CSE 4/587 Project Phase-I and Phase-II

Project Title: *eCommerce behavior data from multi category store.*

Group Members:

Member 1 Harsha Venkateshwara (**hvenkate**, ID:50680830)
Member 2 Saba Minaz Taj (**sabamina**, ID:50681904)
Member 3 Sphoorthy Selvaraj (**selvara3**, ID: 50682811)

This dataset provides user activities logged from a large multi-category eCommerce store during the month of October 2019. It captures page views, cart additions, and purchases to analyze customer behavior at the category level. Each row of the data consists of a user ID, product category, event type, and timestamp, enabling trend analysis at the session level. The dataset features over 4 million events to support both purchase prediction and funnel analysis, as well as session-based recommendations. It is a suitable dataset for large-scale analytics and distributed processing by using Hadoop, Spark, or MapReduce for insights related to purchasing patterns that lead to actionable outcomes.

1 Data Cleaning

In Phase I Pandas was used to clean the data in a way that did not use up too much memory. The large e-commerce dataset was read in chunks of 500,000 rows to avoid running out of memory. Each chunk was cleaned separately by removing rows with missing or null values in important columns such as `event_type`, `product_id`, and `price`. The `price` values were converted to numeric type and timestamps were converted to datetime format. Prices that were not positive or exceeded 20,000 were filtered out to remove outliers. To keep the data complete, the placeholder `"unknown"` was used to fill in missing categorical values such as `category_code` and `brand`. Duplicate rows were removed to ensure data accuracy, and Pandas' `category` datatype was used to optimize categorical columns and reduce memory usage. Finally, all of the cleaned chunks were concatenated into a single DataFrame and saved as `ecommerce_cleaned.csv`, resulting in a structured, consistent, and analysis-ready dataset.

In Phase II, we used PySpark to clean up the data in a way that could be scaled up to handle the full e-commerce events dataset, which had 67,501,979 records and 9 features (`event_time`, `event_type`, `product_id`, `category_id`, `category_code`, `brand`, `price`, `user_id`, and `user_session`). We loaded the dataset into a Spark session with more memory for the driver and executor so they could handle a lot of data. The first

thing that was done was to change the `event_time` field from a string to a proper timestamp type so that it could be used for time-based analysis. We dealt with missing values by removing all rows where `product_id` or `user_id` were null. These are important identifiers for analytics that come later. We used the placeholder "unknown" to fill in missing values for other non-critical fields like `category_code` and `brand`. This made sure that no category or brand information messed up the model inputs. We made sure that the `price` column was a float by filtering out negative prices, which are not valid in transaction-level data. We also got rid of any rows where `event_time` couldn't be parsed correctly, which made sure that the time was always the same. The dataset still had 67,501,979 rows after all the cleaning steps were done. This shows that the original dataset was mostly well-structured and only needed a few small changes. We used `df.printSchema()`, `df.show()`, and before/after row counts to check that the schema was consistent and the data was of good quality throughout the cleaning pipeline. This showed that PySpark was able to handle the large dataset and create a clean data frame that was ready for analysis.

2 Exploratory Data Analysis

In Phase I, we performed an in-depth local exploratory data analysis using Pandas on the cleaned October dataset to understand user behavior, product interaction patterns, and pricing trends. We examined distribution patterns, category and brand popularity, hourly activity trends, conversion behavior, and feature correlations to extract insights that support our problem statements.

Phase I EDA Steps

- **Event Type Distribution** – Most user actions are product views, followed by "add to cart" and few purchases-typical conversion funnel.

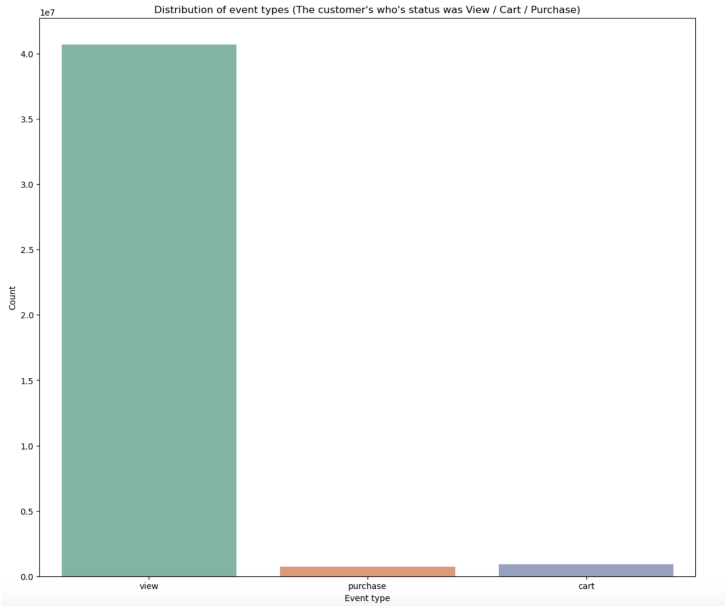

Figure 1: Event Type Distribution

- **Price Distribution** – Reveals the right-skewed nature of product pricing with many low-cost products.

