

## JANUARY 18, 2024

Etl project

Report

etl project report

**T**he project report, encapsulates a comprehensive exploration into the realms of data extraction, transformation, and visualization. Commencing with the extraction phase, data was sourced from three different sources, laying the foundation for a rich and varied dataset. Subsequently, a sophisticated transformation process unfolded, wherein diverse data warehouse and mining queries were applied to refine and enhance the extracted information. The transformed data was then seamlessly loaded into a robust database, ensuring efficiency and accessibility. The final pinnacle of the project involved the creation of a dynamic Power BI report, harnessing the visualization capabilities of the tool to present insights in a visually compelling manner. This report chronicles the journey from raw data extraction to insightful visualization, underscoring the project's commitment to harnessing the power of data for informed decision-making.

# Table of contents

*1. Introduction*

- Project Overview

- Data Set Description

- Problem Statement

*2. Technologies Used*

- Node JS (Backend)

- React JS (Frontend)

- MongoDB (Database)

- PowerBi (Analysis)

*3. ETL Operations*

*- 3.1 Extract*

- Data Structure

- Data Extraction Process (CSV, XLSX, XML)

*- 3.2 Transform*

- Summarize (Sales and Profit Aggregation)

- Categorize (Product Classification by Sales)

- Enrichment (Demand Assessment)

- New Columns Introduction

*- 3.3 Load*

- Data Loading into MongoDB

*4. Visualization*

- Roll Up Visualization

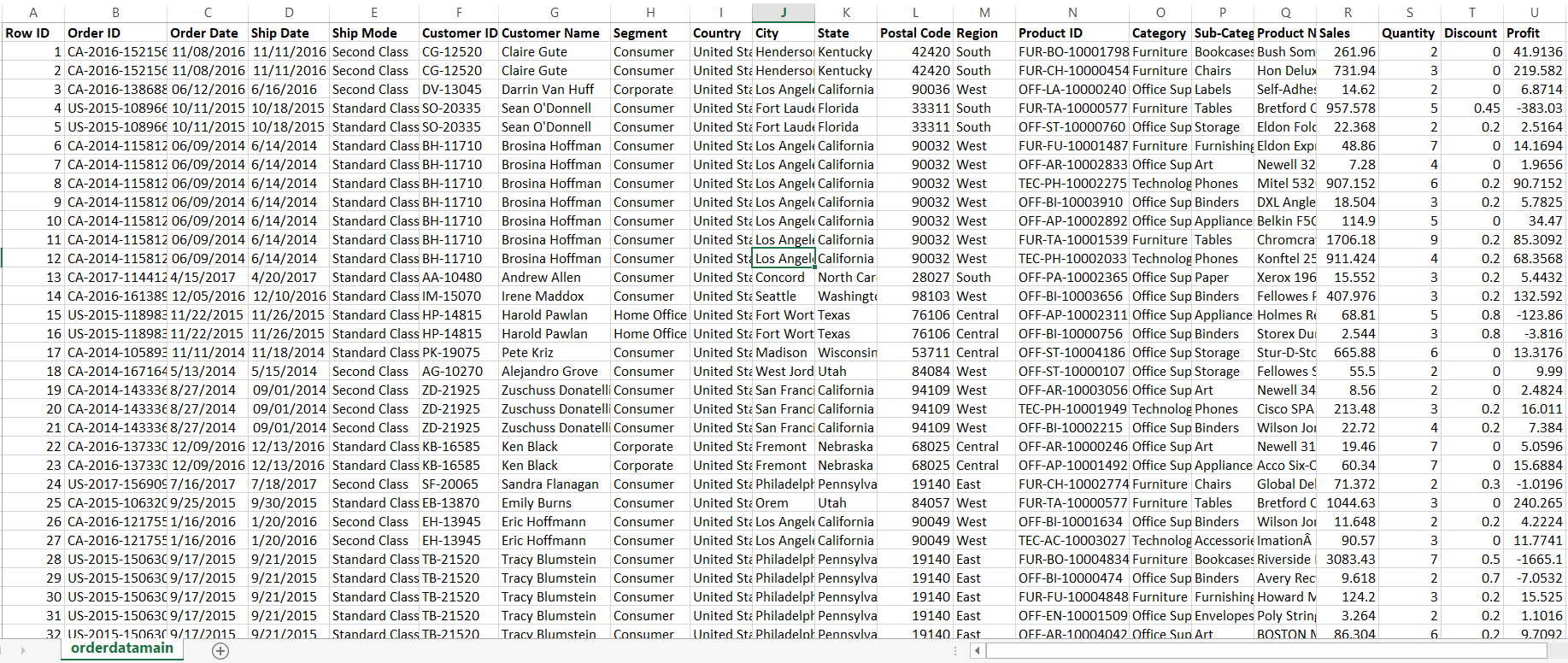
- Drill Down Visualization

*5. Conclusion*

# 1. INTRODUCTION

**Data Set:**

The dataset captures comprehensive sales data in the United States, delving into product specifics such as sales, profit, and more. This rich dataset provides nuanced insights into individual product performance, offering a detailed perspective on market dynamics. Its depth allows for a thorough analysis, uncovering patterns and trends crucial for understanding sales and profitability in the United States.



