Personalized Ecommerce Recommendation

Reya Sadhu
Atharvraj Patil
Arnav Kulkarni
Allen Chen

Date: 4th December 2024

Presentation Overview

Data Overview & Statistics

Dataset's features and key statistics, descriptive summaries, and insights into data distribution.

Seasonal Trends & Patterns

Temporal trends such as daily, weekly, and monthly seasonality to identify patterns in user behavior.

User and Item Level Analytics

Site visiting patterns, view to purchase ratio, engagement frequency, abnormal traffic analysis.

Modelling & Results

Data Manipulation, modelling techniques, evaluation metrics and insights from prediction.

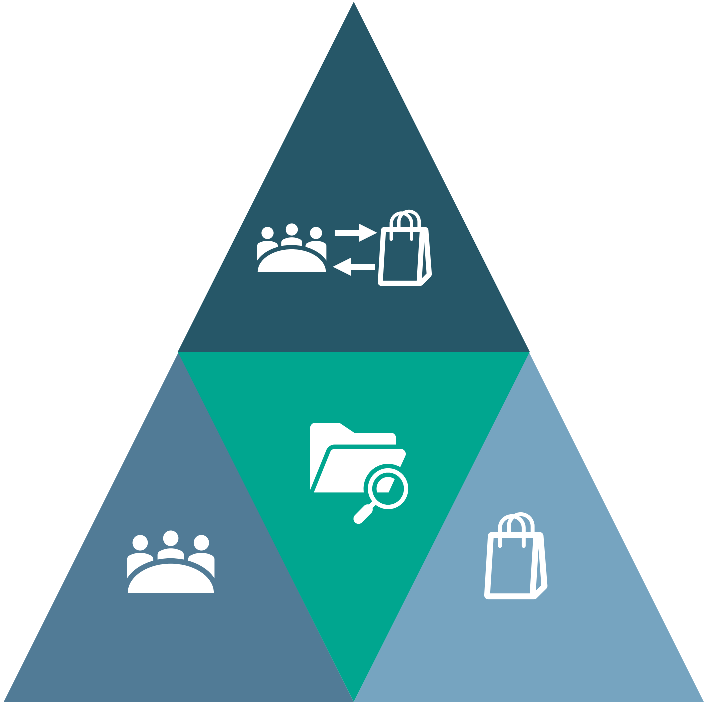
General Ecommerce Analytics Data Structure

User-Item Interaction

Type of interactions user is having with an item: impressions, clicks, views, rating, review, purchases, along with their timestamp.

User Attributes

User demographics, age, gender, location, behavioral pattern, search history.



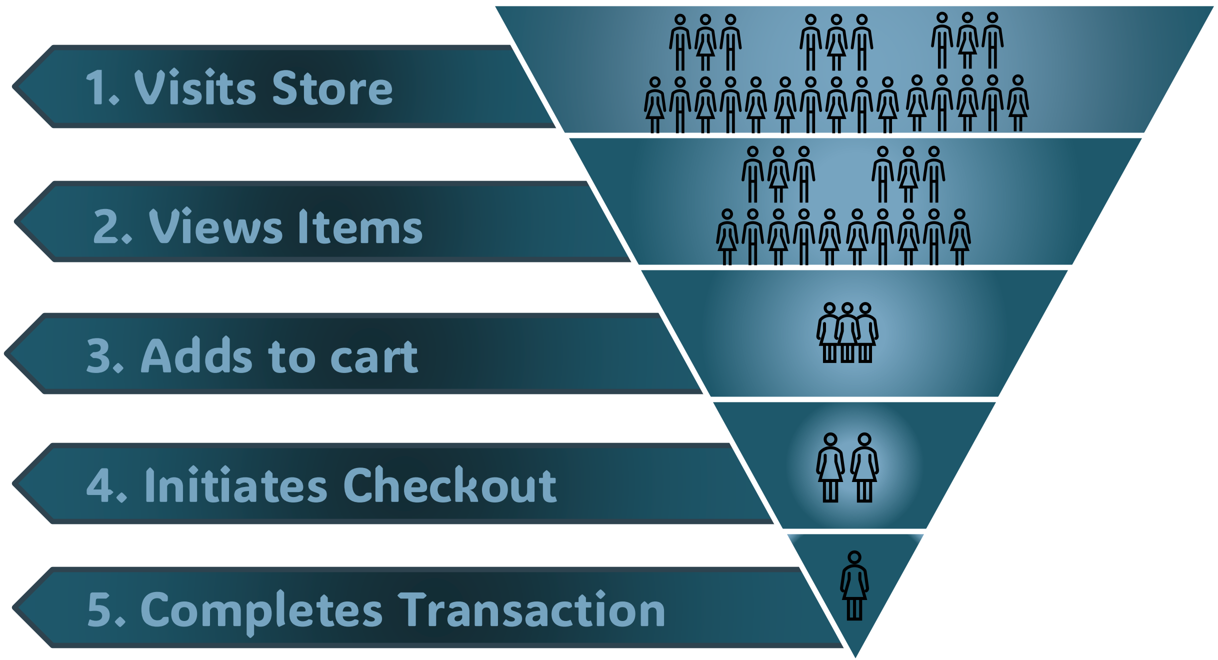
Derived Attributes

Historical interactions features, like number of visits, views, average ratings, time spent, frequency of interaction.

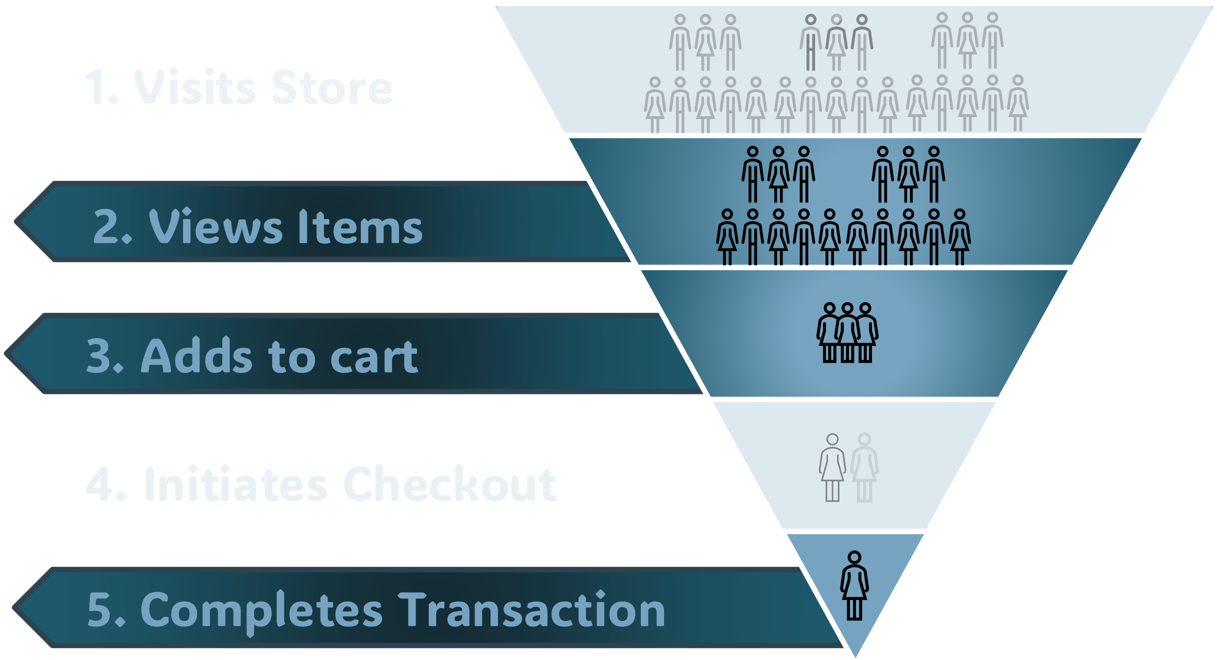
Item Attributes

Item price, tags, title, description, quantity, category, rating, availability, visual features, material etc.

Ecommerce Sales Funnel



Ecommerce Sales Funnel



RetailRocket Ecommerce Data Description

1.4M Users

400K Items

2.7M Interactions

1180 Categories

Three types of interactions

1
- View (click)
- Add-to-cart
- Transaction

Timeline

2

2015-05-02
To
2015-09-17

Item Properties

3

Except availability and category, all other properties are hashed

Events

User
Item
Event
Time

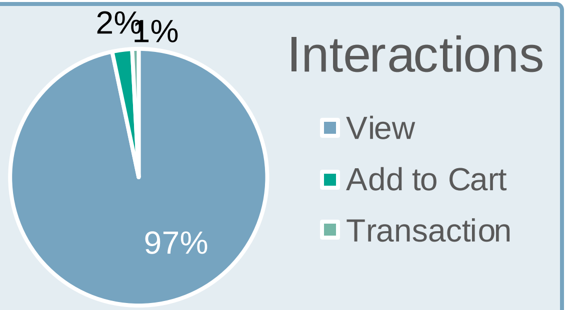
Item properties

Item
Time
Properties

Category tree

Category

Parent category



Data Insights

Trend

How does the seasonal trend look? Can we explain those? Can we focus our marketing or promotional strategy to fit them?

