# Retail Data Analysis

Adrika Shinjini
Under the guidance of Mr Manoj K, EY India

Summer 2025

26/06/2025

# CONTENTS

# Dataset Overview

- **Source:** *cleaned_retail_data.csv* containing 30 columns and 10,000+ records
- **Key fields:** Transaction ID, Customer details, Date, Product Category, Total Amount, Feedback
- **Common issues identified:**
  - Nulls in columns like Ratings, Product_Type, Zipcode
  - Inconsistent types (e.g., float for age and customer ID)
- **Tools used:**
  - Python (Google Colab) for EDA
  - PostgreSQL for data cleaning and modelling
  - Power BI for visualization

# Data Cleaning & Preparation

## Python (pandas)

- Identified and handled missing values (df.isnull().sum())
- Removed duplicate rows
- Converted float columns to integers (e.g., Age, Customer_ID)
- Verified value distributions (value_counts(), describe())
- Exported cleaned dataset as CSV for SQL use

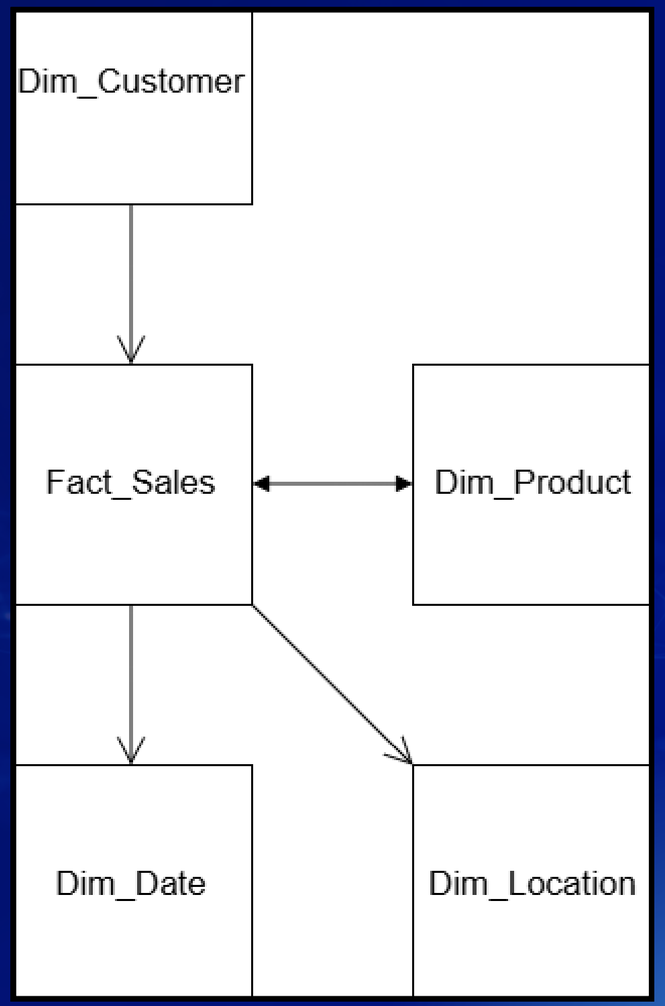
## PostgreSQL (retail_staging table)

- Used ALTER TABLE to clean data types (e.g., Ratings, Phone, Zipcode)
- Filtered null values for fields like Date, Product_Type, Ratings
- Standardised inconsistent formats across fields
- Created cleaned Fact_Sales and Dim_* tables from retail_staging

# Star Schema & SQL Modeling

Designed a star schema with 1 Fact table and 4 Dimension tables to enable normalised, scalable reporting in Power BI.

