to see how discount effectiveness varies.

### 4. Which geographic regions are driving the most and least sales and profit?

\*\*Why it matters:\*\* Business performance is rarely uniform across different markets. Understanding regional differences is key for efficient resource allocation. This analysis helps to:

\* \*\*Optimize Marketing Spend:\*\* Allocate more budget to high-performing regions or run targeted campaigns to boost underperforming ones.

\* \*\*Inform Expansion Strategy:\*\* Identify promising new cities or states based on the performance of similar existing markets.

\* \*\*Tailor Product Mix:\*\* Discover if certain products sell better in specific regions (e.g., office supplies in a business-dense state) and adjust inventory accordingly.

\* \*\*Investigate Supply Chain:\*\* Uncover if poor performance in a region is related to logistics issues or higher shipping costs.

\*\*Columns to Analyze:\*\*

\* `Region`, `State`, & `City` (for geographic segmentation)

\* `Sales` (to identify top markets by revenue)

\* `Profit` (to identify top markets by profitability, which is more important)

\* `Profit Ratio` (calculated as `Profit` / `Sales`) to compare profitability fairly across regions of different sizes.

### 5. What are the sales and profit trends over time, and is there seasonality?

\*\*Why it matters:\*\* Understanding the rhythm of your business is essential for planning and forecasting. Looking at performance over time helps you:

\* \*\*Manage Inventory:\*\* Prepare for predictable peaks (like the holiday season) to avoid stockouts and plan for lulls to avoid overstocking.

\* \*\*Strategic Staffing:\*\* Align staffing levels in stores and warehouses with demand.

\* \*\*Financial Forecasting:\*\* Create more accurate revenue and profit forecasts for budgeting and financial planning.

\* \*\*Evaluate Performance:\*\* Determine if the business is growing year-over-year and measure the impact of major marketing campaigns or strategic changes.

\*\*Columns to Analyze:\*\*

\* `Order Date` (the primary column for all time-series analysis)

\* `Sales` & `Profit` (the metrics you want to track over time)

\* You would extract the year, quarter, month, and week from the `Order Date` to analyze trends at different levels of granularity.”

**Exploration 2: Data Quality Assessment**

Data Quality Issues to Check For

When working with a Superstore dataset, common data quality issues to look for include:

* Completeness: Are there missing values in critical fields like Sales , Profit , Order Date , Customer ID , or Product Name ?
* Validity: Do values adhere to expected formats and ranges? For example:
  + Are Sales and Profit always non-negative?
  + Are Order Date values legitimate dates within a reasonable timeframe?
  + Is Region or Category spelled consistently?
* Consistency: Is the data uniform across the dataset?
  + Are product names or customer names spelled the same way every time they appear?
  + Does the sum of Sales for individual items in an order match the total Sales for that order (if applicable)?
* Accuracy: Does the data reflect the true state of the business? While harder to check without external sources, internal inconsistencies can flag inaccuracies. For instance, a very high negative profit might indicate a data entry error.
* Uniqueness: Are there duplicate records that should be unique (e.g., duplicate Order ID entries unless multiple line items share an order ID)?

Patterns Indicating Problems

Analyzing statistics can reveal underlying data quality issues:

* Unusual Averages or Medians: If the average Sales or Profit is unexpectedly low or high, it could point to erroneous entries.
* High Number of Nulls/Blanks: A significant percentage of empty cells in a column strongly indicates a completeness issue.
* Outliers: Extreme values in Sales , Profit , or Quantity can be legitimate but also could be data entry errors.
* Inconsistent Categorization: If a "Category" field has variations like "Office Supplies", "Office supply", and "office supplies", it's a consistency problem.
* Negative Values in Non-Negative Fields: Negative Sales or Quantity (unless it represents returns/refunds and is explicitly designed as such) is a clear validity issue.
* Dates Far Outside Expected Range: Order Date values in the future or very far in the past suggest entry errors.

Specific Validation Queries

Assuming your Superstore data is in a table named superstore\_data (you'd replace this with your actual table name), here are some example SQL queries:

1. Check for Missing Values (Completeness):

Generated code may be subject to license restrictions not shown here. Use code with care. 

2. Check for Negative Sales/Quantity (Validity):

Generated code may be subject to license restrictions not shown here. Use code with care. 

3. Check for Inconsistent Categorization (Consistency):

Generated code may be subject to license restrictions not shown here. Use code with care. 

*Self-correction: If you see variations like "Office Supplies" and "Office supply" in the output, this query flags it.*

4. Check for Duplicate Order IDs (Uniqueness - assuming Order ID should be unique per transaction/row):

Generated code may be subject to license restrictions not shown here. Use code with care. 