What are the dropout rates between each step to understand potential friction points in the buying process?

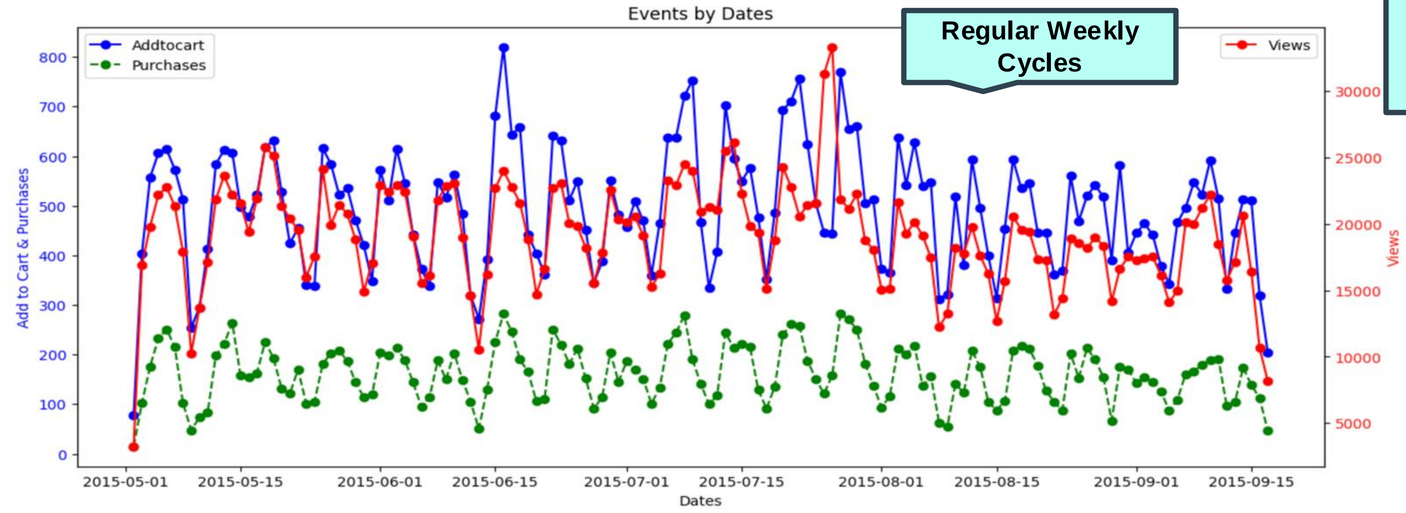
Segmentation

Which are the loyal customers? How often do they purchase in a specified timeframe? Which segment can be potential churners?

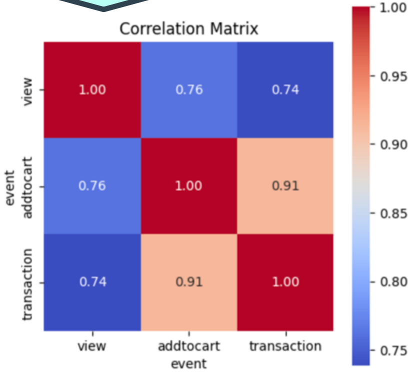
Which are the top selling items? What categories do they belong to? Which products are viewed but not purchased?

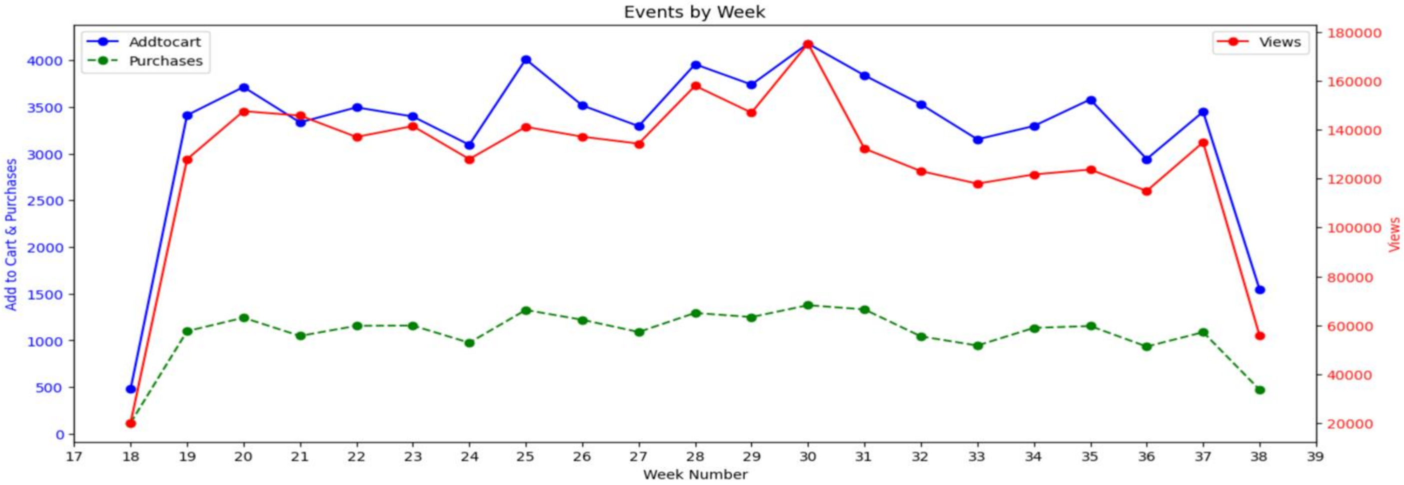
Are there any unusual traffic? How do we detect them? Can they affect the recommendations?

Regular Weekly Cycles

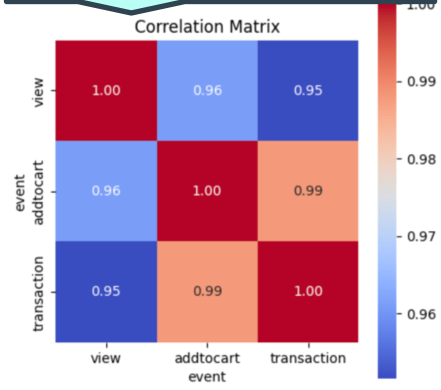


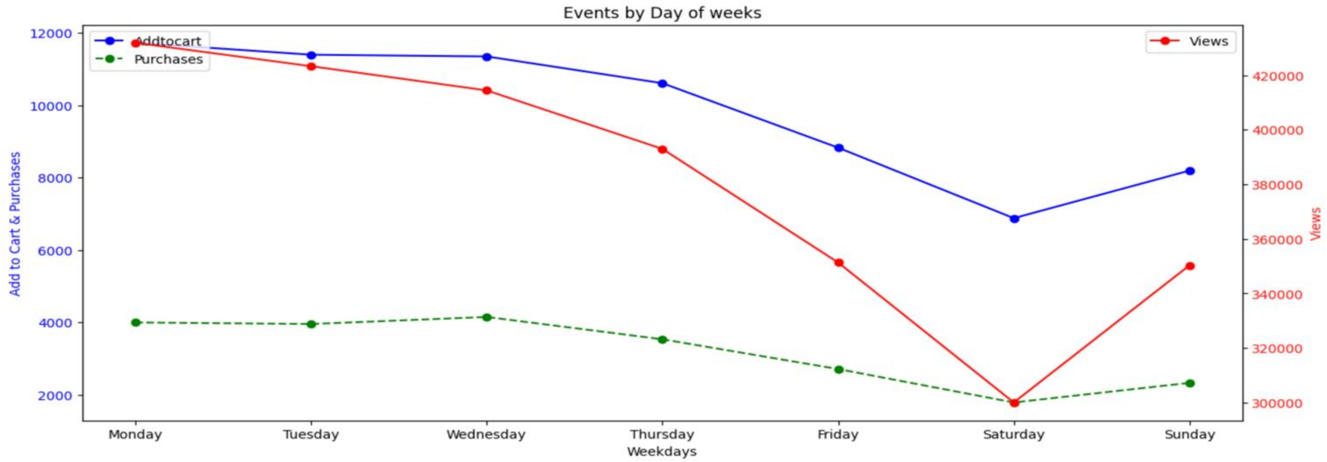
On daily level, there are fluctuations caused by random events such as holidays, promotions, weather etc.



