Project: Purchases Prediction.
Brief: EDA to predict customer purchases.

1. Problem Statement

We need to understand users' purchasing patterns to predict what articles each customer will purchase in the 7-day period immediately after the training data ends.

"7-day target week": we look at the behavior in it to understand what in the past data (features) can explain the purchases.

Plan:

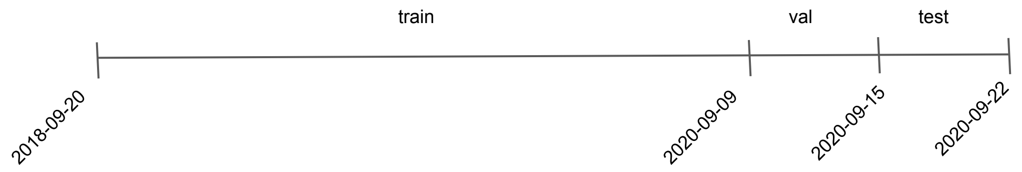
1. Data.
2. Distribution of users & purchases
3. Key Analysis Points.
4. Search for patterns and insights.
5. Conclusions. Next Steps.

2. Data

Data is provided by Kaggle competition (H&M Personalized Fashion recommendations)
https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data

Transactions.csv - data logs, consisting of the purchases each customer for each date.

Info: transactions
Size: 31.8m transactions, 1.37m unique customers, 5 columns.



BUT For Proof of Concept

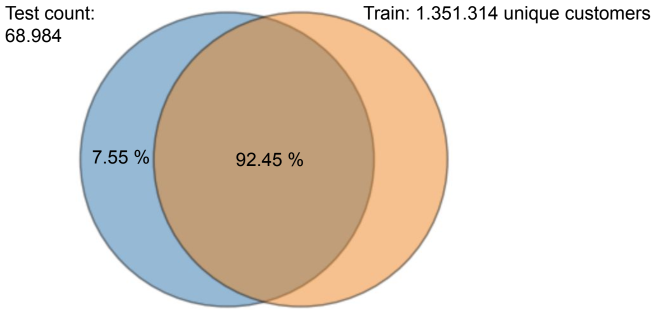
let's take a chunk from test_set and train_set, but keep the proportion of users who
- were found in train_set (users with history),
- were not found in train_set (new users).

```
Users with history of purchases: 92.45%, New users: 7.55%
Info: chunks
```

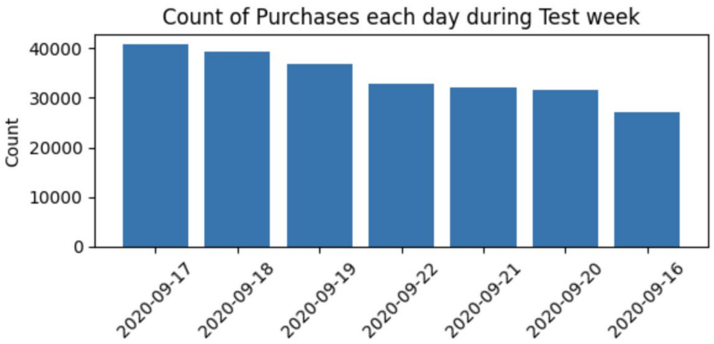
Train_chunk Size: 673k transactions, 8.2k unique customers
Test_chunk Size: 10k transactions, 8.9k unique customers

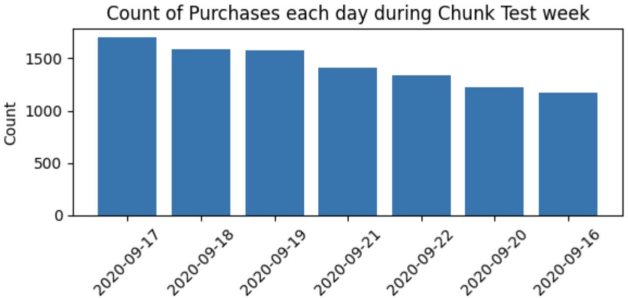
3. Data Overview Full Data

Distribution of Users Full & Chunk Data is the same.

