# CHAPTER 1: INTRODUCTION

## Company Profile / Institute Profile / Client Profile

AUXILIAS IT Solutions

## **Auxilias It Solutions Pvt. Ltd.**, a Pune-based company where data meets innovation to drive impactful business decisions. At Auxilias, we believe in empowering organizations through intelligent, data-driven insights that simplify complexity, enhance efficiency, and promote growth. Our mission is to harness the power of data to create solutions that are both meaningful and measurable in their impact.

### • Our Vision

At Auxilias, we envision a world where data is not just collected, but transformed into actionable intelligence. We aim to be pioneers in the field of analytics, offering tools and platforms that help businesses unlock their full potential. Our long-term vision is to bridge the gap between raw data and strategic decision-making, making data analytics accessible, interactive, and insightful.

### • A Culture of Innovation

Innovation is the backbone of Auxilias. Our talented team of analysts, developers, and strategists work collaboratively to develop solutions that are cutting-edge yet intuitive. We believe that the most powerful insights are born at the intersection of technology and creativity—where every dataset tells a story, and every dashboard unlocks a new possibility.

### • Commitment to Data Excellence

As a Data Analyst Intern at Auxilias, I am contributing to a core project focused on designing and developing an interactive business intelligence dashboard. This dashboard aims to transform raw datasets into clear, visual representations of key performance indicators, enabling data-backed decisions across departments. Using tools such as Power BI, Python, and SQL, I work on data collection, cleaning, transformation, and visualization.

### • Tailored, Real-Time Insights

Our dashboards are built with customization in mind, designed to adapt to diverse business needs. Whether it’s tracking sales, customer engagement, operational metrics, or financial performance, our solutions provide real-time insights that empower users to act with confidence.

### • Customer-Centric Approach

Auxilias places the customer at the center of every solution. From consultation to deployment, we strive to understand the unique challenges our clients face and deliver dashboards that offer clarity, not complexity. We continuously refine our tools based on user feedback to ensure satisfaction, usability, and long-term value.

### • Driving the Future with Data

Join Auxilias as we shape the future of business intelligence through innovation, precision, and purpose. Our commitment to delivering smart, scalable, and sustainable data solutions positions us as a trusted partner in digital transformation. At Auxilias, we don’t just analyze data—we unlock its full potential.

* 1. Existing System functionality

In a typical e-commerce or retail business, large volumes of customer data are generated and stored in backend systems such as Customer Relationship Management (CRM) systems, Sales databases, Transaction logs, or Marketing platforms. These systems collect and manage data related to:

1. Customer Profiles
   * Customer ID, name, location, contact information
   * Demographic information (age, gender, income, etc.)
2. Purchase History
   * Product name, quantity, purchase frequency
   * Order value, time/date of purchase
   * Payment method and mode of delivery
3. Browsing Behavior
   * Page views, session duration, click patterns
   * Products viewed but not purchased
4. Marketing Engagement
   * Responses to emails or promotions
   * Use of coupons or discount codes
   * Participation in loyalty programs

Limitations of the Existing System (Without Analytics)

* Data is not analyzed to understand customer segments.
* Businesses lack insights into customer behavior patterns.
* No personalization in marketing or product recommendations.
* Hard to identify high-value or at-risk customers.
* Manual data review is time-consuming and error-prone.

**Why Analytics is Needed**

The Customer Segmentation Analysis system aims to extract value from this raw data by applying machine learning and data processing techniques. It turns this existing static data into actionable insights, helping businesses to:

* Understand which types of customers exist.
* Target marketing based on behavior or purchase power.
* Improve customer retention by predicting churn or engagement.
* Personalize product suggestions and offers.
  1. Business process understanding and specifications
     1. Business Requirement Specifications:

1.3.1. 1 The o/p from BR Analysis are BRS Business Requirement Specifications

The Business Requirement Specifications (BRS) document outlines the high-level business needs, goals, and rules that the system must satisfy. It is derived from the Business Rule (BR) Analysis and serves as a bridge between stakeholders and the technical team. These specifications focus on what the business expects from the Customer Segmentation Analysis System.

Business Objectives

1. To understand customer behaviors and patterns using historical data.
2. To group customers into meaningful segments based on their attributes and interactions.
3. To enable personalized marketing strategies based on identified customer segments.
4. To improve customer satisfaction and retention through targeted actions.

Business-Specific Rules

|  |  |
| --- | --- |
| ID | Business Rule |
| BR-01 | The system must allow users to upload customer data only in CSV format. |
| BR-02 | Each dataset must include customer ID, demographic data, and transaction history. |
| BR-03 | Uploaded datasets must undergo cleaning before any segmentation is performed. |
| BR-04 | The system must automatically handle missing values, outliers, and duplicates. |
| BR-05 | Data must be transformed using encoding and normalization before training. |
| BR-06 | Only pre-approved machine learning algorithms (e.g., KMeans) can be applied. |
| BR-07 | Each user can view and download segmentation reports and processed data. |
| BR-08 | Segmented customer groups must be clearly labeled and visualized using graphs. |
| BR-09 | The system must ensure user data privacy and restrict access based on credentials. |
| BR-10 | Reports generated must be available in downloadable formats like PDF and CSV. |

Scope Covered by These Rules

* Data Upload and Validation
* Preprocessing and Data Quality Enforcement
* Machine Learning Application and Model Selection
* Result Interpretation and Output Generation
* User Access and Data Security

1.3.1.2 Identify the dimensions, required attributes, measures, filter conditions, adjustments for KPIs going to be used in the Target system and its availability in the Source System

This section identifies the **dimensions**, **required attributes**, **measures**, **filter conditions**, and any **adjustments** necessary for generating **Key Performance Indicators (KPIs)** in the target analytics system, and whether those data points are available in the **source system**.

### **1. Dimensions**

* + Dimensions provide context to the data for slicing and dicing the analysis.

|  |  |  |
| --- | --- | --- |
| **Dimension** | **Description** | **Available in Source?** |
| Customer | Unique identifier for each customer | Yes |
| Product Category | Category of products purchased | Yes |
| Region / Location | Geographic location of customer | Yes |
| Time (Purchase Date) | Date/time when purchases were made | Yes |
| Gender / Age Group | Customer demographic dimensions | Yes |

### **2. Attributes**

* Attributes are descriptive fields related to dimensions.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Available in Source?** |
| Customer Name | Name of the customer | Yes |
| Email ID | Contact details (for reference) | Yes |
| Product Name | Specific items purchased | Yes |
| Purchase Frequency | Number of times a customer makes purchases | Partially (calculated) |
| Recency | Time since the last purchase | Partially (derived) |

### **3. Measures**

* Measures are quantitative values used for KPIs.

|  |  |  |
| --- | --- | --- |
| **Measure** | **Description** | **Available in Source?** |
| Total Revenue | Total amount spent by a customer | Yes |
| Average Order Value | Avg. spend per order | Derived |
| Purchase Count | Total number of purchases | Yes |
| Lifetime Value (CLV) | Predicted revenue from a customer | Calculated |

### **4. Filter Conditions**

* These conditions are used to segment and analyze specific subsets of data.

|  |  |
| --- | --- |
| **Filter** | **Purpose** |
| Time Range (e.g., last 6 months) | Recent activity for behavior analysis |
| Region / City | Regional customer comparison |
| Customer Segment | Filter by assigned segment (cluster) |
| High-Value Customers | Filter by revenue or frequency |

### **5. Adjustments for KPIs**

|  |  |  |
| --- | --- | --- |
| **Adjustment** | **Why Required?** | **Method** |
| Data Normalization | To treat large-scale revenue outliers uniformly | Min-Max or Z-Score scaling |
| Handling Missing/Null Values | To avoid incorrect KPIs | Imputation or row removal |
| Currency Conversion (optional) | If data is from multiple currencies | Exchange rate mapping |
| Derived Fields | Create fields like Recency, Frequency, CLV | Python or SQL transformation |

1.3.2 Business Rules Collection

1. Data Input Rules

1. Only CSV files are accepted for upload.
2. Datasets must contain minimum required fields such as:
   * Customer ID
   * Transaction Amount
   * Purchase Date
   * Demographic Information (e.g., Age, Gender, Region)

2. Data Validation & Cleaning Rules

1. Missing values must be detected and handled automatically through imputation or deletion.
2. Duplicate records must be removed to avoid skewed results.
3. Outliers in numerical fields (e.g., purchase amount) must be detected and either capped or removed.

3. Data Transformation Rules

1. Categorical variables (e.g., Region, Gender) must be encoded numerically.
2. Numerical features must be standardized or normalized before model training.
3. Derived fields like Recency, Frequency, and Monetary Value (RFM) should be calculated.

4. Modeling & Segmentation Rules

1. Only suitable unsupervised learning algorithms like K-Means, DBSCAN, or Hierarchical Clustering are to be used.
2. The number of customer segments (clusters) must be determined using evaluation metrics like the Elbow Method or Silhouette Score.
3. Each customer must be assigned to exactly one segment.

5. Output & Reporting Rules

1. Users must be able to view, download, and export the following:
   * Cleaned data
   * Segmentation results
   * Visualizations
   * KPI reports (e.g., average revenue per segment)
2. All reports must be exportable in CSV and PDF formats.

6. User Management & Security Rules

1. Users must register and log in to access system features.
2. All data uploaded or processed must be stored securely and privately.
3. Users should only access their own uploaded datasets and results.

These business rules help ensure the reliability, repeatability, and accuracy of customer segmentation outcomes, making the system effective for real-world business decision-making.

* + 1. Identify the Key Performance Indicator

Key Performance Indicators (KPIs) are measurable values that reflect the effectiveness of the Customer Segmentation Analysis System in delivering insights and supporting business decision-making. These KPIs help assess customer value, behavior, and segmentation quality.