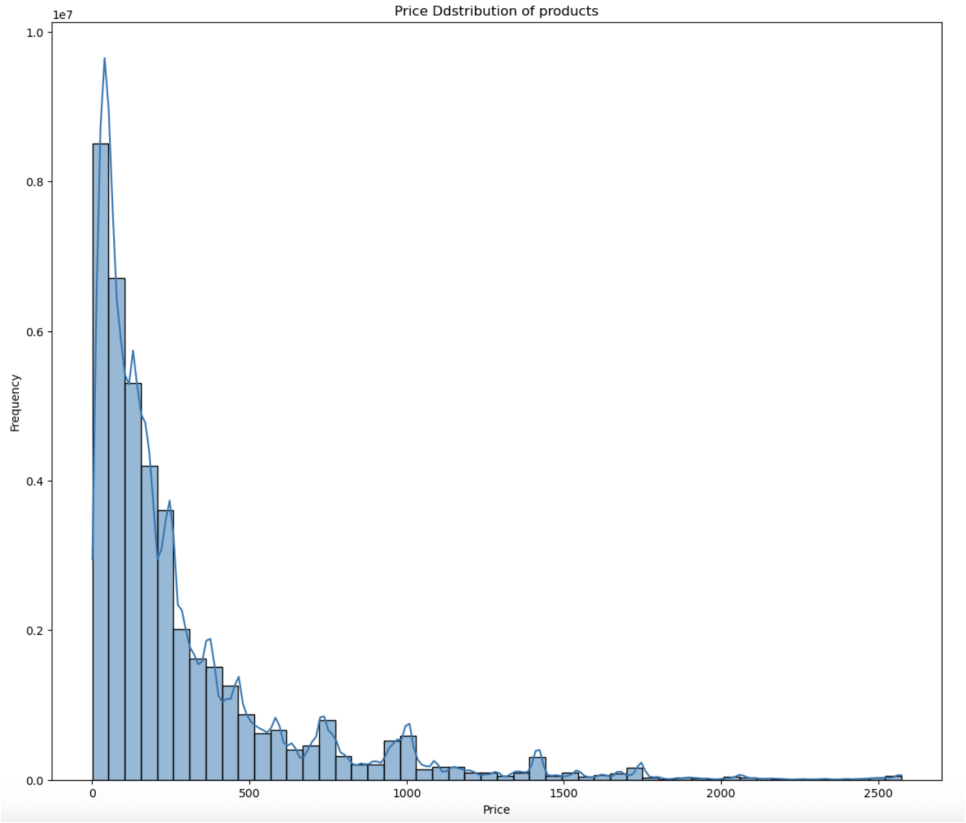

Figure 2: Price Distribution

- **Top Purchased Categories** – Identifies which product categories generate the

highest purchase volume.

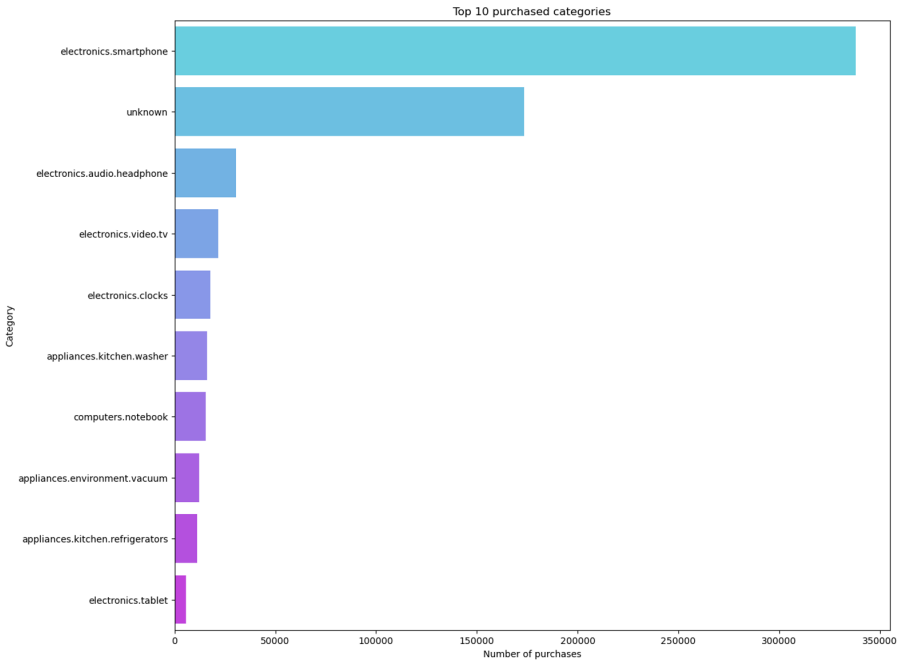

Figure 3: Top 10 Purchased categories

- **Top Purchased Brands** – Highlights brand-level purchase frequency and customer preferences.

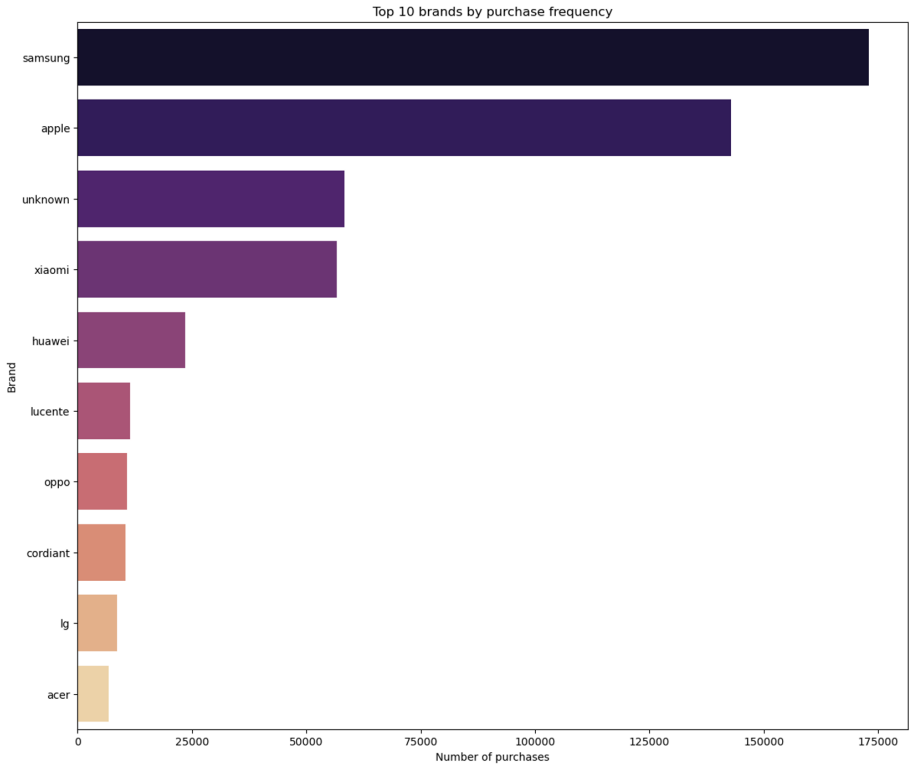

Figure 4: Top 10 brands by purchase frequency

- **Hourly Activity Trend** – Displays peak user engagement hours throughout the day.