- Fact_Sales: metrics like amount, rating
- Dim_Customer: name, id, age, segment, etc
- Dim_Product: category, brand
- Dim_Date: year, quarter, month, date
- Dim_Location: country, state, cities, zip codes



# Data Transformations using SQL

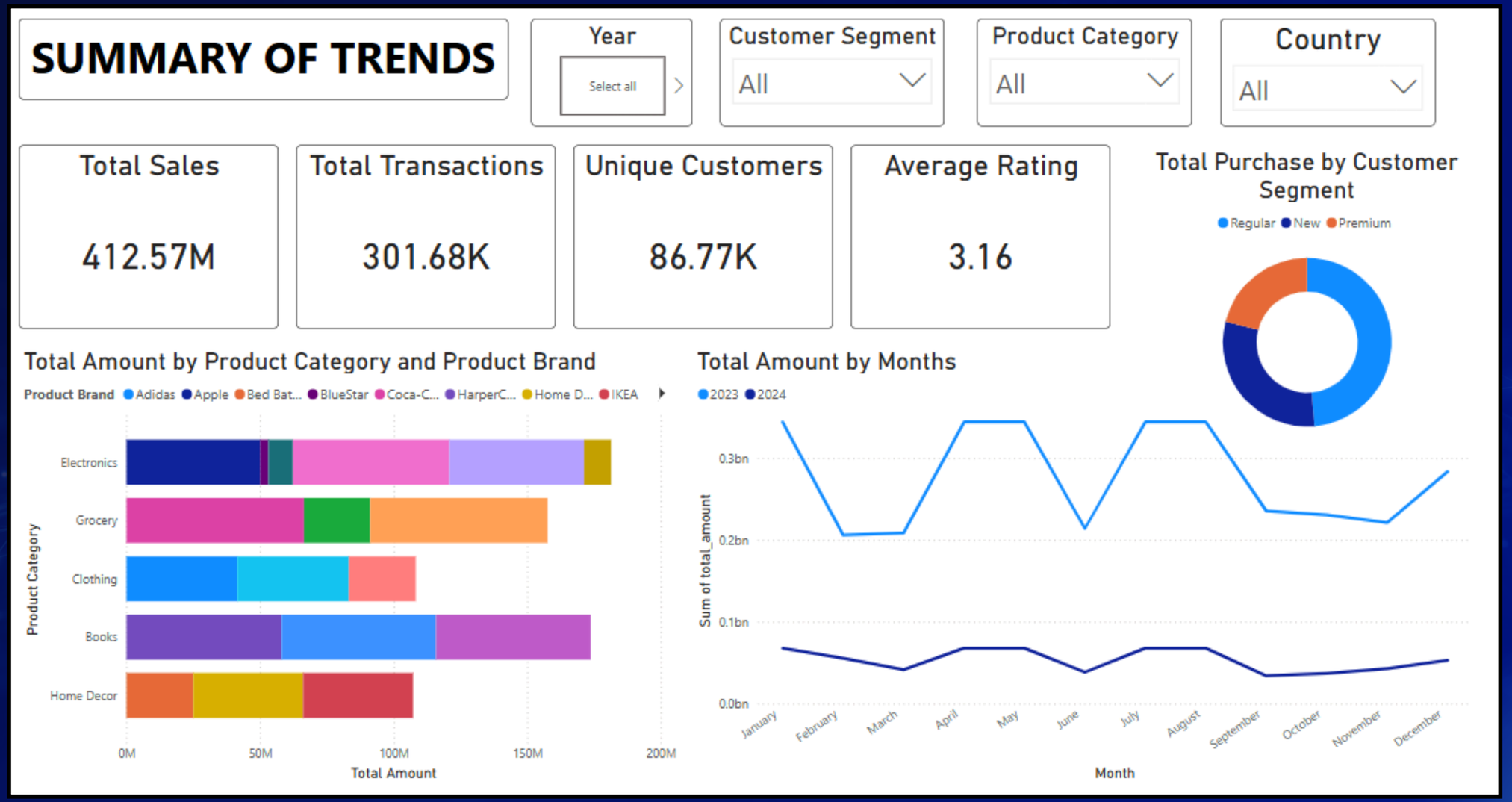
- Cleaned data types (Age, Customer_ID, etc.)
- Created views like Monthly_Sales_Trend, Total_Spend_Per_Customer
- Removed nulls and inconsistent values

| View Name | Purpose |
|---|---|
| Total_Spend_Per_Customer | Total & avg spend by customer |
| Sales_By_Product | Sales by product category & brand |
| Monthly_Sales_Trend | Aggregated sales by month |
| Sales_By_Region | Sales by country & state |
| Avg_Spend_Per_Product | Avg sales per product type |