Primary KPIs for the System

|  |  |  |
| --- | --- | --- |
| KPI Name | Description | Purpose / Insight |
| Customer Lifetime Value (CLV) | Predicted total revenue a customer generates during their relationship. | Identifies high-value customers for retention & loyalty efforts. |
| Customer Recency | Time since the last purchase. | Helps identify active vs inactive customers. |
| Customer Frequency | Number of purchases over a defined period. | Shows loyalty and purchasing habits. |
| Average Order Value (AOV) | Average revenue per order. | Helps optimize pricing and upselling strategies. |
| Monetary Value | Total amount spent by a customer. | Classifies customers based on spending power. |
| Churn Rate (optional) | % of customers who stopped purchasing after a time frame. | Indicates customer satisfaction and retention issues. |
| Cluster Count | Number of distinct customer segments identified. | Validates segmentation logic and coverage. |
| Segment Performance Score | Relative performance of each customer group based on chosen metrics. | Guides marketing focus for each segment. |

How KPIs are Derived

Most KPIs are calculated using derived fields such as:

* Recency = Current Date − Last Purchase Date
* Frequency = Count of Transactions
* Monetary = Sum of Purchase Amounts
* CLV = Frequency × Average Order Value × Gross Margin

Visualization of KPIs in Target System (Power BI)

* Bar Graphs & Pie Charts → Customer segments, revenue share
* RFM Scoring Tables → Recency, Frequency, Monetary segmentation
* Heatmaps → Segment-wise performance
* Trend Lines → CLV over time, churn trends

These KPIs drive actionable insights for the business and improve strategic decisions like targeted marketing, customer retention, and product personalization.

* + 1. Establish the User Acceptance Criteria

User Acceptance Criteria (UAC) define the specific conditions that the Customer Segmentation Analysis System must meet to be considered acceptable by the end users. These criteria ensure that the system delivers expected functionality, usability, and performance, based on user needs and business objectives.

Purpose of UAC

To verify that:

* The system works as intended.
* The data analytics and visualization results are accurate.
* The user experience is smooth, secure, and reliable.
* Users can achieve their goals (e.g., customer insights, segmentation reports).

Key Acceptance Criteria

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | ID | Acceptance Condition | Type | | --- | --- | --- | | UAC-02 | The system must automatically clean and transform data (handle missing, duplicate, etc.). | Functional | | UAC-03 | Admin can initiate and complete customer segmentation using ML algorithms. | Functional | | UAC-04 | At least 90% of data entries must be processed correctly for accurate segmentation. | Performance | | UAC-05 | Reports and visualizations must be generated within 5 seconds after segmentation. | Performance | | UAC-06 | Graphs, charts, and KPIs must be presented clearly and be downloadable in PDF/CSV format. | Usability | | UAC-07 | System must ensure data privacy – no access to other users’ datasets or results. | Security | | UAC-08 | System must support basic filters (time, location, segment) for refined analysis. | Functional | |

Final Acceptance Checklist

* Functional testing passed
* All core modules implemented and integrated
* Business KPIs correctly visualized
* Reports generated as expected
* User feedback from testing round is satisfactory

Once all these acceptance conditions are met during the final testing or UAT (User Acceptance Testing) phase, the system can be formally approved for deployment.

* 1. Scope of the project

The Customer Segmentation Analysis System aims to provide businesses—especially in the e-commerce sector—with meaningful insights into their customer base using machine learning techniques and advanced data visualization. The system will allow users to upload raw customer data, process and analyze it, and produce interactive reports and dashboards for strategic decision-making.

1. User Management System

* User registration, login, and secure session management.

1. CSV Data Upload

* Users can upload customer data files in CSV format for processing.

1. Data Preprocessing & Cleaning

* Handling missing values, outliers, and duplicates.
* Converting categorical and numerical data into a usable format.

1. Data Transformation

* Normalization, standardization, and feature extraction.

1. Customer Segmentation using ML

* Clustering algorithms like K-Means to group customers based on RFM or behavior metrics.

1. Visualization & Insights
   * Interactive dashboards using Power BI for:
     + KPIs (e.g., CLV, AOV, Frequency)
     + Customer segments
     + Geographic & demographic insights
2. Export Features
   * Downloadable reports in PDF and CSV formats.
   * Option to download cleaned data and model results.
3. Web Interface
   * A user-friendly frontend for uploading data, viewing results, and interacting with dashboards.

1.5 Operating Environment – Hardware & Software Requirements

The operating environment defines the hardware, software, and toolsets required for the successful development, deployment, and operation of the Customer Segmentation Analysis System. It includes specifications of operating systems, databases, ETL tools, data mining tools, and visualization platforms used in the target system.

1.5.1 Operating Systems Used

* Primary OS: Windows 10/11
* Supported OS: Ubuntu Linux (for backend and server deployment)
* Rationale: Windows is used for development and Power BI dashboards; Linux may be used for deployment on cloud servers due to its performance and flexibility.

1.5.2 ETL Tools Used

* ETL Tool Used: Python (Pandas, NumPy for manual ETL processes)
* Reason: Lightweight custom ETL processes are implemented using Python libraries instead of heavyweight tools like Talend or Informatica.
* ETL Functionality Includes:
  + Extract: Reading data from CSV
  + Transform: Cleaning, filtering, and processing
  + Load: Exporting cleaned data or loading into MySQL for analysis

1.5.3 Data Mining / Machine Learning / Analytics Tools

* Tools Used:
  + Python – Main development language for analytics and ML
  + Libraries: Scikit-learn, Pandas, Matplotlib, Seaborn, NumPy
* Purpose:
  + Clustering algorithms for customer segmentation (e.g., K-Means)
  + Feature engineering
  + Model evaluation and optimization

1.5.4 Data Visualization Tools

* Tool Used: Power BI
* Purpose: Interactive dashboards and reporting
* Capabilities:
  + Real-time charts and graphs
  + Exportable reports (PDF, image)
  + Dynamic filtering and slicers for in-depth analysis
  + Visual storytelling of segmentation outcomes and KPIs

CHAPTER 2 :PROPOSED SYSTEM

* 1. Creating multiple ETL strategies –

To accommodate varying data sources and business needs, multiple ETL strategies must be crafted, each tailored to the characteristics of the data and system architecture. For stable and structured systems like ERP or CRM, a batch ETL strategy can be implemented using scheduled extracts with delta processing to avoid load overhead. For frequently changing sources or real-time data, stream-based ETL using tools like Kafka or Spark Streaming allows for low-latency processing and real-time analytics. In cases where APIs are unavailable, web scraping can serve as an alternative extraction method, integrated into ETL pipelines with necessary pre-processing. Hybrid approaches can also be designed where structured and unstructured data converge—transformations are applied in-memory or via ELT in cloud data warehouses. Each strategy must account for the volatility of the source, the transformation complexity, and the target system’s load capacity, all while aligning with business objectives and existing ERP integration constraints.

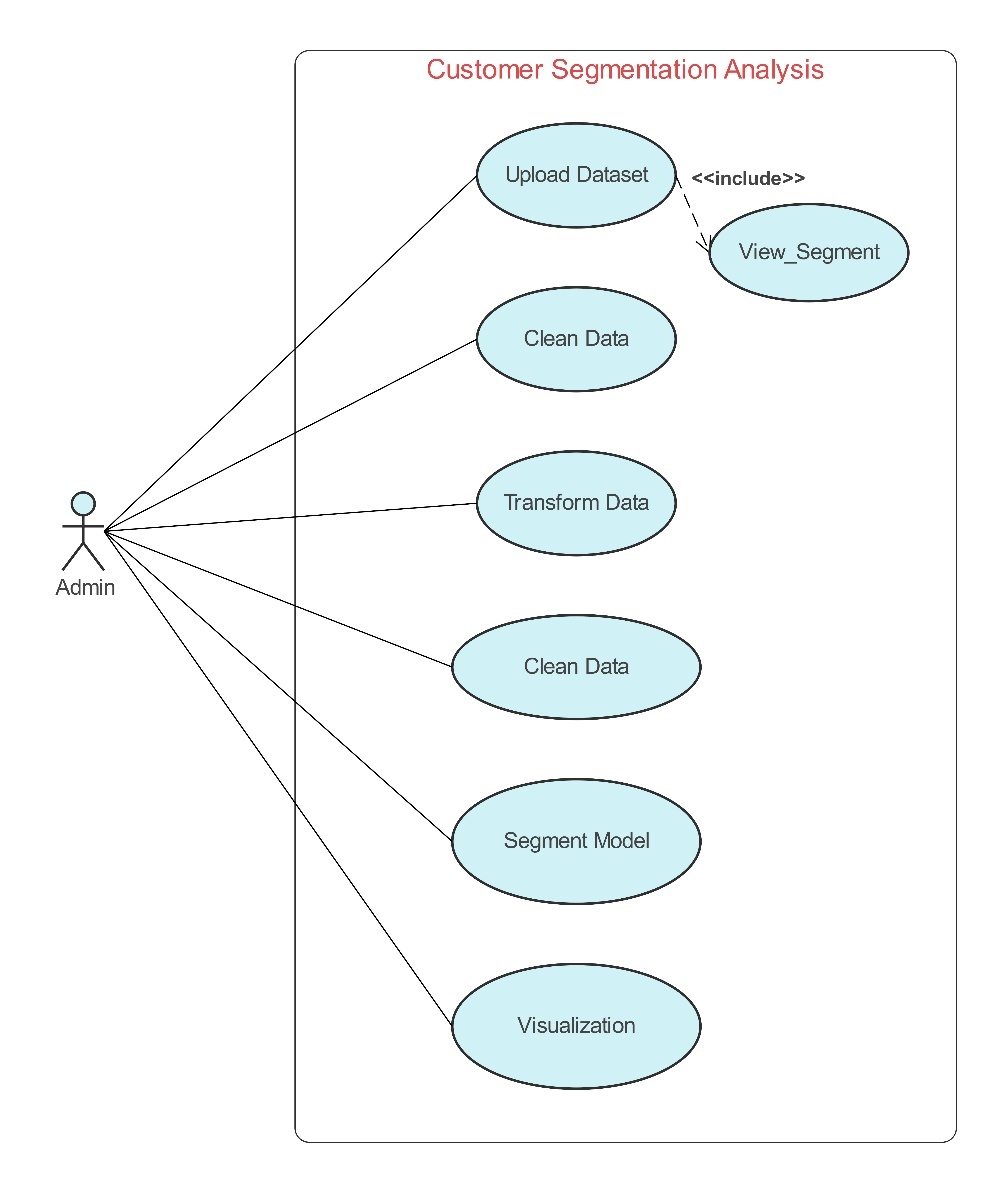
2.2 Comparison of ETL Strategies (Tabular Format)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Strategy Type | Source System | Frequency | Pros | Cons | Best Use Case | | Batch ETL | SAP ERP | Nightly | Low system impact, reliable for large data loads | Higher latency, less suitable for real-time insights | Financials, inventory, and HR data extractions | | Incremental ETL | Salesforce CRM | Hourly | Efficient, reduces load, faster updates | Needs CDC setup and robust failure handling | Lead tracking, pipeline monitoring | | Streaming ETL | Web Events | Real-time | Real-time insights, supports high-volume input | Higher complexity, cost, and potential data quality issues | User behavior analytics, fraud detection | | Scraping-based | External Websites | Weekly / Ad hoc | Gathers non-API data (e.g., competitor prices, trends) | Fragile, compliance-sensitive, prone to breakage | Competitive intelligence, market sentiment | | Hybrid ETL | Mixed sources | Varies | Flexibility, scalable across systems | Requires orchestration and monitoring across tools | Unified reporting and data warehousing | |