*Note: If Order ID can repeat for multiple items within one order, this query might need adjustment based on your schema (e.g., checking Order ID and Product ID combination for uniqueness).*

Initial Findings from Superstore Data: Interpretation and Context

Given your initial findings:

* Total sales: $X
* Profit margin: Y%
* Top category: Z

Let's interpret these:

What do these numbers tell us about the business health?

* Total Sales ($X): This gives you a baseline for the company's revenue generation. A high X suggests strong top-line performance, but without historical data or industry benchmarks, it's hard to say if it's "good."
* Profit Margin (Y%): This is a crucial indicator of efficiency. A positive Y% means the business is profitable. A higher Y% indicates better cost control relative to sales.
  + Healthy Sign: A consistently positive and relatively high profit margin (e.g., 10-20% or more, depending on industry) suggests efficient operations and good pricing strategies.
  + Potential Concern: A very low or negative Y% indicates that the cost of goods sold and operating expenses are too high relative to sales, leading to low or no profitability.
* Top Category (Z): This highlights the primary revenue driver.
  + Healthy Sign: A clear top category indicates specialization or successful product lines.
  + Potential Concern: Over-reliance on a single category could be a risk if demand for that category declines.

What additional metrics would provide more context?

To get a more comprehensive view of business health, consider these metrics:

* Average Order Value (AOV): Total Sales / Number of Orders. This tells you how much customers spend on average per transaction.
* Customer Count / Unique Customers: How many distinct customers made purchases? This indicates customer reach.
* Sales Growth Rate: Compare current sales to previous periods (month-over-month, year-over-year). Is the business growing, stagnant, or declining?
* Profit Growth Rate: Similar to sales growth, but for profit.
* Sales by Region/Segment: Understanding geographical or customer segment performance can reveal strengths and weaknesses.
* Return Rate: Total returns / Total sales. A high return rate impacts profitability.
* Average Profit per Order/Customer: More granular profit insights.
* Inventory Turnover (if applicable): How quickly inventory is sold and replaced.

What might be concerning about these figures?

Without specific values for X, Y, and Z:

* Low Profit Margin (Y%): If Y% is very low (e.g., single digits or even negative), it's a major concern. It implies the business is not efficiently converting sales into profit, potentially due to high costs, aggressive pricing, or discounting.
* Stagnant or Declining Sales (X): If X hasn't grown over time, or worse, has shrunk, it indicates a lack of market penetration or declining demand.
* Over-reliance on a Single Category (Z): While having a top category is good, if Z accounts for an overwhelmingly large portion of sales and profit, and other categories are severely underperforming, it creates a vulnerability. A shock to the demand for category Z could severely impact the entire business.
* Discrepancy between Sales and Profit Trends: If Total Sales are growing but Profit Margin is shrinking, it means the business is selling more but making less profit per sale, which is unsustainable.

Top of Form

**Task 3.2: Create Initial Visualizations**



**Part 4: DIVE Method Application (45 minutes)**

**Task 4.1: Apply the DIVE Framework**

**D - Discover (Basic Finding)**

**I - Investigate (Dig Deeper)**

**V - Validate (Challenge Assumptions)**

**E - Extend (Strategic Application)**

**DIVE Analysis: Product and Category Profitability**

D - Discover (Basic Finding)

Initial Question: What are the top and bottom product categories and sub-categories by total profit?

Basic Answer/Metric (BigQuery Standard SQL Queries): To find the most and least profitable categories:

Generated code may be subject to license restrictions not shown here. Use code with care.

To find the most and least profitable sub-categories:

Generated code may be subject to license restrictions not shown here. Use code with care.

Likely Findings (based on typical Superstore data patterns):

Categories:

Technology will likely be the most profitable category by a significant margin.

Furniture will be profitable but often with a lower profit margin than Technology.

Office Supplies will be the least profitable overall, and may contain some sub-categories with negative profit.

Sub-Categories:

Most Profitable: Copiers , Phones , Chairs . These typically have high per-unit profit.

Least Profitable / Negative Profit: Tables , Bookcases , Supplies , and importantly, Binders . Binders are often characterized by high sales volume but significantly negative profit, largely due to high discounts.

First Impression:

"The initial view confirms that profitability is not uniform across our product portfolio. While Technology is a clear profit driver, several sub-categories, particularly within Furniture and Office Supplies, are significantly underperforming or even losing money. This immediately flags these areas for deeper investigation, especially those showing negative profit."

