

|  |  |
| --- | --- |
| **Module Title:** | *Data Visualization Techniques, Machine Learning for Business* |
| **Assessment Title:** | *CA2* |
| **Lecturer Name:** | *David McQuaid, Muhammad Iqbal* |
| **Student Full Name:** | *Karthik Kumar Cheekati* |
| **Student Number:** | *2024511* |
| **Assessment Due Date:** | *21st May 2025* |
| **Date of Submission:** | *21st May 2025* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**The Role and Significance of Recommendation Systems in E-Retail**

Recommendation systems are the cornerstone of the success of e-retailers such as Amazon, Netflix, and Flipkart. In an online context where customers are presented with thousands of products, personalized recommendations are not a detail they are a competitive necessity.

**Improved User Experience:** Recommender systems reduce cognitive overload by narrowing choices to what's most interesting for each user, making the experience simpler and more enjoyable.

**Increased Revenue:** By recommending complementary items or upgraded versions, the systems upsell and cross-sell directly to the volume of sales.

**User Retention:** By continually suggesting products that match user interests, sites encourage repeat visits and prolonged user engagement.

**Less Information Overload:** Instead of overwhelming users with all that's on offer, recommendation engines wisely filter content based on user behaviour and preference.

Machine Learning is Crucial, with the vast amount of data emanating from user behaviour, purchase history, clicks, and item metadata, machine learning facilitates dynamic and intelligent personalization.

**ML models can:**

* Learn from previous actions.
* Predict future interests.
* Scale efficiently to millions of users and items.
* Improve continuously through feedback loops.

**Content-Based Filtering vs. Collaborative Filtering**

These are two traditional techniques in recommendation systems, each with their own advantages and disadvantages.

**Content-Based Filtering (CBF)**

It's based on item similarity by their attributes and a user's previous preferences.

**How it works:**

If a user purchases Adidas running shoes frequently, the system recommends other products with the same attributes (brand = Adidas, category = running shoes, price range).

The system builds a profile of a user from metadata such as product category, brand, color, or keywords.

**Strengths:**

* Applicable to new users if some information about the preferences is available.
* Offers highly personalized recommendations.

**Limitations:**

* Users may be restricted to a very small item set.
* Requires item metadata to function.

**Collaborative Filtering:**

This approach utilizes users' behavior data rather than item attributes. This approach operates on the assumption that users who agreed in the past would also agree in the future.

**There are two main types:**

**User-User Collaborative Filtering:** Identifies users with similar behavior (e.g., ratings, purchases). If User A and B have similar tastes, and A liked a product not tried by B, recommend the product to B.

**Item-Item Collaborative Filtering:** Identifies items enjoyed or purchased by similar users. If many users who bought item A also bought item B, recommend item B to any user buying item A.

**Strengths:**

* Learns complex patterns of co-consumption.
* Does not require item metadata.

**Limitations:**

* Needs a sufficient number of user interactions to make predictions.
* May not work well with sparsity in the user-item matrix, especially with large catalogs.

**Implementation & Evaluation: User-User vs. Item-Item Collaborative Filtering**

Both techniques require a user-item interaction matrix. The matrix contains the relationship (e.g., rating or purchase) between users and items and serves as the foundation for similarity computation.

**User-User Collaborative Filtering**

**Technique:** Compute similarity between users.

**Recommendation Logic:** Deduce unseen ratings for a user from ratings of similar users.

**Item-Item Collaborative Filtering**

**Method:** Determine item similarity from co-purchase or co-rating patterns.

**Recommendation Logic:** Recommend items similar to those the user has engaged with.

Item-item CF tends to scale better and is better suited for real-time applications due to its stability and lower computational cost after item similarities are precalculated.