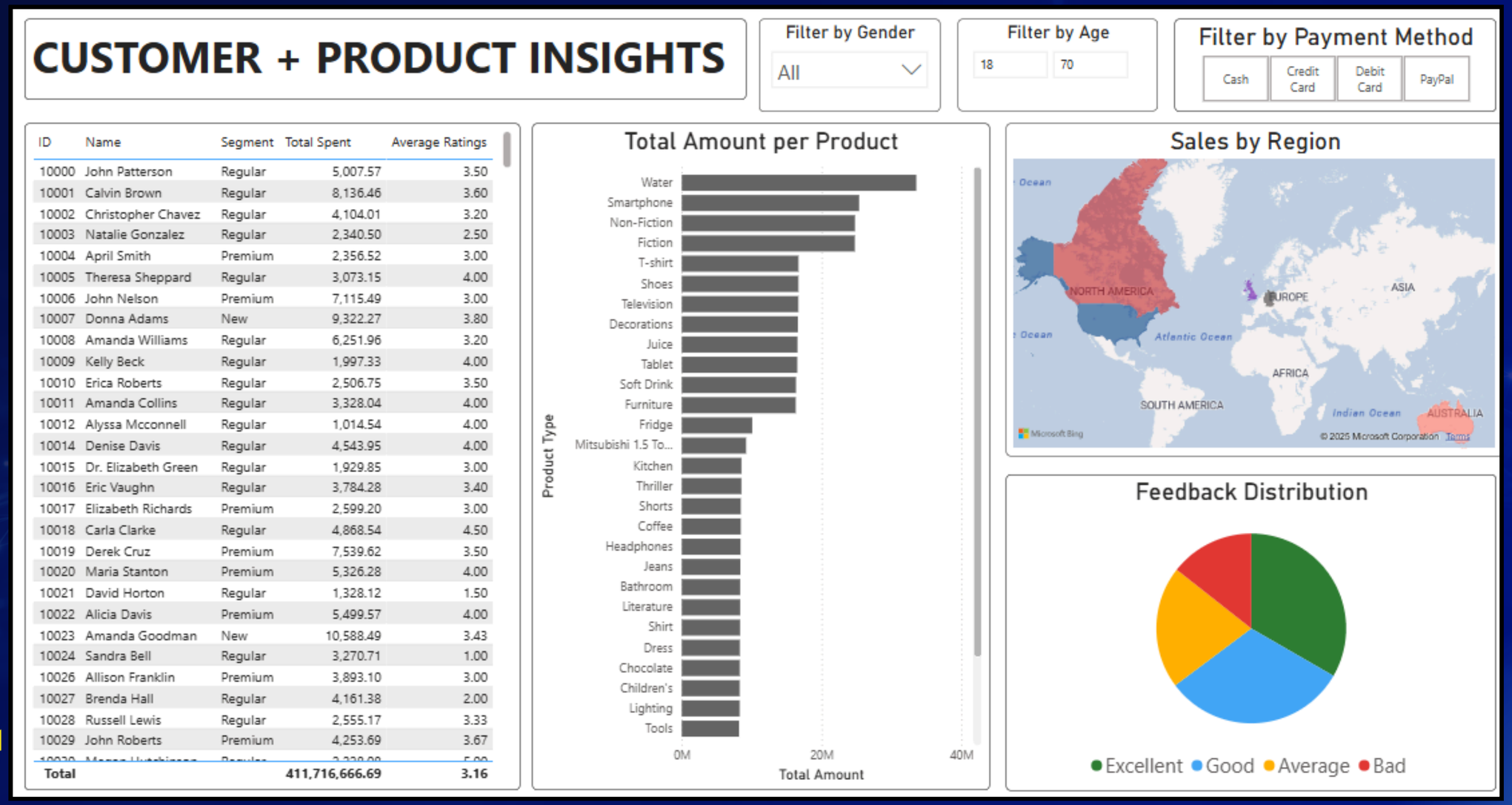
# Data Visualisation

- Creating dashboards is an iterative process.
- Multiple dashboards were made, after taking inputs and performing requirement analysis after every meeting.
- The final 2-pager dashboard has been shown here for reference.
- This is a live dashboard that can show specific data based on various data points, KPI cards, etc.