**Problem Statement:**

This study seeks to determine the financial viability of the sales dataset within the United States. It endeavors to discern the presence of profitability and, if affirmed, aims to elucidate the underlying causative factors. Conversely, in instances of financial loss, the investigation is oriented towards comprehending the contributory factors shaping the adverse financial outcome.

# 2. Technologies used

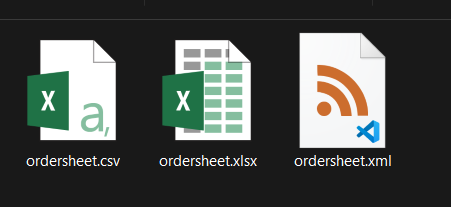
1. Node JS (Backend)
2. React JS (Frontend)
3. MongoDB (Database)
4. PowerBi (Analysis)

# 3. ETL OPERATIONS

3.1 Extract

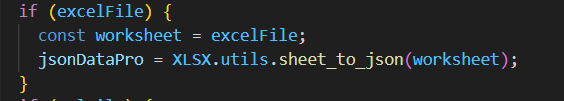
Data Structure

The data in the files exhibits unique rows across the three file types: CSV, XLSX, and XML. While the data differs between files, the structure remains consistent.



Data extraction involves uploading three distinct file types: CSV, XLSX, and XML. The data from these files will be extracted and converted into JSON format for further processing.

* 1. **Excel File:**

****

* 1. **Xml File:**

****

* 1. **Csv File:**

****

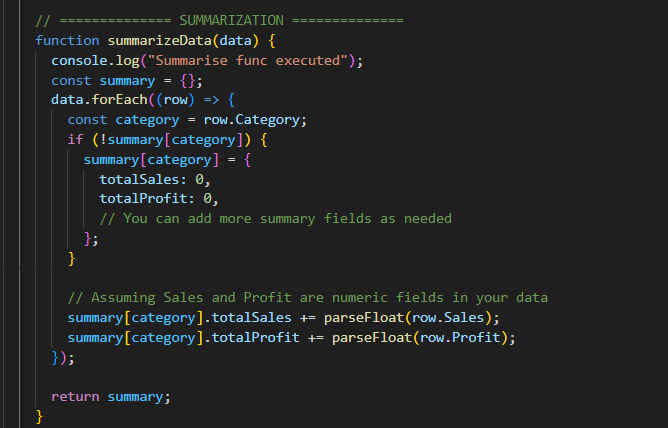
3.2 transform

The transformation phase comprises three main functions:

**Summarize:**

Aggregates sales and profit based on product categories.

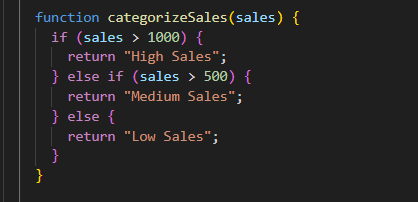
* + Total sales and profit for the "Technology" category.
  + Total sales and profit for the "Office Supplies" category.



**Categorize:**

Classifies products into three categories based on sales:

* + Low Sale: Sales < 500
  + Medium Sale: 500 ≤ Sales < 1000
  + High Sale: Sales ≥ 1000



**Enrichment:**

Determines demand for a specific quantity. If the product’s quantity is greater than 4, it is marked as high demand; otherwise, it is labeled as medium demand.

