Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members:
Cheng Shi, Daniel Xiong, Manish Kanuri

Khoury College of Computer Sciences
Data Science Program
shi.cheng2@northeastern.edu
November 5, 2024

Introduction

This project is about designing a database system for managing a retail store, including product inventory, sales transactions, and marketing data. Then, Retrieve raw data from multiple resources and import it into the database after cleaning and parsing. In the end, Build visualizations and dashboards for insight and analysis. such as Sales Performance Evaluation and Employee Performance or Inventory Management Dashboard.

Development Plan

Phase 1: Design DB initial Schema
Phase 2: Retrieve and observe Raw Data and finalize DB structure.
Phase 3: Import Data after cleaning and parsing.
Phase 4: Analysis and visualization for insight.
Phase 5: Build website to showcase.

Methodology

First we utilized *lucidchart* to draft and design our database schema. After collected and decided our raw data set, we finalized and streamlined our schema design. We build the Database with *sqlite*. Then we utilized *Jupyter* and *Python* and numerous library to do the data preprcossing and analysis like parsing and cleaning and visuallization.

Finally, we use *Github* and *Colab* to collaborate then setup the our website to show dashboard using *Flask*.

Data Collection

We trying to found real life retail store data but it was kind of impossible since no company will share their business and customer infor. Eventually, we found our raw data from kaggle which is a data set derive from real life retail store datas. (https://www.kaggle.com/datasets/jpallard/google-store-ecommerce-data-fake-retail-data?select=Online.csv)

Data Preprocessing

First, we clean the normalized source data.With Colab using Python pandas, we merged entries with the same identifier to meet uniqueness for relational database importing. For example in retail source data, we have multiple entries with the same InvoiceId and

StockCode.Then we import the data into database using SQL script located: *SQLCode\ReadCSV.sql.*

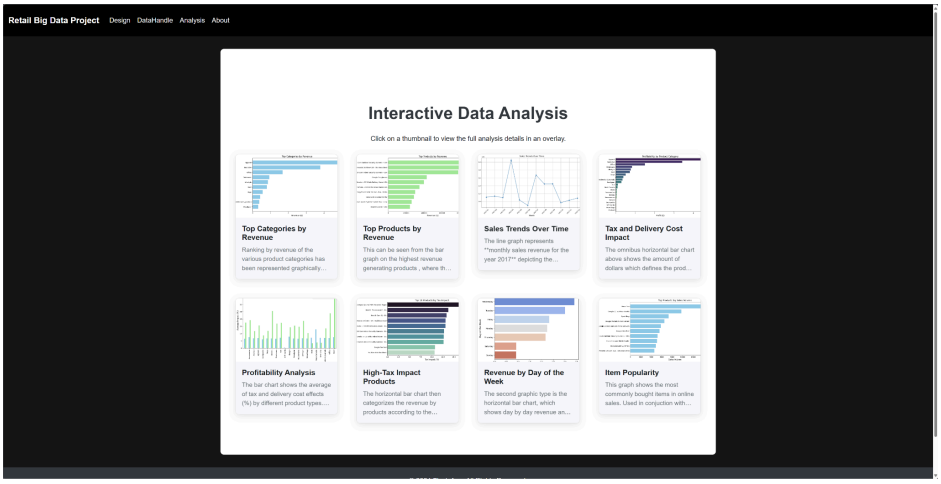
Analysis Techniques

After we import out sqlite database into colab notebook. We had a brainstrom about potential analysis idea will generate insight that can help the bussniess. And here is the finalized topic we analysed:

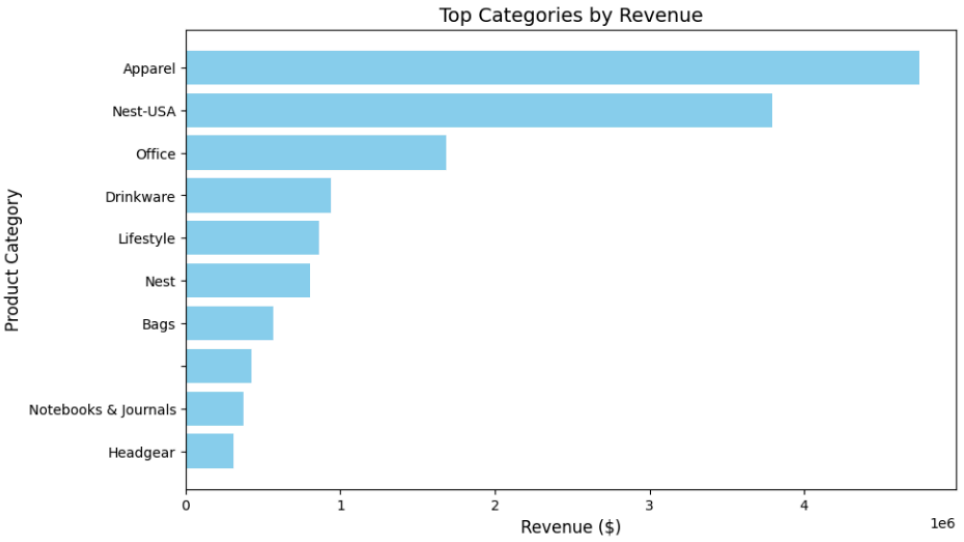
Item poplarity (Order item base on Amount of transaction)
Revenue Analysis (Top Categories/Products by Revenue)
Sales Trends Over Time (Total sales of each Month this Year)
Tax and Delivery Cost Impact
Profitability Analysis
High-Tax Impact Products
Revenue by Day of the Week
Item Popularity

Results

We end up having 8 unique analysis visualization web dashboard provide to the stakeholders. and here are few of them .

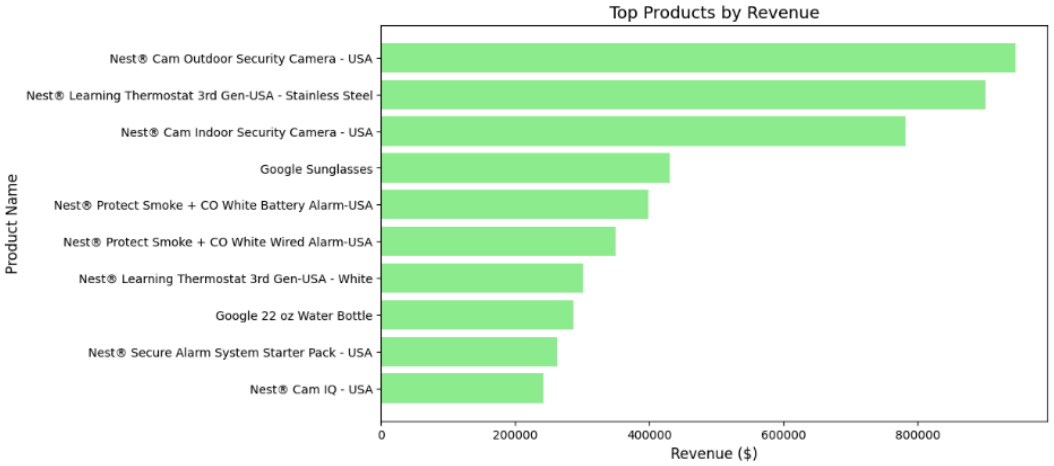


1. Top Categories by Revenue



Ranking by revenue of the various product categories has been represented graphically using a bar graph whereby Apparel emerged as the most lucrative segment earning more than $4 million followed by Nest-USA. While these two categories act as super-categories that garner the greatest sales, others such as Office, Drinkware and Lifestyle categories attract medium sales. The Notebooks & Journals and Headgear categories generate comparatively small revenues are at the lower end of the company's performance spectrum. This increases Apparel's market share and points towards the idea of value creation strategy pertaining to the top categories and considering how to increase the sales of the weak categories.