I - Investigate (Dig Deeper)

Why does this pattern exist? What factors contribute to this? How does it vary across dimensions? Hypotheses:

Discounting: The most common reason for low/negative profit in Superstore is aggressive discounting. Products with high sales but low profit might be routinely sold at steep discounts.

Shipping Costs: Bulky items (e.g., Tables , Bookcases ) can incur high shipping costs, especially over long distances or to less accessible regions, which might erode profit.

Cost of Goods Sold (COGS) / Supplier Issues: Some products might inherently have very low margins due to high purchasing costs from suppliers.

Returns: While not directly in the standard Superstore dataset, high return rates for certain products could contribute to profit loss through reverse logistics costs.

BigQuery Queries for Investigation: Query 1: Examine Average Discount by Sub-Category

This query helps test the discounting hypothesis.

Generated code may be subject to license restrictions not shown here. Use code with care.

Query 2: Profitability by Sub-Category and Region

This query helps test the shipping cost hypothesis and identify regional variations.

Generated code may be subject to license restrictions not shown here. Use code with care.

Likely Findings from Investigation:

Discounting is a major factor: You'll likely see a strong inverse correlation between average\_discount\_rate and profit\_margin\_percent . Binders often have extremely high discount rates (e.g., >30-50%) leading to negative profit margins. Tables and Bookcases also show high discounts.

Regional Impact: Tables often exhibit negative profit, particularly in the Central region, which might suggest higher shipping costs or less efficient logistics to that area.

Supplies tend to have consistently low profit margins even with moderate discounts, suggesting an underlying COGS or pricing issue.

V - Validate (Challenge Assumptions)

What could make this conclusion wrong? What data limitations exist? Are there alternative explanations?

Hidden Costs/Missing Data: Our dataset simplifies "profit." It doesn't explicitly include all operational costs (e.g., marketing spend per product, warehousing, specific shipping costs per item, customer service for returns). The profit column is typically Sales - (Cost of Goods Sold + Discount) . If Superstore's Profit definition is Sales - COGS , then discounts are an additional drain. If Profit is Sales - COGS - Discount , then the issue is directly with pricing/COGS or the discount itself. For Superstore data, the latter is usually assumed.

Strategic Loss Leaders: Some products might be intentionally sold at a loss or very thin margin to attract customers who then buy higher-margin complementary products. For example, cheap Binders might bring customers in to buy expensive Printers . Our analysis doesn't capture this halo effect directly.

Data Quality: Are there any data entry errors? For instance, extremely high negative profit values for a few records could skew the overall sub-category profit. (Initial data quality checks would mitigate this).

Time Sensitivity: Profitability can change over time due to market shifts, new suppliers, or evolving discount strategies. Our aggregate view might mask recent improvements or deteriorations.

Perceived Value: Are we missing qualitative data? Customers might value the convenience of buying all supplies from one place, even if some items are low-profit for the store.

E - Extend (Strategic Application)

What should the business do? How can we measure impact? What are the risks?

Implement Discount Control Policies:

Action: For sub-categories with negative profit (e.g., Binders , Tables ), set strict minimum profit margin thresholds for applying discounts. Train sales staff on these new policies. Consider dynamic pricing strategies.

Measure Impact: Track the average discount rate and the profit margin of these sub-categories month-over-month. Monitor the overall gross profit margin.

Risks: Potential reduction in sales volume for these specific items. Risk of customer dissatisfaction if they're accustomed to heavy discounts.

Re-evaluate Pricing & Suppliers for Chronically Low-Margin Items:

Action: For items like Supplies that consistently show low profit despite moderate discounts, engage with suppliers to negotiate better costs or explore alternative, cheaper suppliers. If supplier costs cannot be reduced, consider a slight price increase.

Measure Impact: Monitor COGS (if available), new supplier pricing, and profit margins. Track sales volume post-price adjustment.

Risks: Price increases can lead to decreased sales volume. Switching suppliers requires due diligence and can have supply chain risks.

Optimize Logistics for Bulky/Heavy Items:

Action: For Tables and Bookcases , especially in regions like Central where profit is severely negative, investigate specific shipping costs. Explore options like regional distribution centers, partnerships with different logistics providers for oversized items, or adjusting shipping surcharges for these products.

Measure Impact: Track the profit margin of these items by region. Monitor shipping costs (if granular data becomes available).

Risks: Increased shipping costs passed to customers might deter purchases. New logistics arrangements require upfront investment and time.

Prioritize Marketing & Inventory for High-Profit Items:

Action: Shift marketing budget and promotional efforts to highlight highly profitable items (e.g., Copiers , Phones ). Ensure adequate stock levels for these "winners." Train sales teams to cross-sell and upsell these items.

