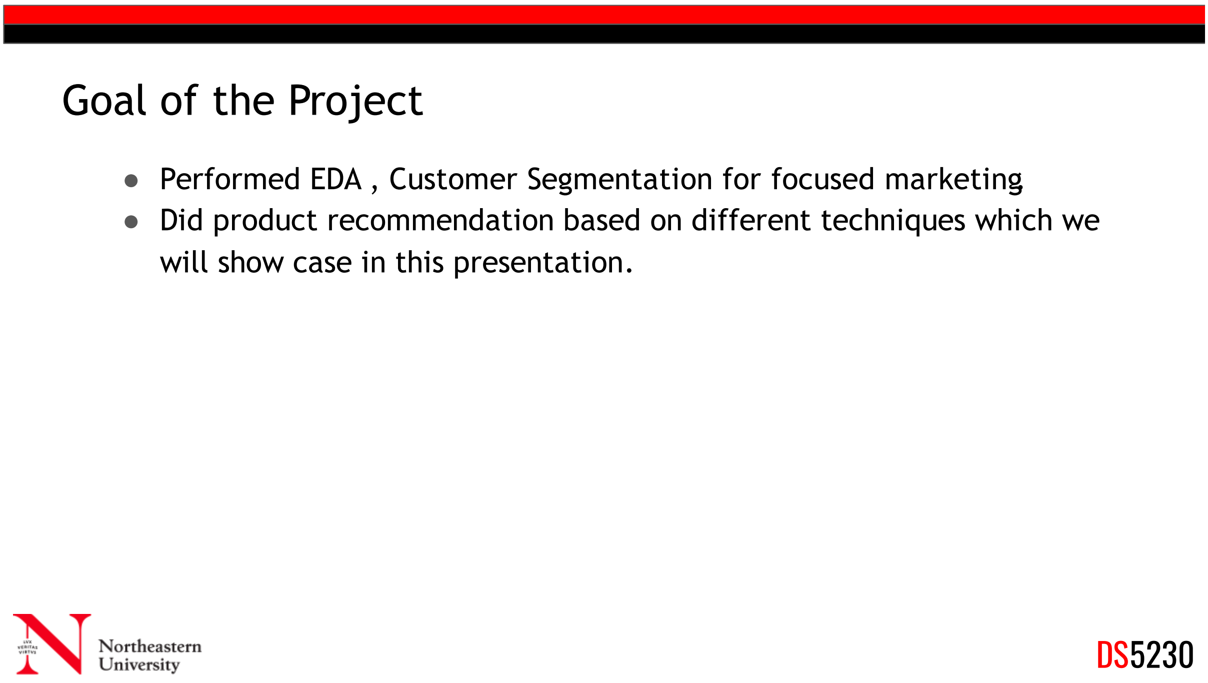
FINAL PROJECT REPORT DS 5230

H&M Product Recommendation System

By

Farhan Chughtai

Tannishtha Mandal



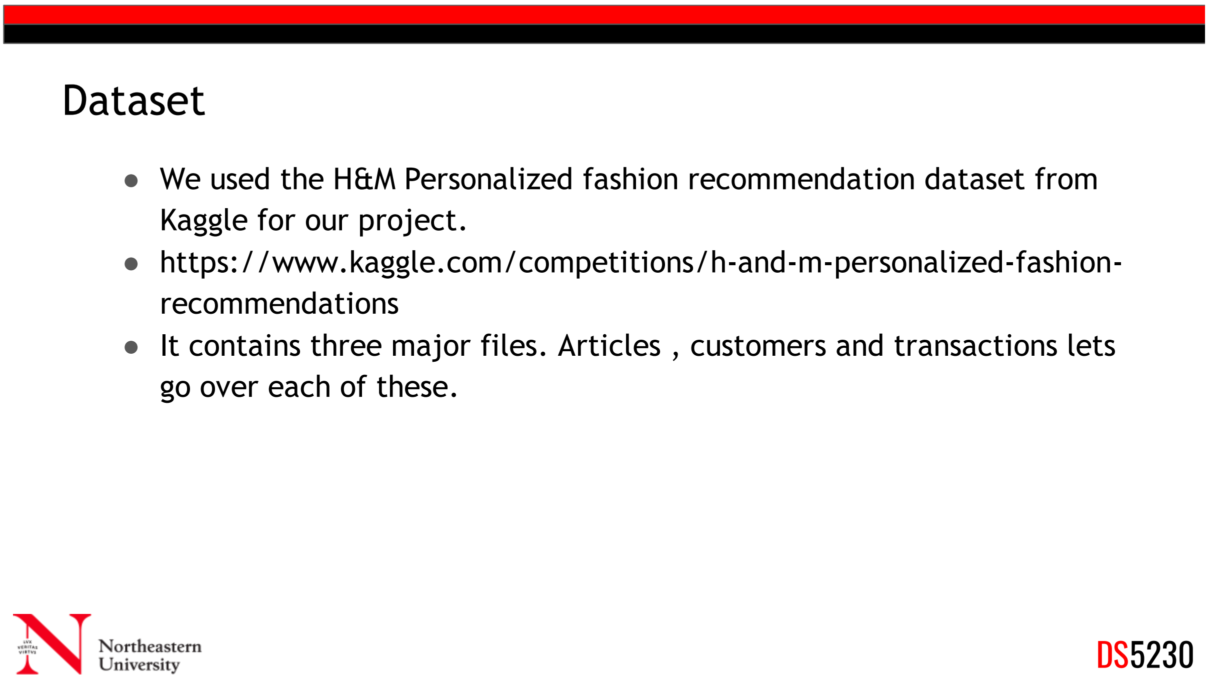
Goal of the Project

- Performed EDA , Customer Segmentation for focused marketing
- Did product recommendation based on different techniques which we will show case in this presentation.

Northeastern University

DS5230

The main goal of the project is to perform Exploratory data analysis and perform customer segmentation so that H&M can do focused marketing depending on the customer behavior. Furthermore, we are going to do product recommendation based on different techniques which we will show case in this presentation.