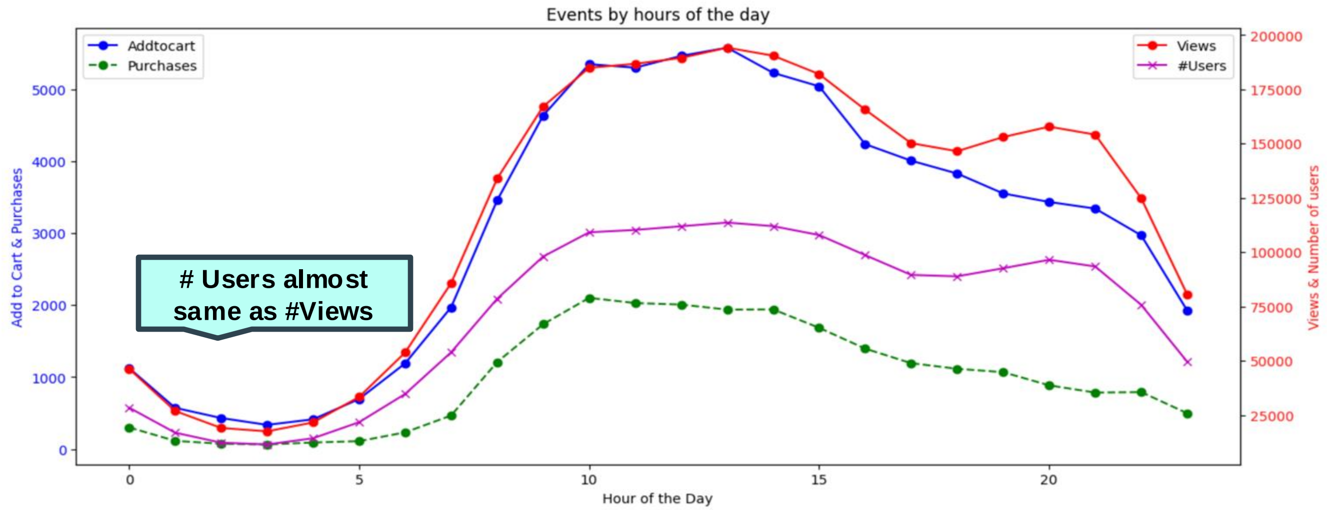


The overall trend looks more apparent. We have strong correlations.



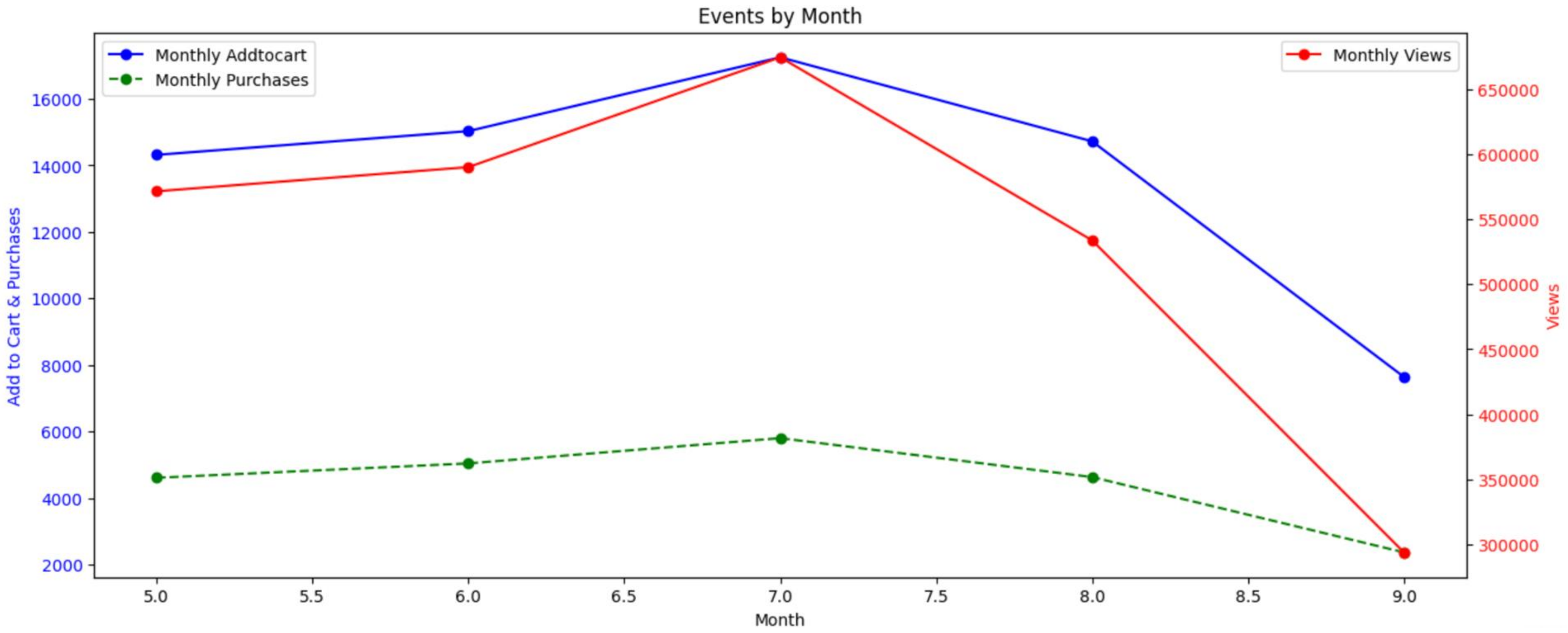


- **Explains the weekly cycles!**
- **This is a normal phenomena in e-commerce: lower engagement and sales on weekends!**
- **Many e-commerce businesses focus their marketing efforts on weekdays, often launching promotions and advertisements that pique interest during work hours.**



- **Most activities happen between 8 am to 8 pm.**
- **Not only the number of users increase during this peak period, activities per user also increase.**

- **July shows the highest engagement and sales.**
- **Mostly due to back-to-school sales**



Data Insights

How does the seasonal trend look? Can we explain those? Can we focus our marketing or promotional strategy to fit them?

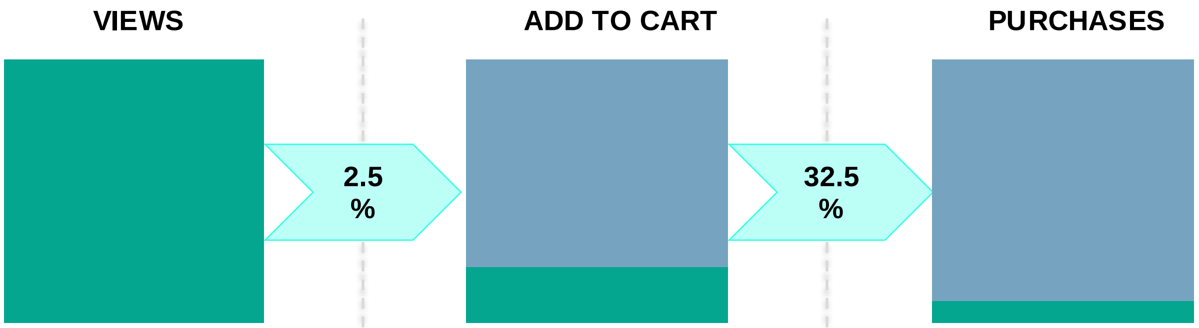
Bottlenecks

What are the dropout rates between each step to understand potential friction points in the buying process?

Which are the loyal customers? How often do they purchase in a specified timeframe? Which segment can be potential churners?

Which are the top selling items? What categories do they belong to? Which products are viewed but not purchased?

Are there any unusual traffic? How do we detect them? Can they affect the recommendations?



Add-to-Cart rate: 2.5%
Average: 7-8%

Cart abandonment Rate: 68%
Average: 70%

Data Insights

How does the seasonal trend look? Can we explain those? Can we focus our marketing or promotional strategy to fit them?

