Fundamentals of Big Data Analytics and BI [2017 E.C]

COLLEGE OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

NAME	ID_NO
Biruk Damte	0344/13

SUBMITTED TO: Mr. DERBEW F.

SUBMITTED DATE: FEB 13, 2025

Contents

End-to-End Data Pipeline Project Report

Overview

This document provides a comprehensive overview of the findings, design choices, and assumptions made during the development of an end-to-end data pipeline for electronics purchase data set downloaded from Kaggle. The pipeline focused on extracting, cleaning, transforming, storing, and visualizing a large dataset to derive actionable business insights.

1. Project Objective

- **Goal:** To build a scalable and efficient data pipeline to analyze a large e-commerce dataset.
- **Dataset Source:** Kaggle dataset with over 2 million rows.
- **Scope:** The pipeline covers data extraction, transformation, storage in DuckDB, and visualization in Power BI.
- **Metrics:** Focused on sales trends, customer segmentation, and category-wise performance.

2. Dataset Details

The dataset was sourced from Kaggle and consists of transactions related to electronics purchases. It includes the following columns:

- **event_time:** Timestamp of the order.
- **order_id:** Unique identifier for each order.
- **product_id:** Unique identifier for each product.
- **category_id:** Unique identifier for product categories.
- **category_code:** Hierarchical structure of product categories.
- **brand:** Brand of the product.
- **price:** Price of the product.
- **user_id:** Unique identifier for users.
- **category_encoded:** Added column with encoded category values.

3. Design Choices

The dataset was ingested into DuckDB using the following schema:

- event_time: TIMESTAMP
- order_id: STRING
- product_id: STRING
- category_id: STRING
- category_code: STRING
- brand: STRING
- price: FLOAT
- user_id: STRING
- category_encoded: INTEGER

Data Cleaning Process

The objective of this process is to ensure data quality and usability for analysis. Steps Taken are:

1. Removed rows with missing or null values in critical fields (e.g., price, category_code).
2. Encoded category_code using one-hot encoding to create the category_encoded column.
3. Converted event_time to a standard TIMESTAMP format for consistent time-based analysis.
4. Filtered out erroneous data, such as negative prices or duplicate rows.

Transformation Choices

I Used **PySpark** for efficient handling of the large dataset to:

- Aggregated data for category-wise performance.
- Calculated metrics such as average price, total sales, and user counts.
- Derived day-of-week insights from event_time.

Transformed cleaned data into CSV and stored in DuckDB for optimized querying.

4. Storage and Visualization

For storing cleaned data, I used duck dB for a reason

- Lightweight, high-performance analytical database.
- Easy integration with Python and Power BI.

For Visualization I used Microsoft Power BI for advance analysis and visualization

- **Dashboards Created:**
    - Sales trends across days of the week.
    - Category-wise average prices.
    - User segmentation by brand and category.
- **Features:**
    - Interactive charts with filters for brands, categories, and time periods.
    - Interactive dashboard with day of the week filter and additional tab.

5. Analysis Assumptions

- Dataset reflects actual e-commerce transactions without significant outliers.
- All time-based insights are derived from the cleaned event_time column.
- Price values represent actual transaction amounts and are consistent across the dataset.
- Categories like "others" group less-defined or miscellaneous items.

