Rocket Retail Dataset Exploratory Data Analysis

Contents
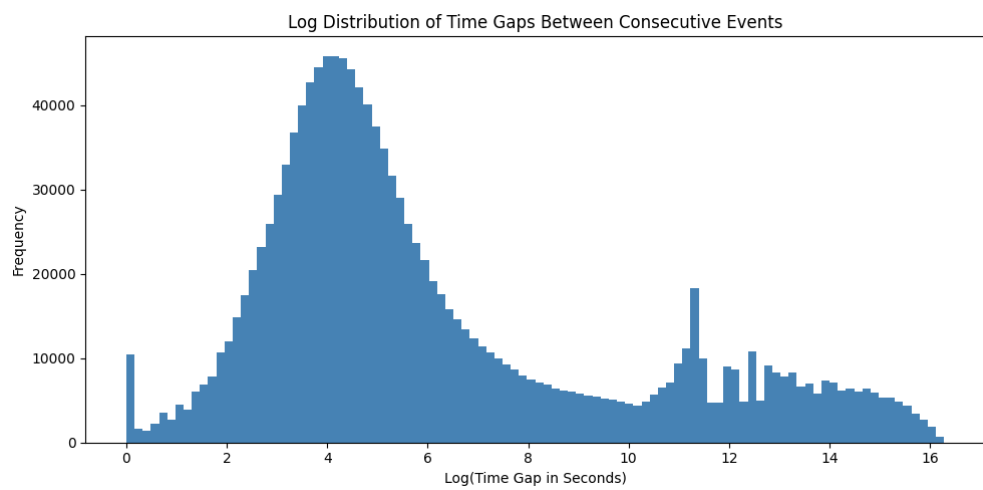
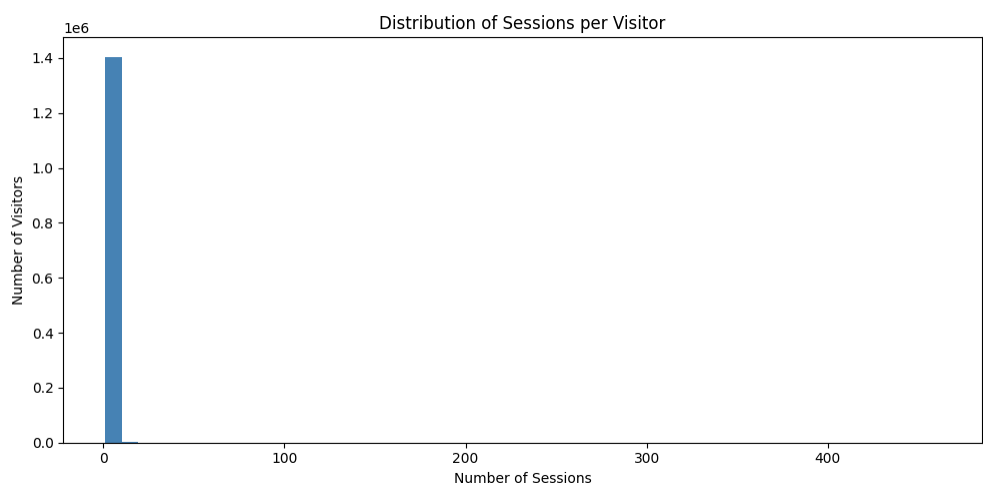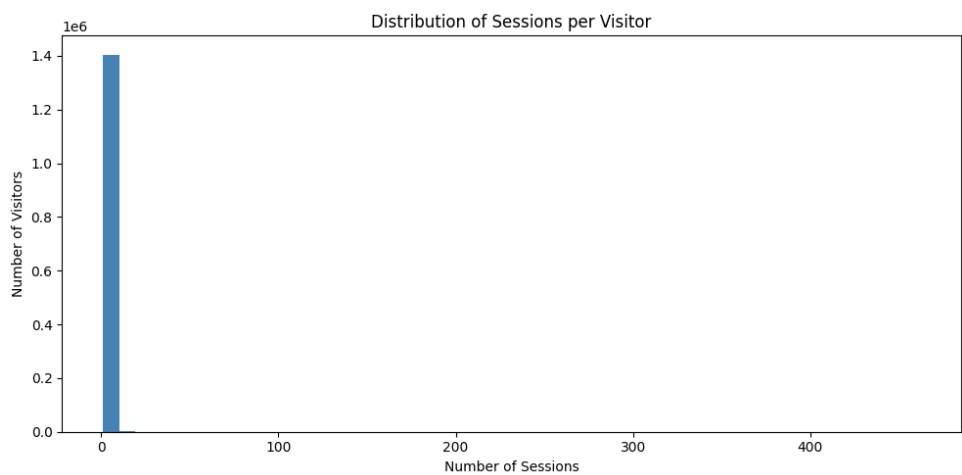# Rocket Retail EDA – Executive Summary