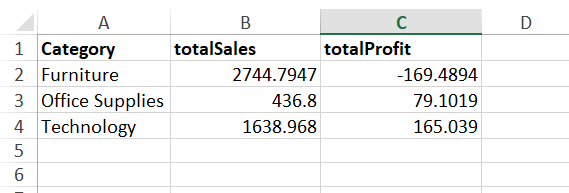


# New Columns added:

After transformation, new columns are introduced:

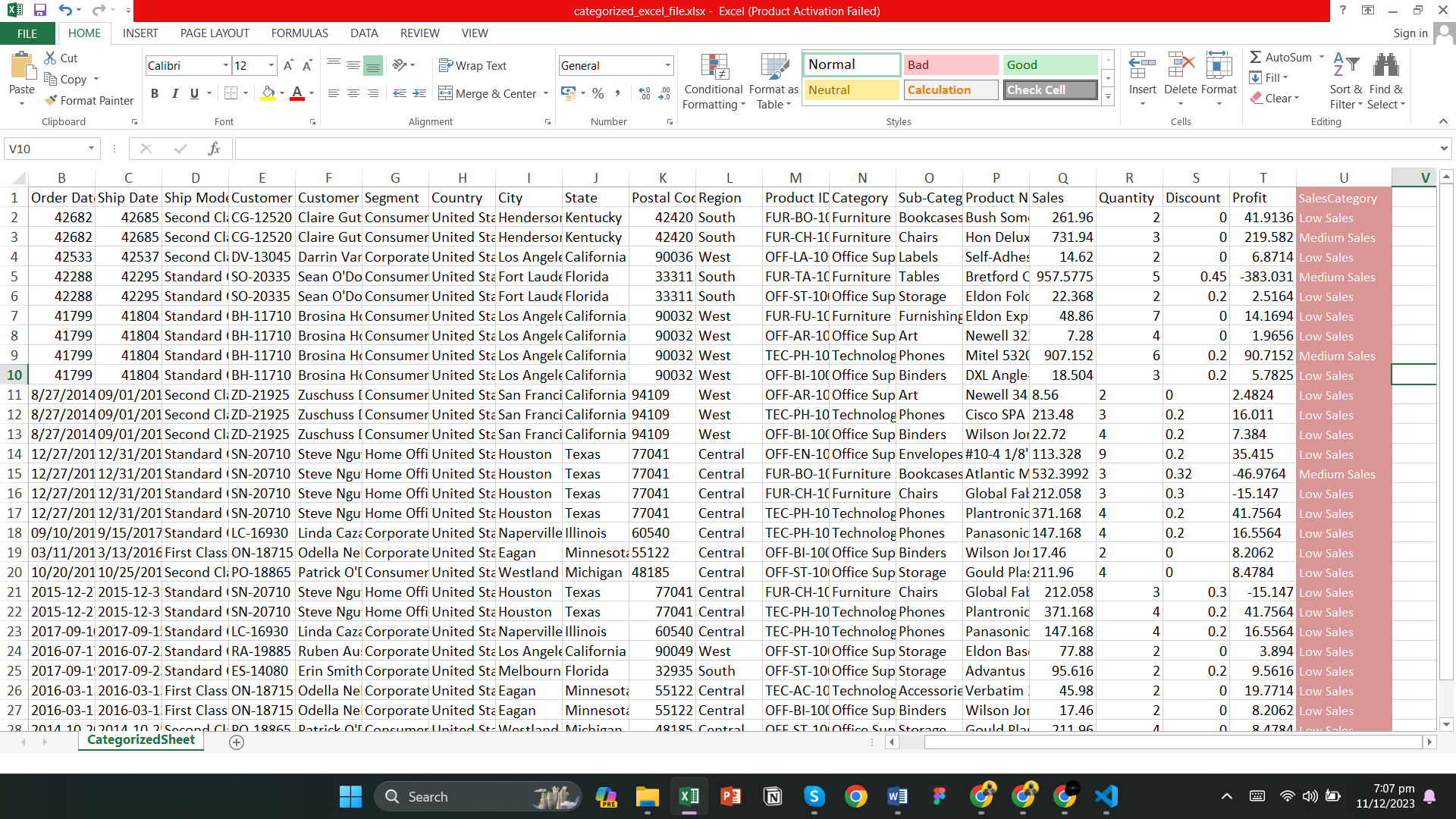
**Summarize:**

* + Total Sales