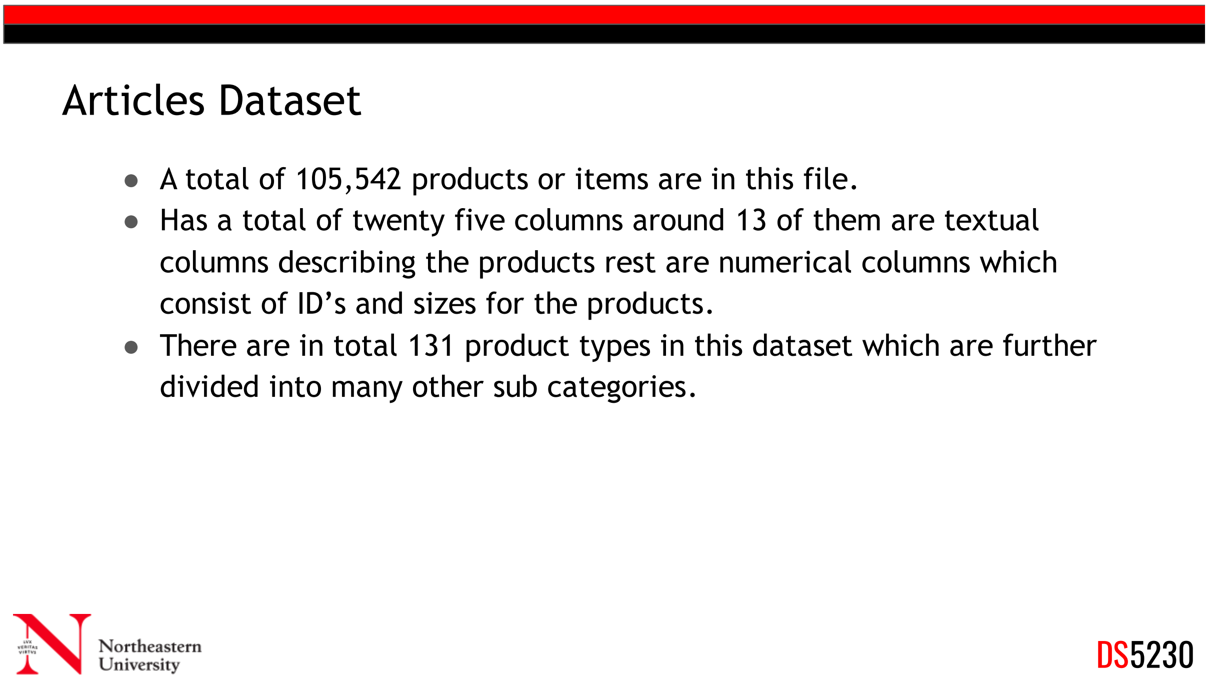
Dataset

- We used the H&M Personalized fashion recommendation dataset from Kaggle for our project.
- https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations
- It contains three major files. Articles , customers and transactions lets go over each of these.

Northeastern University

DS5230

We used the H&M personalized fashion recommendation dataset from Kaggle for our project which is available publicly online. It contains three major files the articles or the products csv which contains all the information about the items and the customer csv file that contains information regarding the H&M customer base and finally a transaction dataset consisting of all the transactions the customers did in the store online and instore.

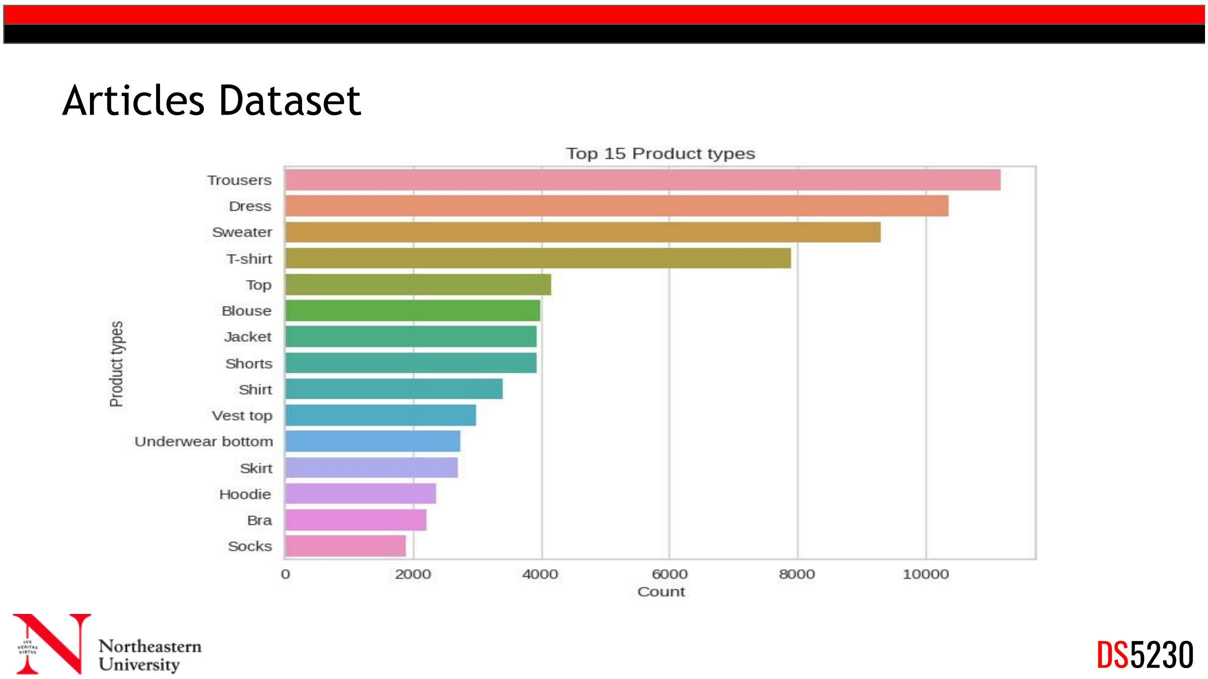


Articles Dataset

- A total of 105,542 products or items are in this file.
- Has a total of twenty five columns around 13 of them are textual columns describing the products rest are numerical columns which consist of ID’s and sizes for the products.
- There are in total 131 product types in this dataset which are further divided into many other sub categories.

DS5230

The articles dataset contains a total of 105542 products or items. It has a total of twenty-five columns around thirteen of them are textual columns describing the products rest are numerical columns which consist of article ID(unique for each product) and sizes of the product. There are in total 131 product types in that file.



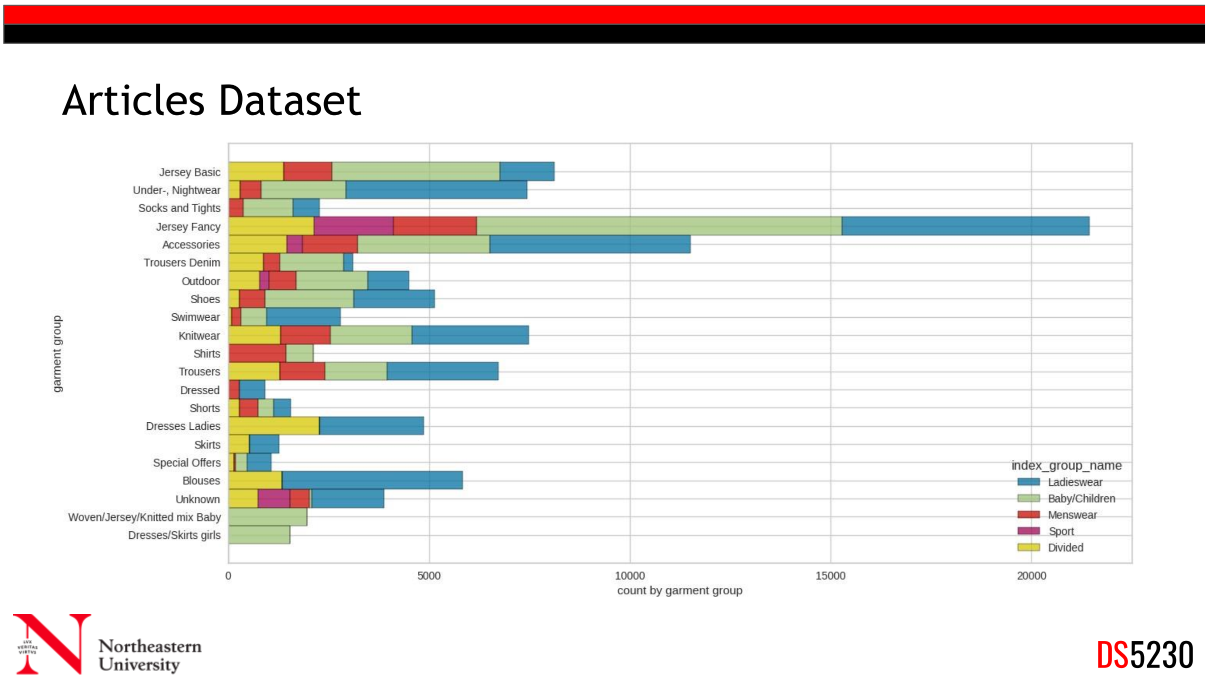
Articles Dataset

Northeastern University

DS5230

Insights of the given data analysis:

1. Identify Popular Product Categories: Which product types have the highest count can indicate the most popular categories among consumers. For instance, trousers and dress have a significantly higher count compared to categories like socks, hoodie, suggesting a potential higher demand for these items.
2. Fashion Trends and Preferences: A consistent increase or decrease in specific product types might reflect changing fashion trends or shifts in consumer preferences. Monitoring these trends can assist in making informed decisions regarding product development, marketing strategies, and stock replenishment.