This report provides a detailed behavioural overview of the Retail Rocket ecommerce dataset, based on millions of real-world customer interactions. The analysis follows the customer journey across key stages, from passive browsing to purchase behaviour, to highlight insights for business strategy and machine learning segmentation.

1. Event Type Distribution
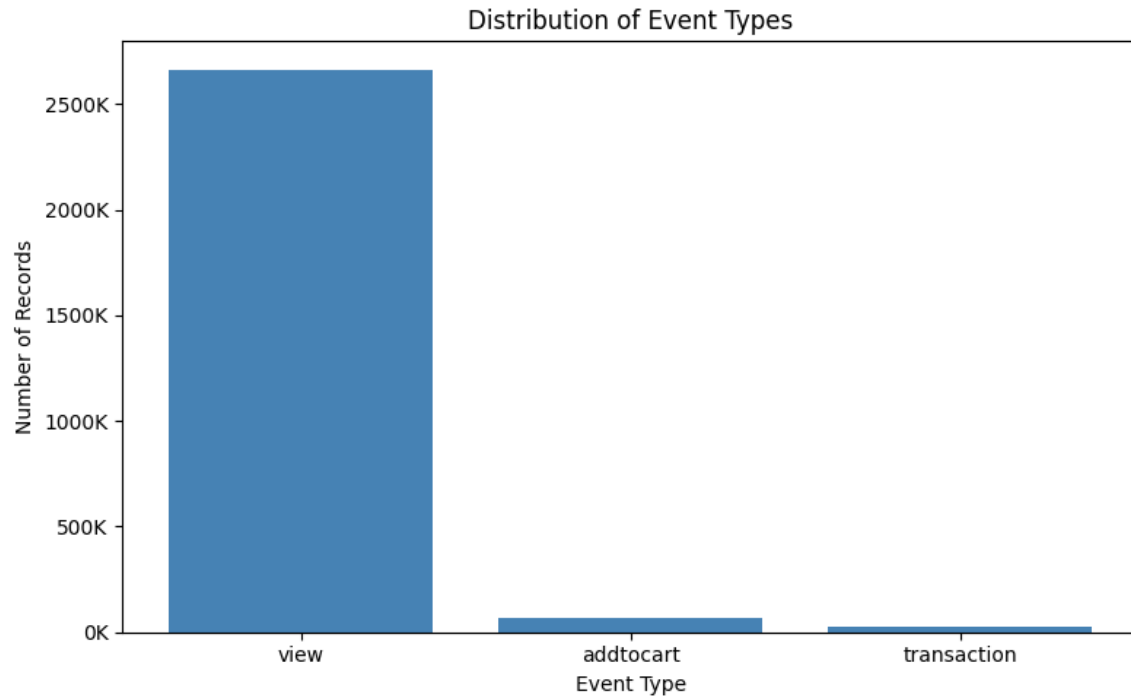
The initial event type distribution reveals a significant skew towards viewing events, which dominates the dataset. Add-to-cart and transaction events form a much smaller proportion, showing natural friction in the funnel — particularly in early stages.