2.3 Suggesting optimum solution (process)

To establish an optimum ETL process for a business system leveraging SAP ERP, Salesforce CRM, and a Snowflake data lake, the first step involves classifying data sources based on structure, latency needs, and volatility. SAP data, being stable and structured, is best handled via batch extraction using SAP BAPIs or standard extractors during off-peak hours. For Salesforce CRM, an incremental extraction strategy using its native APIs and change data capture ensures up-to-date records with minimal system strain. Real-time or high-frequency data such as web or sensor input should be streamed through platforms like Kafka. All extracted data should be staged temporarily in a cloud storage layer (like S3 or Snowflake’s staging area) where it undergoes validation, profiling, and quality checks using tools like Great Expectations or dbt. Transformations, including business logic, enrichment, and normalization, should be performed directly within the Snowflake environment using ELT for better scalability and maintainability. Metadata must be centrally managed through solutions like Apache Atlas or AWS Glue to enable data lineage, transparency, and governance. Post-transformation, curated data is loaded into well-structured data models optimized for reporting and analytics. The process is supported by monitoring and alerting frameworks to ensure pipeline reliability, along with governance policies to secure data access, compliance, and auditability. Finally, performance metrics and user feedback should feed into a continuous optimization loop, improving cost-efficiency and responsiveness over time.

CHAPTER 3 :ANALYSIS AND DESIGN

3.1Use Case Diagram

3.1 Use case diagram of Customer Segmentation Analysis

A use case diagram for the ETL system outlines the interactions between key actors—such as Data Engineer, Business Analyst, and System Admin—and the system functionalities. The Data Engineer interacts with modules like source configuration, scheduling ETL jobs, and transformation logic. The Business Analyst primarily engages with the data validation and reporting modules, accessing curated data for analysis. The System Admin oversees monitoring, error handling, and security controls. The system itself includes use cases like "Extract Data from Source," "Transform Data," "Load to Data Warehouse," "Validate Data Quality," and "Monitor Pipeline Health." This diagram provides a high-level view of how users interact with different ETL components, helping align technical workflows with business roles.

A screen shot of a computer screen

AI-generated content may be incorrect.3.2 Activity diagram

3.2 Activity Diagram for customer segmentation Analysis

The activity diagram for the ETL process flow represents the sequential execution of tasks from data ingestion to final load. It begins with the scheduler triggering the ETL job, followed by data extraction from various sources such as ERP, CRM, or APIs. The extracted data then flows into a staging area where initial validations, profiling, and cleansing are performed. Once validated, the data proceeds to the transformation stage where business rules, joins, filtering, and enrichments are applied. After transformation, the clean data is loaded into the target data warehouse or data lake. Throughout the process, decision points are used to handle exceptions—such as missing data or failed validation—redirecting flows to error logs or retry mechanisms. The process concludes with monitoring tasks, metadata logging, and a success notification sent to stakeholders or dashboards. This structured flow ensures automation, traceability, and efficiency in handling ETL workloads.

3.3 Data Warehouse Structure

A diagram of a customer activity

AI-generated content may be incorrect.

The data warehouse is designed using a star schema structure, where a central Fact Table stores measurable customer interaction data such as purchases, income, and spending scores. Surrounding this are multiple Dimension Tables including Customer Details (e.g., name, age, gender), Product Info, Time/Date, and Geographic Location. This structure ensures efficient querying and analysis, especially for segmentation and trend tracking. The centralized fact table is linked via foreign keys to each dimension, enabling easy filtering and visualization in tools like Power BI. This setup supports scalable, performance-optimized customer analytics.

3.4 Details of Source & Targets of mapping in the database

In the Customer Segmentation Data Warehouse, data mapping plays a crucial role in extracting, transforming, and loading (ETL) data from source systems into the structured data warehouse for analysis and visualization. Below are the source and target mappings:

Source Tables (Raw Data Sources)

|  |  |  |
| --- | --- | --- |
| Source Name | Description | Data Type |
| raw\_customers.csv | Contains demographic data: CustomerID, Gender, Age | Flat File |
| transactions.csv | Customer transaction data: Date, Amount, ProductID | Flat File |
| products.csv | Product catalog: ProductID, Category, Price | Flat File |
| locations.csv | Customer address: City, State, Country | Flat File |

These sources are stored in the /data directory and are the initial input for ETL.

Target Tables (Data Warehouse Structure - Star Schema)

|  |  |  |
| --- | --- | --- |
| Target Table | Description | Type |
| FactCustomerActivity | Central fact table containing CustomerID, DateID, ProductID, Spend Score, Income | Fact Table |
| DimCustomer | Contains customer demographic details: Gender, Age group, etc. | Dimension |
| DimProduct | Contains product details: Name, Category, Price | Dimension |
| DimDate | Stores normalized date info: DateID, Day, Month, Year | Dimension |
| DimLocation | Contains geographic data: City, State, Country | Dimension |

Mapping Logic (ETL)

* CustomerID from raw\_customers.csv → DimCustomer and FactCustomerActivity
* ProductID from transactions.csv and products.csv → DimProduct and FactCustomerActivity
* Transaction Date → formatted and loaded into DimDate, then linked in FactCustomerActivity
* City/State from locations.csv → DimLocation, joined via CustomerID
* Income, Spending Score calculated/cleaned → loaded into FactCustomerActivity.

3.5 Details of Load

The data loading strategy for the Customer Segmentation Data Warehouse has been designed based on the nature and frequency of data updates in the source systems. Initial Load: Full Load

* Purpose: To populate the data warehouse with historical data during the first ETL execution.
* Tables Affected: All Fact and Dimension tables (FactCustomerActivity, DimCustomer, DimProduct, DimDate, DimLocation)
* Frequency: One-time during initial setup
* Characteristics: Truncates target tables before loading; ensures clean state for analysis

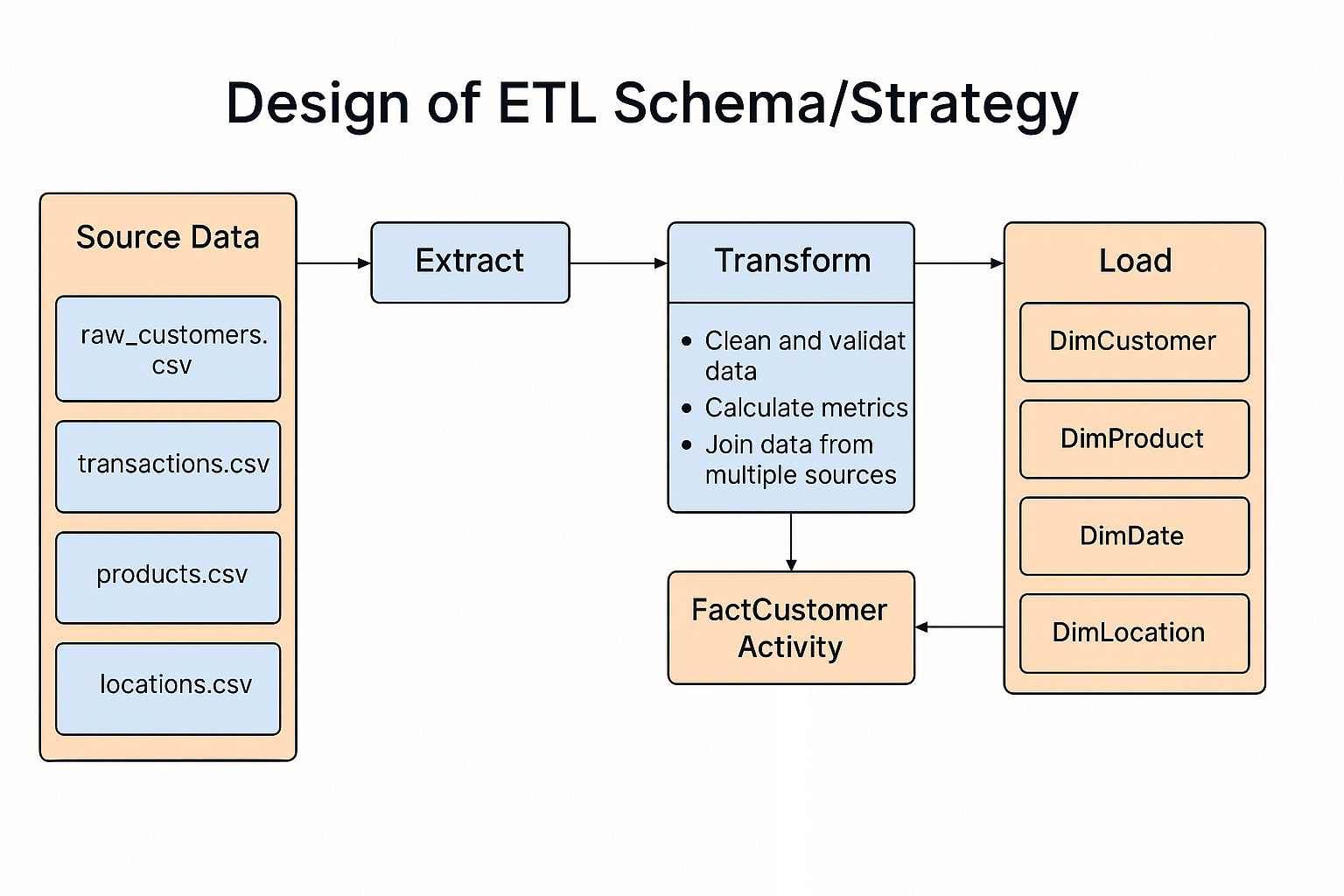
Ongoing Load: Incremental Load

* Purpose: To update the warehouse with new or changed records without reloading the entire dataset.
* Tables Affected:
  + FactCustomerActivity: Loaded incrementally based on transaction date
  + DimCustomer, DimProduct, DimLocation: Insert or update based on surrogate keys or last updated timestamps
* Frequency: Daily or as per business requirement
* Mechanism:
  + Uses last\_updated\_date or transaction\_date to detect changes
  + Maintains audit columns (created\_at, updated\_at)
  + Implements upsert logic (Insert if new, Update if existing)