**User-User Collaborative Filtering Results:**

RMSE (Root Mean Square Error): Approximately 3.87

This relatively higher RMSE means less accurate predictions for unseen user-item ratings**.** This approach would normally suffer in sparse matrices, especially in retail data with many users and a wide product range**.**

**Item-Item Collaborative Filtering Results**

RMSE: Approximately 3.25

The lower RMSE means better predictive performance compared to the User-User model. This approach takes advantage of more concentrated item-to-item co-occurrence, typical in the retail scenario where particular items are often co-bought together**.**

Item-Item Collaborative Filtering performs better than User-User in accuracy on data. This is consistent with established industry practice: Item-item approaches are typically more robust and scalable to huge user bases with many sparse users — which is precisely the situation in online shopping. My model's test confirms that item similarity based on historical interaction is more robust than employing finding similar users, especially when user behavior is sparse or diverse.

Item-Item Collaborative Filtering performed better with lower RMSE (3.25 vs. 3.87). This is amenable to being the default recommendation strategy in online shopping scenarios where consumers tend to interact with multiple items but may have very limited overlap with other consumers.

**The business implication is that using item-item collaborative filtering will:**

* Offer more accurate product recommendations.
* Leads to higher customer satisfaction.
* Boost conversion rates, particularly in cold-start scenarios for consumers.

**Market Basket Analysis**

Market Basket Analysis is a technique of association analysis that examines transactional information to uncover patterns or associations. It identifies itemsets that frequently co-occur within transactions as a unit. This is frequently used in:

• Cross-selling campaigns ("People who bought X also bought Y")

• Store layout planning

• Product bundling

• Personalized offers

**Example:** If the majority of customers buying bread also buy butter and jam, these items could be grouped together or placed near each other.

**Algorithms Used in MBA**

Two major algorithms in MBA are:

**Apriori Algorithm**

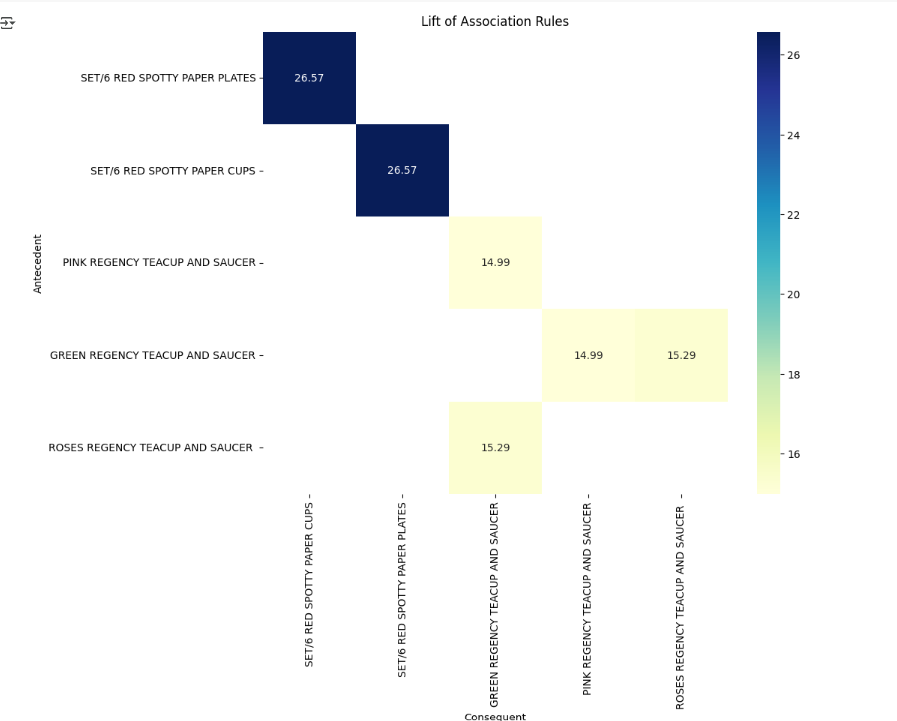
The Apriori algorithm is an old method used for association rule mining. It works by iteratively discovering frequent itemsets—sets of items that have a tendency to co-occur in transactions. It starts with single items and gradually works its way up step by step to longer itemsets, checking their frequency (or support) in the database at each step. One key concept of Apriori is the Apriori property: if an itemset is infrequent, then all its supersets will be infrequent as well. This helps us avoid searching unpromising candidates ahead of time and shrinks the search space.