  + Total Profit



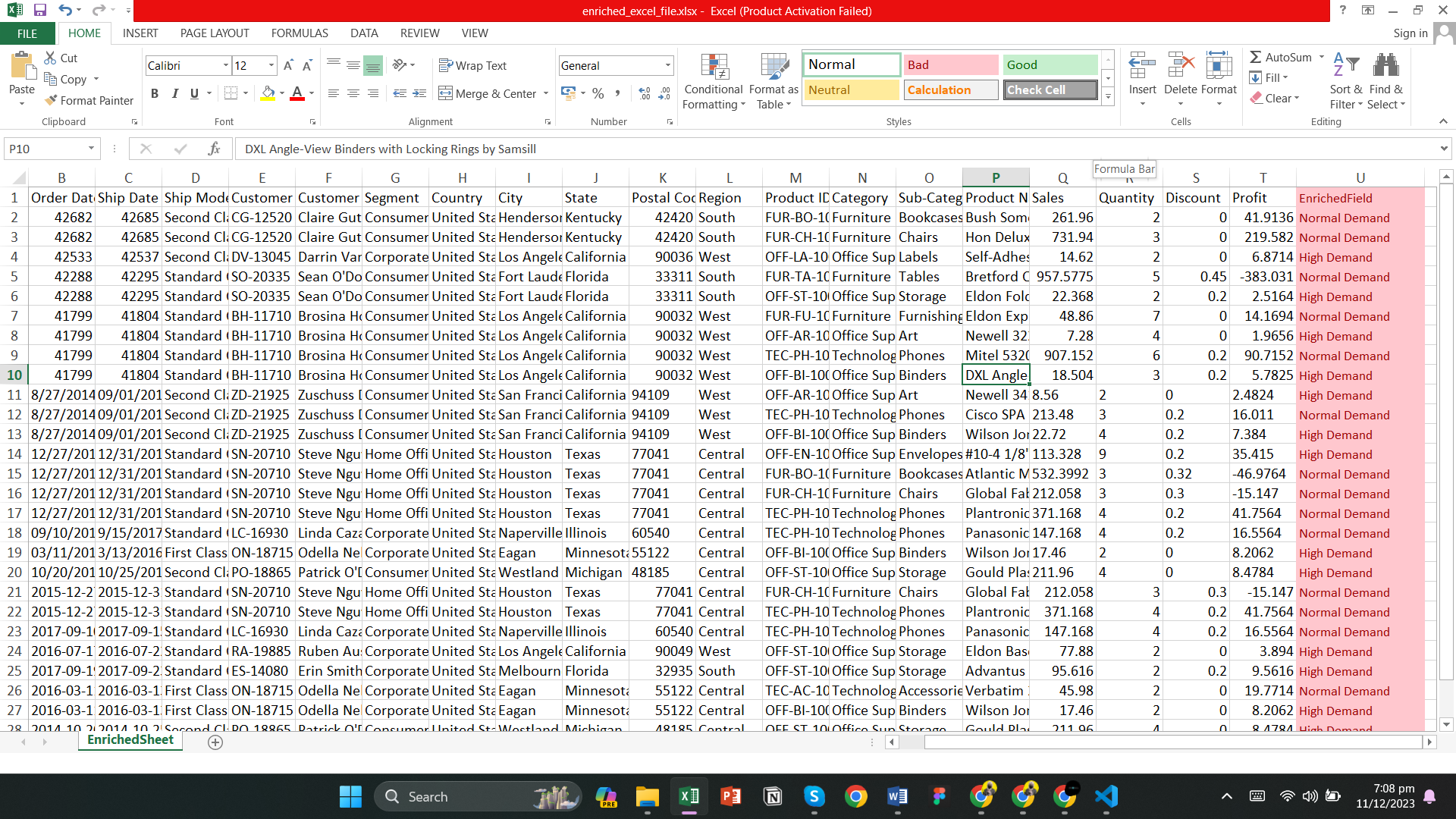
**Categorize:**

* + Sales Category



**Enrichment:**

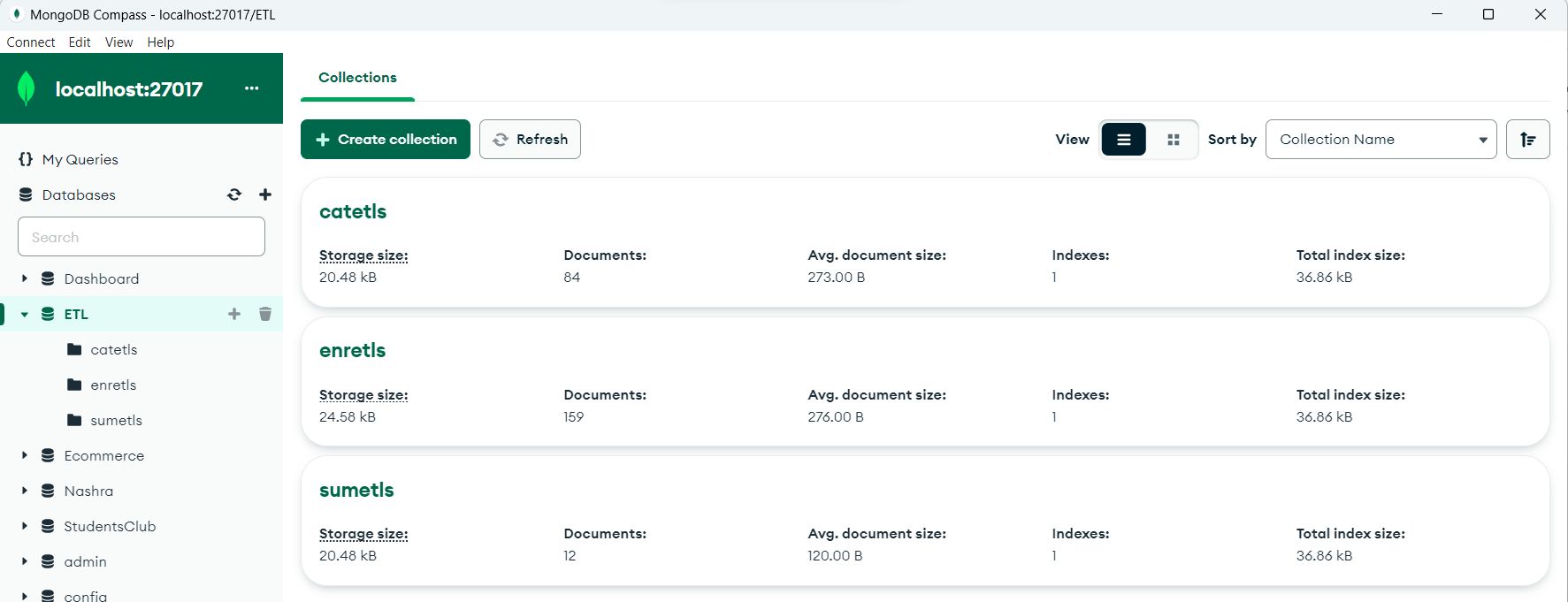