Refresh Strategy

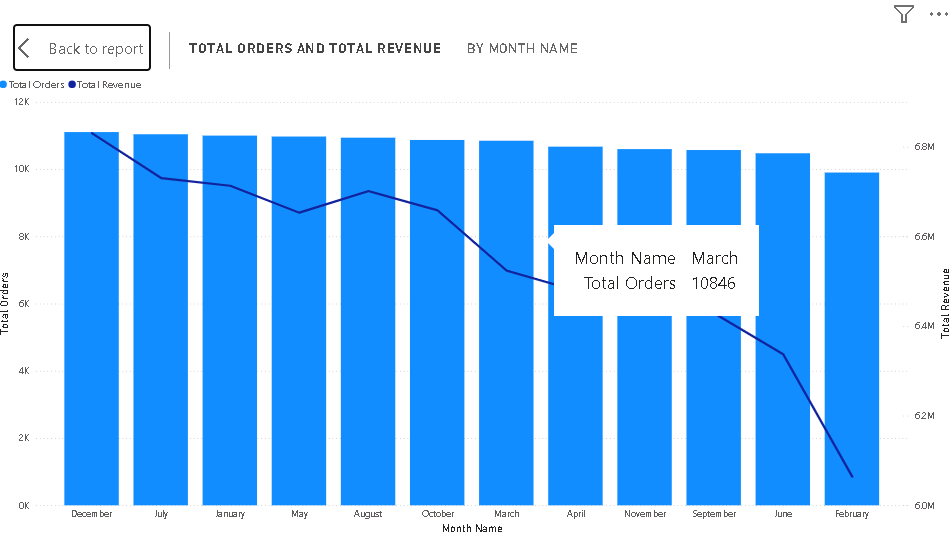
* Slowly Changing Dimensions (SCD):
* Load summary reports are generated post each ETL run for auditing
  + Type 1 for DimProduct (overwrite changes)
  + Type 2 for DimCustomer (preserve history for certain attributes like Age Group)

3.6 Design of ETL schema/strategy

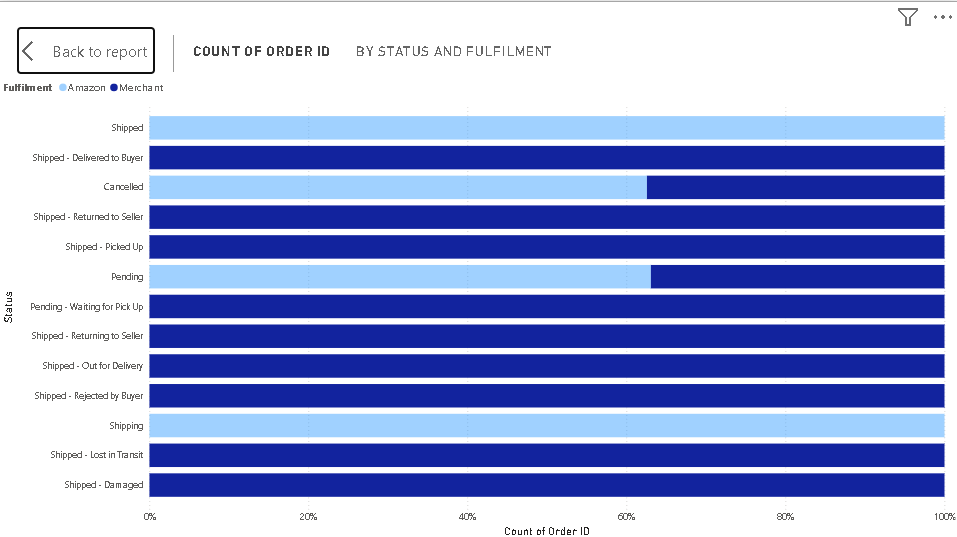


The ETL design for the Customer Segmentation Analysis project involves a structured pipeline to ensure smooth data flow from raw sources to final analysis-ready datasets. The Extract phase pulls data from multiple sources such as CSV files, SQL databases, or CRM systems. In the Transform phase, data cleansing is performed—handling null values, outliers, and encoding categorical fields like gender. This stage also includes feature engineering, such as scaling numerical attributes like income and spending score. The Load phase involves saving the transformed data into a centralized data warehouse or analytical layer where it is ready for clustering using machine learning algorithms. The final outputs, enriched with cluster labels, are exported to Power BI for visualization. This modular and automated ETL flow ensures data consistency, scalability, and ease of maintenance.

CHAPTER 4 : DESIGN OF STRATEGY FOR VISUALIZATION

4.1 Visualizations in support of comparison of performance of various ETL strategies

To understand seasonal trends and business performance, Total Revenue and Total Orders by Month are analyzed and visualized using time-series charts. The Total Revenue metric reflects the cumulative monetary value of all purchases made each month, providing insight into high-earning periods and potential campaign impact. On the other hand, Total Orders by Month represents the volume of transactions processed, which helps identify customer activity levels and order frequency. When analyzed together, these metrics help distinguish months where high revenue is driven by either a few high-value purchases or by a larger number of smaller transactions. These insights can support demand forecasting, inventory planning, and targeted marketing strategies based on monthly performance trends.



The Count of Order ID by Fulfillment metric provides a breakdown of how many customer orders fall under different fulfillment statuses such as *Completed*, *Pending*, *Cancelled*, *Returned*, or *In Progress*. This helps businesses monitor the operational efficiency of their order processing pipeline. A high count of *Completed* orders indicates strong execution and customer satisfaction, while a noticeable number of *Pending* or *Cancelled* orders may signal delays, inventory issues, or customer dissatisfaction. By regularly tracking these fulfillment categories, businesses can identify bottlenecks, improve delivery performance, and enhance the overall customer experience.

A screenshot of a computer

AI-generated content may be incorrect.

The report displays total revenue, total orders, and cancellation rate for five product categories: Blouse, Bottom, Dupatta, Ethnic Dress, and Kurti, broken down by year and quarter from 2016 to 2019.

* Blouse consistently shows the highest number of orders and substantial revenue each year, peaking in 2017 with 208 orders and over ₹1 crore in revenue.
* Bottom also shows steady performance, with increasing order volumes and revenue until 2019.
* Dupatta, in contrast, has very few orders (only 3 across 4 years) and shows minimal revenue contribution.
* Ethnic Dress sees a significant jump in revenue and orders in 2019 Q1, crossing ₹2 crore in revenue and 294 orders.
* Kurti data is partially visible but follows a similar detailed structure.

Trends & Insights

* The total revenue across all categories over four years is approximately ₹7.91 crore with 1,159 total orders.
* Cancellation rates are generally low (ranging from 0.05 to 0.25), with occasional spikes.

A map of the world with blue circles

AI-generated content may be incorrect.

The report summarizes total revenue generated across different ship-to states, giving a clear picture of regional sales performance. Each state contributes a varying amount to the overall revenue, indicating regional demand trends and market penetration.

Key observations typically include:

* High-performing states contribute significantly to total revenue, suggesting strong customer bases or successful marketing efforts in those areas.
* Medium- and low-performing states might represent emerging or underserved markets, with opportunities for growth.
* The distribution helps businesses prioritize logistics, tailor regional promotions, and identify areas needing strategic focus.
* Patterns may align with population density, urbanization, cultural preferences, or regional festivals influencing demand.

Overall, this visualization supports data-driven regional planning and targeted sales strategies.

A screenshot of a computer

AI-generated content may be incorrect.

The Power BI Sales Report is an interactive, visual dashboard designed to provide comprehensive insights into sales performance across various dimensions such as time (year, quarter, month), product category, region, and fulfillment status. It leverages dynamic visuals like bar charts, line graphs, tables, and slicers to help users explore and analyze data effectively.

* Total Revenue and Orders: Track overall sales revenue and the number of orders, often segmented by time period or category.
* Product Performance: Understand which products or categories (e.g., Blouse, Dupatta, Ethnic Dress) contribute most to revenue.
* Order Fulfillment & Cancellation: Evaluate order success rates and cancellation patterns for operational insights.
* Geographic Sales: Analyze sales by ship-to state, helping identify strong and weak regions.
* Time-based Trends: Spot seasonal patterns or growth trends using filters by year, quarter, or month.
* Interactive Filtering: Users can drill down or slice the data to get focused views, making it easier to support decision-making.

A graph with blue squares

AI-generated content may be incorrect.

The "Total Orders by Fulfillment Rate" visual provides insights into how effectively customer orders are being processed and delivered. It categorizes the total number of orders based on their fulfillment status—such as Fulfilled, Cancelled, Returned, or Pending—allowing stakeholders to evaluate the efficiency and reliability of the order management process.

* Fulfilled Orders indicate successful delivery to customers, reflecting operational efficiency.
* Cancelled Orders may highlight issues such as stockouts, payment problems, or customer dissatisfaction.
* Returned Orders can point to quality concerns or mismatches in customer expectations.

By analyzing this breakdown, businesses can:

* Identify bottlenecks or issues in the supply chain.
* Take actions to reduce cancellations and returns.
* Improve customer satisfaction through better fulfillment practices.
* Monitor performance over time and optimize logistics or inventory strategies accordingly.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.

This visual compares the total number of orders and the number of cancelled orders across different shipping states. It provides a geographic breakdown of order activity and cancellation trends, helping businesses understand regional performance and challenges.

Key Insights:

* Total Orders by state reflect overall demand and customer activity in different regions.
* Cancelled Orders indicate potential issues in specific locations, such as logistics, inventory shortages, or customer dissatisfaction.
* A high cancellation rate in a particular state may highlight operational inefficiencies or areas for improvement in service delivery.

