DETAILED REPORT FOR RETAIL INSIGHTS LTD

1. Executive summary

This project builds an end-to-end analytics solution for Retail Insights Ltd using the Global Superstore transactional dataset.

The goal is to move from ad-hoc reporting to a repeatable, automated view of business health, with clear actions on:

- Where we are making / losing money (regions, categories and sub-categories)
- Which customers should be protected, grown or re-activated (RFM segments)
- How discounts are eroding profit
- What the next 3–6 months of revenue are likely to look like

Using Python, Azure SQL and Power BI, we created:

- A clean, governed data model loaded daily into Azure SQL via an automated Python ETL job
- A five-page Power BI report:
    1. At-a-glance business health
    2. Sales & profit trends (time and geography)
    3. RFM-based customer segmentation
    4. Product & discount performance
    5. Revenue trend analytics and forecasting
- Supporting BA documentation (epic, user stories, requirements, assumptions, constraints, risks, BPMN and a Jira-style board) stored in the project repo, so the analysis can realistically plug into a delivery squad.

At a high level, the analysis shows:

- Overall profit margin ~12–13% on ~2.3M revenue – reasonable but fragile.
- West and East regions drive the largest revenue and profit; South lags both in revenue and in margin.
- Three sub-categories – Tables, Bookcases and Machines – generate high revenue but very low or negative profit and must be fixed or de-emphasised.
- RFM analysis shows that “Lost Customers” currently contribute the largest revenue and profit, indicating we are too reliant on one-off / lapsed spend rather than nurturing Champions and Loyal customers.
- Revenue has been trending upward year-on-year, but the last 3 months are ~41% lower than the last 6-month average, signalling recent softening.

- Power BI’s built-in forecast and a separate Python ARIMA(1,1,1) model both point to moderate revenue growth but not a step-change, unless we address the above product and customer issues.

The remainder of the report explains the pipeline, findings and recommended actions.

2. Data pipeline and solution architecture

2.1 Source data

- Input: Global Superstore orders file (CSV) with order-level data from 2014–2017, including:
    - Order dates, ship dates, region, country, city
    - Customer and segment
    - Category, sub-category and product
    - Sales, quantity, discount, profit and postal code

2.2 Python data cleaning

A Python script was used to:

- Read the raw CSV (handling non-UTF-8 characters and inconsistent encodings).
- Standardise column types:
    - Convert dates to proper datetime
    - Ensure numeric fields (Sales, Profit, Discount, Quantity) are numeric
- Remove duplicate rows and obvious corrupt records.
- Create an RFM base table (Recency, Frequency, Monetary) at customer level.
- Export a clean file superstore_clean.csv.

2.3 Azure SQL data model

- A database RetailInsightsDB was created on Azure SQL.
- The main fact table: dbo.superstore_orders, loaded from the clean CSV.
- Key fields include:
    - Order_ID, Order_Date, Ship_Date
    - Region, State, City, Postal_Code
    - Customer_ID, Customer_Name, Segment

    - Category, Sub_Category, Product_ID, Product_Name
    - Sales, Profit, Discount, Quantity
- SQL scripts were written and stored in /sql to generate:
    - Overall KPIs
    - Monthly revenue and profit
    - Product and sub-category performance
    - Regional revenue and margin
    - RFM segmentation view

2.4 Automation (ETL job)

- A reusable Python ETL script:
    - Reloads the latest CSV
    - Reapplies cleaning rules
    - Connects to Azure SQL using ODBC Driver 18
    - Truncates and reloads superstore_orders in bulk.
- The script is scheduled in Windows Task Scheduler to run daily, providing an automated refresh of the data model.

2.5 Reporting layer (Power BI)

Power BI Desktop connects directly to Azure SQL and uses star-like modelling (single fact table with dimensions via columns). DAX measures were defined for:

- Total Revenue, Total Profit, Profit Margin
- Average Order Value
- Average Discount
- 3-month and 6-month revenue windows for the forecasting page.

3. Business problem and success criteria

3.1 Epic

As the Head of Commercial at Retail Insights, I want a single, reliable view of revenue, profit, customers and products so I can quickly identify where to grow, where to fix margin leaks and how our revenue is likely to behave in the next few months.