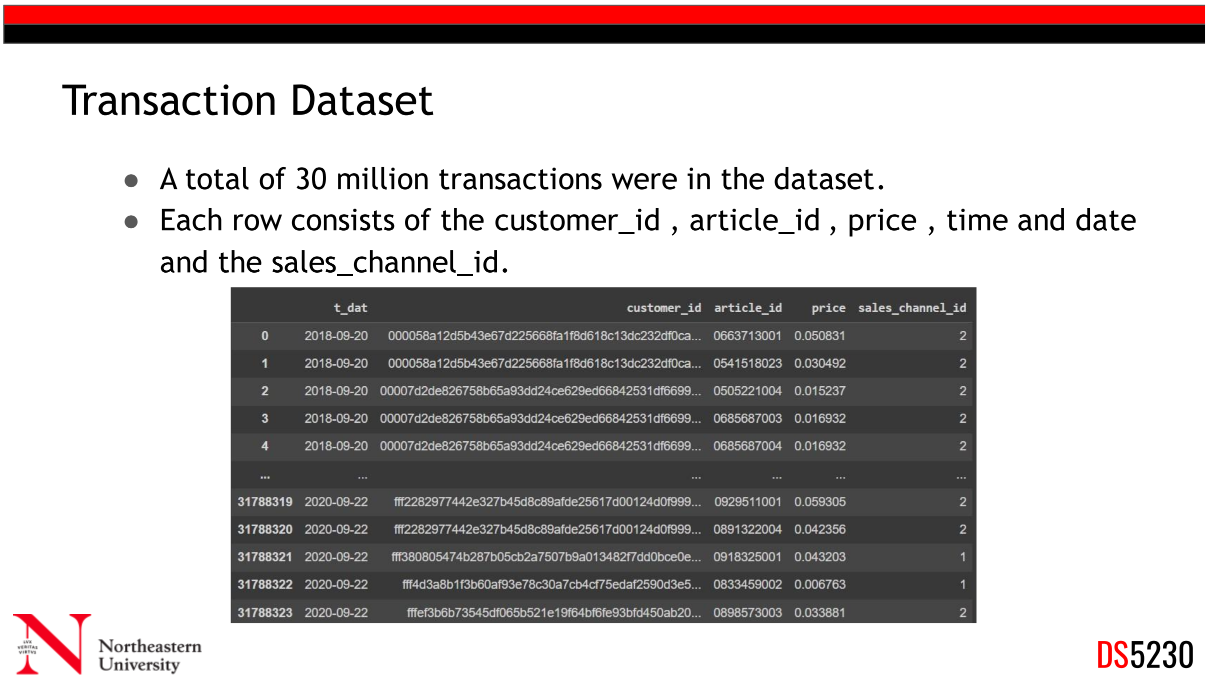


Articles Dataset

DS5230

Another visualization we can look at is garment groups broken up by index group names which are ladies wear, menswear, sports, baby / children and divided. This visualization gives an insight to:

1. which demograpgy to target to boost more sales.For example, for popular index group like female products ,norifications about new products, offers for free delivery cancan be offered to increase revenue. Similarly, less popular demography like menswear, sports offer like buy one get one, 50% off first purchase, etc can be used.
2. Partnerships and Collaborations: Identifying the most popular indices can guide businesses in forming partnerships or collaborations with brands or influencers within those specific segments, fostering strategic alliances that can drive sales and brand visibility.



Transaction Dataset

- A total of 30 million transactions were in the dataset.
- Each row consists of the customer_id , article_id , price , time and date and the sales_channel_id.

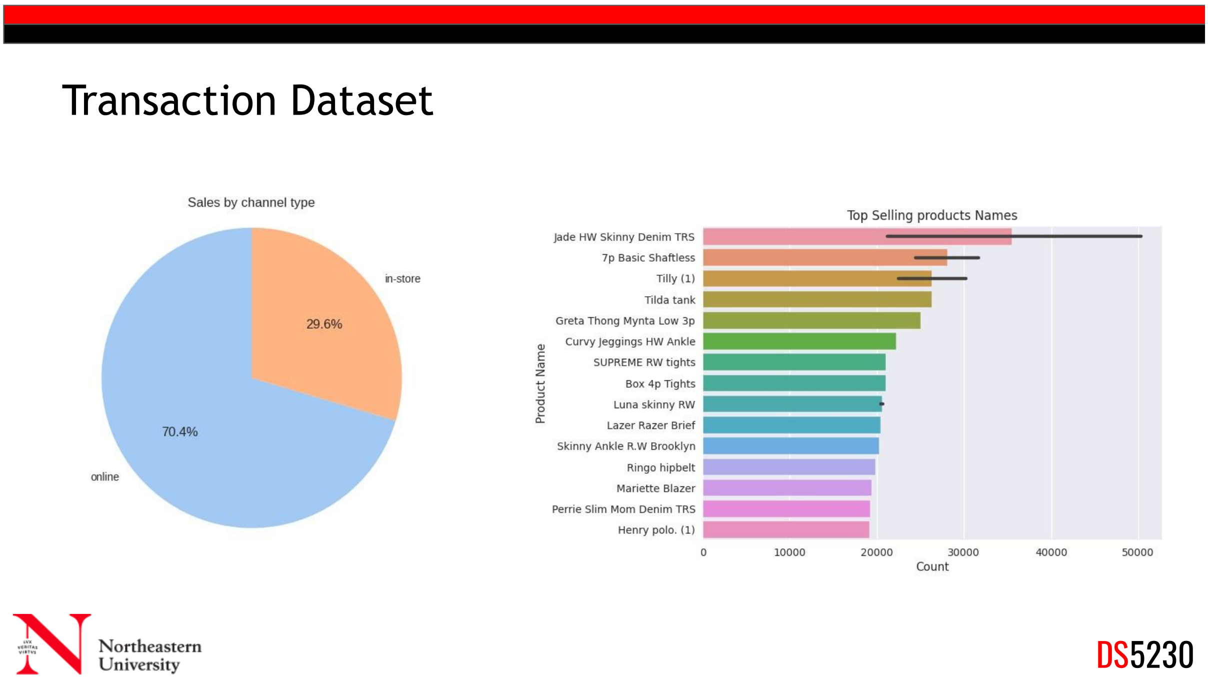
	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	0663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	0541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	0505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	0685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	0685687004	0.016932	2
...	...	...	...	...	...
31788319	2020-09-22	fff2282977442e327b45d8c89afde25617d00124d0f999...	0929511001	0.059305	2
31788320	2020-09-22	fff2282977442e327b45d8c89afde25617d00124d0f999...	0891322004	0.042356	2
31788321	2020-09-22	fff380805474b287b05cb2a7507b9a013482f7dd0bce0e...	0918325001	0.043203	1
31788322	2020-09-22	fff4d3a8b1f3b60af93e78c30a7cb4cf75edaf2590d3e5...	0833459002	0.006763	1
31788323	2020-09-22	fffef3b6b73545df065b521e19f64bf6fe93bfd450ab20...	0898573003	0.033881	2

Northeastern University

DS5230

The transaction data consists of 30 million records. Where each row tells us the customer id, product id, the price of the product and the date and time it was bought along with the sales channel ID.