* + Enriched Field



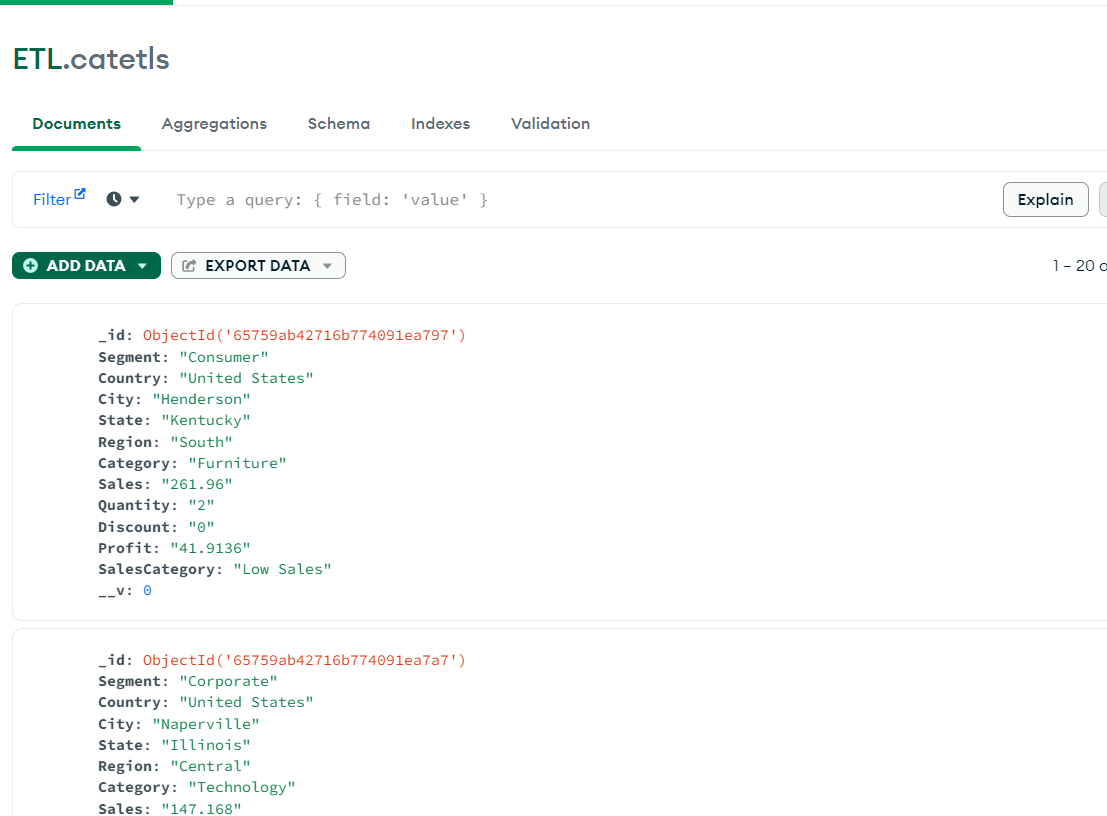
These columns enhance the dataset's analytical capabilities and provide insights into sales performance and demand patterns.