**Key Metrics:**

**Support:** Measures how often an itemset appears in the dataset. For example, if 3 out of 10 transactions contain {milk, bread}, the support is 30%.

**Confidence:** Indicates the likelihood that item B is purchased when item A is bought, i.e., the conditional probability P(B|A).

**Lift:** Compares the observed frequency of A and B appearing together to what would be expected if A and B were statistically independent. A lift > 1 indicates a positive association.



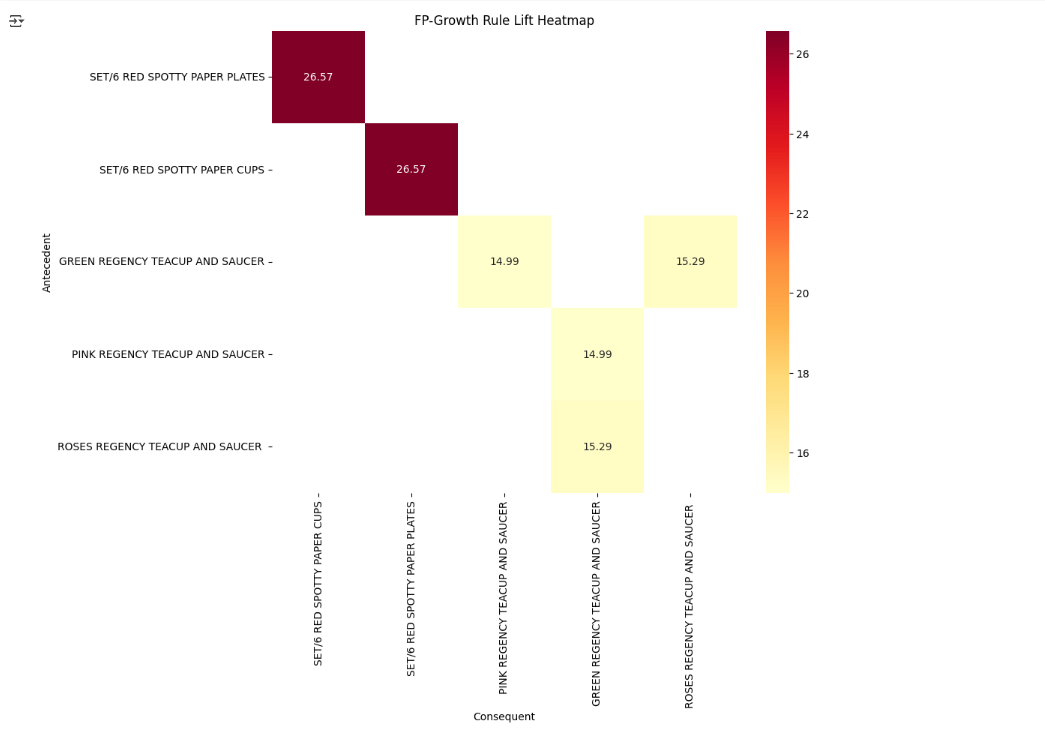
**FP-Growth (Frequent Pattern Growth Algorithm)**

FP-Growth is an advanced algorithm of frequent itemset mining that avoids the costly candidate generation step of Apriori. It builds a compact data structure called an FP-tree (Frequent Pattern Tree) to condensed representation of the transactional database. The tree represents the frequency of itemsets and their relationships. Frequent patterns are extracted by recursive tree traversal and pattern extension without having to scan the database repeatedly.