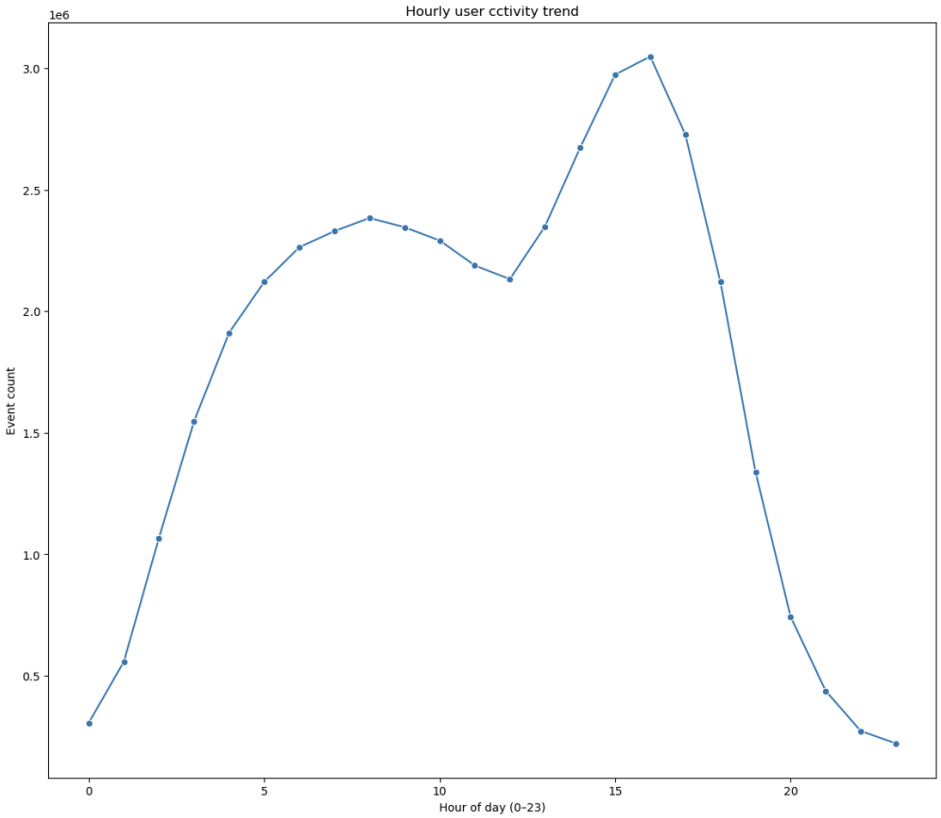

Figure 5: Hourly user activity trend

- **Category Conversion Rate** – Measures the efficiency of each category through purchase-to-view ratios.

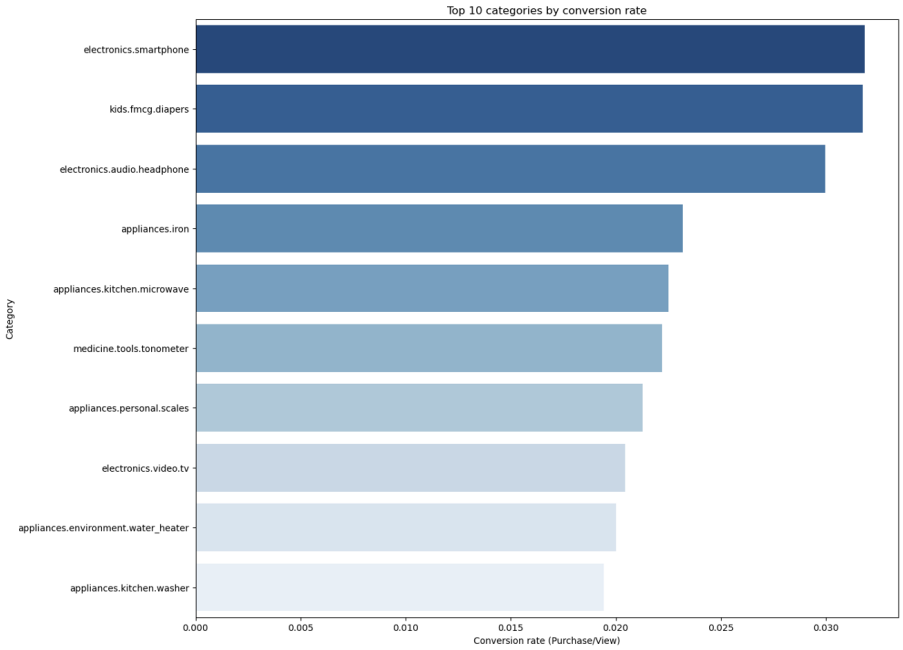

Figure 6: Top 10 categories by conversion rate

- **Correlation Heatmap** – Examines relationships between numeric features such as price and product identifiers.