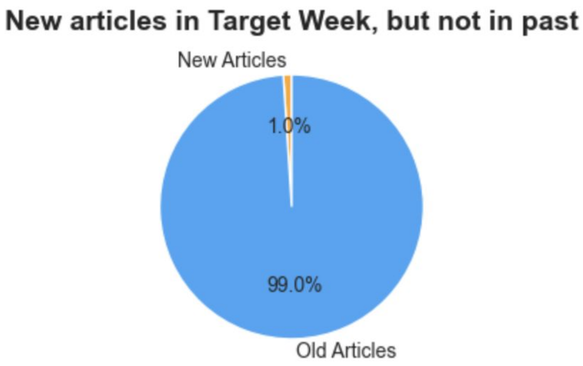


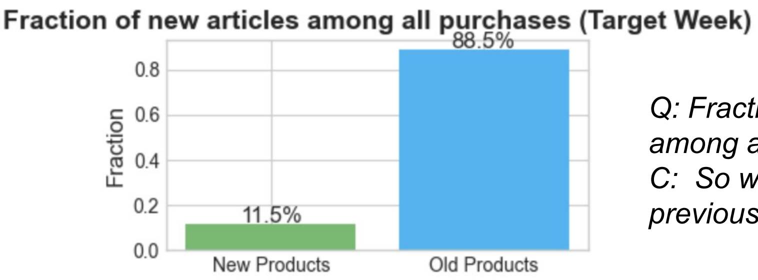
~ 7.55 % new customers during test week
(who doesn't have a purchases history)
~ 92.45% customers have purchases before.





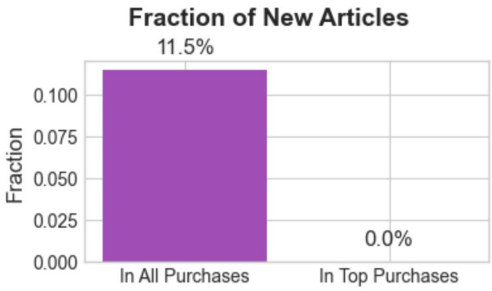
4. Key Analysis Points Test week





Q: Fraction of purchases of new articles among all purchases target week?
C: So we see that users mostly buy that previously purchased.

Q: how many new articles appeared in target week but not in past?
C: the new ones are only a small count.



Q: Fraction of new articles among target week's top articles?
C: But these articles didn't make top week's top articles.

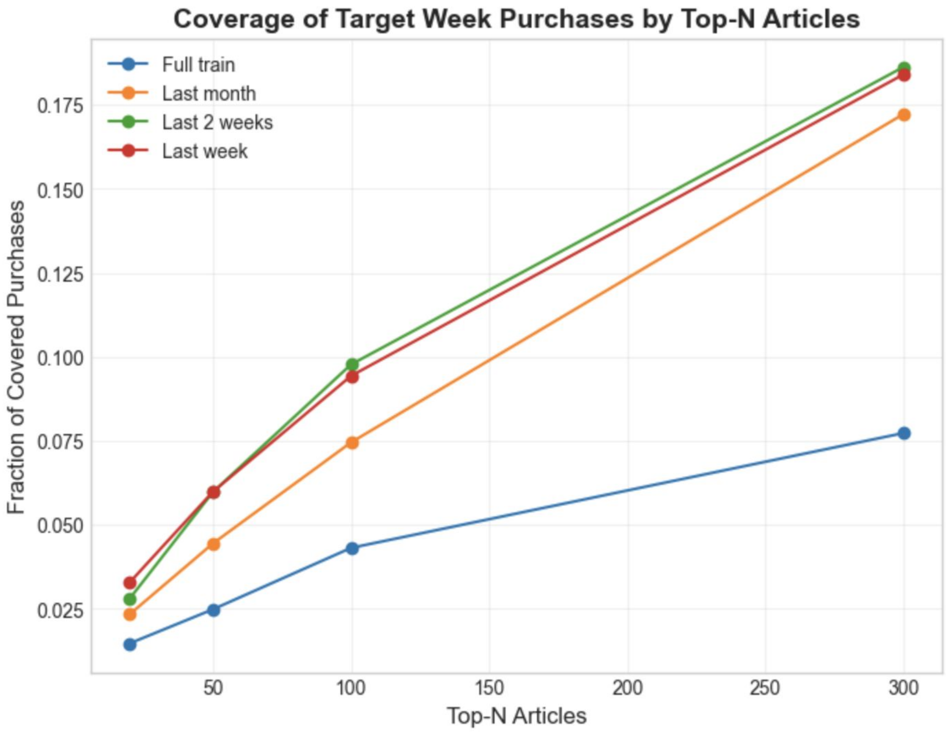
<u>User behaviour conclusion</u>: Despite the fact that new unique articles account for only 1%, they attract 11% of all purchases target week although not in top articles of this target week. We will handle this with a separate analysis, but for now let's focus only on transactions of articles that are both in history and in test week.