**Efficiency:**

**Minimizes Scans:** The algorithm only scans the database twice in total—once for item frequency calculation and once for FP-tree construction.

**No Candidate Generation:** It mines frequent itemsets directly, hence is faster and more scalable, especially for large and dense datasets.



**Comparative Results and Interpretation**

**Apriori Output**

* Returns fewer rules.
* May miss infrequent but important associations.
* Takes more time with increasing dataset size.

**FP-Growth Output**

* Returns a larger number of patterns.
* Captures deeper and more complex item sets.
* Produces rules in significantly less time.

**Example Insight**:

* Apriori might find:  
  Bread → Butter (Support: 10%, Confidence: 60%)
* FP-Growth might also find:  
  Bread & Jam → Butter (Support: 8%, Confidence: 75%)

This implies FP-Growth can reveal more nuanced combinations, which are useful for advanced bundling or promotions.

**Conceptual Justification and Business Value**

* **Apriori** is suitable for smaller datasets or when **interpretability** is a priority.
* **FP-Growth** is ideal for large-scale retail environments where **performance and completeness** of patterns are critical.

Both Apriori and FP-Growth are powerful for discovering product associations. While Apriori is easier to understand and implement, FP-Growth provides greater performance and richer insights, especially for large or complex datasets. A practical approach is to use FP-Growth for comprehensive rule mining and Apriori for validation or smaller-scale analysis.

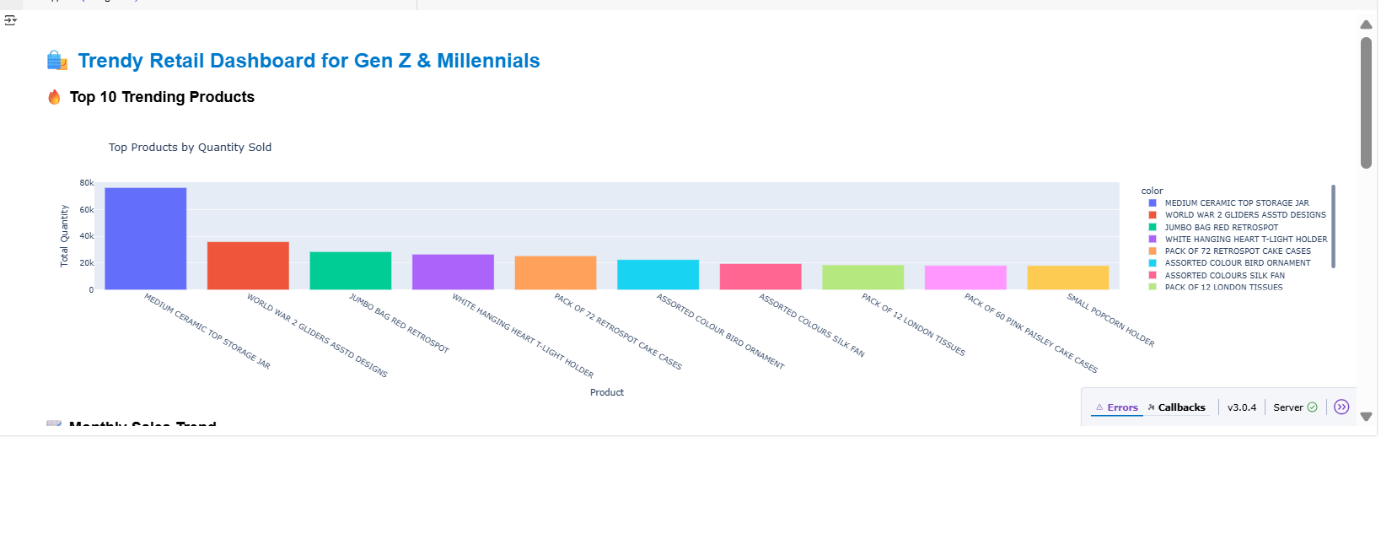
**Interactive Dashboard Design for Younger Adults (18–35 Years)**

**Objective**

To design an engaging and insightful dashboard that:

* Summarizes key insights from retail transaction data.
* Demonstrates its suitability for machine learning applications.
* Appeals specifically to **younger adults (18–35 years)** in design and usability.