2. Visitor Activity Patterns

Distribution of Sessions per Visitor

Distribution of Sessions per Visitor

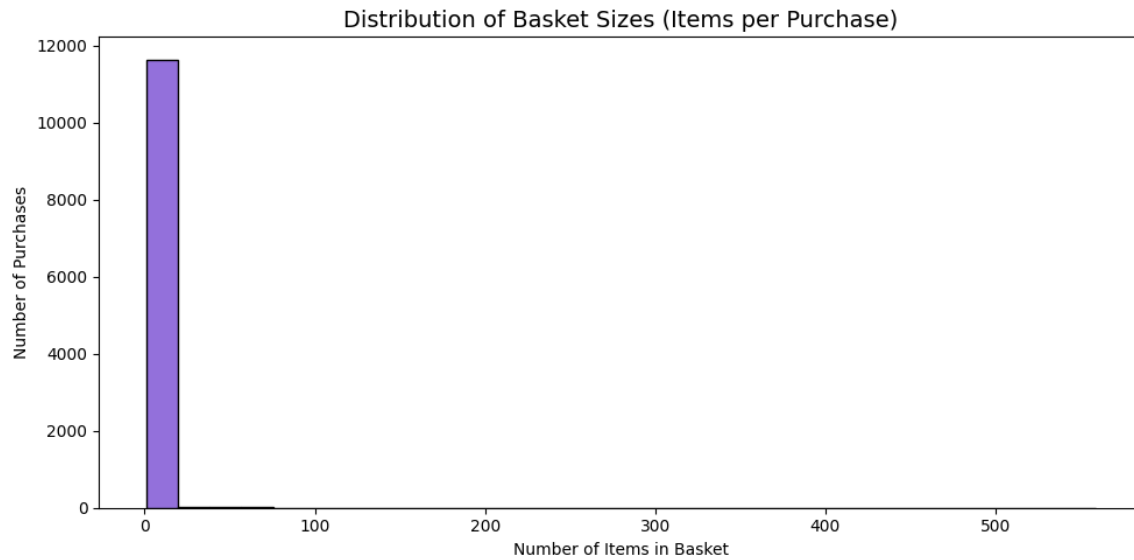**Log Distribution of Time Gaps Between Consecutive Events**

User engagement is similarly imbalanced. While many users only appear once, a small cohort interacts frequently and across sessions. This pattern aligns with the Pareto principle in ecommerce. Sessions were constructed using a 30-minute inactivity threshold, supported by analysis of event time gaps. These sessions represent meaningful behavioural windows that serve as input to downstream sequence modelling.

3. Conversion Funnel
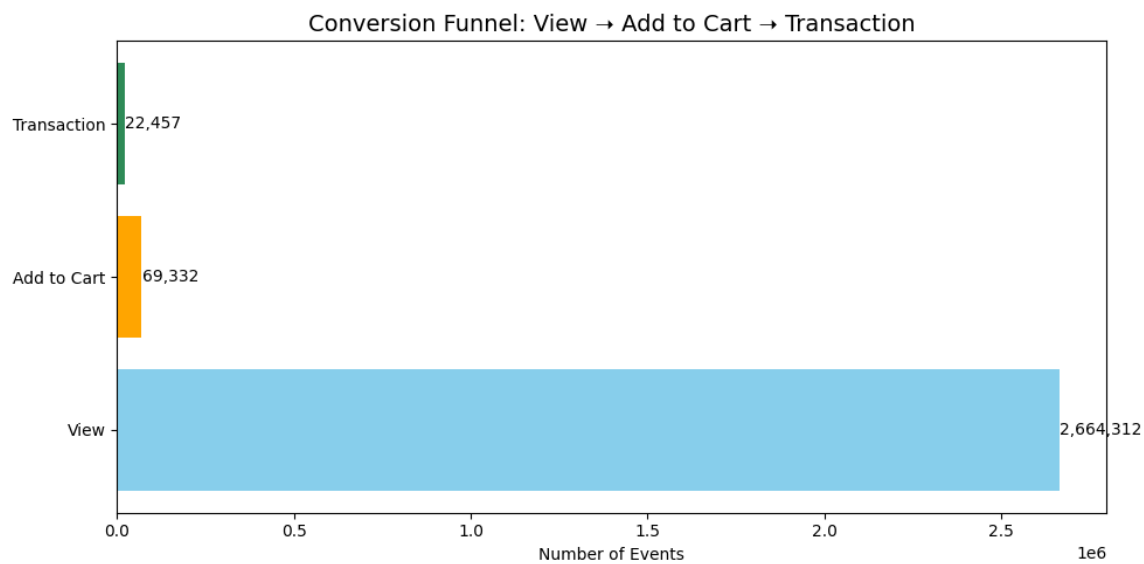
The conversion funnel shows that only ~2.6% of views lead to cart additions, and ~32.4% of that result in purchases. This funnel visualisation quantifies drop-off at each stage and underscores where ecommerce interventions (like retargeting or UX improvements) are most needed.