# Power BI Dashboard: Overview



# Power BI Dashboard: A Deep Dive




| ID | Name | Segment | Total Spent | Average Ratings |
|---|---|---|---|---|
| 10000 | John Patterson | Regular | 5,007.57 | 3.50 |
| 10001 | Calvin Brown | Regular | 8,136.46 | 3.60 |
| 10002 | Christopher Chavez | Regular | 4,104.01 | 3.20 |
| 10003 | Natalie Gonzalez | Regular | 2,340.50 | 2.50 |
| 10004 | April Smith | Premium | 2,356.52 | 3.00 |
| 10005 | Theresa Sheppard | Regular | 3,073.15 | 4.00 |
| 10006 | John Nelson | Premium | 7,115.49 | 3.00 |
| 10007 | Donna Adams | New | 9,322.27 | 3.80 |
| 10008 | Amanda Williams | Regular | 6,251.96 | 3.20 |
| 10009 | Kelly Beck | Regular | 1,997.33 | 4.00 |
| 10010 | Erica Roberts | Regular | 2,506.75 | 3.50 |
| 10011 | Amanda Collins | Regular | 3,328.04 | 4.00 |
| 10012 | Alyssa Mcconnell | Regular | 1,014.54 | 4.00 |
| 10014 | Denise Davis | Regular | 4,543.95 | 4.00 |
| 10015 | Dr. Elizabeth Green | Regular | 1,929.85 | 3.00 |
| 10016 | Eric Vaughn | Regular | 3,784.28 | 3.40 |
| 10017 | Elizabeth Richards | Premium | 2,599.20 | 3.00 |
| 10018 | Carla Clarke | Regular | 4,868.54 | 4.50 |
| 10019 | Derek Cruz | Premium | 7,539.62 | 3.50 |
| 10020 | Maria Stanton | Premium | 5,326.28 | 4.00 |
| 10021 | David Horton | Regular | 1,328.12 | 1.50 |
| 10022 | Alicia Davis | Premium | 5,499.57 | 4.00 |
| 10023 | Amanda Goodman | New | 10,588.49 | 3.43 |
| 10024 | Sandra Bell | Regular | 3,270.71 | 1.00 |
| 10026 | Allison Franklin | Premium | 3,893.10 | 3.00 |
| 10027 | Brenda Hall | Regular | 4,161.38 | 2.00 |
| 10028 | Russell Lewis | Regular | 2,555.17 | 3.33 |
| 10029 | John Roberts | Premium | 4,253.69 | 3.67 |
| **Total** | | | **411,716,666.69** | **3.16** |

# Key Business Insights

| Category | Insights |
|---|---|
| • Sales and Reach | The business generated ₹412.57M in total sales from over 301K transactions, reaching 86.77K unique customers across key global markets like USA, Canada, UK, and Australia. |
| • Product Performance | Electronics and Books are the top-performing product categories, with brands like Apple, BlueStar, and IKEA leading in total sales. |
| • Customer Segments | The Regular segment dominates purchases, but Premium customers show higher individual spending — signaling strong lifetime value potential. |
| • Regional Trends | North America leads in sales, followed by Europe and Australia — indicating concentrated demand in developed markets. |
| • Customer Feedback | Over 70% of feedback is positive (Excellent/Good), but a portion of Average and Bad ratings highlights areas for service or product improvement. |
| • High-spending Customers | A small group of Premium customers spent over ₹10,000 each, making them key revenue drivers despite being a minority — ideal for loyalty or VIP programs. |

# Conclusions & Learnings

This project successfully analysed a large-scale retail dataset using a full-stack data pipeline involving Python, SQL, and Power BI. Key business insights were extracted regarding sales performance, customer behaviour, product demand, and geographic trends. A fully interactive, two-page dashboard was built to enable dynamic data exploration for decision-makers.

- Gained hands-on experience with data cleaning using pandas and SQL, focusing on type corrections, missing values, and standardization.
- Understood and implemented data modeling using a star schema in PostgreSQL to structure dimensional analysis.
- Learned to create SQL views for reusable business logic like aggregations, trend breakdowns, and customer profiling.
- Developed a strong foundation in Power BI dashboarding, including slicers, interactivity, sorting logic, and storytelling through visuals.
- Explored principles of dashboard UX/UI, including layout design, color theory, and Figma wireframing for future design planning.

EY

# THANK YOU