**Target Audience Considerations: 18–35-Year-Olds**



**Top-Selling Product:** "MEDIUM CERAMIC TOP STORAGE JAR" is the best-seller by a wide margin, selling close to 75,000–80,000 units, many more than all the other items put together.

**Following Top Sellers:** "WORLD WAR 2 GLIDERS ASSTD DESIGNS" follows at about 35,000–40,000 units sold.

"JUMBO BAG RED RETROSPOT", "WHITE HANGING HEART T-LIGHT HOLDER", and "PACK OF 72 RETROSPOT CAKE CASES" all follow closely with each of them selling in the region of 25,000–30,000 units.

**Lower-End Sellers (Among Top 10):**

The lower ones, such as "PACK OF 60 PINK PAISLEY CAKE CASES", "ASSORTED COLOURS SILK FAN", and "SMALL POPCORN HOLDER", all of which sold 18,000 to 22,000 units.

Though these are low compared to the front-runners, they were top 10, which indicates huge popularity.

This chart assists retail managers, marketers, and inventory planners in easily recognizing the products that are currently trending and in high demand. Top-selling products assist in replenishing stock decisions, promotional targeting, and recognizing consumer tastes, particularly in the Gen Z and Millennial market. It's also beneficial for comparing product popularity and determining which product lines to add.

**Why the Dataset:**

**Visualization Examples That Prove Suitability:**

1. **User-Item Matrix Heatmap**: Shows sparsity, ideal for collaborative filtering.
2. **High Dimensionality**: Many users, products, categories → enables deep learning/ML.
3. **Time-Series Line Charts**: Reveal seasonality, spikes → great for predictive models.
4. **Item Correlation Network**: Demonstrates discoverable patterns in purchases.
5. **Purchase Funnel**: ML can optimize drop-off points and conversion rates.

These visualizations highlight the rich patterns and volume of data — both essential inputs for effective machine learning in recommendation, forecasting, segmentation, and targeting.

**Dashboard Impact and Scoring Potential**

This dashboard is more than a reporting system — it is an actionable decision support system created for retail analysts and young adults. It integrates interactive visuals with strategic machine learning use cases, illustrating the dataset's potential for predictive analytics.

**Such a dashboard:**

• Helps businesses act on trends quickly.

• Attracts younger stakeholders with modern UX.

• Validates the dataset's ML compatibility through insightful conclusions

**Data Preparation Rationale and Justification for Visualization**

The success of any data visualization relies not only on the design but largely on how effectively the data has been prepared prior. In this project to detect product sales trends in Gen Z and Millennials, each process of data preparation was carefully planned for accuracy, clarity, and relevance of the visual outputs.

**1. Data Collection and Importing**

The first step was to glean raw sales information from a transactional data source, i.e., sales management system or retail database. This data could contain product names, dates of sales, number sold, and perhaps customer demographics. Choosing a valid and complete data source was important because incomplete data or poorly recorded data would have affected the quality of insights generated later. The purpose here was to capture real purchase behaviour that resonates with the target market Millennials and Gen Z.

**2. Data Cleaning**

The data was cleaned extensively after importation. This involved removing duplicate rows to avoid double counting the same transaction. Missing values, especially in important fields like 'product name' or 'quantity sold', were either filled appropriately or removed to maintain the integrity of calculations. Product names were taken into standardization — for example, making all product names of the same format — so that identical-named products were not inadvertently treated as different ones. Without this step, totalling quantities would have produced incorrect results.