The sales channel id has two values: 1 and 2, indicating online purchase and in-store purchase.



Transaction Dataset

Northeastern University

DS5230

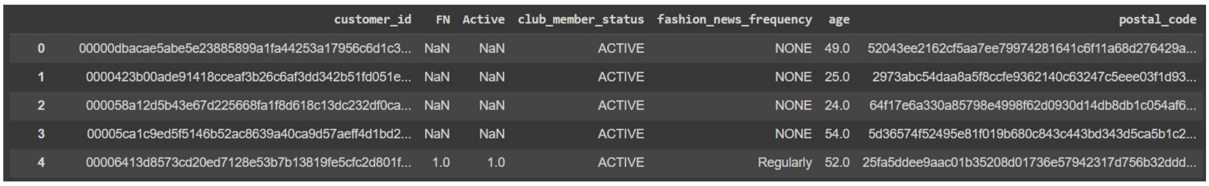
Most of the transactions are done online as you can see from the graph a total of 70 percent of the products were purchased online. While only 30 percent were bought in store. Moreover, we got the list of the top selling items from the transaction data set and the top items sold were the Jade HW skinny Denim TRS, 7p Basic Shaftless.

The EDA for transaction data above helps us to identify:

1. Popular channel: Online platform has earned significantly more revenue than in-store purchases, suggesting in investment of the online platform for more potential sales.
2. Inventory and Stock Management: Understanding the distribution of counts across indices assists in inventory management. It aids in ensuring adequate stock levels for high-demand segments while optimizing inventory for less popular categories, reducing carrying costs and stock wastage.

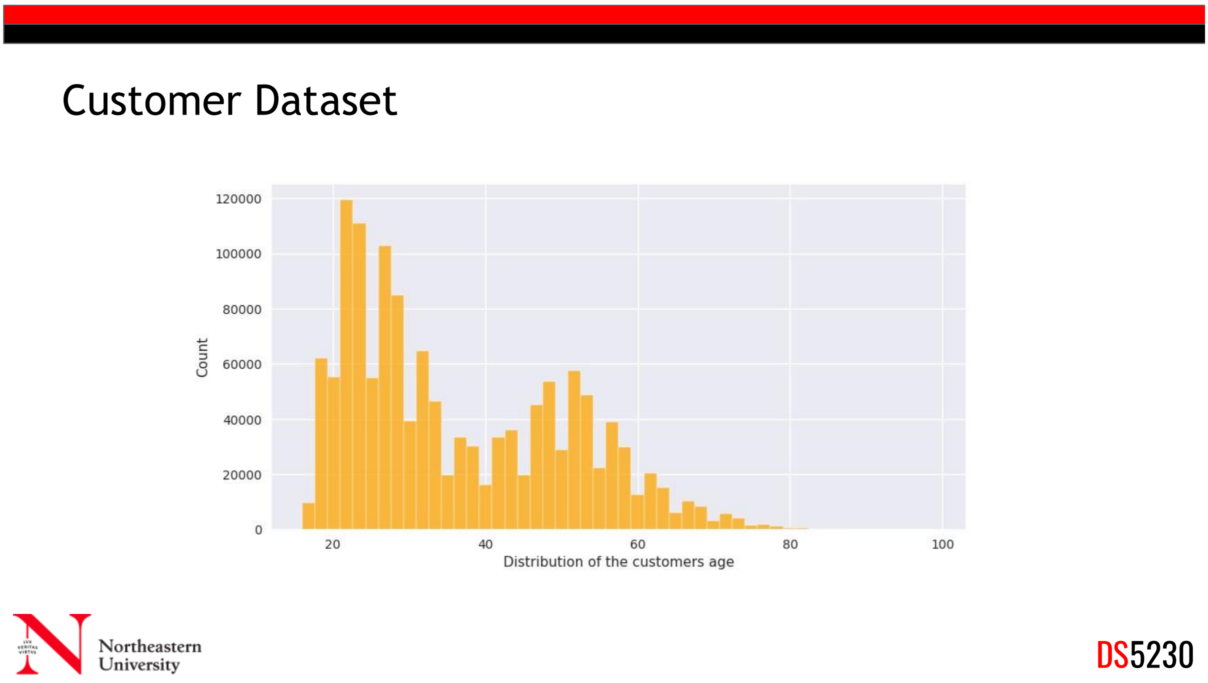
Customer Dataset

- This dataset has around 1.3 million customers.
- Each row has seven columns notably , unique id for each customer their subscription to the fashion news, age and there communication activity along with their postal code.



	customer_id	FN	Active	club_member_status	fashion_news_frequency	age	postal_code
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	NaN	NaN	ACTIVE	NONE	49.0	52043ee2162cf5aa7ee79974281641c6f11a68d276429a...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	NaN	NaN	ACTIVE	NONE	25.0	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	NaN	NaN	ACTIVE	NONE	24.0	64f17e6a330a85798e4998f62d0930d14db8db1c054af6...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	NaN	NaN	ACTIVE	NONE	54.0	5d36574f52495e81f019b680c843c443bd343d5ca5b1c2...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	1.0	1.0	ACTIVE	Regularly	52.0	25fa5ddee9aac01b35208d01736e57942317d756b32ddd...

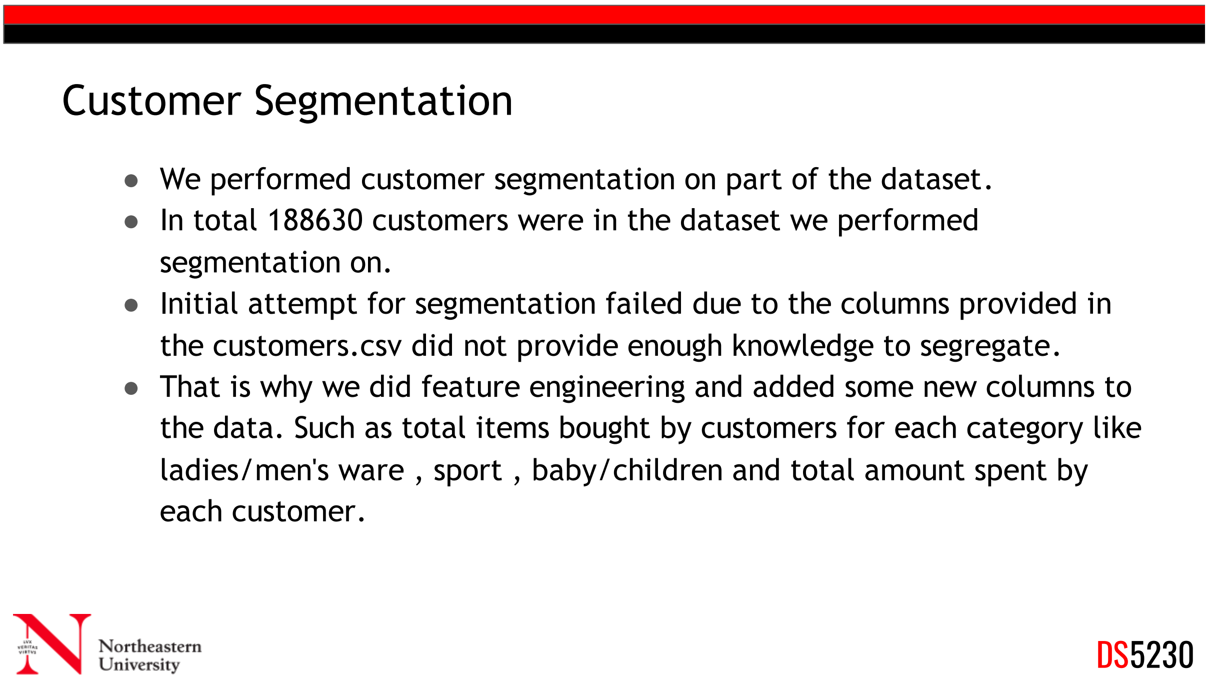
The customer dataset has around 1.3 million customers. It has their ID's their age and info about their activity, club status and fashion news frequency along with their postal code.