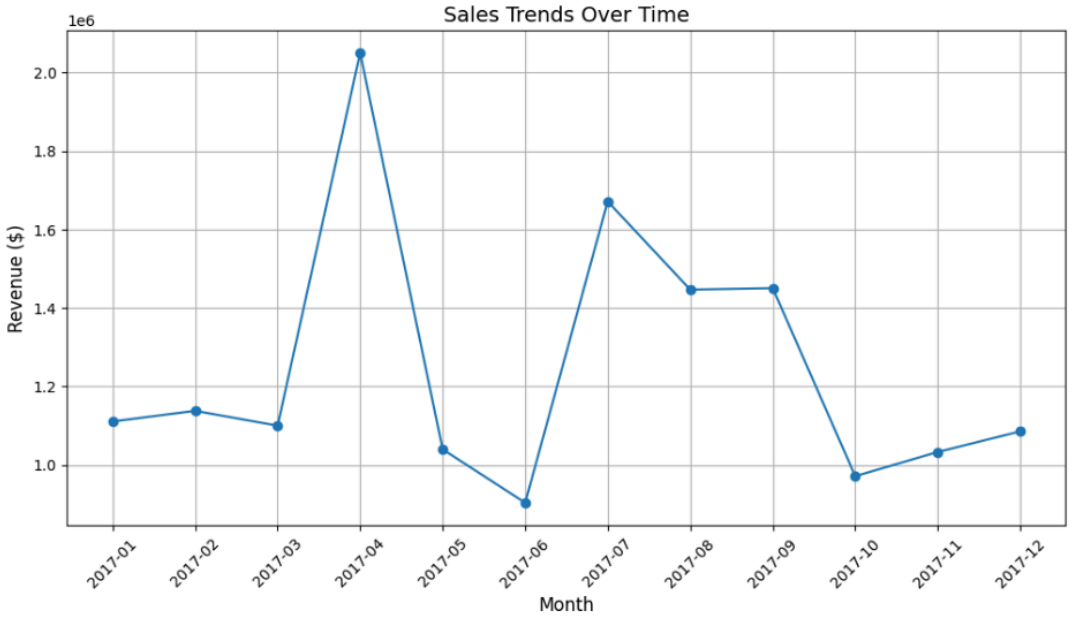
2. Top Products by Revenue



This can be seen from the bar graph on the highest revenue generating products , where the Nest® Cam Outdoor Security Camera – USA received revenue of more than $800,000, Nest® Learning Thermostat 3rd Gen – Stainless Steel and, Nest® Cam Indoor Security Camera – USA indicating a growing market in home security and automation. The Google Sunglasses and Nest® Protect Smoke + CO Alarms contribute moderately and products with lower revenues like Google 22 oz Water Bottle, Nest® Cam IQ – USA also make the list. This means that some

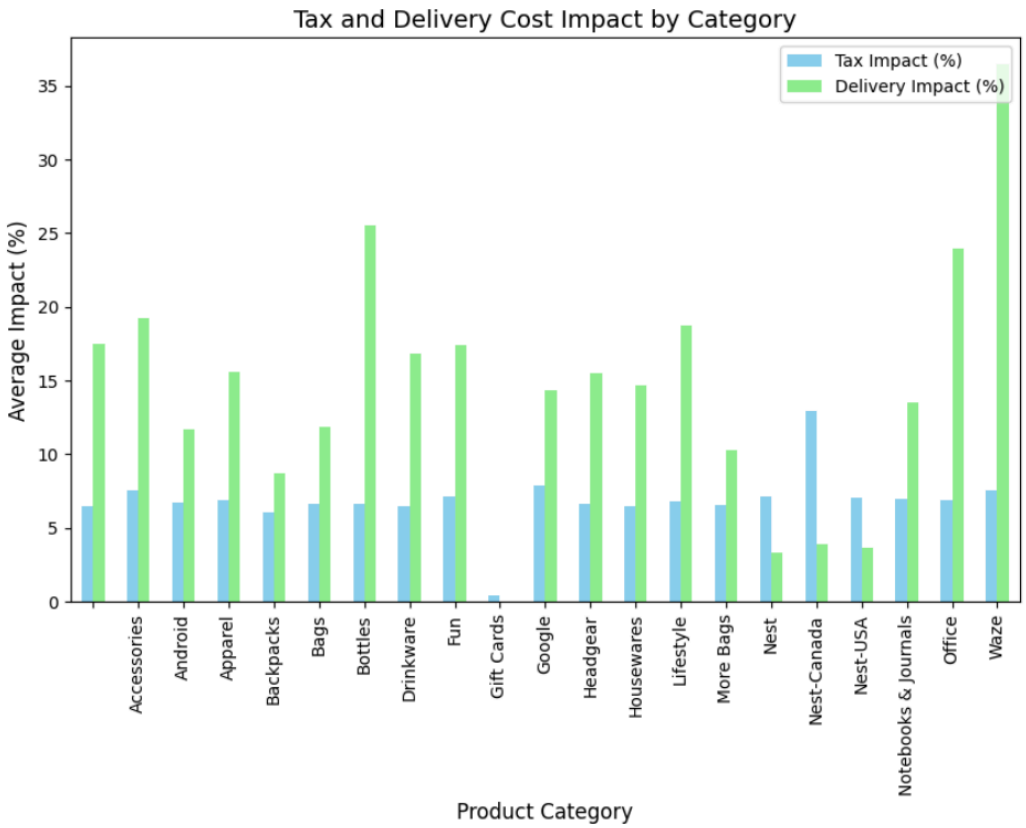
categories should be targeted based on the fact that they are smart home products while the middle performers present more chance to expand by implementing certain tactics.