4. Key Analysis Points Test week

"What % of all transactions of target week will we cover with the top?"

Conclusions:

We tested different options, and it turned out that last-week top provides the best coverage.
Top for the whole train does not catch articles that were bought during the target week well.



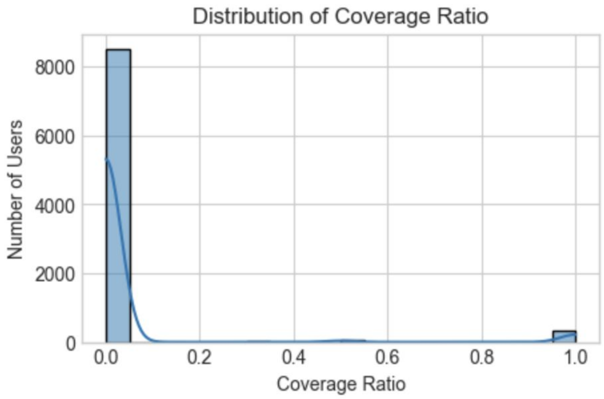
N = [20, 50, 100, 300]

5. Insights. Does personal history cover user purchases during Target Week?

Conclusions:

Only 4.6% users their history completely match the purchases.
But before we saw 85% coverage all purchases in target week.

So this gives us an understanding that most users don't buy the same articles repeatedly, and gives us an important signal to pay attention to overall top articles rather than personal purchase history.



	customer_id	article_id_test	article_id_history	coverage_ratio	overlap_ratio
0	00077dbd5c4a4991e092e63893ccf29294a9d5c46e8501...	[915529005]	[676387001, 685687003, 662980003, 529953001, 7...	0.0	0.0
1	0026ebdd70715d8fa2befa14dfed317a1ffe5451aba839...	[759465001]	[700761011, 680262001, 859105003, 714790017, 8...	0.0	0.0
2	003ca8034fe32b9bab8e1c03d74c972abd80dccf84a302...	[895376004]	[613246001, 579865003, 624645006, 623347003, 6...	0.0	0.0
3	00465ec96dd32dca19f85108cbce142de6667a7ace8208...	[862103001]	[688873001, 688873004, 641620001, 600274006, 6...	0.0	0.0
4	004c3751ed6f9dfc98b870291c95be6702d3afa97d9467...	[891763004]	[772988001, 490113011, 612481007, 726537001, 5...	0.0	0.0

6. Conclusions. What we got so far? Next Steps.

Based on analysis, personal history works only for a very small segment, and global tops of recent weeks do not repeat with tops of the test week. But globally, the articles have been encountered in history → we saw “85% of articles are repeated”, but individually each user buys almost new combinations for themselves.

It is necessary to look for global or contextual signals that will help predict purchases.
Possible global/contextual signals:

1. Global popular articles cover purchases test week. Check how many exactly? (10 / 4573 articles covered.)
2. Categories/subcategories
Users often buy certain categories, even if the articles_id themselves are different.
Like "user's favorite categories" → you can recommend products from these categories.
3. Related purchases (co-purchase).