6. Key Findings

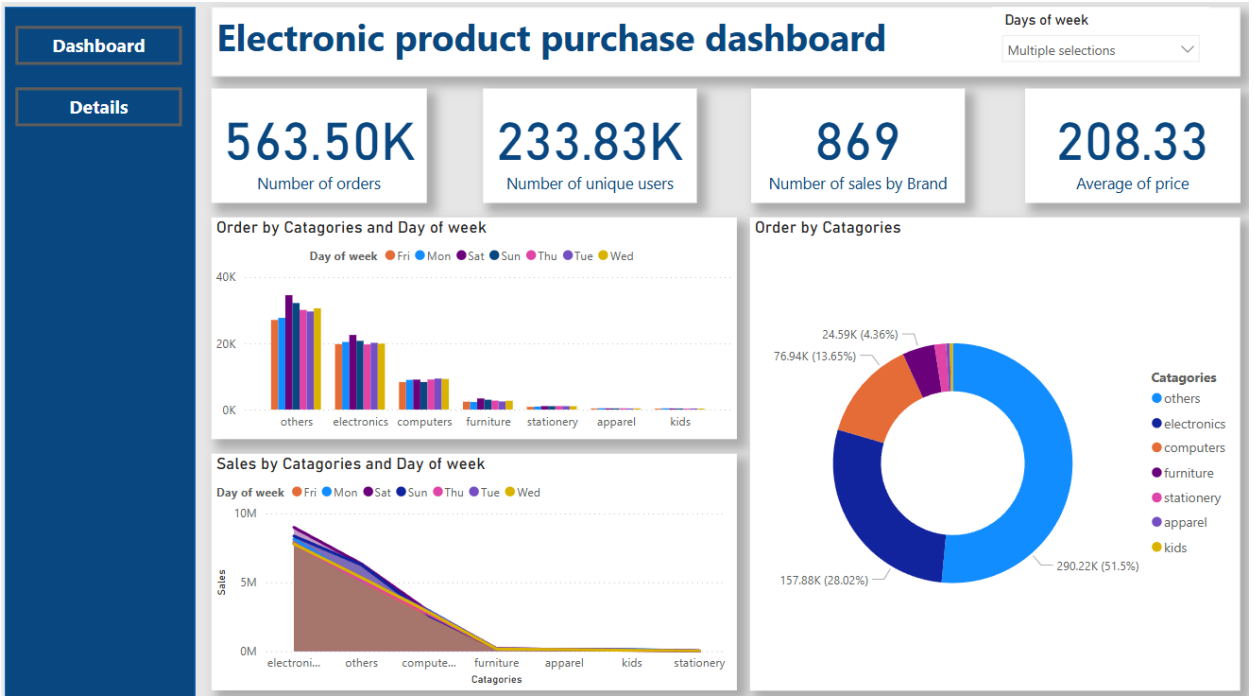
General Insights

- The dataset reveals trends in sales, user behavior, and category preferences.
- Brands like Samsung and Ava dominate in user engagement and sales metrics.
- The day of the week significantly impacts total sales volume.

➢ TOGET DETAIL VIEW CHECKOUT THE FILE “step_4_power_bi_analysis.pbix” IN MY GITHUB REPO.

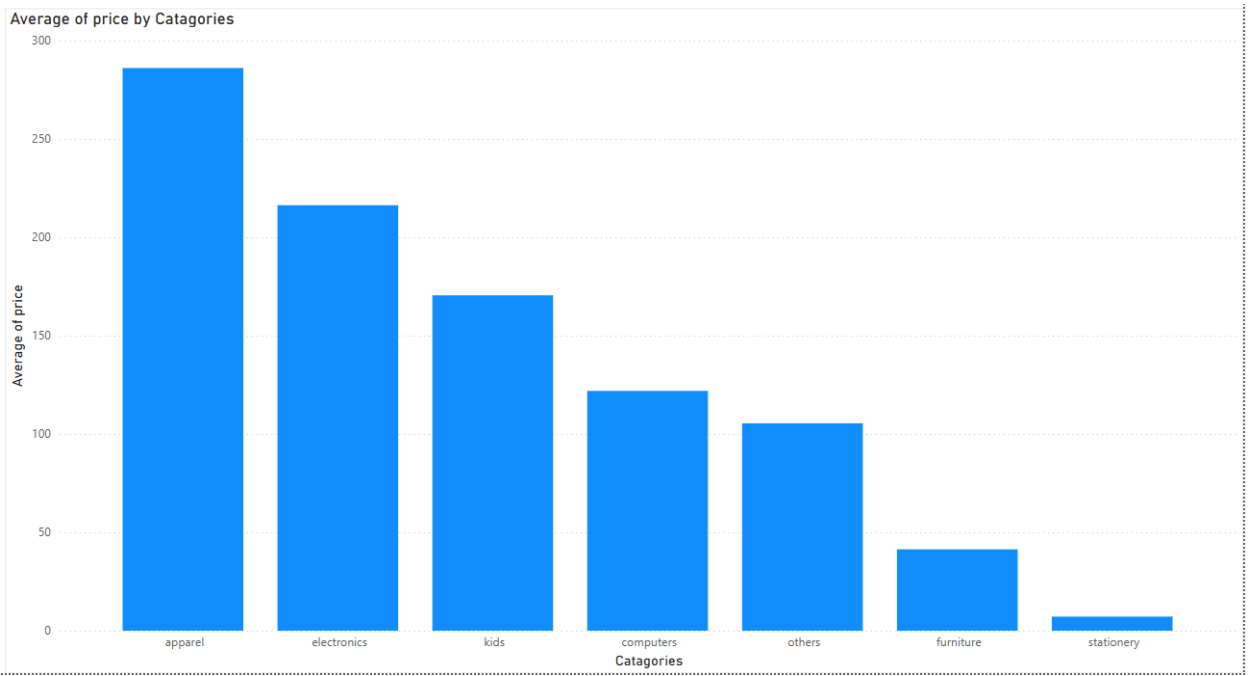
Results:

1. **Day of the Week Analysis:**



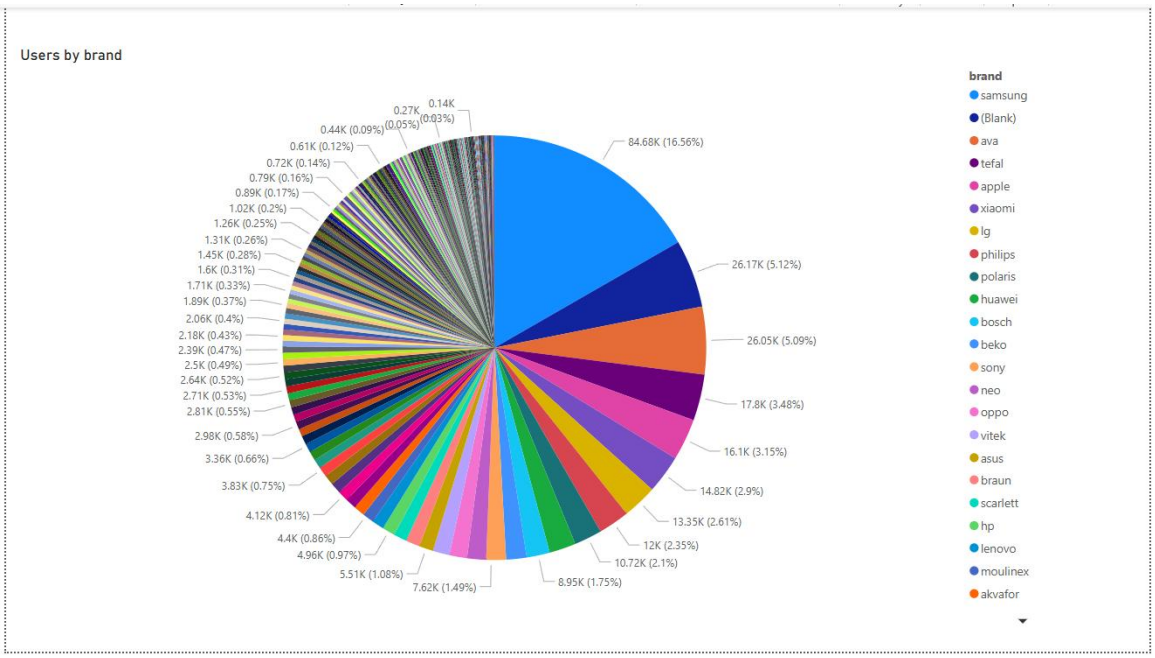
- "Others" in the category had 7.72% of orders on Saturday.
- Total sales were higher on Sunday ($17,585,571.36) compared to Monday ($16,603,691.42).