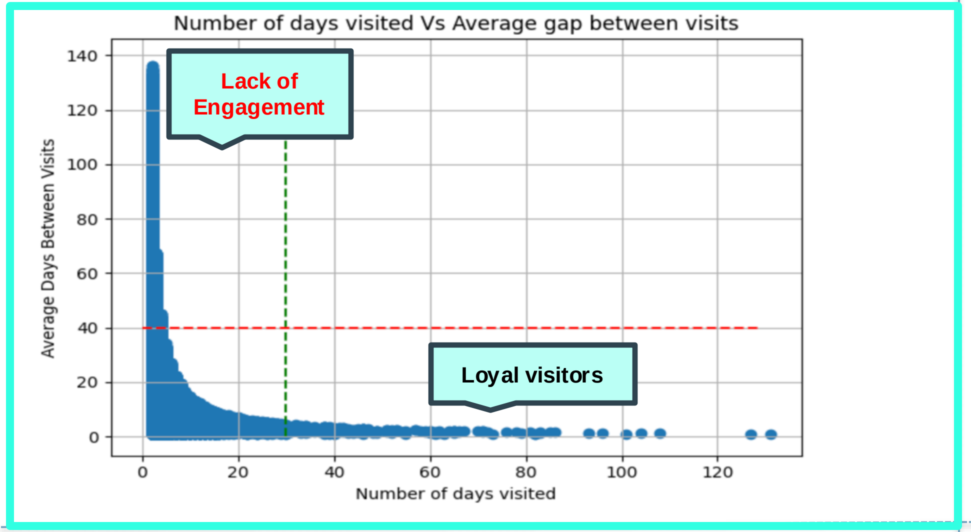
What are the dropout rates between each step to understand potential friction points in the buying process?

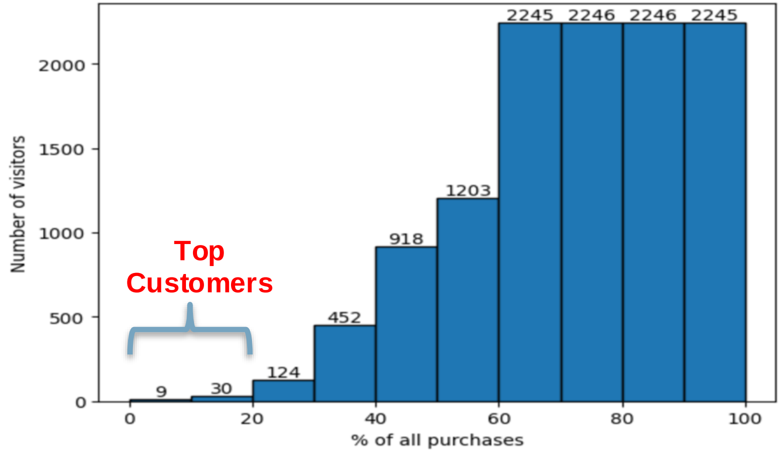
Customer Segmentation

Which are the loyal customers? How often do they purchase in a specified timeframe? Which segment can be potential churners?

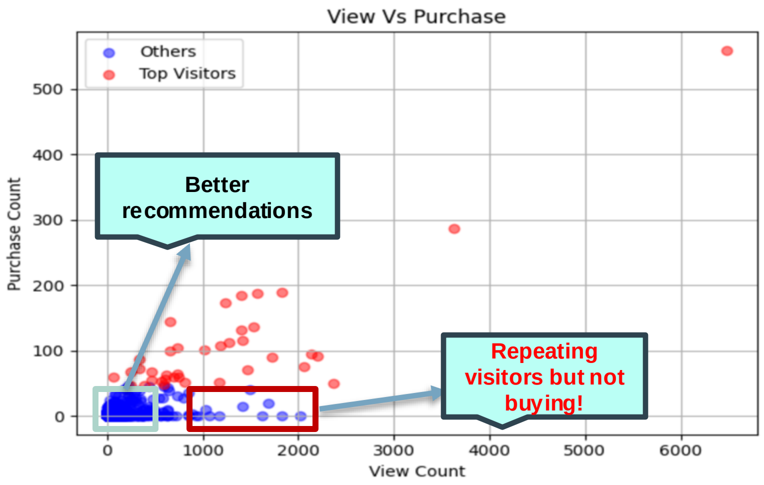
Which are the top selling items? What categories do they belong to? Which products are viewed but not purchased?

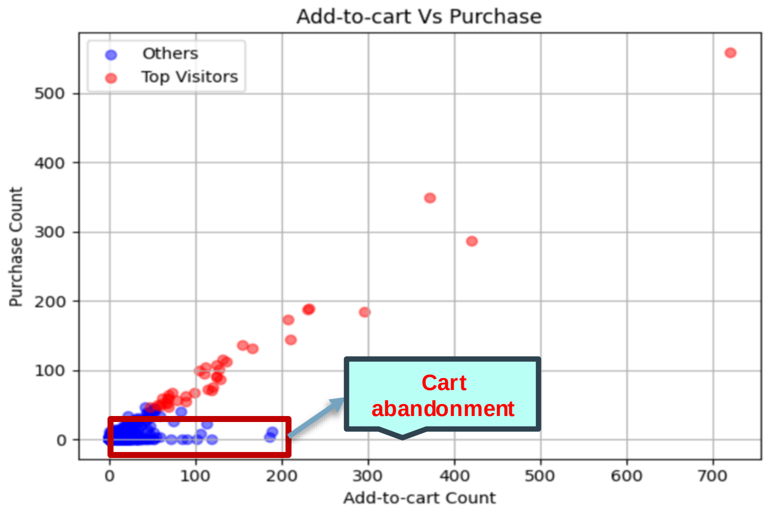
Are there any unusual traffic? How do we detect them? Can they affect the recommendations?

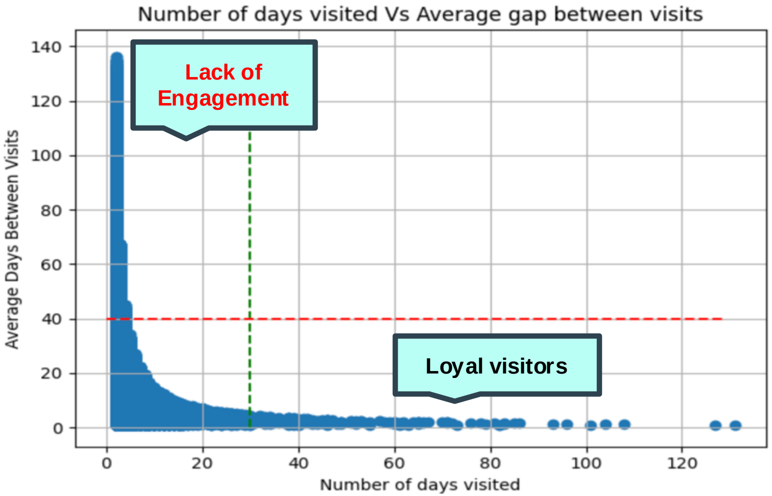


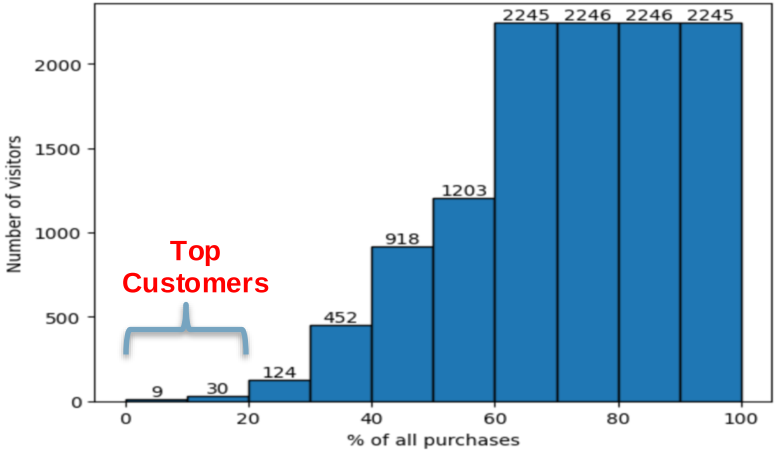


Only 39 users among 1.4M users are responsible for 20% of total sales!