4. Basket Size Analysis

Distribution of Basket Sizes (Items per Purchase)

Basket size analysis shows that most transactions consist of just one item, but some purchases exceed 500 items. These outliers suggest either bulk buying or automation. Understanding these behaviours helps identify opportunities for cross-selling or flagging anomalous activity.

5. Segmentation

Visitor Interaction Frequency Distribution (Log Scale)

Segmentation analysis highlights that most users are one-time visitors, while a much smaller segment — power users — are responsible for most of the engagement. This validates the need for behavioural clustering and differential targeting.

6. Category-level insights

Top 10 Categories by Transactions
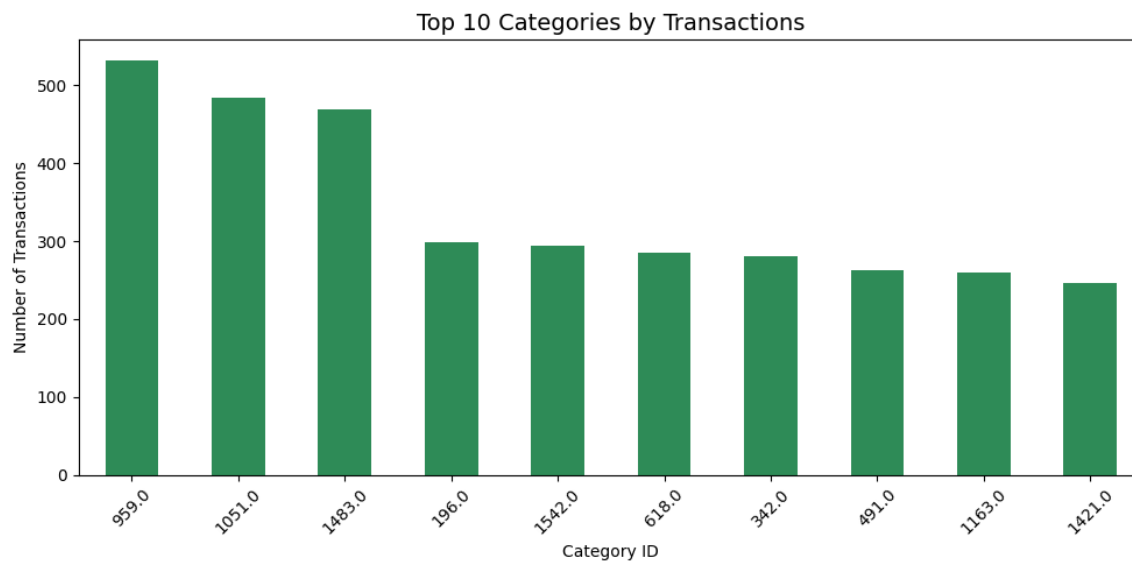
Category-level insights reveal that a few specific categories dominate transactions. These patterns inform inventory management and promotional design. Time-lag analysis between view, cart, and transaction shows that many actions are fast and impulsive, while others span longer durations — possibly across sessions. This justifies the use of both real-time and delayed recommendation strategies.

7. Sunburst of Raw Category IDs

Sunburst of Transactions by Raw Category ID

Lastly, the sunburst chart provides a visually intuitive summary of transactions across raw `categoryid`s. No assumptions were made about category hierarchy, maintaining analytical transparency.

Together, these insights form a coherent picture of Rocket Retail user behaviour — supporting more personalised, efficient, and effective customer engagement strategies.