Customer Dataset

DS5230

This is the Customer Dataset distribution of age we can see that the most common ages are from 20 to 40 while very few customers are aged between 60 to 80.



Customer Segmentation

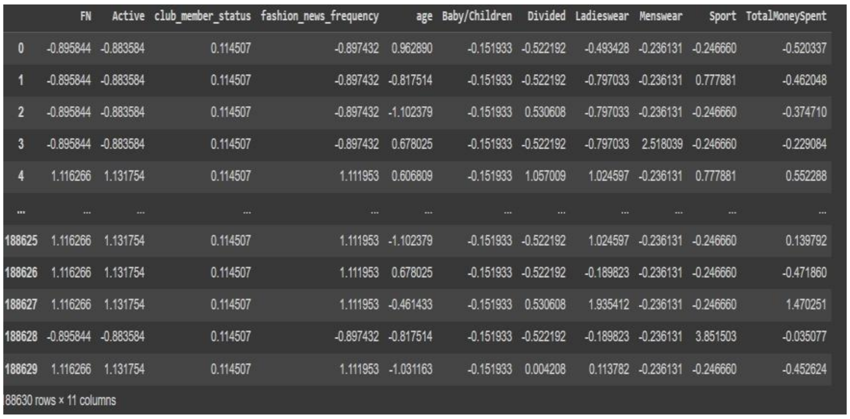
- We performed customer segmentation on part of the dataset.
- In total 188630 customers were in the dataset we performed segmentation on.
- Initial attempt for segmentation failed due to the columns provided in the customers.csv did not provide enough knowledge to segregate.
- That is why we did feature engineering and added some new columns to the data. Such as total items bought by customers for each category like ladies/men's ware , sport , baby/children and total amount spent by each customer.

Northeastern University

DS5230

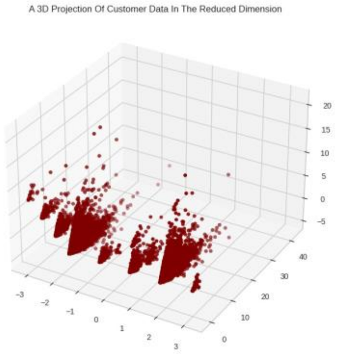
We wanted to perform customer segmentation on the customers.csv file we got from the dataset to give a more in-depth insight of the customers and how H and M should make marketing campaigns tailored to specific customers. Which customers they should focus on more and with what type of products they should push to the customers. For that we took a portion of the dataset for this segmentation in total we took 188630 customers from the data to perform segmentation on. Our initial attempt to segment them failed because the columns in the customers.csv dataset did not provide much knowledge to segregate them that is when we decided to leverage the transactional data set and feature engineer to add some new columns for each customer. This is how we added the total money spent column for each customer and the No of products they bought in each category like ladies' ware, men's ware, sport and baby/children. Addition of these new columns helped us a lot to segregate the customer data into clusters.

Customer Segmentation



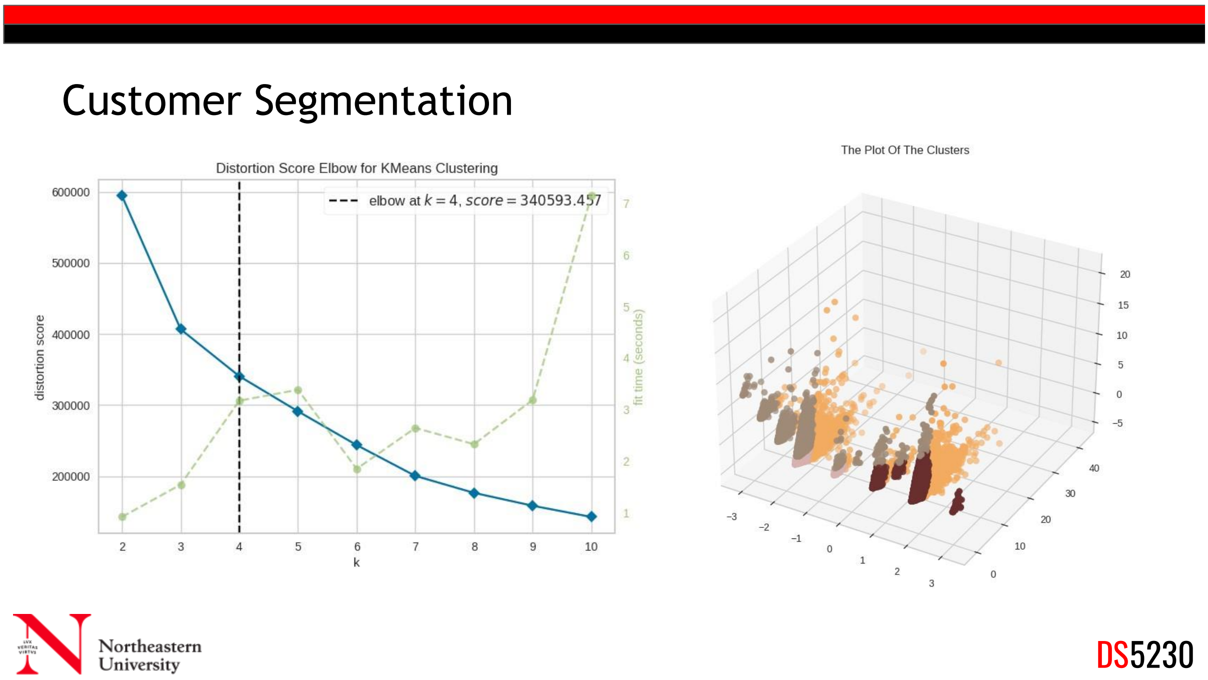
	FN	Active	club_member_status	fashion_news_frequency	age	Baby/Children	Divided	Ladieswear	Menswear	Sport	TotalMoneySpent
0	-0.895844	-0.883584	0.114507	-0.897432	0.962890	-0.151933	-0.522192	-0.493428	-0.236131	-0.246660	-0.520337
1	-0.895844	-0.883584	0.114507	-0.897432	-0.817514	-0.151933	-0.522192	-0.797033	-0.236131	0.777881	-0.462048
2	-0.895844	-0.883584	0.114507	-0.897432	-1.102379	-0.151933	0.530608	-0.797033	-0.236131	-0.246660	-0.374710
3	-0.895844	-0.883584	0.114507	-0.897432	0.678025	-0.151933	-0.522192	-0.797033	2.518039	-0.246660	-0.229084
4	1.116266	1.131754	0.114507	1.111953	0.606809	-0.151933	1.057009	1.024597	-0.236131	0.777881	0.552288
...	...	...	...	...	...	...	...	...	...	...	...
188625	1.116266	1.131754	0.114507	1.111953	-1.102379	-0.151933	-0.522192	1.024597	-0.236131	-0.246660	0.139792
188626	1.116266	1.131754	0.114507	1.111953	0.678025	-0.151933	-0.522192	-0.189823	-0.236131	-0.246660	-0.471860
188627	1.116266	1.131754	0.114507	1.111953	-0.461433	-0.151933	0.530608	1.935412	-0.236131	-0.246660	1.470251
188628	-0.895844	-0.883584	0.114507	-0.897432	-0.817514	-0.151933	-0.522192	-0.189823	-0.236131	3.851503	-0.035077
188629	1.116266	1.131754	0.114507	1.111953	-1.031163	-0.151933	0.004208	0.113782	-0.236131	-0.246660	-0.452624