3.2 Key questions (user stories simplified)

1. Overall health – How is the business performing in terms of revenue, profit, margin and discounting?
2. Regional performance – Which regions are driving growth and which require intervention?
3. Product & discount – Which categories/sub-categories contribute most to revenue and where are margins being destroyed by discounting?
4. Customer portfolio – Which customer segments are Champions, Loyal, At Risk or Lost, and how valuable are they?
5. Trends & future – What does revenue look like over time, and what is the expected trajectory over the next 3–6 months?

Acceptance criteria are met when each of the above questions can be answered from the dashboard in under a few minutes, with drill-down to table level.

4. Analytical findings

4.1 Overall sales and profit KPIs

From the Sales & Profit KPI SQL:

- Total Revenue: ~2.30M
- Total Profit: ~286K
- Average Order Value: ~230
- Average Discount: ~15–16%
- Profit Margin: ~12.5%

Interpretation

- The business is profitable but operates on thin margins – a 2–3 point erosion in margin would materially impact profit.
- Average discount of ~16% is high; given margin is only 12–13%, some orders are clearly loss-making.

Implication

- Discount governance is a priority: we must understand where discounts create volume and where they simply destroy profit.

4.2 Regional performance

From the Regional revenue and profitability query and dashboard:

- West has the highest revenue (~725K) and highest profit, with a solid margin.
- East is the second highest in revenue (~679K) and profit, also healthy margin.
- Central is mid-pack (~501K revenue), acceptable margin.
- South is the lowest revenue (~392K) and has the weakest profit profile.

Implication

- West and East should be treated as growth engines – good candidates for targeted upsell, cross-sell and experimentation with premium products.
- South needs diagnostic work: is it a demand issue, pricing issue, logistics cost, or mix of low-margin products?
-

4.3 Product and discount performance

4.3.1 Category & sub-category

From the category & sub-category performance output:

- Overall, Technology and Furniture drive the highest revenue, followed by Office Supplies.
- At the sub-category level, items like Phones, Chairs, Storage, Tables sit at the top in revenue.

However, when we examine profit margin by sub-category:

- Tables show high revenue but negative profit and the lowest margin in the entire portfolio.
- Bookcases and Machines also show low or borderline margins despite substantial revenue.

The “low margin & high revenue” SQL (Sales > 50,000 and margin < 5%) confirms that:

- Tables, Bookcases and Machines are the three major problem children:
    - They collectively contribute a large share of revenue.
    - Their margins range from slightly positive to significantly negative.

4.3.2 Discount vs profit scatter

The scatter plot of Sum of Discount vs Total Profit by Product_ID shows:

- A dense cluster of products with modest discounts and positive profit (healthy zone).

- Several outliers with very high discounts and negative profit, especially in Furniture.

Implications

- Retail Insights is over-discounting specific sub-categories, particularly Furniture items like Tables and Bookcases, which then drive negative profit.
- Without interventions, higher sales in these lines will actually reduce overall profit.

4.4 Customer segmentation – RFM

Using RFM SQL logic, customers are segmented into:

- Champions – recent purchases, high frequency, high monetary value
- Loyal Customers – frequent, good spend, slightly less recent
- At Risk – historically good value but haven't purchased recently
- Lost Customers – long time since last purchase

The RFM dashboard shows:

- Lost Customers surprisingly contribute the largest share of revenue and profit historically – they bought heavily in the past but have not returned.
- Champions and Loyal Customers together also contribute significant revenue, but are smaller in count compared to Lost.
- The scatter of Frequency vs Monetary coloured by segment visually confirms that Champions and Loyal occupy the top-right (high value), while At Risk and Lost spread across mid to high monetary but low recency.

Implications

- The business has not built a stable base of recurring customers; it is overly dependent on customers who have already lapsed.
- There is a strong opportunity to:
    - Re-activate Lost customers with targeted campaigns.
    - Protect Champions and Loyal through better service, loyalty programs and tailored offers.