2. **Category Performance:**



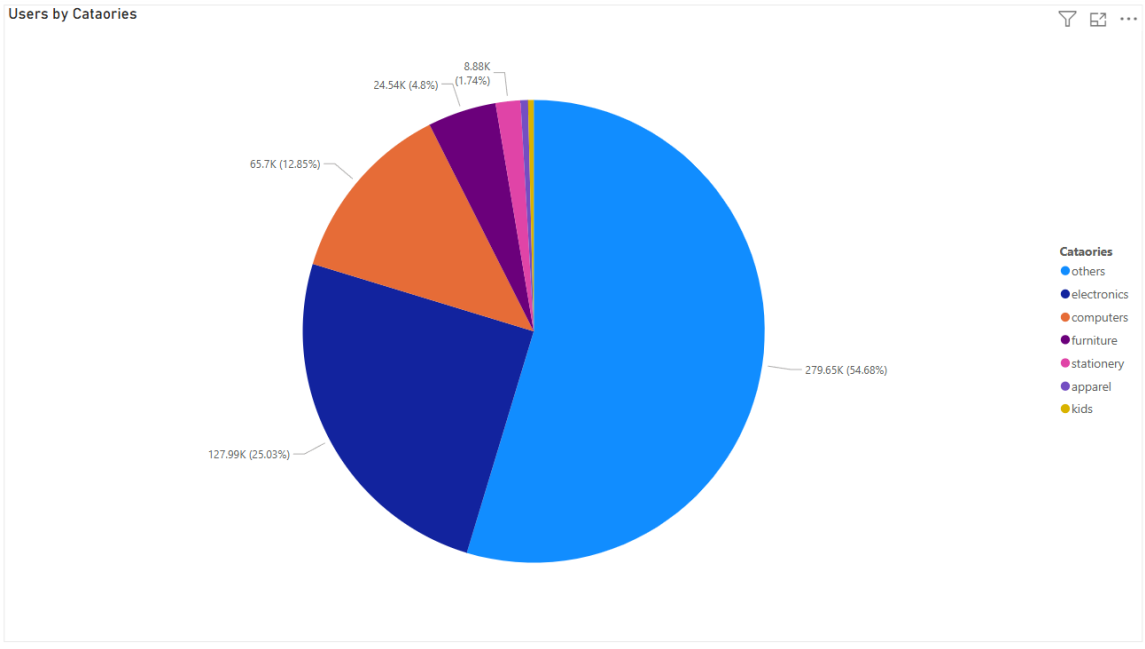
- Apparel had the highest average price at $286.03, which was 3,894.55% higher than stationery ($7.16).
- Electronics and furniture were the most popular product categories.
- Across all categories, the average price ranged from $7.16 to $286.03.

3. **Brand Engagement:**



- Samsung had the highest user count at 84,682, accounting for 16.56% of all users.
- Ava followed closely, contributing significantly to user engagement.

4. **User Distribution:**



- The "others" category accounted for 54.68% of all users, highlighting its broad appeal.

7. Challenges and Resolutions

Challenges

- Configuring PySpark with Hadoop for data transformation was challenging.
- Optimizing the visualization of large datasets in Power BI.

Resolutions

- Resolved PySpark configuration by downloading version 3.2.2 and setting up environment variables.
- Used Power BI's data reduction strategies to optimize dashboard performance.

8. Conclusion

This project demonstrated the successful implementation of an end-to-end data pipeline using modern tools and technologies. The findings provide actionable insights for e-commerce businesses, highlighting key trends in sales, pricing, and user engagement. The combination of DuckDB for efficient storage and Power BI for visualization proved highly effective in handling large datasets and delivering interactive dashboards. These insights can help businesses improve their marketing strategies and product offerings.