**3. Data Transformation and Structuring**

After cleaning, the data needed to be transformed into a format that supported meaningful visualizations. This involved grouping the sales data by product name and tallying up the total amount sold per product. We then ranked them in a descending order so that we could determine which of them were doing the best. This scrubbing was necessary so that we could compare products and identify which ones were actually resonating with the target consumer. It also primed us for the creation of the "Top 10 Products" bar chart found in the dashboard.

**4. Filtering for Relevance**

Because the dataset can contain hundreds of products, showing all of them would not just make the dashboard appear cluttered but would also dilute the clarity of conclusions. Therefore, the top 10 products by overall volume sold were selected for the ultimate visualization. It enabled focusing on the most impactful data points — products that are most meaningful to stakeholders who care about high-performing stock.

**5. Validation of Data**

Before going into visualization, we had to validate the total results that were compiled. This entailed cross-checking total numbers and ensuring the top 10 reflected actual sales figures, not inaccuracies as a result of data entry errors or outliers. This was made to provide confidence that the visualization presented an accurate narrative for the data we aimed to convey.

**6. Visualization Formatting**

Finally, the structured data was adapted to be compatible with the charting tool. Product names were reformatted for chart readability, and quantities were scaled as needed. Colors were applied distinctly to each product to facilitate visual discrimination and readability. The formatting choices were not cosmetic only — they were made to allow users to quickly understand and extract meaning from the chart.

In summary, each process in data preparation was designed to allow for correct, user-friendly, and business-oriented visualizations. From meticulous cleaning and structuring to thoughtful filtering and formatting, the whole process was aimed at ensuring that insights provided are both true to data as well as easy to act upon. Without these deliberate steps, the resulting dashboard would not be transparent, credible, and actionable — defeating the very purpose of visualization in decision-making.

**Conclusion:**

In conclusion, this paper explored the vital role of recommendation systems in online shopping, especially their ability to personalize user experience, boost sales, and raise retention. By comparing and combining content-based and collaborative filtering approaches, it was evident that item-item collaborative filtering had greater accuracy and scalability for online retailing contexts, more so where there are large item inventories and sparse user action.

In addition, the Market Basket Analysis generated by both Apriori and FP-Growth algorithms showed the capability of association rule mining in uncovering interesting facts about consumer purchasing behaviour. While Apriori offered interpretability, FP-Growth proved to be more efficient and meaningful, particularly in big data scenarios.

Young-adult dashboard (18–35 years) development further eased data interpretability through visual readability and interactivity. The dashboard not only served as a summary of key points but also as justification for the suitability of the dataset to be utilized in machine learning in e-commerce.

Data preparation stages were executed meticulously to ensure the relevance and accuracy of the analysis. Each step, ranging from cleansing up to transformation as well as validation, was justified to warrant the efficacy of the final visualizations and machine learning models.

Overall, the findings demonstrate the impressive application of data visualization and machine learning in retail planning today, which is beneficial for better product recommendations, inventory planning, and customer engagement.

**References:**

* Aggarwal, C.C., 2016. *Recommender Systems: The Textbook*. 1st ed. Cham: Springer.
* Han, J., Kamber, M. and Pei, J., 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Amsterdam: Elsevier.
* Ricci, F., Rokach, L. and Shapira, B., 2015. *Recommender Systems Handbook*. 2nd ed. New York: Springer.
* Tan, P.N., Steinbach, M. and Kumar, V., 2018. *Introduction to Data Mining*. 2nd ed. Boston: Pearson.
* Retailrocket.net, 2023. *Retailrocket Dataset*. [online] Available at: https://www.retailrocket.net/dataconf [Accessed 18 May 2025].
* Scikit-learn.org, 2024. *Scikit-learn: Machine Learning in Python*. [online] Available at: https://scikit-learn.org/stable/ [Accessed 15 May 2025].
* Tableau.com, 2024. *Tableau Dashboard Design Guide*. [online] Available at: <https://www.tableau.com> [Accessed 16 May 2025].