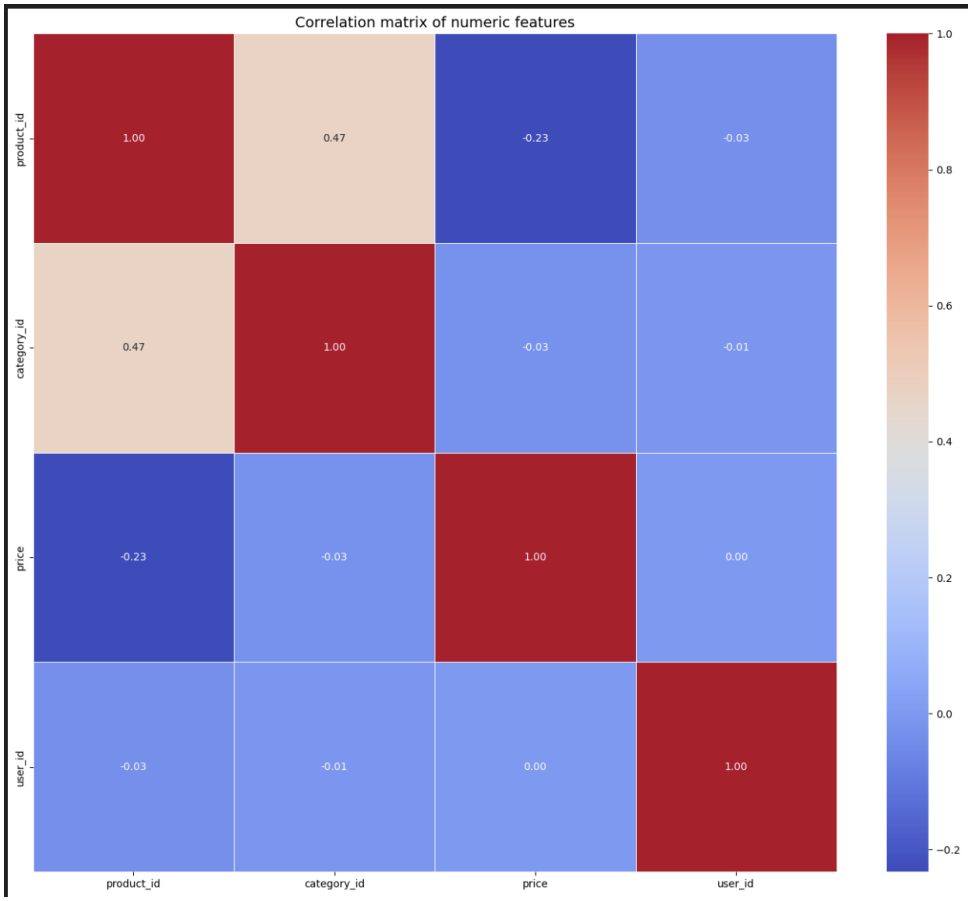

Figure 7: Correation Heatmap

- **Summary Statistics** – Provides descriptive metrics for price, user activity, and product interactions.
- **Missing Value Analysis** – Verifies that the cleaning pipeline effectively handled null or invalid values.

In Phase II, the exploratory data analysis was scaled using PySpark to process the complete 67 million–record e-commerce dataset, ensuring fast computation and reliable large-scale insights. This phase validated the trends identified during Phase I while extending the analysis to include deeper behavioral patterns such as session-level activity, daily purchase fluctuations, and detailed category-level interactions across the entire dataset.

Phase II EDA Steps

- **Event Type Frequency** – Confirms large-scale distribution of view, cart, and purchase events.

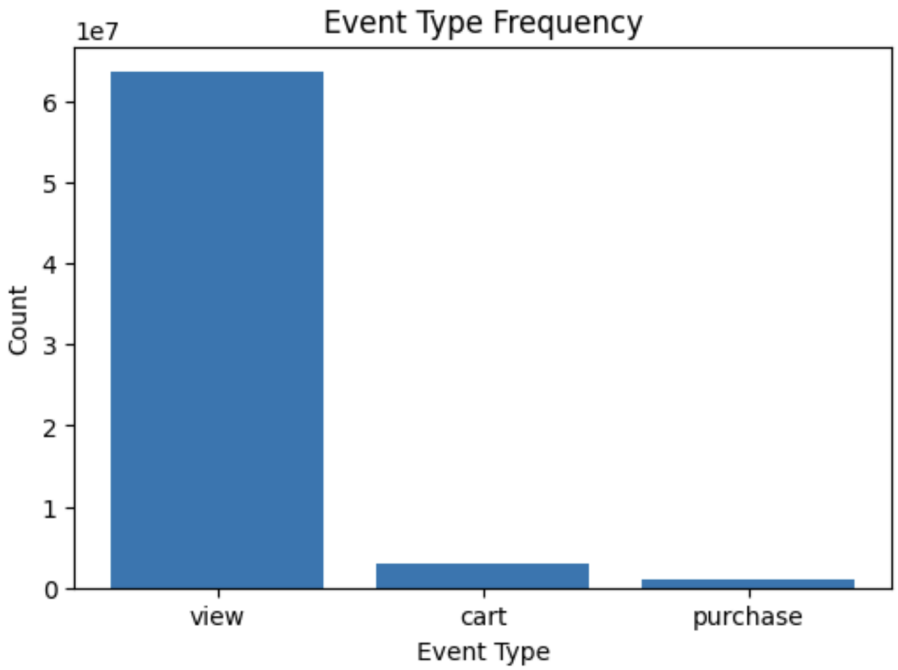

Figure 8: Event Type Frequency

- **Top Viewed Brands** – Determines the brands with highest user interaction across the full dataset.