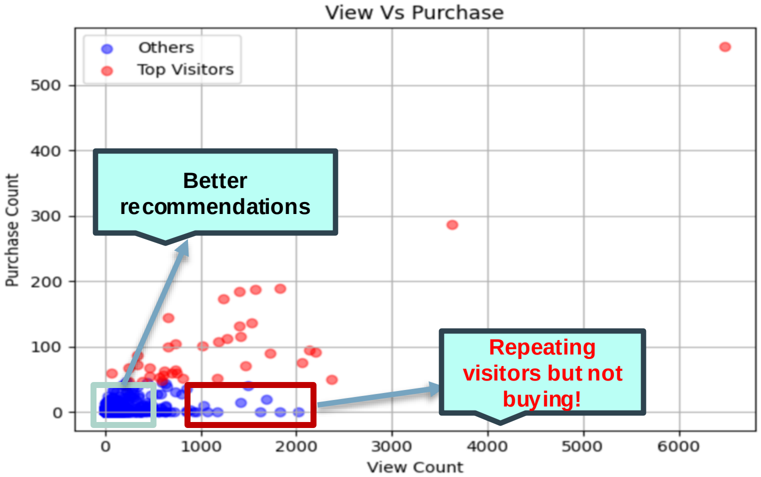


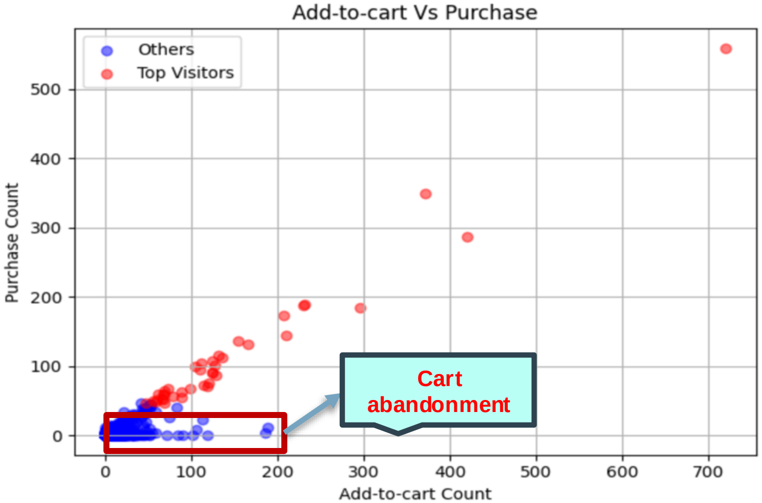


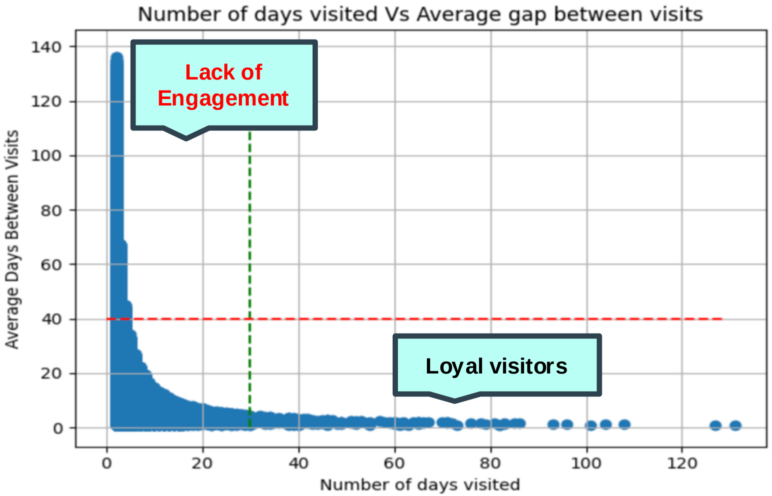


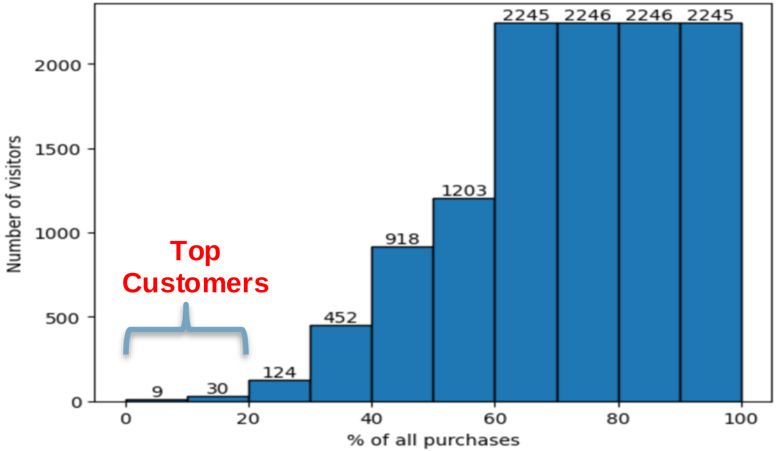


Only 39 users among 1.4M users are responsible for 20% of total sales!

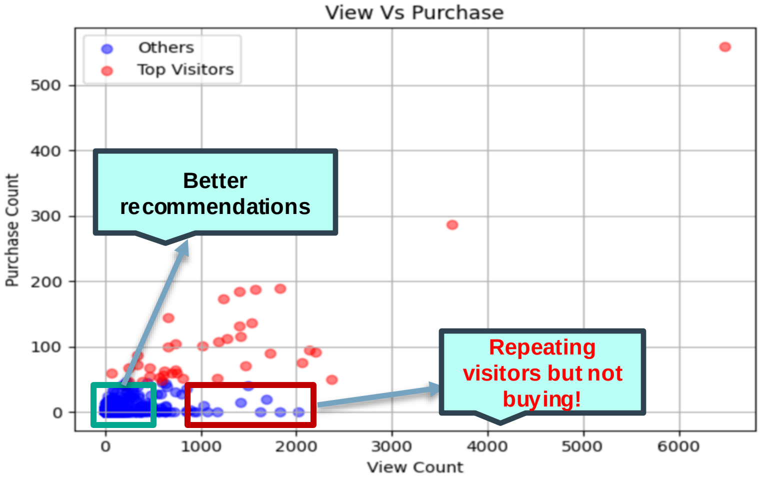


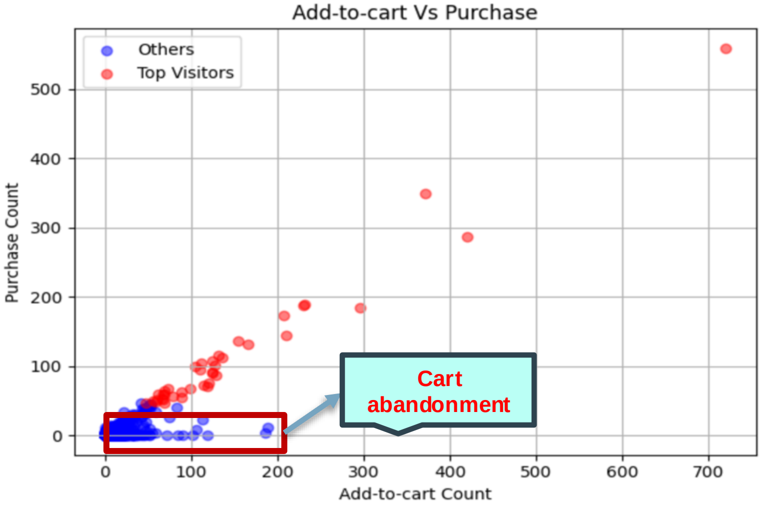






Only 39 users among 1.4M users are responsible for 20% of total sales!





Data Insights

How does the seasonal trend look? Can we explain those? Can we focus our marketing or promotional strategy to fit them?

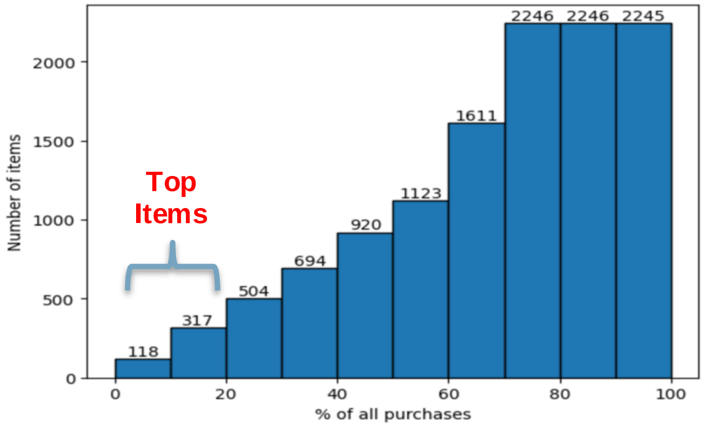
What are the dropout rates between each step to understand potential friction points in the buying process?

Which are the loyal customers? How often do they purchase in a specified timeframe? Which segment can be potential churners?