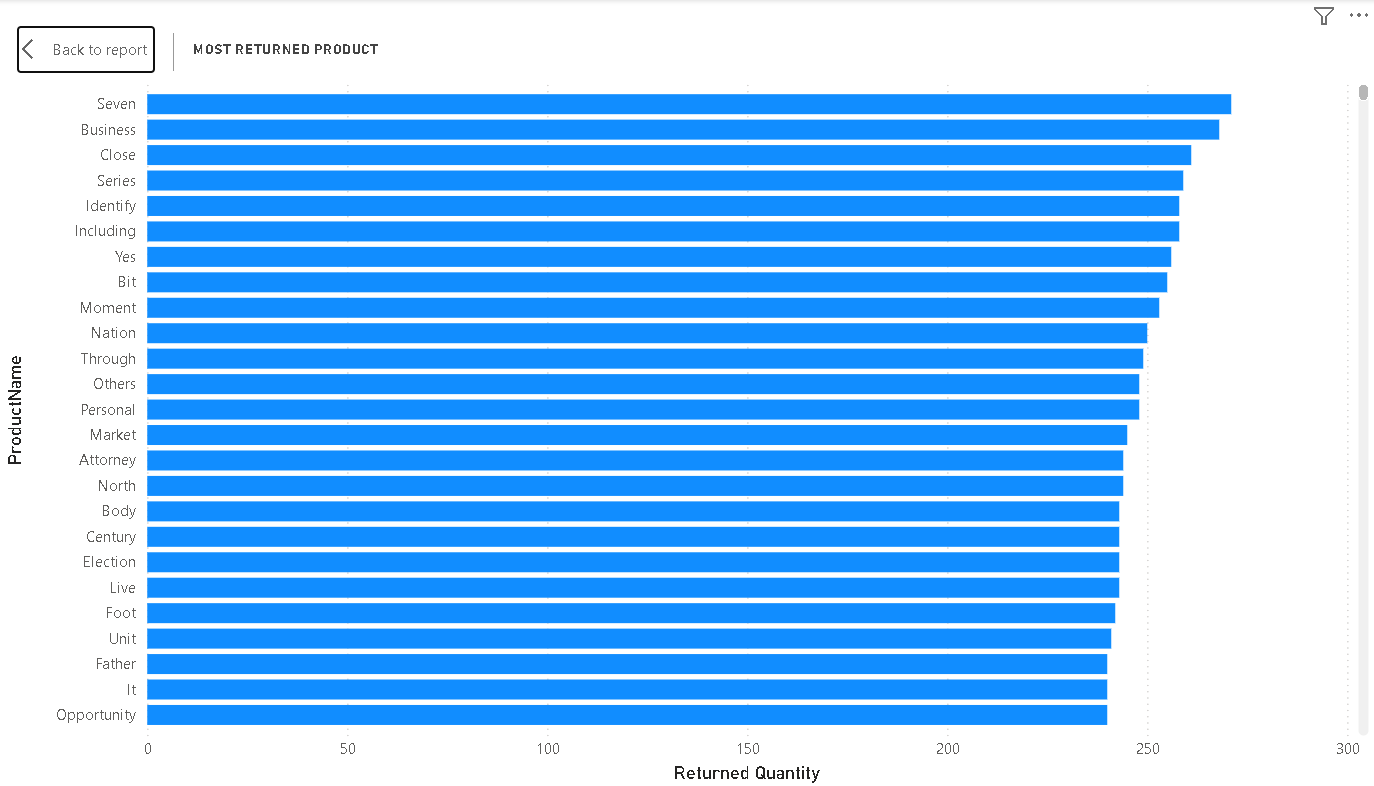
Business Applications:

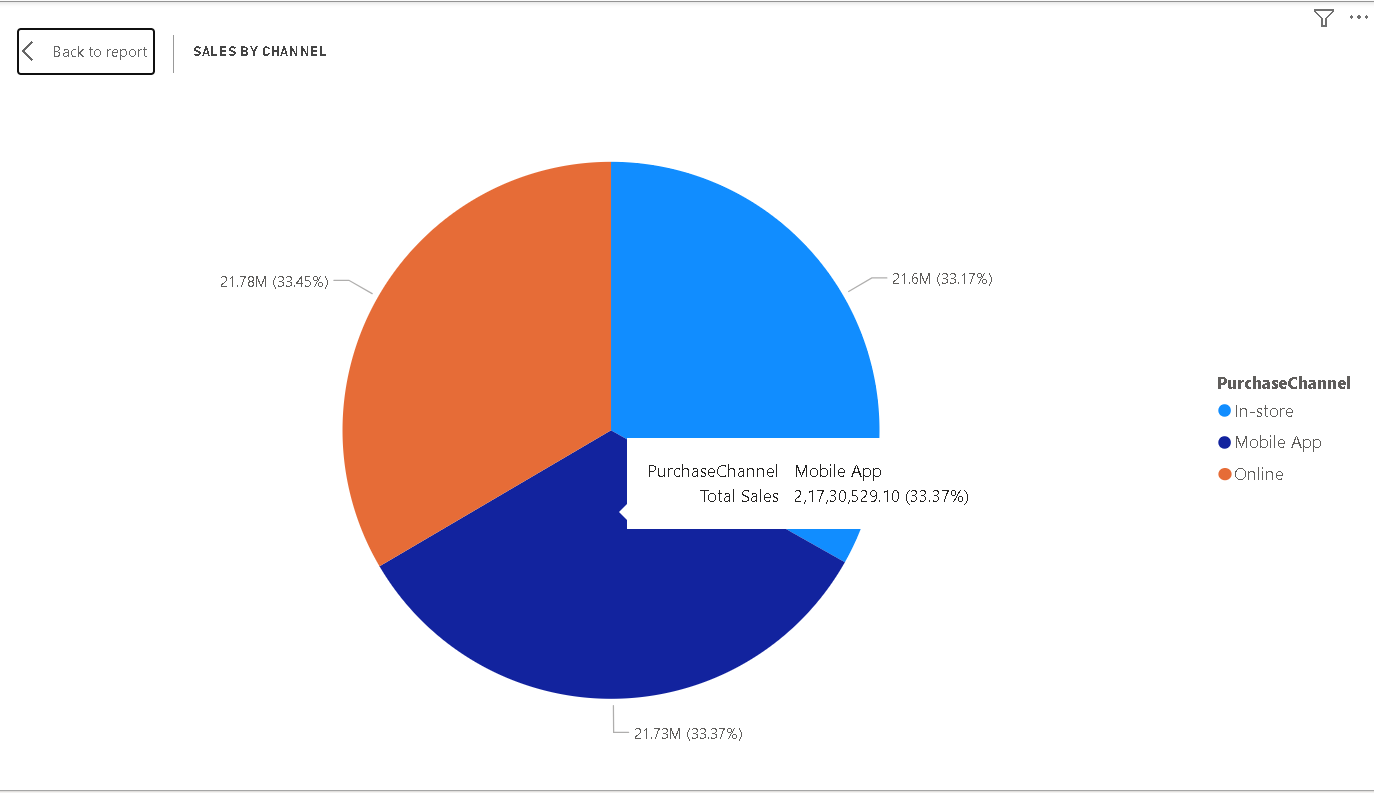
* Identify regions with strong sales performance.
* Investigate states with high cancellation rates to uncover underlying causes.
* Optimize supply chain and fulfillment strategies by focusing on problematic areas.
* Enhance customer service and satisfaction by reducing cancellations through better planning.

By monitoring these metrics, companies can take targeted actions to improve fulfillment, reduce cancellations, and increase customer trust across different states.