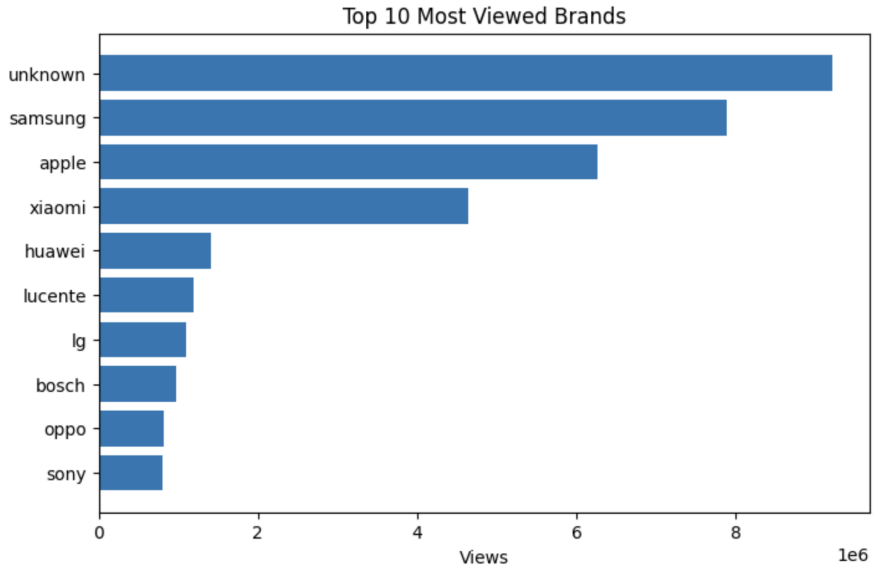

Figure 9: Top 10 Most viewed Brands

- **Top Viewed Categories** – Shows the most frequently accessed product categories platform-wide.

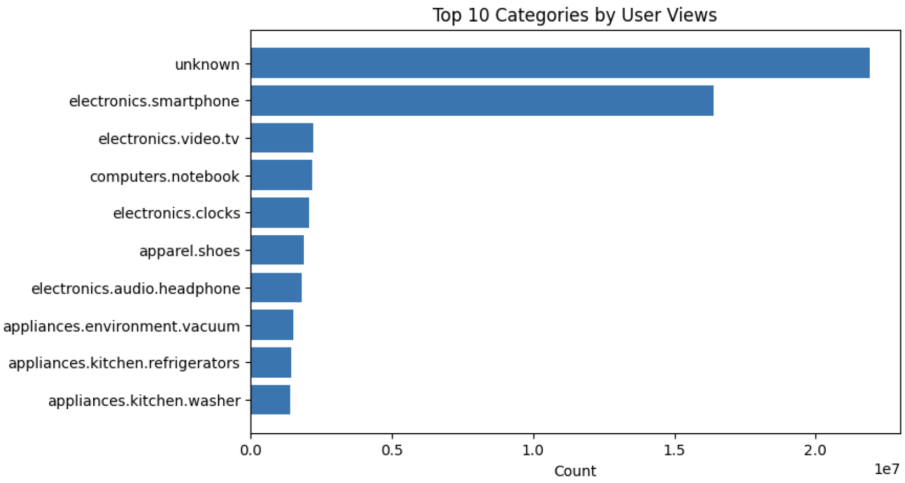

Figure 10: Top 10 Categories by User Views

- **Price Distribution** – Uses a representative sample to visualize global price patterns efficiently.

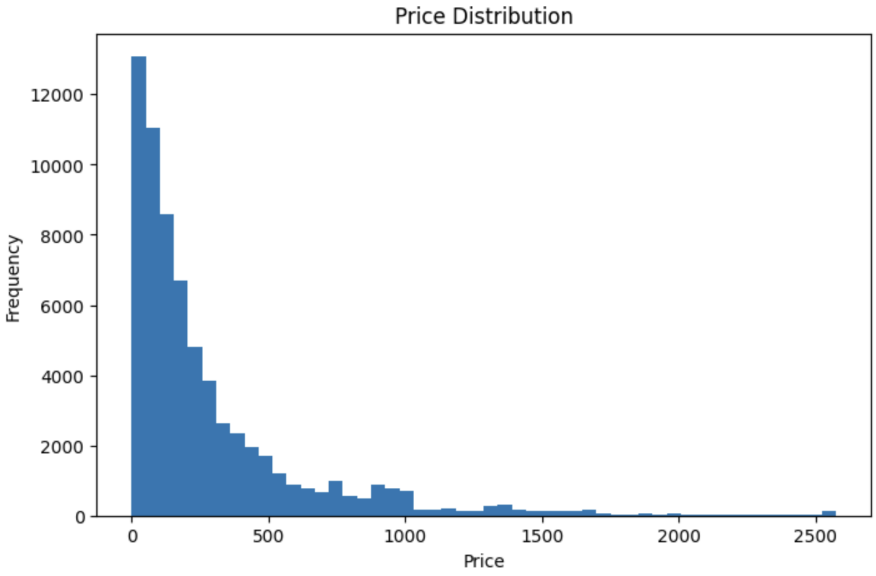

Figure 11: Price Distribution

- **Daily Purchase Trend** – Reveals temporal changes in consumer purchase behavior.

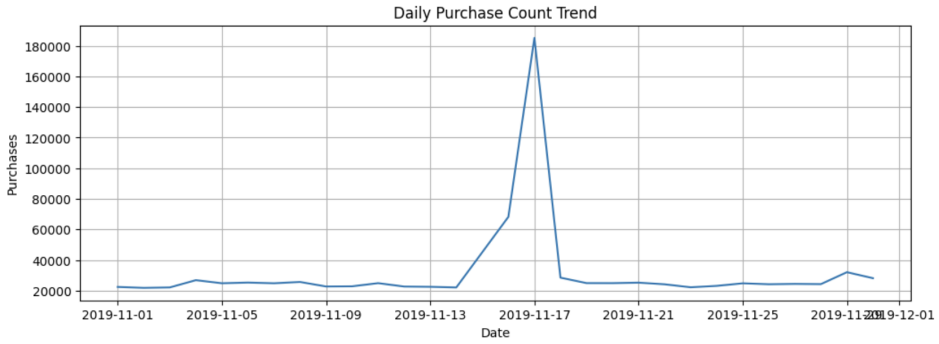

Figure 12: Daily Purchase Count Trend

- **Active User Sessions** – Highlights the most active sessions exhibiting large numbers of events.