Measure Impact: Track the sales growth and profit contribution of these high-margin items. Monitor marketing ROI.

Risks: Potential for customer fatigue if only high-margin items are promoted. May neglect brand diversity if focus is too narrow.

Strategic Review of True Loss Leaders:

Action: For items that remain persistently unprofitable even after interventions, a strategic decision needs to be made: are they truly necessary as "loss leaders" to attract customers for other purchases? If not, consider discontinuing them to free up resources and capital.

Measure Impact: Assess the overall profit improvement and resource reallocation after discontinuation.

Risks: Customer dissatisfaction if a popular (even if unprofitable) item is removed. Potential loss of associated sales from customers who bought the "loss leader" first.

**Task 4.2: Document Your DIVE Journey**

Create a markdown document (dive\_analysis.md) with:



**Part 5: Lab Submission (30 minutes)**

**Task 5.1: Compile Your Work**

Organize all deliverables in your GitHub repository:

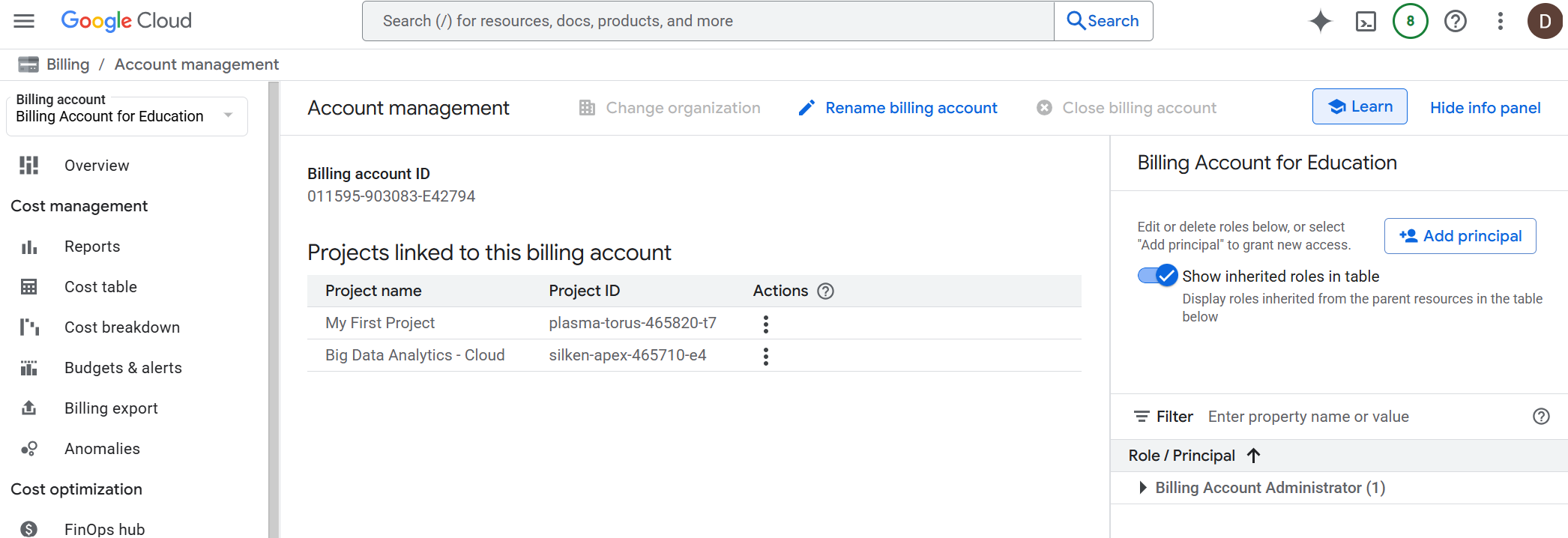
Labs/Lab1/

├── screenshots/

A screenshot of a computer

AI-generated content may be incorrect.

│ ├── gcp\_credits.png



A screenshot of a computer

AI-generated content may be incorrect.

│ ├── github\_repo.png

│ ├── storage\_bucket.png

│ └── bigquery\_dataset.png

├── Lab1\_AI\_Analysis.ipynb

├── dive\_analysis.md

└── lab1\_summary.md

**Task 5.2: Create Lab Summary**

Create lab1\_summary.md with:



**Task 5.3: Submit to D2L**

1. Ensure all files are pushed to GitHub
2. Submit to D2L:
   * Your GitHub repository URL
   * Direct link to your Lab1 folder