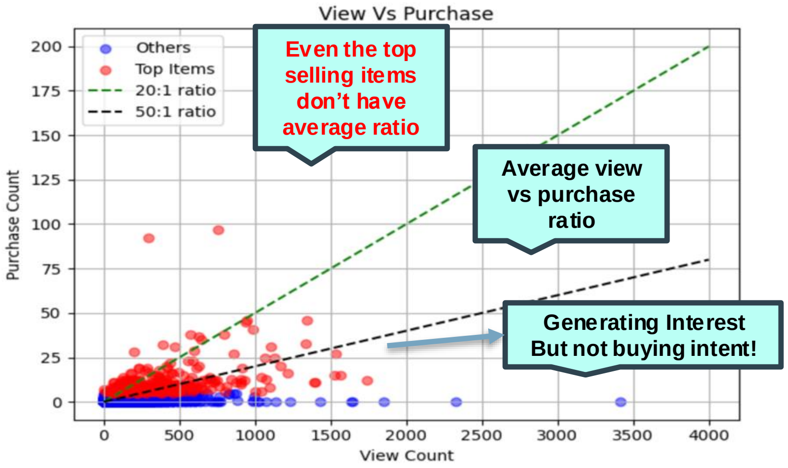
Item Segregation

Which are the top selling items? What categories do they belong to? Which products are viewed but not purchased?

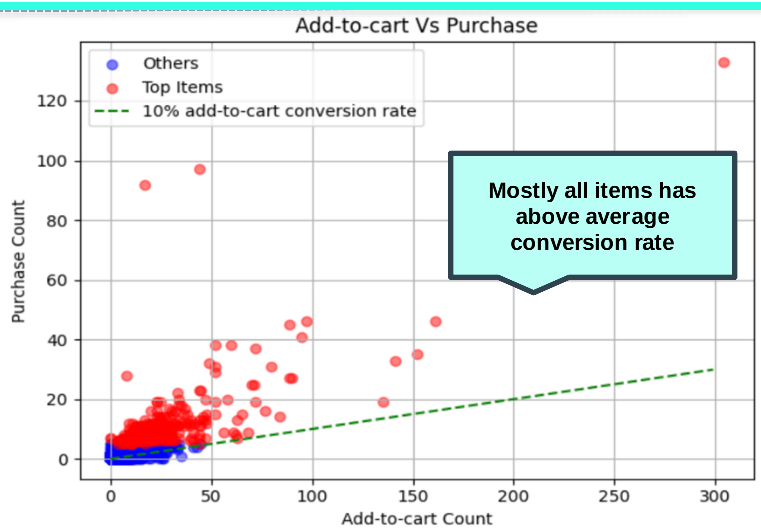
Are there any unusual traffic? How do we detect them? Can they affect the recommendations?

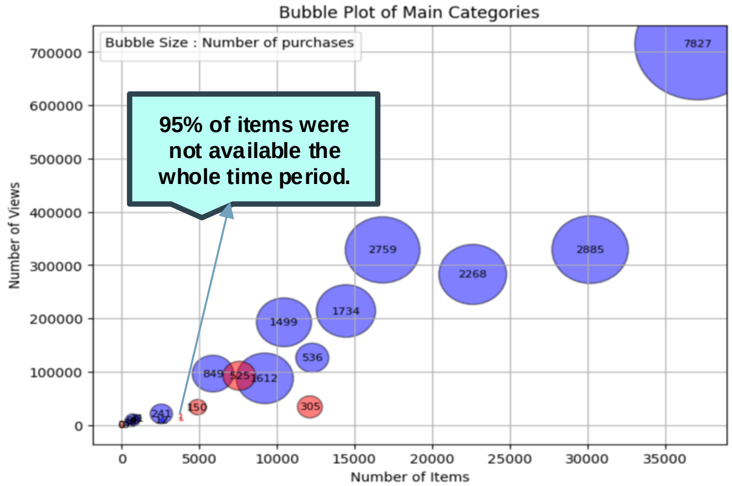


Only 435 items among 400K are responsible for 20% of total sales!

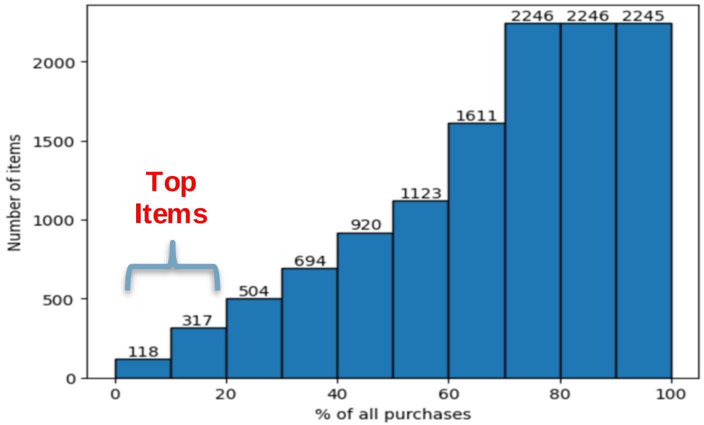


Reasons might include unclear product descriptions, poor quality image.

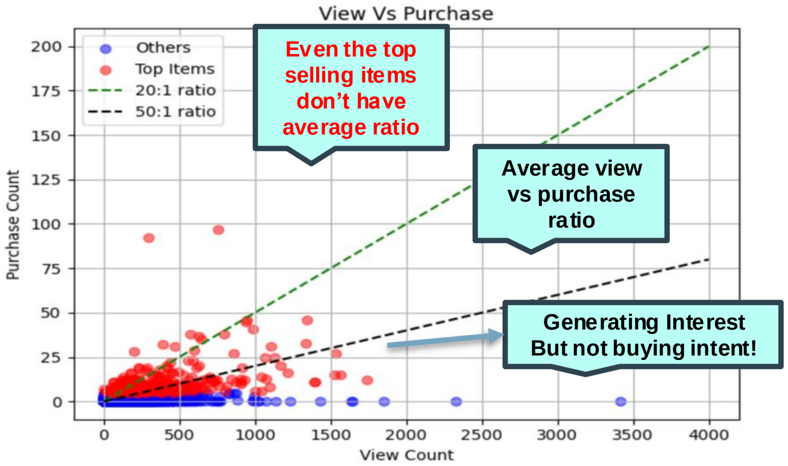




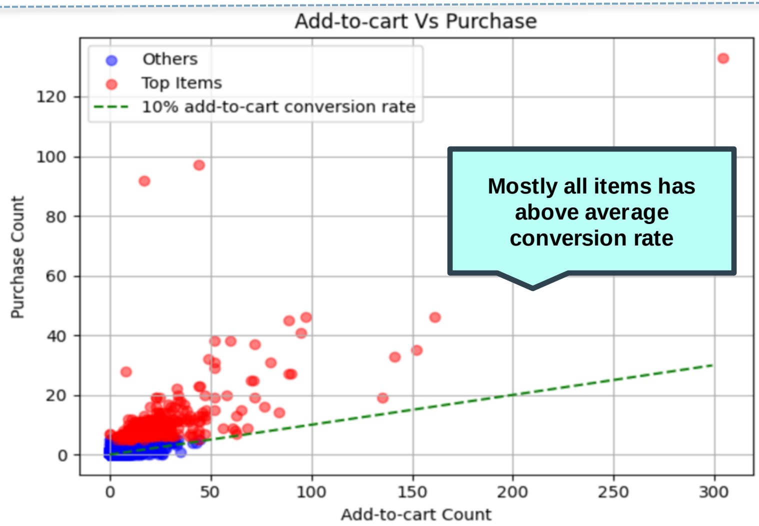
*Availability data was recorded on a weekly level, so this is an approximation

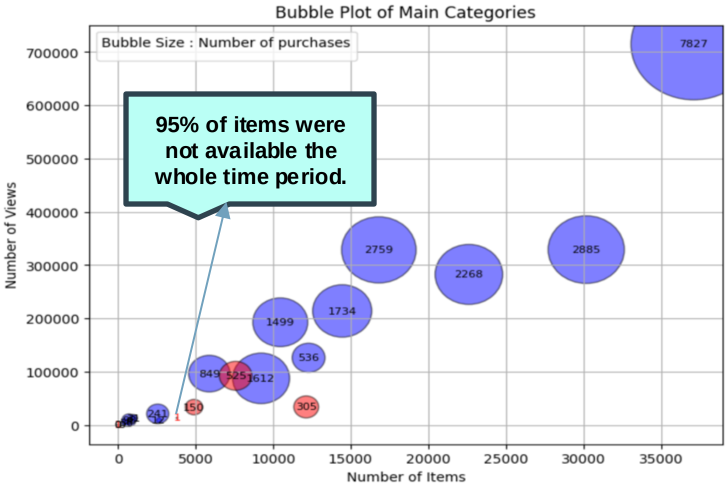


Only 435 items among 400K are responsible for 20% of total sales!