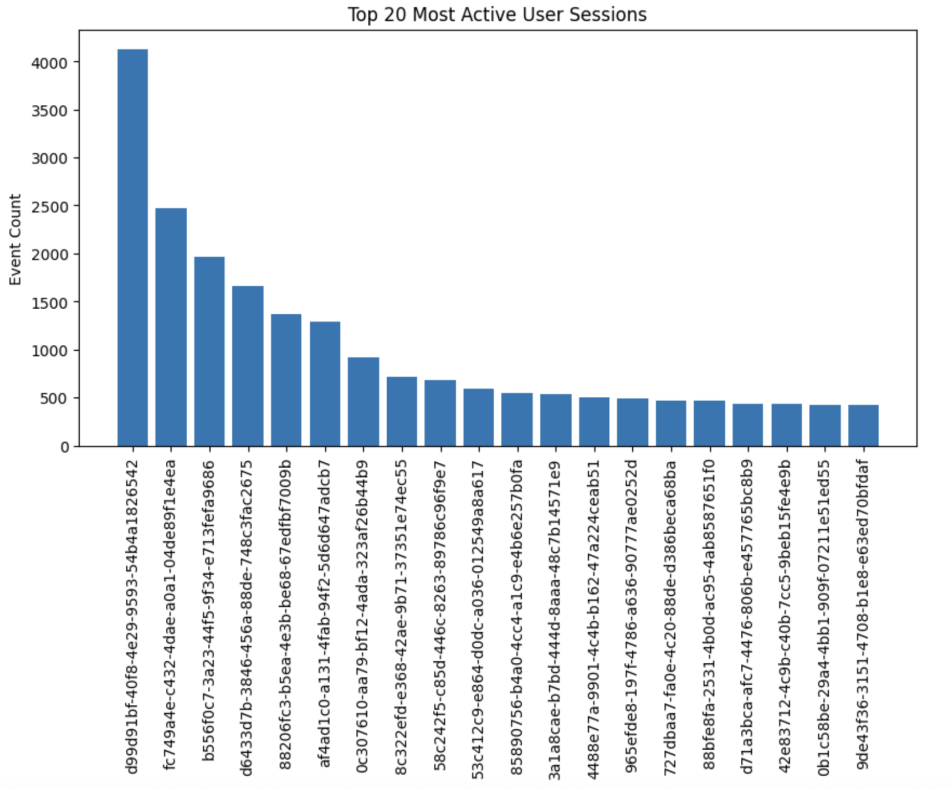

Figure 13: Top 20 Most Active User Sessions

- **Session Behavior Aggregates** – Summarizes per-session metrics such as total events and unique items viewed.

user_session	total_events	total_views	unique_products	unique_categories	avg_price	unique_brands	label
00006325-76f4-4a9...	1	1	1	1	163.4499969482422	1	0
00044634-a460-438...	1	1	1	1	202.5800018310547	1	0
00054e4b-7145-4b3...	1	1	1	1	231.63999938964844	1	0
00060054-401d-4d3...	1	1	1	1	501.9200134277344	1	0
00078b94-99b3-4c9...	1	1	1	1	41.189998626708984	1	0

Figure 14: Session Behavior Aggregates

- **Category-Level Feature Summary** – Aggregates views, purchases, pricing, and user counts per category.
- **K-Means Category Clustering** – Identifies behavioral clusters among product categories.
- **Silhouette Score Evaluation** – Assesses the separation quality of the clustering model.

3 Data Analysis Objectives

In this project, we primarily concentrate on three main issues that were initially outlined in our Phase I report. These problems were chosen by considering the eCommerce behavioral dataset structure, the multi-event nature of user interactions, the rich temporal information embedded in the timestamps, and the existence of category-level and session- level patterns. Apart from being different from each other, each problem is firmly supported by the dataset's features and offers significant analytical worth when moving from Phase I to Phase II.

1. **User Conversion Prediction**
   The data collection is made up of user sessions in the millions, with each session having several events like views, cart additions, and purchases. As a result of the fact that just a tiny portion of sessions end in purchases, figuring out if a certain session is going to convert is a valuable task. The task here is to create a classification model that will utilize behavioral features (event frequencies, unique products viewed, category diversity, and average price) to predict the probability of a purchase. This is in perfect harmony with the work done in Phase II where session-level feature engineering and classification models were used.

2. **Product Category Clustering**
   Clustering categories based on user interactions is very informative about product affinities and demand patterns when there are hundreds of product categories and large differences in engagement levels. Unsupervised learning to group categories according to metrics like total events, views, purchases, average price, and unique users is the main focus of this work. The Phase II K-Means clustering and silhouette score evaluation provide more evidence for this issue by showing that category groups are distinctly separable.

3. **Sales Forecasting**
   The dataset comprises events showing purchases with the time and date, thus making it possible to examine daily sales as well as sales patterns over time. The objective here is to predict the number of purchases made daily in the future through the use of regression methods that are based on temporal features like the day of

the week and the day of the month. In the second phase, it goes further to creating these models with linear and random forest regression, thereby checking the usefulness and figuring out the significance of forecasting short-term sales trends for planning operations.

4 Problem Statements

1. **User Conversion Prediction**
   Develop a classification model to predict whether a user session will result in a purchase based on prior browsing actions and event frequency.

   - **Goal 1:** Session-level behavioral features, such as total events, total views, unique products, unique categories, and average price, were derived from the Phase II EDA to capture user engagement patterns.
   - **Goal 2:** A classification modeling pipeline was developed to train and evaluate multiple models, including Logistic Regression, Random Forest, and Gradient Boosted Trees. Model performance was assessed using accuracy and F1-score, providing measurable evidence of how effectively session-level purchases can be predicted based on user interaction features.