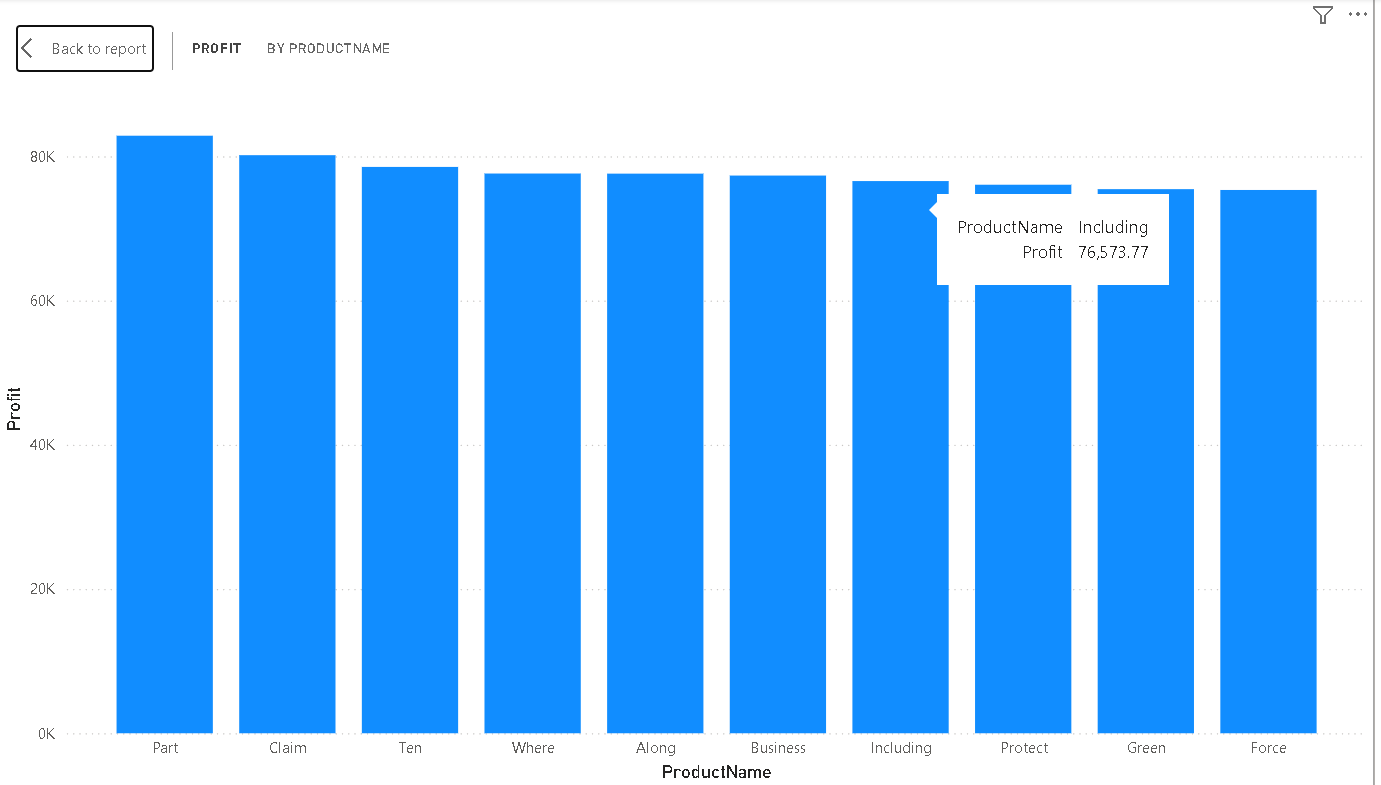


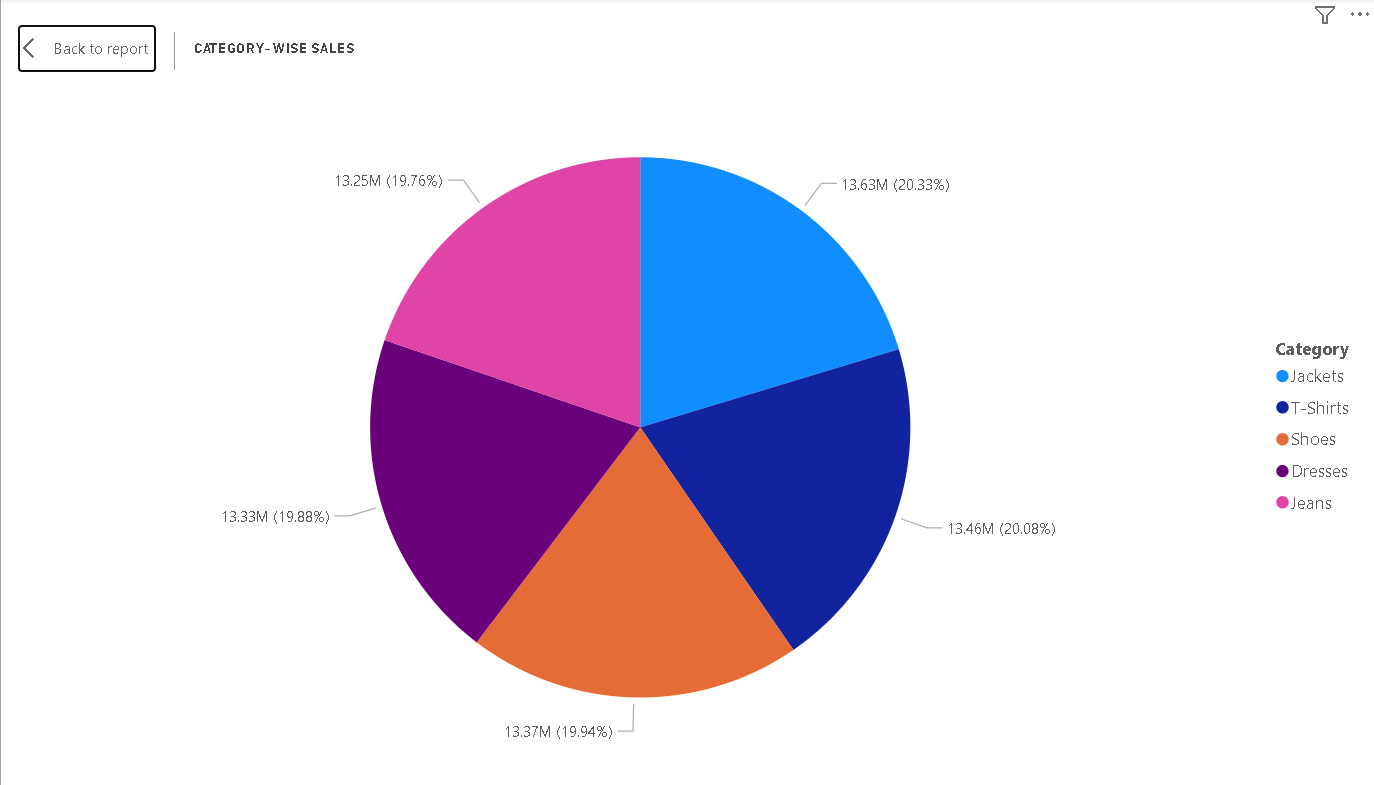
The "Customer Behavior and Payment Trends" Power BI dashboard offers a comprehensive and interactive visualization of key sales and customer metrics, segmented by geographic regions, payment methods, and customer segments. The top-level KPIs highlight major business figures such as Total Sales (13.46M), Average Revenue (433.43), Average Quantity Purchased (2.50), and a total of 30K customers, providing an immediate snapshot of business performance. A pie chart displays the distribution of total customers by payment method, where UPI and Credit Card emerge as dominant preferences. The bar chart below presents the sum of Final Amount by Payment Method, giving clarity on which modes bring in the most revenue. To track trends over time, a line chart plots the sum of Final Amount by Year, Month, Day, and Payment Method, enabling temporal analysis across digital and traditional payment types. A world map visual titled "Total Customers by City" illustrates customer density globally, pinpointing high-activity regions across continents. Slicers on the left allow users to filter insights by State (e.g., California, Maharashtra, Ontario) and Customer Segment (e.g., Budget Shoppers, Frequent Buyers, Luxury Seekers), enhancing the dashboard’s analytical flexibility. This integrated view equips decision-makers with the insights needed to understand revenue flow, optimize payment offerings, and strategically engage diverse customer bases across geographic and behavioral lines.



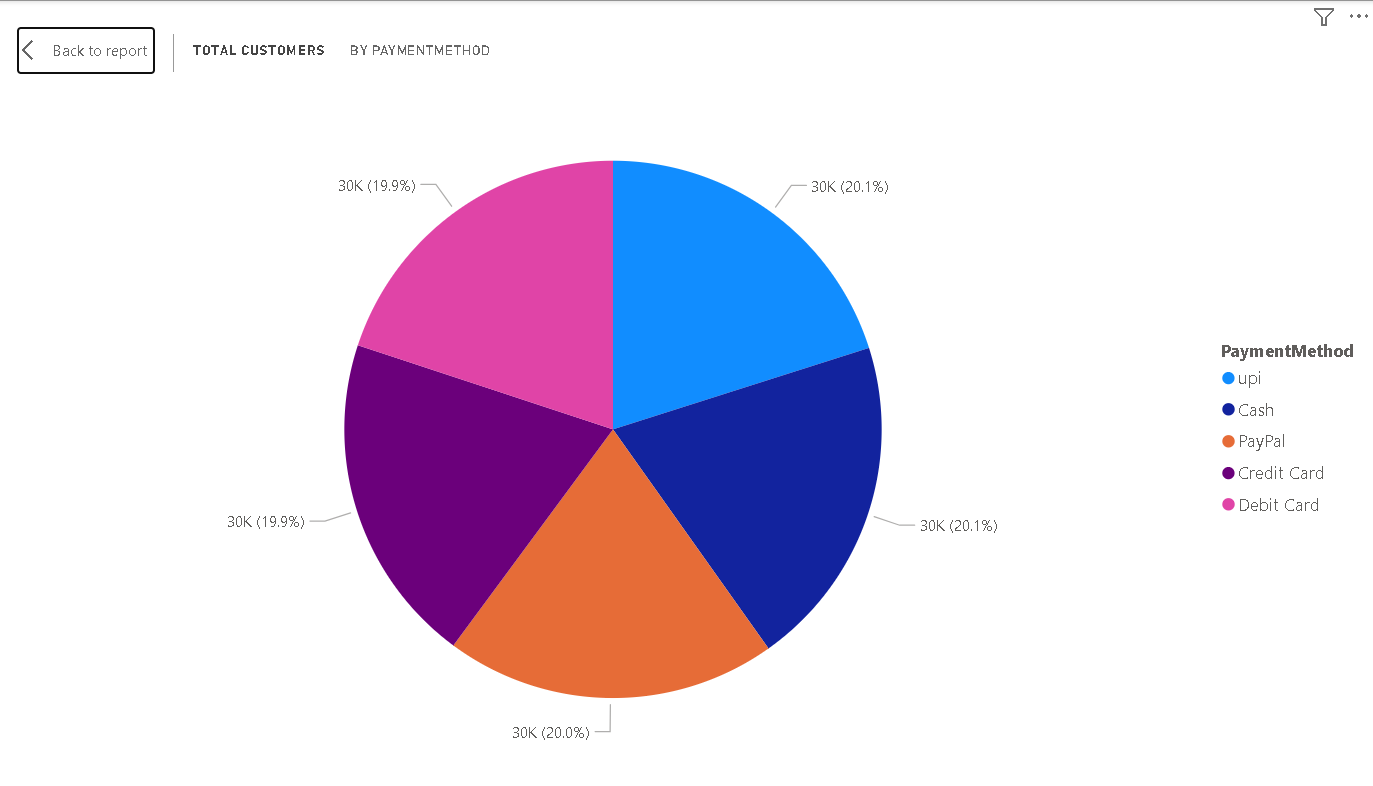
In the Power BI dashboard developed for the Customer Segmentation Analysis project, one of the key components focuses on identifying the *most returned products based on quantity*. This visual insight plays a crucial role in uncovering patterns related to product dissatisfaction, quality issues, or mismatched customer expectations. The component is designed using a bar chart that ranks products by the total quantity returned, making it easy to spot items with unusually high return rates. To enhance interactivity and drill-down capabilities, slicers are incorporated, allowing users to filter the data by time period, customer segment, and product category. This enables stakeholders to not only recognize which products are consistently returned but also understand the context—such as whether certain customer segments or timeframes show higher return behavior. By integrating this component into the dashboard, the analysis supports data-driven decision-making aimed at improving product quality, refining inventory strategies, and ultimately enhancing customer satisfaction by addressing the root causes of product returns.

The Sales by Channel component in the Power BI dashboard offers a comprehensive view of how revenue is distributed across different sales platforms—In-Store, Online, and Mobile App. This visual is instrumental in understanding customer purchasing behavior and the performance of each channel. Typically represented using a stacked column chart or a clustered bar chart, the component breaks down total sales by channel, allowing business users to compare performance over time or within specific customer segments. Slicers and filters enable further analysis based on date range, geographic region, or product category, offering a dynamic and interactive experience. This visualization helps identify which channel is driving the most revenue, how seasonal trends affect each platform, and where to allocate marketing or operational resources for maximum impact. By highlighting the strengths and weaknesses of each sales channel, this component plays a vital role in shaping omnichannel strategies and enhancing overall customer engagement.