88630 rows × 11 columns



DS5230

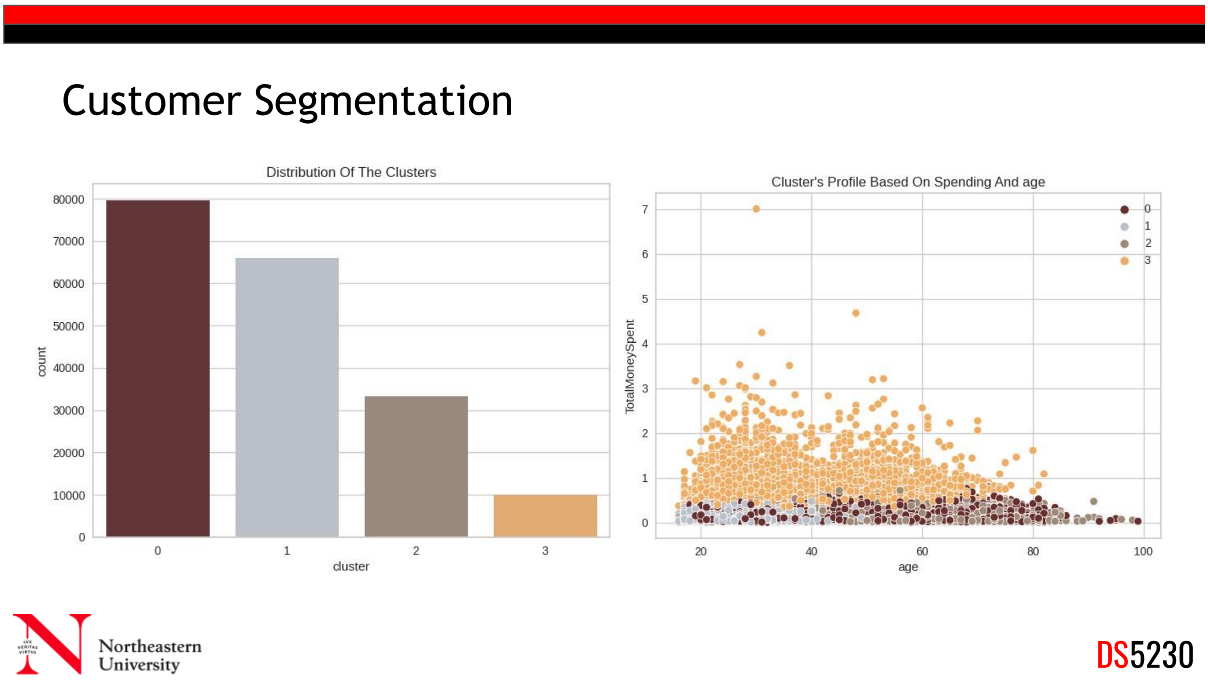
This is the customer dataset after we added some more features using feature engineering and applied a Standard Scaler to the dataset. In total we had 11 columns. So, we decided to perform dimeson reduction using PCA or T-SNE to reduce the dimension down to three columns. After that we plotted the data to see its structure. We also performed T-SNE but for our dataset it was taking a lot of time and resources to complete and we could not do it in the allotted time so we decided to go with PCA instead.



Customer Segmentation

DS5230

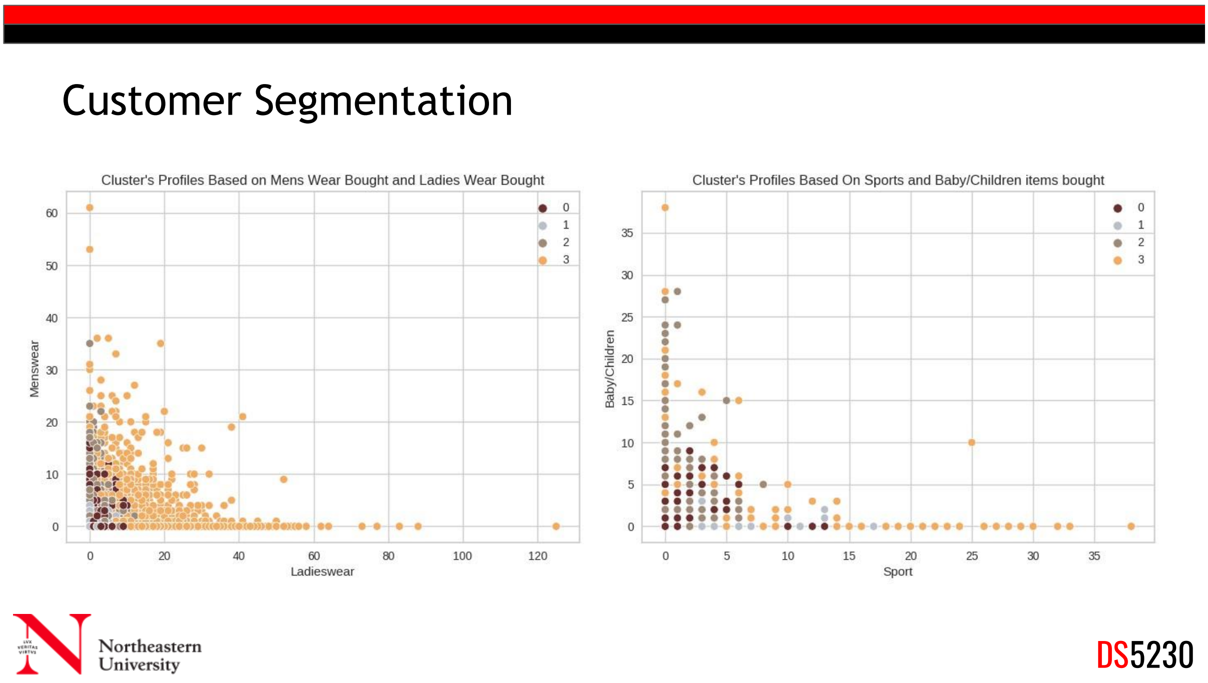
After we performed dimension reduction using PCA we used the knee elbow technique to find the optimal number of clusters in the dataset. And according to the plot we found that the elbow was at k = 4 so we did k-means clustering using four clusters and the this is the end result we got in 3d space. In addition to k-means clustering we also tried hierarchal clustering and spectral clustering but both of them failed as hierarchal cluster does not really suite our dataset and spectral clustering was again taking too much resources and time to complete.



Customer Segmentation

DS5230

Looking at the distribution cluster 0 is the most dominant one and has the greatest number of customers in it while cluster 3 has the least number of customers. Next, we plotted the clusters with respect to the age and total money spent by each customer and we started to see a pattern. Cluster 3 is the one with the most buying power customers reside that is customer in cluster 3 spend the most and are in the age range of 20 to 60 years old. While clusters 2 and 3 are made of customers that are older but do not spend as much as cluster 3 and cluster 1 is made up of customers with young ages.