2. **Product Category Clustering**
   Utilize unsupervised learning to group product categories according to user interaction patterns and identify commonalities in customer interest.

   - **Goal 3:** Using Phase II EDA, build category-level metrics—total events, views, purchases, average price, and unique users—that reflect the behavioral characteristics of each category.
   - **Goal 4:** Perform K-Means clustering and verify cluster quality using the Silhouette score, confirming that the resulting clusters are meaningful and well separated.

3. **Sales Forecasting**
   Forecast daily purchase counts based on historical patterns and time-based trends using regression techniques.

   - **Goal 5:** Consolidate daily purchases records and derive time features (weekday, day of month) guided by the fluctuating patterns observed in Phase II EDA.
   - **Goal 6:** Create and test regression models (Linear Regression, Random Forest Regression) with RMSE and MAE as quantifiable forecasting metrics.

5 Machine Learning Models

The machine learning models deployed in this work align with the three issues highlighted in Phase I, and their selection was validated through the data characteristics uncovered during both Phase I and Phase II EDA. Phase I served as an initial modeling stage using Scikit-Learn on the cleaned October subset, while Phase II extended these models into distributed PySpark pipelines operating on the full dataset. For each issue, we describe the models implemented, provide justification for their selection, outline the training and hyperparameter tuning process, and present key performance metrics along with the most suitable visualizations such as ROC curves, learning curves, and confusion matrices.

5.1 Classification Models for User Conversion Prediction

The goal here is to figure out if a user purchase would be the outcome of a session relying on the behavioral features that we derived from the event log. Phase I was more about the trials done in a quick way with the help of Scikit-Learn, while in Phase II, models for classification that were scalable and written in PySpark were used for training on the samples of session-level aggregates taken from the whole dataset.

Models Implemented

- Logistic Regression (LR)
- Random Forest Classifier (RF)
- Gradient Boosted Trees Classifier (GBT)

Justification

- **Logistic Regression** works as an understandable baseline for binary classification and is a good performer with feature sets of moderate scale.
- **Random Forest** identify nonlinear user behavior patterns and the interactions among features such as unique categories, price levels, and view frequency.
- **Gradient Boosted Trees** are capable of delivering excellent results on high-dimensional structured data and can represent complex decision boundaries.

Training and Hyperparameter Tuning

- Data was aggregated at the session level using PySpark, generating features including total events, total views, unique products, unique categories, average price, and unique brands.

- Train/test split: 80–20.
- PySpark's `VectorAssembler` and `StringIndexer` were used for feature preparation.
- Hyperparameters tuned:
  - LR: regularization parameter, elasticNet mixing parameter.
  - RF: number of trees, max depth.
  - GBT: number of iterations, learning rate, max depth.
- A 3-fold `CrossValidator` was used for tuning Logistic Regression.

Performance Summary (Phase II)

Model	Accuracy	F1-score
Logistic Regression	0.978	0.9792
Random Forest	0.9858	0.977
Gradient Boosted Trees	0.977	0.9787

Plots

- ROC curves for LR, RF, and GBT (AUC $\approx$ 0.978 for all models).