3.3 Load

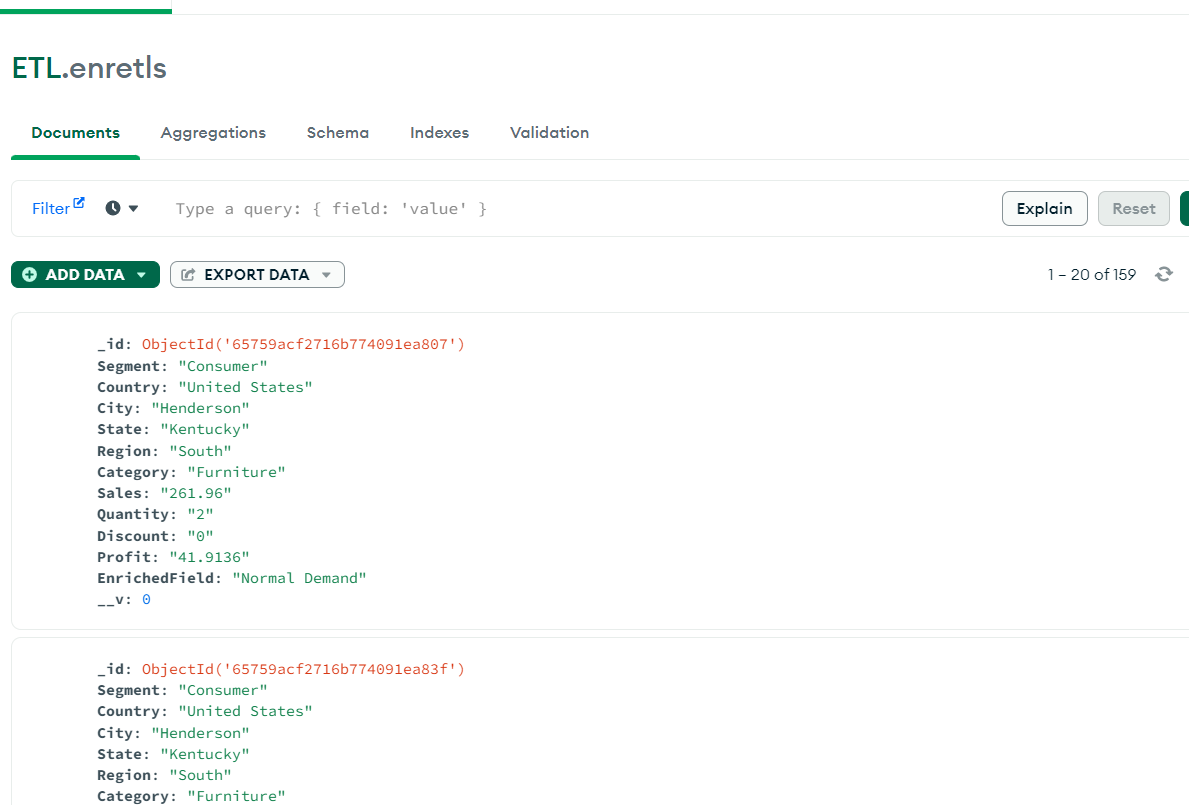
The final step involves loading the transformed data into the desired destination for analysis, reporting, or further integration with other systems. We have used Mongo DB for the storage of data.