Customer Segmentation

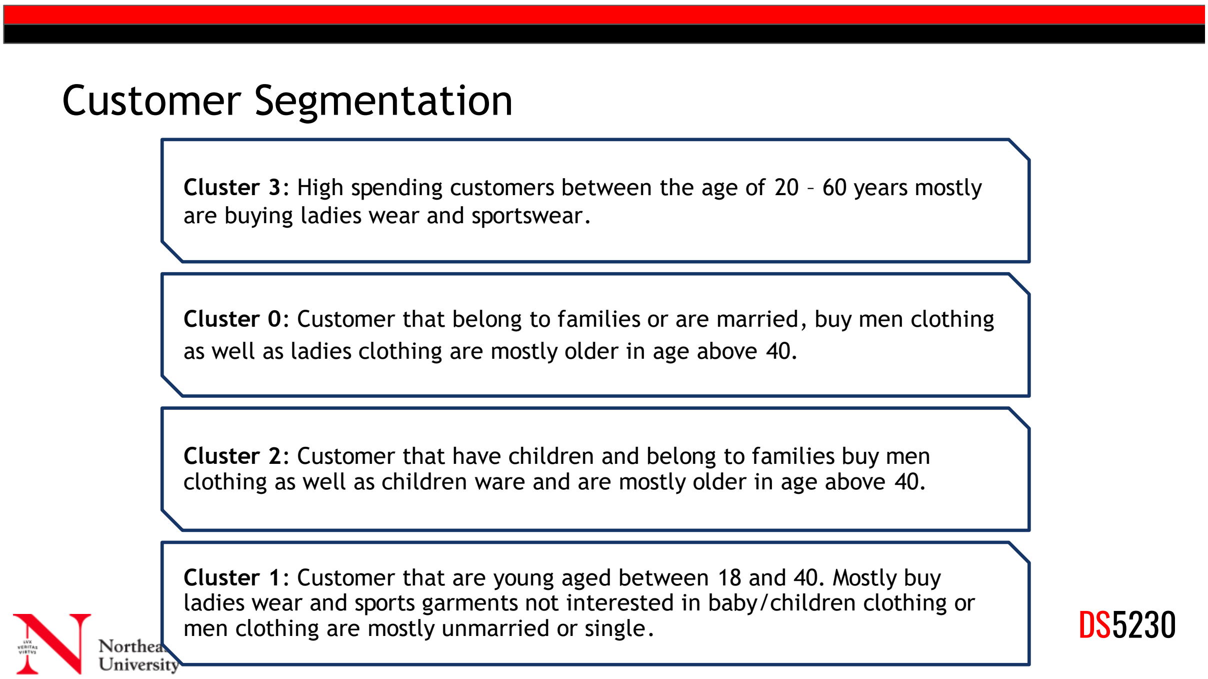
Cluster's Profiles Based on Mens Wear Bought and Ladies Wear Bought

Cluster's Profiles Based On Sports and Baby/Children items bought

Northeastern University

DS5230

Looking at some more graphs. We can see that cluster 3 customer buy mostly ladies wear garments and cluster 0,2 are mostly customer that buy both men's ware and ladieswear while the cluster 1 mainly buys ladies wear so can deduct from this that cluster 0, 2 belong to customer that belong to families or have families. Similarly, this trend is again seen if we look at the sports and baby/children wear graph. Again cluster 3 is buying more sports goods than babies and children clothing and cluster 0,2 are mostly buying children/baby ware clothing as opposed to sports ware and cluster 1 is mostly buying sportswear.



Customer Segmentation

Cluster 3: High spending customers between the age of 20 – 60 years mostly are buying ladies wear and sportswear.

Cluster 0: Customer that belong to families or are married, buy men clothing as well as ladies clothing are mostly older in age above 40.

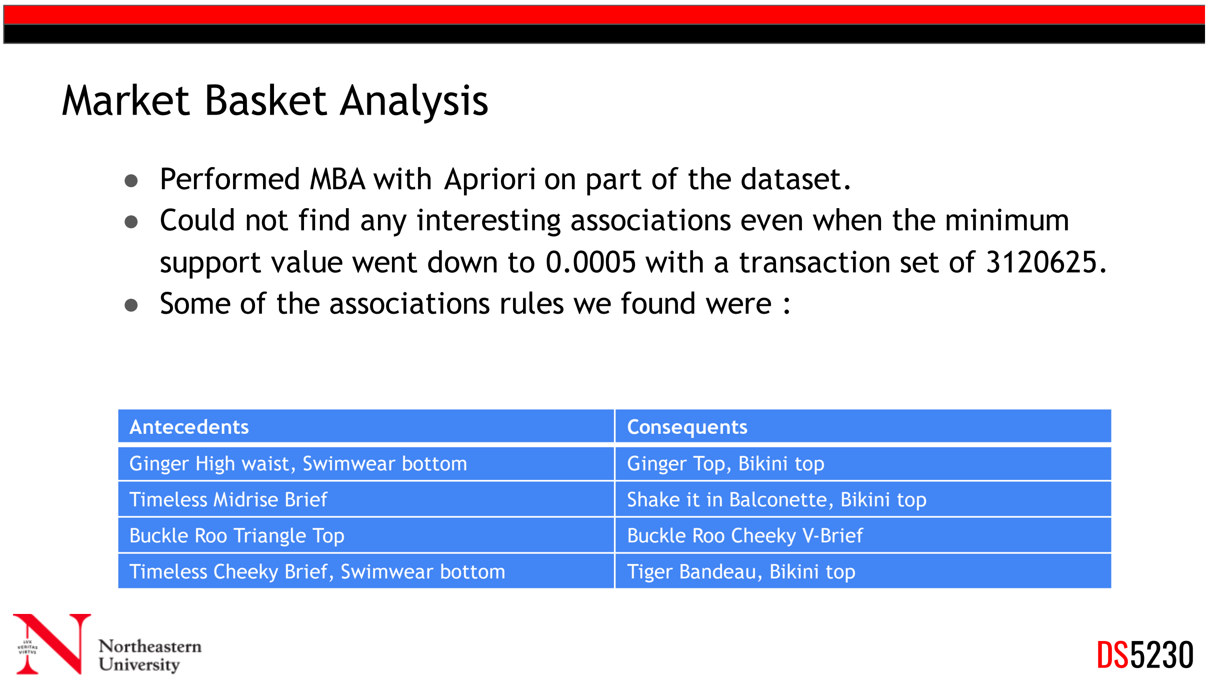
Cluster 2: Customer that have children and belong to families buy men clothing as well as children ware and are mostly older in age above 40.

Cluster 1: Customer that are young aged between 18 and 40. Mostly buy ladies wear and sports garments not interested in baby/children clothing or men clothing are mostly unmarried or single.

Northea
University

DS5230

Cluster 3 are the highest spending customers and aged between 20 and 60 years of age and spend the most out of all the other clusters they mostly buy ladies wear and sportswear. This the customer base that H and M should target the most in order to keep their revenue High. Next comes cluster 0 and cluster 2 they are similar in nature mostly are of older age group above 40 and spend less than cluster 3 but are the high in numbers so they make a big chunk of H and M customer base. Cluster 0 mostly belong to families or are married because they buy men’s clothing as well as ladies wear while cluster 2 also buys children clothing that means customer in cluster 2 have children as well and should be marketed with this in mind. Lastly the cluster 1 which is made of customers that are very young in age and are mostly unmarried with no children as they mostly buy ladies wear or sports good. They expenditure is also less than cluster 3.