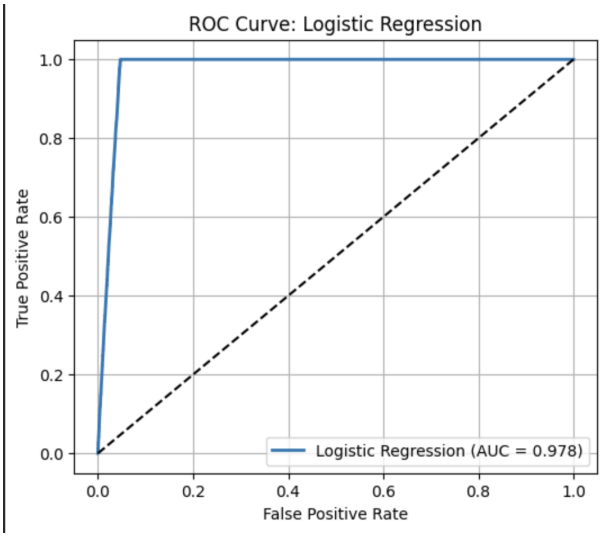

Figure 15: ROC Curve- Logistic Regression

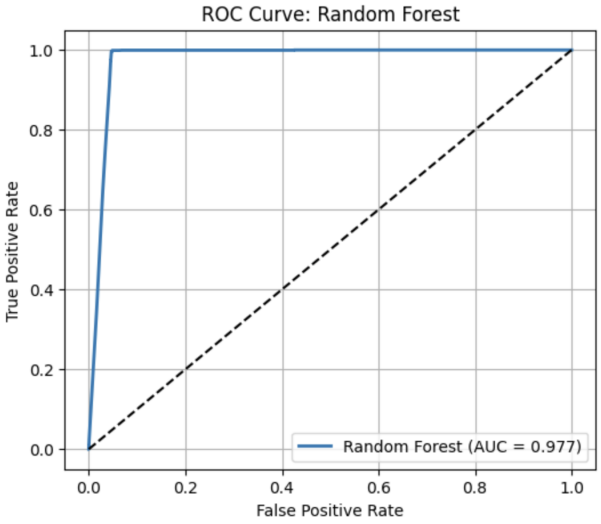

Figure 16: ROC Curve - Random Forest

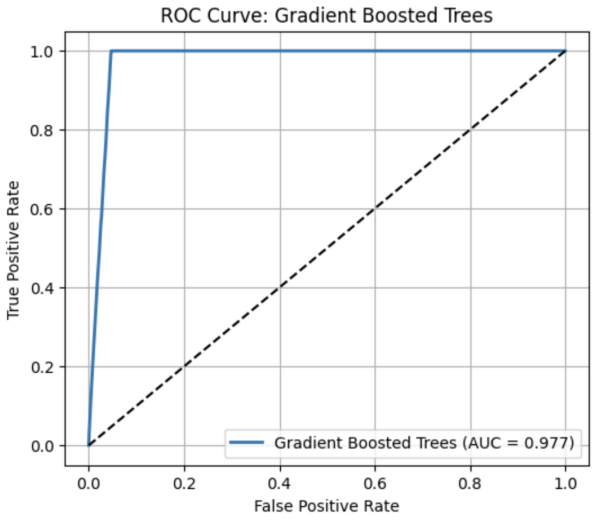

Figure 17: ROC Curve - Gradient Boosted Trees

5.2 Unsupervised Learning for Product Category Clustering

This task focuses on categorizing product groups based on patterns of user interactions, allowing us to identify natural clusters of behavioral similarity. In Phase I, strong differences between categories were observed, and Phase II formalized the approach by using PySpark's K-Means clustering on category-level metrics.

Model Implemented

- K-Means Clustering

Justification

- Category-level metrics (events, views, purchases, average price, unique users) are well-suited for centroid-based clustering.
- EDA showed clear separation across product categories, motivating the use of K-Means.
- K-Means is efficient and scales well in PySpark.

Training Details

- Features were engineered per category using grouped PySpark aggregations.
- Standardized feature vectors using `VectorAssembler` and `StandardScaler`.
- K-Means was trained with $k = 5$ after testing multiple values of k.

Performance Summary

- The clustering quality was evaluated using the Silhouette Score.
- Phase II achieved a strong Silhouette Score of **0.8454**, indicating well-separated clusters.

5.3 Regression Models for Sales Forecasting

The goal of this task is to forecast daily purchase volume using historical temporal patterns. Phase II implemented two regression models using aggregated daily purchase data.

Models Implemented

- Linear Regression (LR)
- Random Forest Regression (RFR)

Justification

- **Linear Regression** provides a simple and interpretable baseline for modeling linear time-based trends.
- **Random Forest Regression** captures nonlinear variations and interactions in daily purchasing behavior.

Training and Evaluation

- Daily purchase counts were aggregated using PySpark and enriched with temporal features (day of week, day of month).
- Data was vectorized using `VectorAssembler`.
- Models were evaluated using RMSE and MAE.

Performance Summary (Phase II)

Model	RMSE	MAE
Linear Regression	695.41	311.49
Random Forest Regression	663.74	296.58

Overall,using classification, clustering, and regression models in both phases gave us the ability to respond to all the three problem statements completely. The models were effective to a great extent when measured by accuracy, F1-score, clustering quality, and regression error metrics, thus, they serve as a confirmation that the selected modeling techniques are appropriate for extensive e-commerce behavioral analysis.

6 Key Findings and Recommendations

This project delved into the analysis of user behavior, product category patterns, and temporal purchasing trends using a large-scale eCommerce event dataset. Through detailed Phase I EDA and scalable Phase II modeling, we effectively addressed the three issues raised: user conversion prediction, product category clustering, and sales forecasting. The main findings, along with the corresponding recommendations, are summarized in the table below.

Key Findings

- **User behavior is heavily skewed toward product views.** Phase I EDA showed that most of the events are "view" actions, with a significantly smaller number of cart additions and purchases. The funnel drop-off observed in Phase I was also present in Phase II, and it was the primary motivation behind designing the conversion prediction model.
- **Session-level behavioral features strongly correlate with purchase likelihood.** Several features like total events, total views, unique products, unique categories, and average price were identified as powerful predictors. The Phase II classification models reached a high level of performance (Accuracy ≈ 0.986, F1-score ≈ 0.985), thus supporting the idea that the intensity of user interaction can be inferred with a high degree of certainty.

- **Product categories form distinct behavioral clusters.** The K-Means model in Phase II (Silhouette Score = 0.8454) showed that categories vary consistently in terms of engagement, price, and volume. These clusters can inform marketing segmentation, personalized recommendations, and inventory management.
- **Daily purchasing activity shows consistent temporal patterns.** Both Phases I and II showed strong daily cycles as well as moderate weekly variations. The regression models reflected these patterns with a fair degree of accuracy (RF Regression RMSE = 663.74), thereby indicating that sales can be partly predicted from time-related features.
- **All three problem statements were achieved successfully.**
  - Conversion prediction models achieved high predictive accuracy.
  - Category clustering produced meaningful groupings.
  - Sales forecasting models captured key temporal dynamics.

Recommendations and Future Work

- **Expand modeling beyond session-level features.** If available, incorporating user demographics, device information, or session duration could significantly enhance the accuracy of conversion prediction.
- **Enhance model interpretability.** While Random Forests and GBT gave good results, an analysis by SHAP or LIME might enlighten feature importance and decision paths to a greater extent.
- **Improve category clustering with additional semantic features.** Besides using purely numeric engagement metrics, one may obtain more significant clusters by utilizing product descriptions, textual metadata, or hierarchical category embeddings.
- **Develop multivariate and sequence-based forecasting models.** The current regression models capture linear and nonlinear temporal trends, but recurrent architectures (LSTM, GRU) or Prophet could improve long-range forecasting.

Essentially, the project was successful in demonstrating how large-scale user interaction data can be used to forecast conversions, recognize behavioral patterns among product categories, and predict shifts in demand. While the models achieved strong results, the recommendations provided above highlight the remaining challenges as well as opportunities for more advanced analytics in future versions of this work.

7 Video Presentation

A recorded project presentation video is available at the following link. All group members appear in the video and describe their respective contributions to the project.

https://buffalo.box.com/s/pecpdn3p0cnr6avkn1uq9ccyxevdc85u