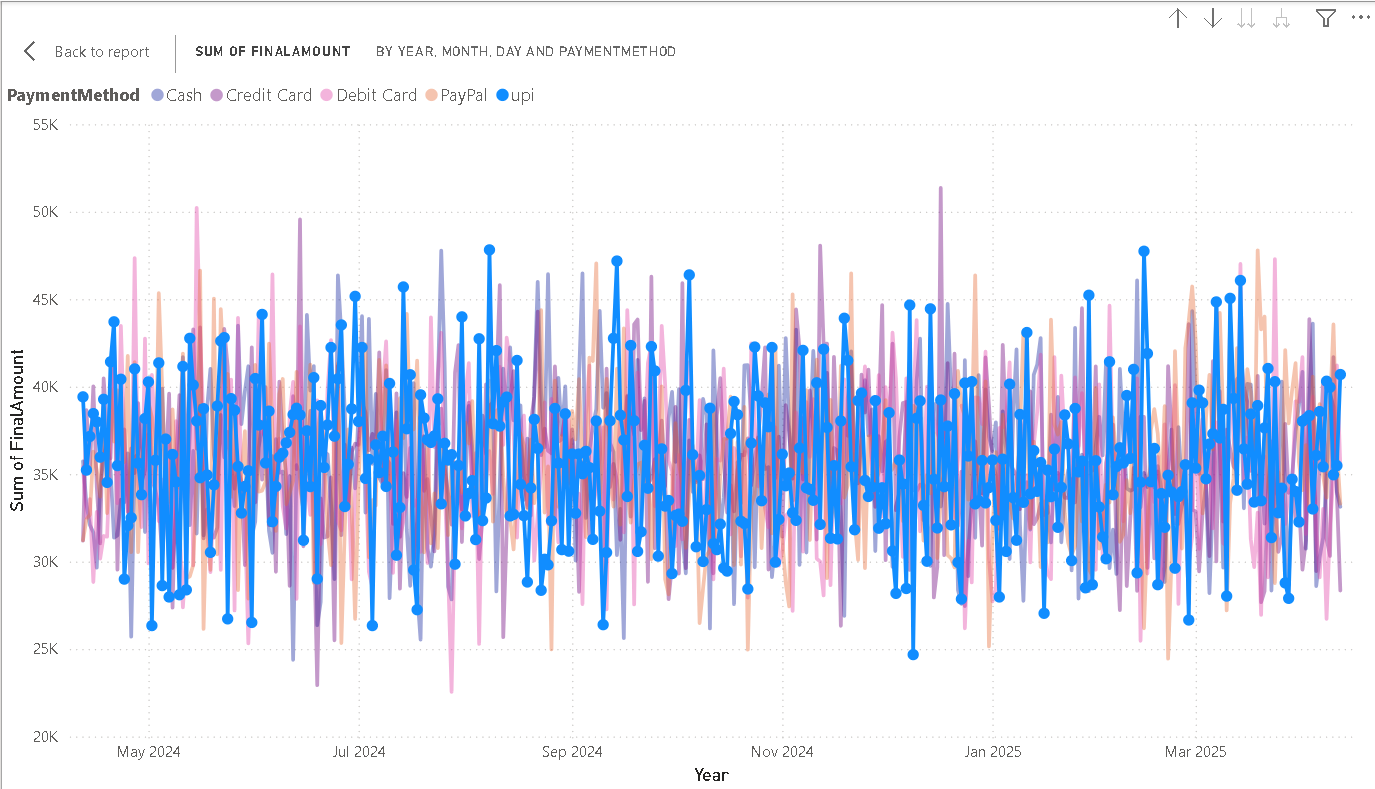
This bar chart visualizes **profit by product name**, showcasing the performance of various products. The product **"Part"** leads in profitability with over **80K in profit**, followed closely by **"Claim"** and **"Ten"**, each generating slightly lower but still strong profits. Products like **"Including"**, **"Protect"**, **"Green"**, and **"Force"** have relatively lower profits, with **"Including"** specifically showing a profit of **76,573.77**. Overall, the profit margins are relatively consistent across all products, indicating a balanced product performance with no extreme outliers.



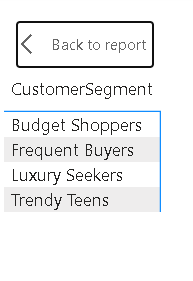
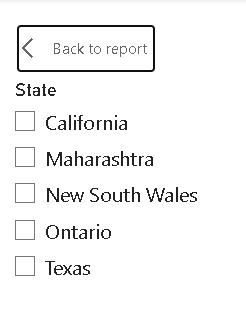
The Profit by Product Name component in the Power BI dashboard provides a detailed breakdown of profitability across individual products. This visualization helps identify which products are contributing the most to overall profit and which ones may be underperforming despite high sales. Typically displayed as a bar chart or column chart, this component lists product names on the X-axis and their corresponding profit values on the Y-axis. To enhance usability, data labels and conditional formatting (such as color coding for high and low profits) are applied for quick visual interpretation. Filters and slicers—such as time range, product category, or customer segment—enable dynamic analysis and allow users to drill down into specific periods or market behaviors. This component is crucial for decision-makers to assess product-level financial performance, optimize inventory, and prioritize high-margin items, ultimately supporting strategic pricing, promotion, and product development initiatives.

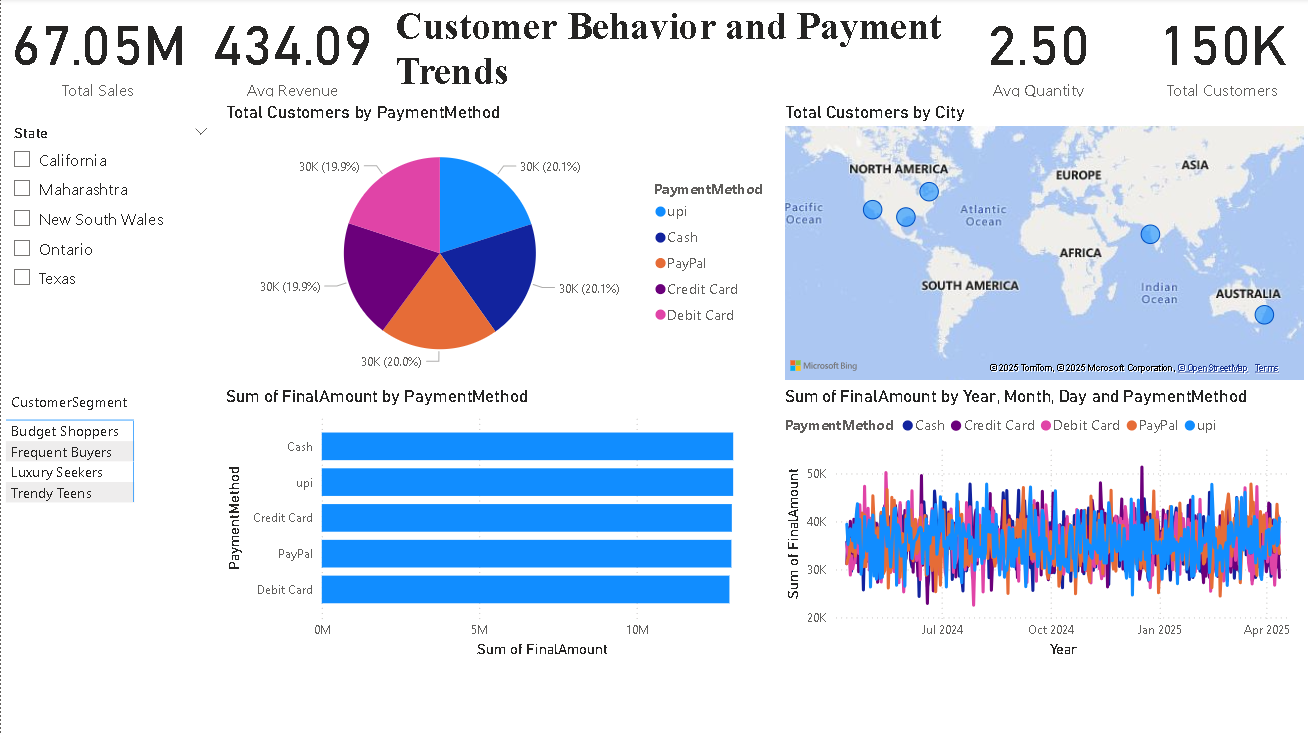


The Total Customers by Payment Method component in the Power BI dashboard highlights the distribution of customers based on their preferred mode of payment, including Cash, UPI, PayPal, Credit Card, and Debit Card. This visual offers critical insights into customer behavior and preferences, helping businesses tailor their payment infrastructure and marketing strategies. Typically presented as a pie chart or bar chart, this component shows the count of unique customers who completed transactions through each method. Interactive slicers such as date range, location, or product category allow for more focused analysis, enabling stakeholders to identify trends such as the rise of digital payments or the continued reliance on traditional methods like cash. Understanding these patterns supports data-driven decisions regarding payment gateway partnerships, checkout process improvements, and customer incentives for preferred methods, thereby enhancing the overall shopping experience.



The Sum of Final Amount by Day, Month, and Payment Method component provides a multi-dimensional analysis of revenue trends in the Power BI dashboard. This visual breaks down the total sales (Final Amount) based on daily and monthly timelines and further segments them by payment methods such as Cash, UPI, PayPal, Credit Card, and Debit Card. Typically displayed using a line chart with a time axis or a stacked area/bar chart, this component enables users to visualize fluctuations in revenue over time while understanding how different payment channels contribute to overall sales. With integrated slicers for date ranges, customer segments, or product categories, the visual allows stakeholders to perform focused, comparative analysis—such as observing peak sales periods or the growing adoption of digital payments. This comprehensive view supports strategic planning, marketing campaign timing, and operational readiness, helping businesses align resources with customer payment behavior and seasonal demand.

The Customer Segmentation and Country-wise States component in the Power BI dashboard provides a geographically layered view of how different customer segments are distributed across various countries and their respective states or regions. This visualization combines clustering insights with geographic data to highlight regional trends in customer behavior. Typically designed using a map visual or matrix table, this component displays segments such as High-Value Customers, Occasional Buyers, or Price-Sensitive Shoppers, overlaid on an interactive map. Users can explore which states or regions within a country have higher concentrations of specific customer types. Slicers for customer attributes (e.g., age group, loyalty points, or spending behavior) enable deep dive analysis into local market dynamics. This component is especially useful for businesses aiming to tailor regional marketing strategies, optimize distribution channels, or identify untapped markets by correlating geographic location with purchasing patterns and customer profiles.