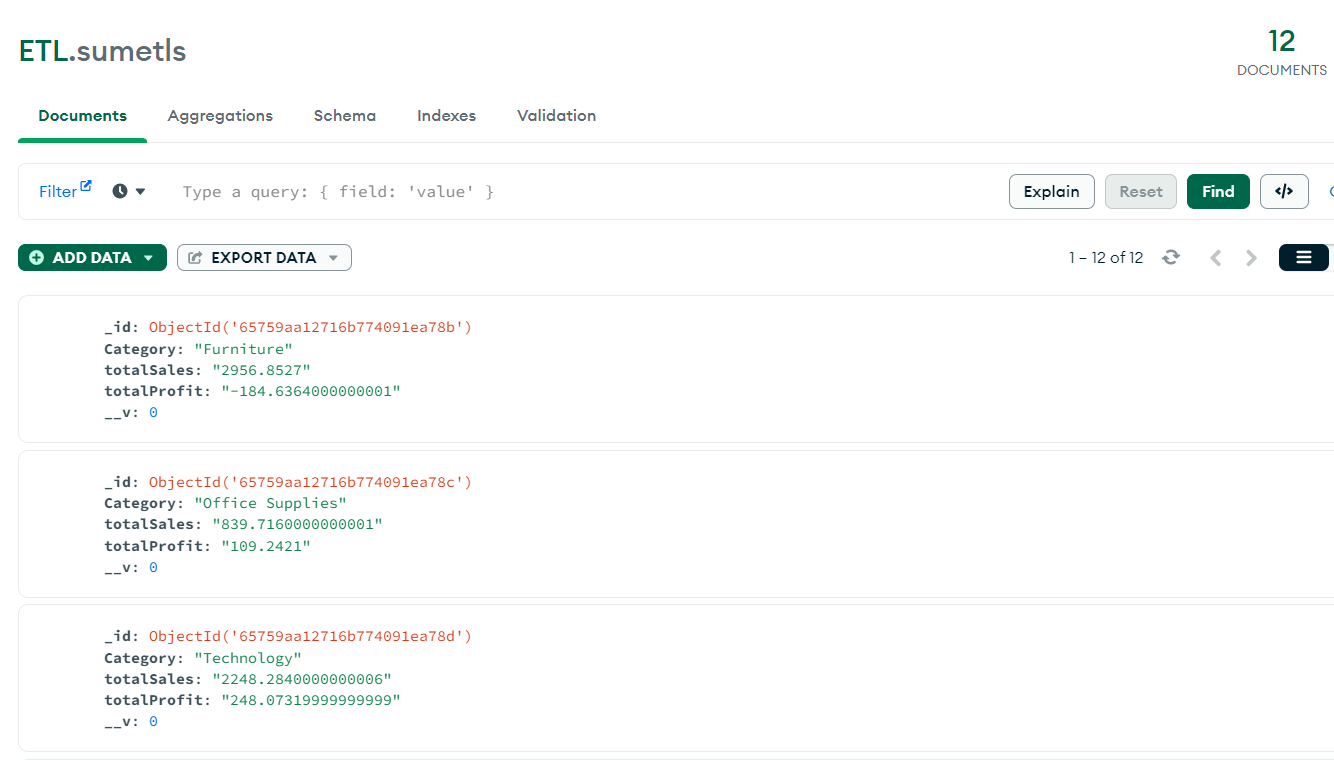
1. **Categorized**:



1. **Enrichment**:

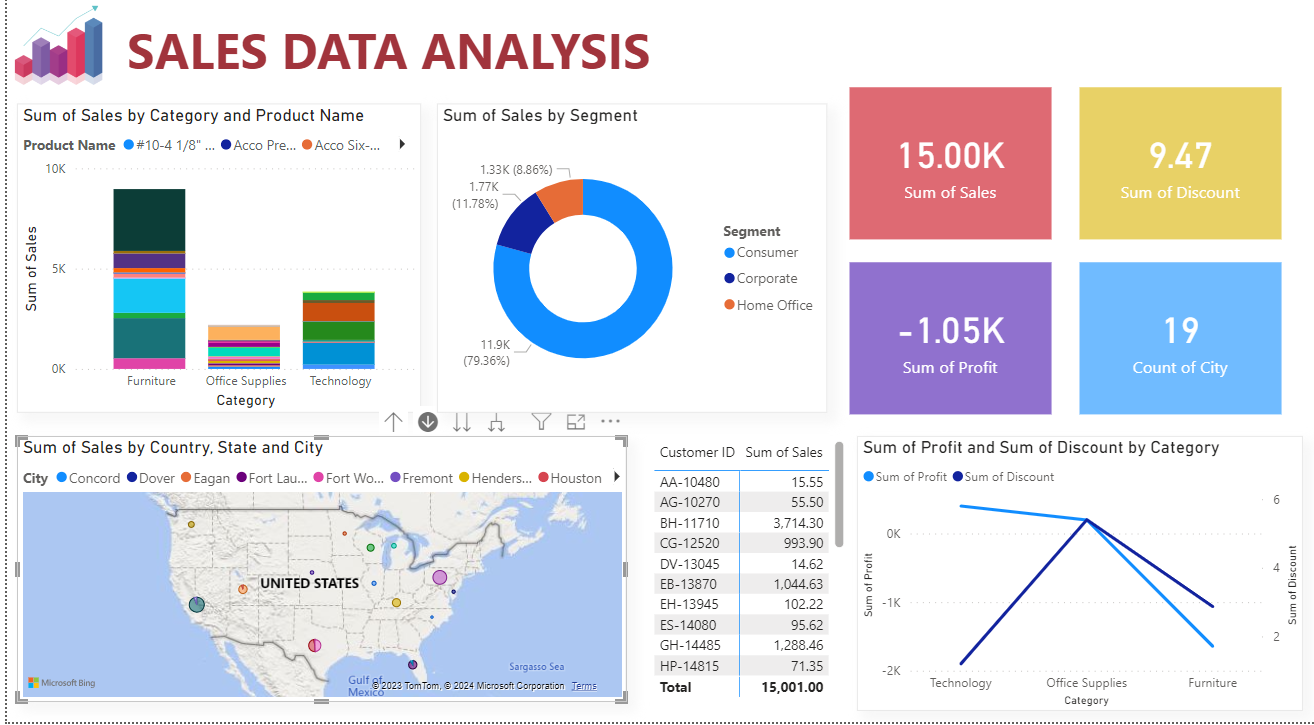


1. **Summarized**:

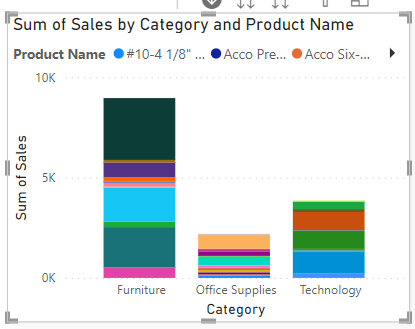


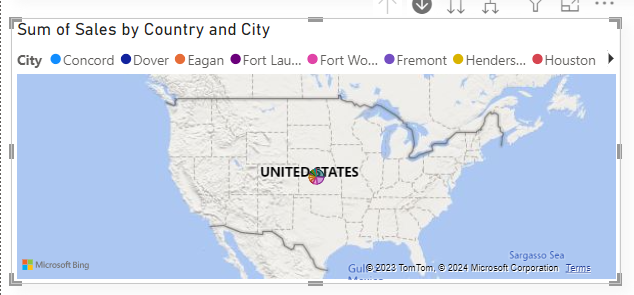
# 4. Visualization

Upon successful transformation and loading of the data into MongoDB, the next pivotal step involved importing this enriched dataset into Power BI for comprehensive visualization. Leveraging Power BI's dynamic features, we meticulously crafted a dashboard(more of a report in Power BI's Pov) that encapsulates key insights derived from the dataset. The dashboard presents an intuitive interface, offering a visually compelling representation of the transformed data. Through the strategic use of charts, graphs, and other visualization elements, our Power BI dashboard provides a user-friendly platform for stakeholders to gain actionable insights into sales patterns, sales regions, and other critical metrics. This visualization section not only enhances the interpretability of the dataset but also fosters a more informed decision-making process by presenting complex information in an accessible and digestible format.

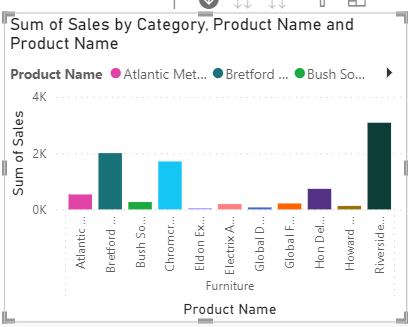


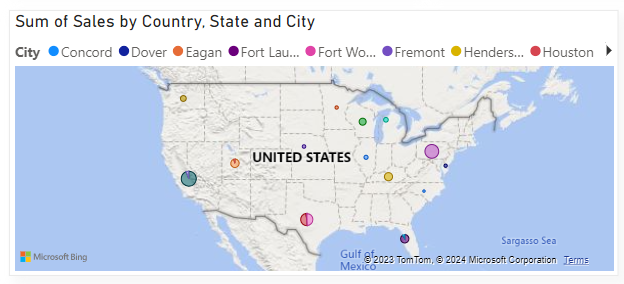
Roll Up Visualization





Drill Down Visualization





CONCLUSION

In conclusion, this documentation serves as a comprehensive guide to the project, elucidating key components such as data structure, extraction methodologies, and transformation processes. The introduction of new columns enhances the dataset's analytical depth, laying the groundwork for meaningful insights. The overarching goal of this project is to empower informed decision-making through the thorough enrichment and categorization of data, providing a solid foundation for strategic analysis and planning.