4.5 Sales & profit trends over time

The Sales & Profit Trends page (line charts by Year–Month) shows:

- A clear upward trend in revenue over 2014–2017, with normal monthly volatility.
- Profit tracks revenue but with more spikes, driven by discounting, product mix and perhaps large deals.
- The stacked bar chart by Category confirms that Technology revenue has grown strongly over time, while Furniture and Office Supplies are more stable.

4.6 Revenue trend analytics and forecasting

The Revenue Trend Analytics and Forecasting page combines:

- A line chart of historical monthly revenue (2014–2017).
- Power BI's built-in time-series forecast for 3–6 months ahead.
- Cards showing:
    - Revenue last 3 months: ~280K
    - Revenue last 6 months: ~476K
    - Growth last 3 vs 6 months: -41.2% (a recent drop vs the prior half-year).

Separately, a Python ARIMA(1,1,1) model was built using aggregated monthly revenue:

- The forecast for the next 6 months shows a flattening / modest growth trajectory, with revenue stabilising around the mid-70K per month range (within a 95% confidence band).
- ARIMA and Power BI's forecast are directionally consistent – neither predicts a crash, but neither indicates explosive growth without changes in strategy.

Implications

- Recent months are weaker than the prior half-year, and the model expects only moderate growth, essentially "more of the same".
- To materially improve the trajectory, we need to:
    - Remove or fix low-margin products.
    - Improve conversion and retention among Champions, Loyal and At Risk segments.
    - Optimise discounts.

5. Assumptions, constraints and risks

Key assumptions

- The Global Superstore dataset is a reliable proxy for actual retail performance.
- Cost of goods sold and profit fields are correctly calculated in the source.
- Historic patterns are reasonably indicative of near-term behaviour (assumption underlying ARIMA and Power BI forecasts).

Constraints

- Historical window limited to 2014–2017; no external drivers (marketing spend, macro conditions, competitor activity) included in the model.
- Discount data is at the order line level; we do not observe underlying negotiated terms (e.g. rebates).
- The solution is currently built for a single dataset; multi-country or multi-brand scaling will require additional modelling and governance.

Risks

- Model risk: If business mix or pricing changes fundamentally, time-series forecasts will be less reliable.
- Margin risk: If we push volume aggressively in low-margin sub-categories without fixing pricing, total profit could decrease even as revenue rises.
- Customer risk: Over-reliance on one-off high-value customers (Lost segment) leaves the business vulnerable to sudden drops.

All of this is captured in the Risk Register and Jira-style tickets in the /docs/ba folder for future squads.

6. Recommended actions

6.1 Fix loss-making products

1. Immediate margin review for Tables, Bookcases and Machines:
    - Re-price or reduce maximum discount levels.
    - Consider removing persistent loss-makers or renegotiating supplier terms.
    - Introduce “good-better-best” bundles where high-margin products are paired with these items.

2. Add a monitoring view in Power BI filtered to *“High Revenue, Low Margin”* sub-categories so commercial teams see this risk every week.

6.2 Strengthen regional strategy

1. West and East:
    - Treat as growth regions – pilot cross-sell campaigns to Champions and Loyal customers there first.
    - Use their stronger performance as a benchmark for South and Central.
2. South:
    - Run a deep-dive analysis combining product mix, discounting and logistics cost.
    - If South is structurally low margin, adjust expectations or reshape the portfolio there.

6.3 Build a customer lifecycle program (from RFM)

1. Champions
    - Exclusive offers, early access to new products, and personal account management.
    - Protect these customers; aim to increase their frequency without heavy discounting.
2. Loyal Customers
    - Upsell campaigns focused on high-margin Technology and Office Supplies.
    - Introduce loyalty points or tiered benefits to keep them from slipping to At Risk.
3. At Risk & Lost Customers
    - Targeted win-back campaigns with limited, controlled discounts on profitable categories.
    - Measure response by segment and feed back into the dashboard.

6.4 Operationalise the analytics asset

- Keep the Python ETL and Azure SQL database under version control (GitHub), with clear deployment instructions.
- Use the existing Jira board and Confluence-style docs in /docs/ba as the base for a real agile delivery stream:

    - Each recommendation above can be framed as a feature with user stories, acceptance criteria and tasks.
- Schedule Power BI refresh aligned with the ETL schedule so business users always see yesterday-complete data.