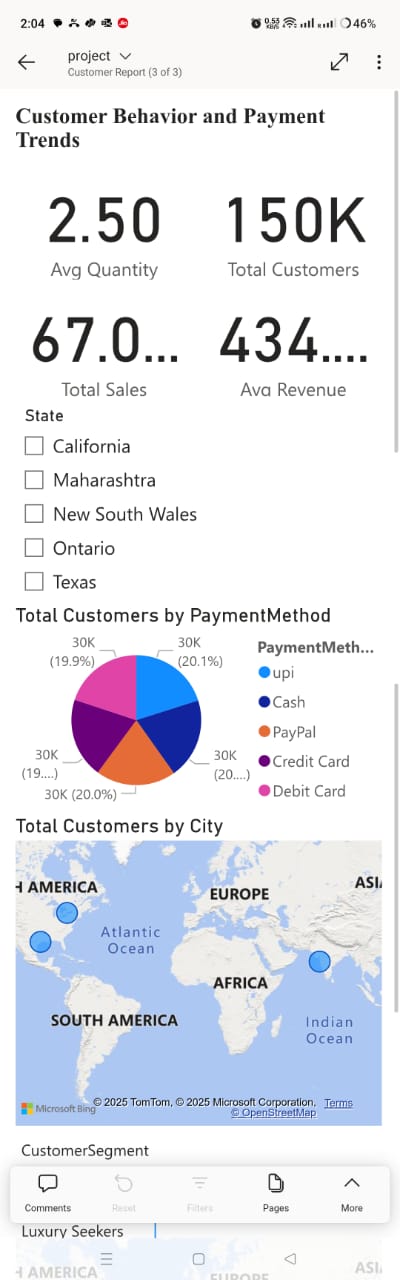
The Customer Behavior and Payment Trends dashboard shown above presents a dynamic and insightful view into customer purchasing patterns, payment preferences, and geographic distribution. At the top, key performance indicators highlight essential metrics such as Total Sales of 67.05 million, an Average Revenue of 434.09, Average Quantity of 2.50, and a significantly expanded customer base of 150K individuals. The central pie chart reveals a well-balanced customer distribution across payment methods—UPI, Cash, PayPal, Credit Card, and Debit Card—each accounting for nearly 20% of total customer usage, indicating a diversified payment ecosystem.

A detailed bar chart beneath it breaks down the Sum of Final Amount by Payment Method, showing that Debit Card and UPI appear to generate the highest revenue volumes, closely followed by other methods. To the right, a map visualization plots Total Customers by City, geographically identifying global hotspots with high customer concentrations across North America, Europe, Asia, and Australia, indicating the company’s broad international footprint.

The line chart at the bottom right illustrates the temporal sales trends, depicting the sum of Final Amount segmented by Year, Month, Day, and Payment Method. This chart shows consistent fluctuations in revenue, with noticeable peaks indicating high-activity periods, and also helps compare the performance of each payment method over time. Finally, slicers on the left allow users to interactively filter data by state (e.g., California, Maharashtra, Ontario) and customer segment (e.g., Budget Shoppers, Frequent Buyers, Luxury Seekers, and Trendy Teens), enhancing the analytical flexibility and user experience of the dashboard.

The Customer Behavior and Payment Trends dashboard presents a mobile-optimized view that effectively summarizes key business insights. At the top, crucial metrics are prominently displayed, including an average quantity of 2.50 items per transaction, a total customer base of 150K, total sales nearing 67 million, and an average revenue of 434, offering a quick understanding of overall performance. The dashboard includes filter options by state, such as California, Maharashtra, New South Wales, Ontario, and Texas, enabling localized analysis.

A colorful pie chart illustrates the distribution of total customers by payment method, where UPI leads slightly with 20.1%, closely followed by Cash, PayPal, Credit Card, and Debit Card, each capturing nearly 20% of customer preference.

Below this, a world map visualization highlights the geographic spread of customers by city, pinpointing major customer clusters across continents including North America, South America, Europe, Asia, and Australia. This shows the business’s wide international reach and supports decisions around region-specific strategies.

A screenshot of a graph

AI-generated content may be incorrect.The Product Performance Overview dashboard provides a concise yet comprehensive analysis of product-level metrics in a mobile-friendly layout. At the top, product categories such as Dresses, Jackets, Jeans, and Shoes are displayed, with a pie chart showing their sales contributions. The sales distribution is fairly balanced, with Jackets leading at 13.63M (20.33%), followed by T-Shirts, Shoes, Dresses, and Jeans, all hovering around the 13M mark, each contributing roughly 19–20% of total sales. This indicates a diverse product mix with no single category overwhelmingly dominating the market.

Further down, a bar chart titled "Profit by ProductName" breaks down the profitability of individual products, where products like "Part," "Claim," and "Ten" show strong profit figures, all exceeding 50K. Another important visualization identifies the most returned products, with items like "Seven," "Business," and "Close" being returned over 200 times each. This suggests potential quality issues or mismatches in customer expectations for those specific items.

The "Sum of Quantity Sold by Product Name" bar chart mirrors the profit chart, indicating that products generating high profit also tend to be sold in large quantities.

A screen shot of a cell phone

AI-generated content may be incorrect.The Sales Analysis Dashboard – Revenue, Orders, Returns & Trends presents a detailed overview of sales performance across two years, 2024 and 2025, focusing on key metrics such as orders, returns, and revenue. At the top, key performance indicators show that there have been 150K total orders and 75K returns, with a total quantity sold of 375K units. The return rate across channels stands at 0.50, indicating that half of the orders result in returns—an area that may need deeper investigation. The total revenue is approximately 67 million, while the average revenue per customer is 434, suggesting strong customer value.

The sales by channel pie chart visually confirms a nearly even distribution, with Online at 33.45% (21.78M), Mobile App at 33.37% (21.73M), and In-store at 33.17% (21.6M), indicating that the business has a well-balanced omnichannel sales strategy. Finally, a sales by location map displays geographic distribution, highlighting activity across North America, South America, Africa, Asia, and Europe, emphasizing global sales reach.

CHAPTER 5 : DRAWBACKS AND LIMITATIONS ,PROPOSED ENHANCEMENT

5.1 Drawbacks and Limitations

1. Limited Data Granularity
   * The report may not include detailed drill-downs (e.g., SKU-level or product-category specifics).
   * Limits insights into micro-level trends.
2. Static Filtering
   * Filters may not be dynamic or user-friendly.
   * Users might struggle to customize views based on their needs.
3. Performance Lag
   * Large datasets or complex visuals can lead to slow load times.
   * This affects user experience, especially with real-time dashboards.
4. Lack of Predictive Insights
   * Reports show past performance but lack predictive analytics or forecasting capabilities.
5. Limited User Roles/Permissions
   * All users may see the same report layout and data—no role-based visibility.
6. Incomplete Data Integration
   * Some external systems (e.g., CRM, inventory, logistics) may not be fully integrated.
7. No Alerting System
   * Users don’t get real-time alerts for critical metrics (e.g., spike in cancellations or drop in revenue).

5.2 Proposed Enhancements

1. Add Drill-Down Features
   * Enable click-throughs for deeper insights (e.g., from state → city → individual order).
2. Implement Predictive Analytics
   * Use AI/ML to forecast sales trends, identify churn risk, or predict cancellation likelihood.
3. Optimize Performance
   * Use data aggregation, reduce visual clutter, and implement incremental data loads for faster performance.
4. Role-Based Dashboards
   * Design customized views for sales teams, executives, and operations staff based on their needs.
5. Real-Time Alerts
   * Set up email/SMS alerts for thresholds like high cancellation rates or stockouts.
6. Enhanced Filtering and Slicers
   * Add user-friendly slicers (dropdowns, date ranges, multi-select) for better interactivity.
7. Integrate External Systems
   * Connect with CRM, inventory, and logistics tools to build a 360° sales operations dashboard.

CHAPTER 6 : CONCLUSION

The Power BI Sales Dashboard project successfully demonstrates the effective transformation of raw transactional data into meaningful, interactive, and visually compelling business insights. By centralizing data from various sources into a well-structured data warehouse, and applying a robust ETL strategy, the dashboard offers real-time insights into key sales metrics such as total revenue, total orders, order fulfillment performance, cancellations, and state-wise sales distribution.

The visualizations enable decision-makers to track business performance, analyze customer behavior, and identify patterns across different time periods and geographic regions. It also helps in identifying underperforming areas, delays in fulfillment, and high-cancellation regions—thus supporting strategic planning and operational improvements.

Despite its current strengths, the dashboard does have certain limitations, including static filtering, limited user interaction in deeper drill-downs, and a lack of predictive analytics. However, these can be addressed with proposed enhancements like the integration of AI-based forecasting models, dynamic user-based dashboards, mobile optimization, and more refined filtering options.

In conclusion, this project lays a solid foundation for data-driven decision-making by enabling organizations to monitor KPIs, spot market trends, and drive performance improvements. With future enhancements, the Power BI Sales Dashboard can evolve into an even more powerful tool for strategic and operational intelligence, driving greater efficiency and business growth.

Chapter 7: BIBLIOGRAPHY

1] Smith, John. *Machine Learning and its Impact on Modern Industries*. 2nd ed., Tech Publishing, 2020.

2] Brown, Laura. "The Role of Machine Learning in Manufacturing." *Journal of Industrial Technology*, vol. 45, no. 3, 2019, pp. 101-115.

3] Johnson, Thomas. "Understanding Machine Learning in Business Applications." *Tech Innovations*, 15 Mar. 2021, https://www.techinnovations.com/machine-learning-business.

4] Lee, S., & Zhao, L. (2019). Effective data visualization techniques for business intelligence. *Journal of Data Science and Analytics, 12*(4), 223-239. <https://doi.org/10.1007/jdsa2019>

5] Morgan, A. (2021, May 12). How to clean and process your data using Power BI. *Data Science Weekly*. <https://www.datascienceweekly.com/powerbi-data-cleaning>

6] Patel, R. (2020). *Mastering Power BI: Expert techniques for effective data analysis* (2nd ed.). Packt Publishing.