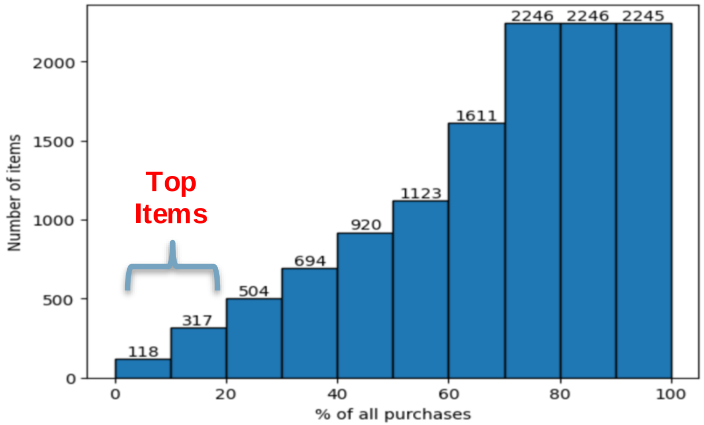


Reasons might include unclear product descriptions, poor quality image.

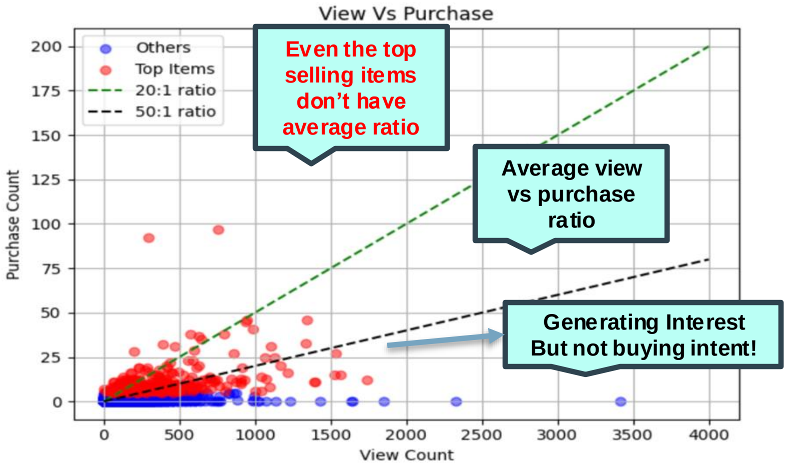




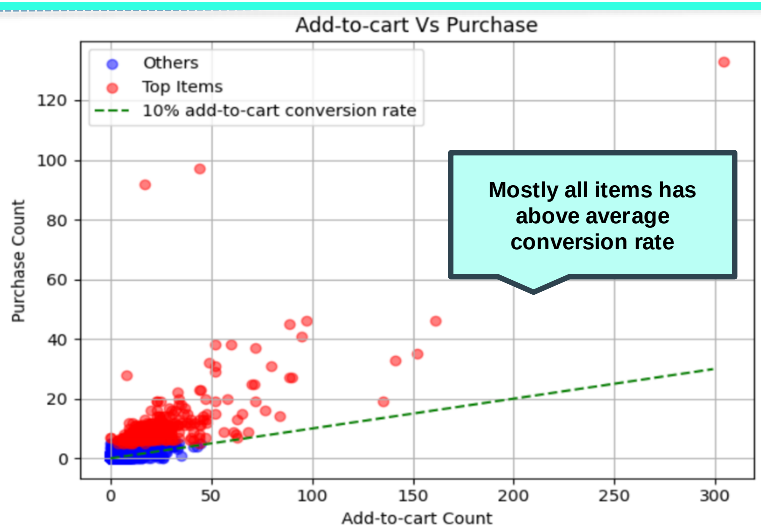
*Availability data was recorded on a weekly level, so this is an approximation

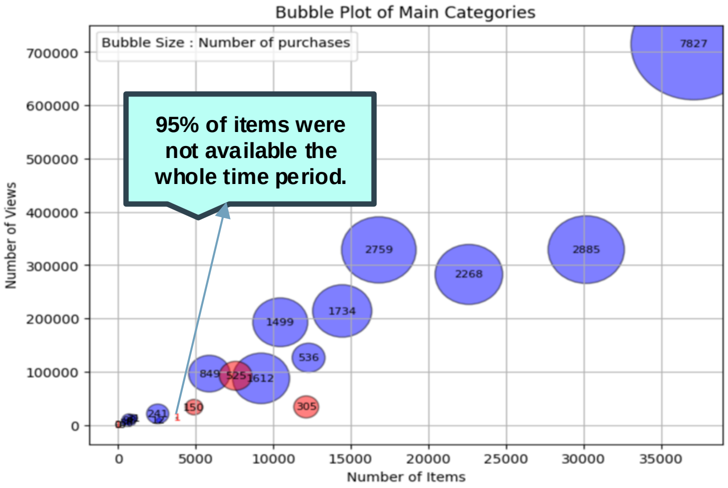


Only 435 items among 400K are responsible for 20% of total sales!

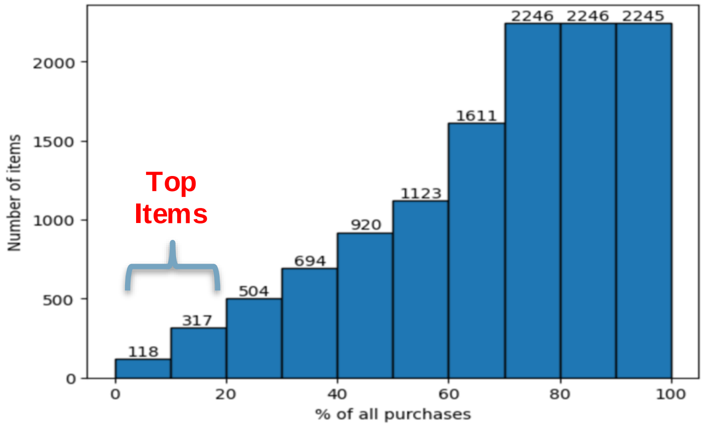


Reasons might include unclear product descriptions, poor quality image.

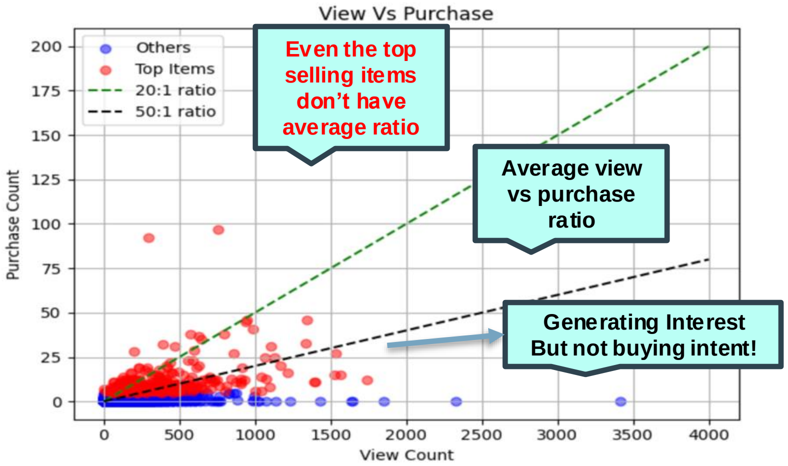




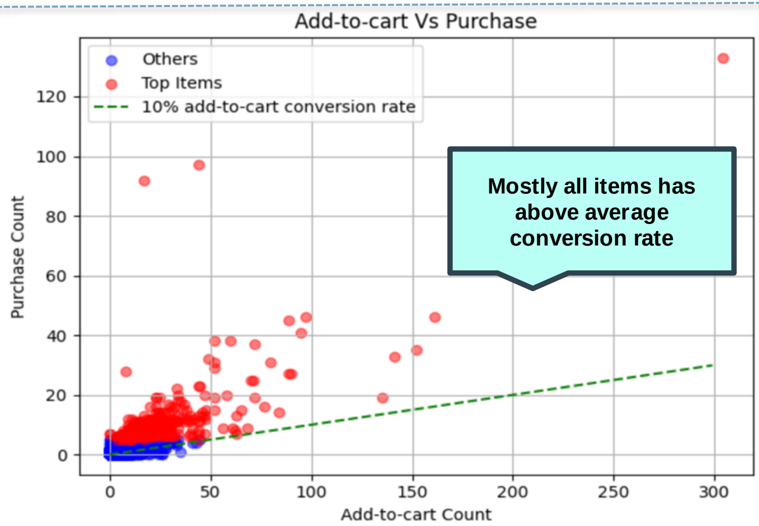
*Availability data was recorded on a weekly level, so this is an approximation

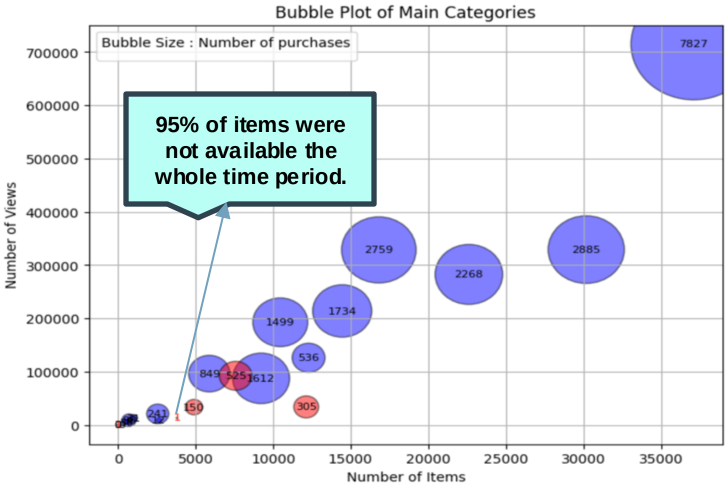


Only 435 items among 400K are responsible for 20% of total sales!