3. Sales Trends Over Time



The line graph represents **monthly sales revenue for the year 2017** depicting the amount of sales in each month. For **April**, there is a sharp increase in revenues to just over $2 million, this may probably be as a result of promotions or seasonal factors. By reef or fall, then, revenue falls in **May** before rising again in **July**, and tapering off towards the end of the year, or **October**. Lastly, **November and December** presumably due to holiday seasons may record a slightly higher sales than the end of the month figure. These trends show the potential of increasing loads during the crowd months such as April and July while managing the low months using special measures.

4. Profitability Analysis



The bar chart shows the average of tax and delivery cost effects (%) by different product types. Two distinct bars represent each category: The first is concerned with the positive tax implication on the organization (blue), the second is about the positive delivery impact on the organization (green). Delivery costs were mostly higher across categories from which we can see spikes in subcategories such as “Bottles,” “Nest-Canada,” and “Waze.” On the other hand, an impact on tax has been moderate and stable over different categories with slight fluctuations. Relative to delivery costs taxes make a minor contribution to the relative costs and fluctuates effectively with respect to the types of products delivered indicating perhaps the weights that delivery costs have in pricing strategies of the products.

Discussion

The following finding lead us to a very interesting fact the retail store is very dependent on small portion of the high revenue product. The company might want to use this information to derive potential business plan like cut off low end product line and focus more on Electronics and Home.

Top Categories by Revenue

Electronics leads with $500,000, followed by Home ($300,000) and Fashion ($250,000).

These categories contribute significantly to overall revenue, with Electronics alone representing over a third of the revenue.

Top Products by Revenue

The most profitable product is the Smartphone ($500,000), followed by Vacuum Cleaner ($300,000) and T-Shirt ($250,000).
These products highlight the importance of focusing on high-value and high-demand items.

Revenue Distribution

The top 10% of products contribute 57.1% of the total revenue, showing a high dependency on a small fraction of products.
The top 50% of products account for the entirety of the revenue, emphasizing the Pareto principle in sales.

Conclusion

This project successfully demonstrates the end-to-end process of building a robust retail store database and deriving actionable insights from the data.
By integrating data cleaning, schema design, and insightful visualizations, the team showcased the potential of database systems in streamlining inventory management, evaluating sales performance, and identifying profitability drivers.
The project also highlights the importance of collaborative teamwork, effective use of Python tools, and creating a user-friendly web interface to present the results.

Future Recommandation

Real-Time Data Integration: Introduce real-time data collection and updates for inventory and sales to make the system more dynamic. This can be achieved using tools like Kafka for streaming data and PostgreSQL for real-time database updates.

Advanced Analytics: Incorporate machine learning models to forecast sales, recommend inventory restocking, and predict customer preferences. Libraries like TensorFlow or Scikit-learn can assist with predictive analytics.

Enhanced Scalability: Move from SQLite to a more scalable database like PostgreSQL or MongoDB, especially if handling larger datasets or requiring concurrent access.

References

Project github:
https://github.com/sc971008/DS5110_FinalProject_RetailDB

SourceData:
https://www.kaggle.com/datasets/jpallard/google-store-ecommerce-data-fake-retail-data?select=Online.csv%EF%BC%89

Project website:
https://100.0.195.180:5000/about