Reasons might include unclear product descriptions, poor quality image.





*Availability data was recorded on a weekly level, so this is an approximation

Data Insights

How does the seasonal trend look? Can we explain those? Can we focus our marketing or promotional strategy to fit them?

What are the dropout rates between each step to understand potential friction points in the buying process?

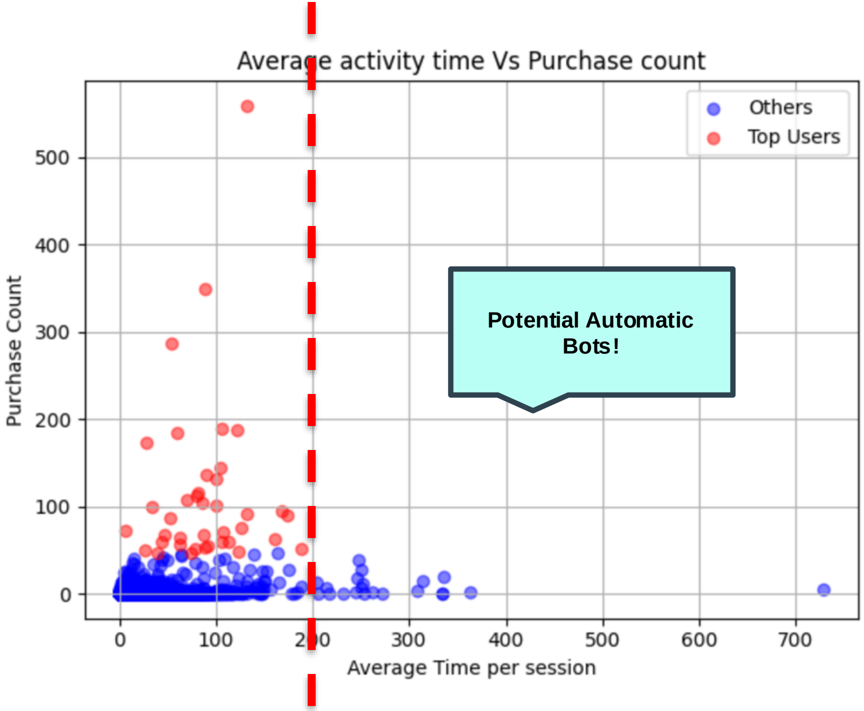
Which are the loyal customers? How often do they purchase in a specified timeframe? Which segment can be potential churners?

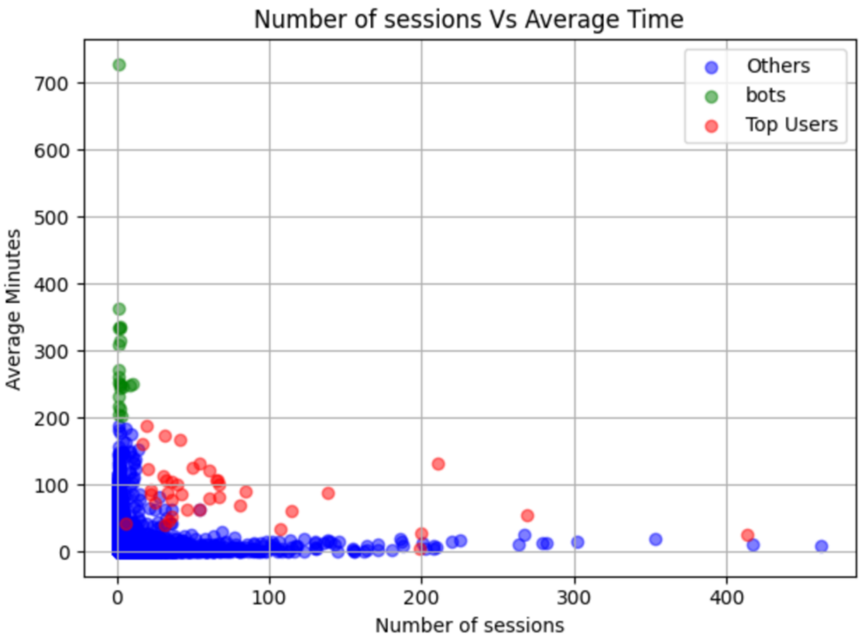
Which are the top selling items? What categories do they belong to? Which products are viewed but not purchased?

Abnormality

Are there any unusual traffic? How do we detect them? Can they affect the recommendations?

We only have view(clicks) timestamp, not visit timestamp. So, average activity time is higher in reality!





We should drop these users, as they are creating abnormal traffic which might affect the recommendation process!

Modeling

- As the original data was huge, we take a subset of the data.
- We sample 10000 items from all the items that were mostly available.
- We do a time splitting for train and validation data. The last 30 days data were kept as validation.

Modeling

- As the original data was huge, we take a subset of the data.
- We sample 10000 items from all the items that were mostly available.
- We do a time splitting for train and validation data. The last 30 days data were kept as validation.

Modeling Purpose: Recommending Items to Users based on their history.

Metrics: Recall@k = $\frac{1}{\#Users} \sum_{user=u} \frac{\#relevant\ items\ in\ top\ K\ \ of\ recommendations}{\#relevant\ items\ for\ user\ u}$

Relevant items: Items user have interacted with for the whole time period (both the training and validation data)

Modeling

- As the original data was huge, we take a subset of the data.
- We sample 10000 items from all the items that were mostly available.
- We do a time splitting for train and validation data. The last 30 days data were kept as validation.

Modeling Purpose: Recommending Items to Users based on their history.

Metrics: Recall@k = $\frac{1}{\#Users} \sum_{user=u} \frac{\#relevant\ items\ in\ top\ K\ \ of\ recommendations}{\#relevant\ items\ for\ user\ u}$

Relevant items: Items user have interacted with for the whole time period (both the training and validation data)

Strategy: Recommend the items that the user has recently interacted with. (Recently viewed items on Amazon)

	K=1	K=10	K=50
Recall@k	**0.52**	**0.67**	**0.69**

These are some very high values. Because in ecommerce, people take time before purchasing something, and they repeatedly interact with them before buying! But we don’t need a recommender system for that!

Modeling

We want to see how the recommender system performs in terms of recommending new items.
Because in e-commerce that's the goal!

Recall@k on unseen items = $\frac{1}{\#Users}\sum_{user=u}\frac{\#unseen\ relevant\ items\ in\ top\ K\ \ of\ recommendations}{\#unseen\ relevant\ items\ for\ user\ u}$

Modeling

We want to see how the recommender system performs in terms of recommending new items.
Because in e-commerce that's the goal!

Recall@k on unseen items = $\frac{1}{\#Users} \sum_{user=u} \frac{\#unseen\ relevant\ items\ in\ top\ K\ \ of\ recommendations}{\#unseen\ relevant\ items\ for\ user\ u}$

User-user CF: Recommends items those are viewed by similar users.
Item Item CF: Recommends similar items that user has interacted with.
SVD: Popular matrix factorization technique that estimate the interaction between user and item

Interaction Matrix Type	Collaborative Filtering	Recall@k – all relevant items			Recall@k – unseen relevant items		
How we quantify the interactions to calculate similarity?		**K=1**	**K=10**	**K=50**	**K=1**	**K=10**	**K=50**
Count of interaction between user and item	User - user	0.5	0.66	0.69	0	0.006	0.006
	Item - item	0.51	**0.68**	**0.7**	0.006	**0.042**	**0.048**
	SVD	0.07	0.14	0.23	0.007	0.017	0.037
Weighing type of interaction between user and item (purchase:3, add to cart: 2, view:1)	User - user	0.51	0.67	0.69	0.005	0.01	0.012
	Item - item	0.52	0.68	0.7	0.008	0.034	0.038
	SVD	0.08	0.17	0.29	0.006	0.014	0.031

Any questions?