Market Basket Analysis

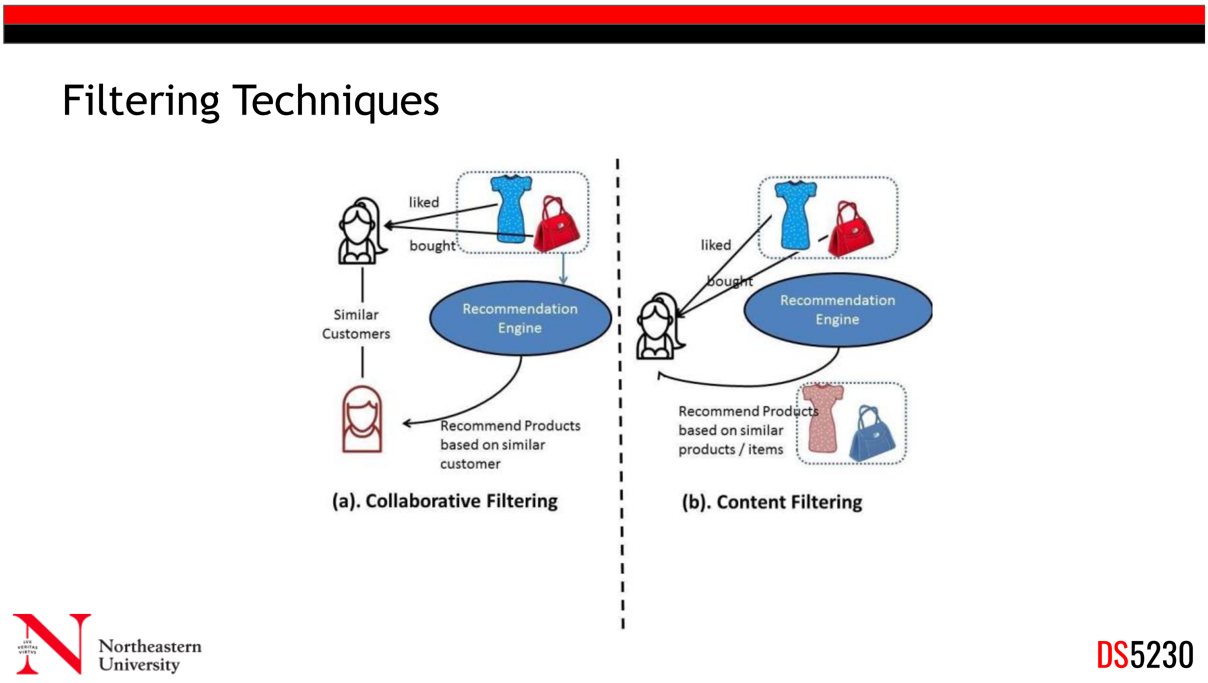
- Performed MBA with Apriori on part of the dataset.
- Could not find any interesting associations even when the minimum support value went down to 0.0005 with a transaction set of 3120625.
- Some of the associations rules we found were :

Antecedents	Consequents
Ginger High waist, Swimwear bottom	Ginger Top, Bikini top
Timeless Midrise Brief	Shake it in Balconette, Bikini top
Buckle Roo Triangle Top	Buckle Roo Cheeky V-Brief
Timeless Cheeky Brief, Swimwear bottom	Tiger Bandeau, Bikini top

Northeastern University

DS5230

We performed MBA using Apriori algorithm to find interesting association rules. Even with a transaction set of 300,000 rows and a support value of 0.0005 we were barely able to see some association rules. We think this happened largely because of how huge the dataset is and since we are looking at a portion of the dataset getting interesting association rules or frequent item sets was not that effective. We have listed down some of the association rules we found like if someone buys Ginger High waist, swim wear bottom they are likely to buy the ginger top, bikini top as well. Another one is that if you buy the timeless midrise Brief customer more likely buy the shake it in Balconette, Bikini top with it as well so that is one of the interesting associations rules, we could find but overall, we did get much success to mine other rules.



Filtering Techniques

liked
bought
Similar Customers
Recommendation Engine
Recommend Products based on similar customer
liked
bought
Recommendation Engine
Recommend Products based on similar products / items

(a). Collaborative Filtering

(b). Content Filtering

Northeastern University

DS5230

Next, we come to the filtering techniques:

Within the spectrum of filtering techniques, two primary categories emerge:

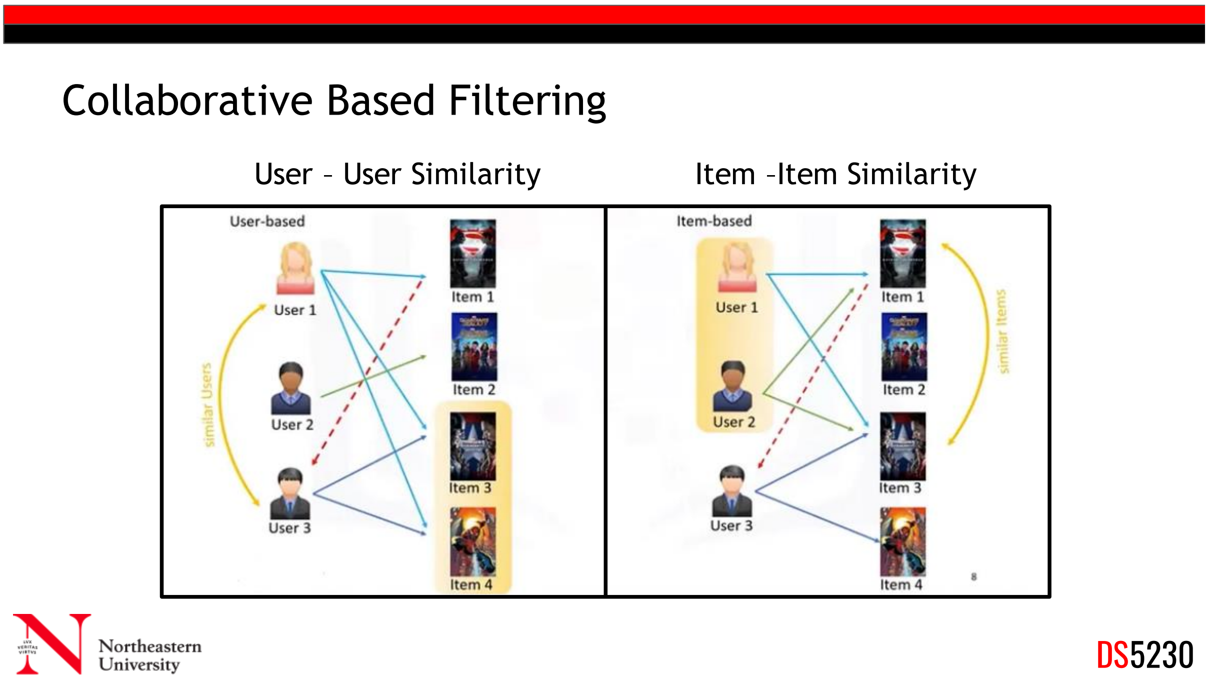
1. Collaborative Filtering
2. Content based Filtering.

Collaborative Filtering:

It is a **memory-based** approach which directly works with the values of recorded interactions or data and independent of item features.

Content-Based Filtering:

This approach requires a good amount of information about **items' features**, rather than using the user's interactions.



Collaborative Based Filtering

User – User Similarity Item –Item Similarity

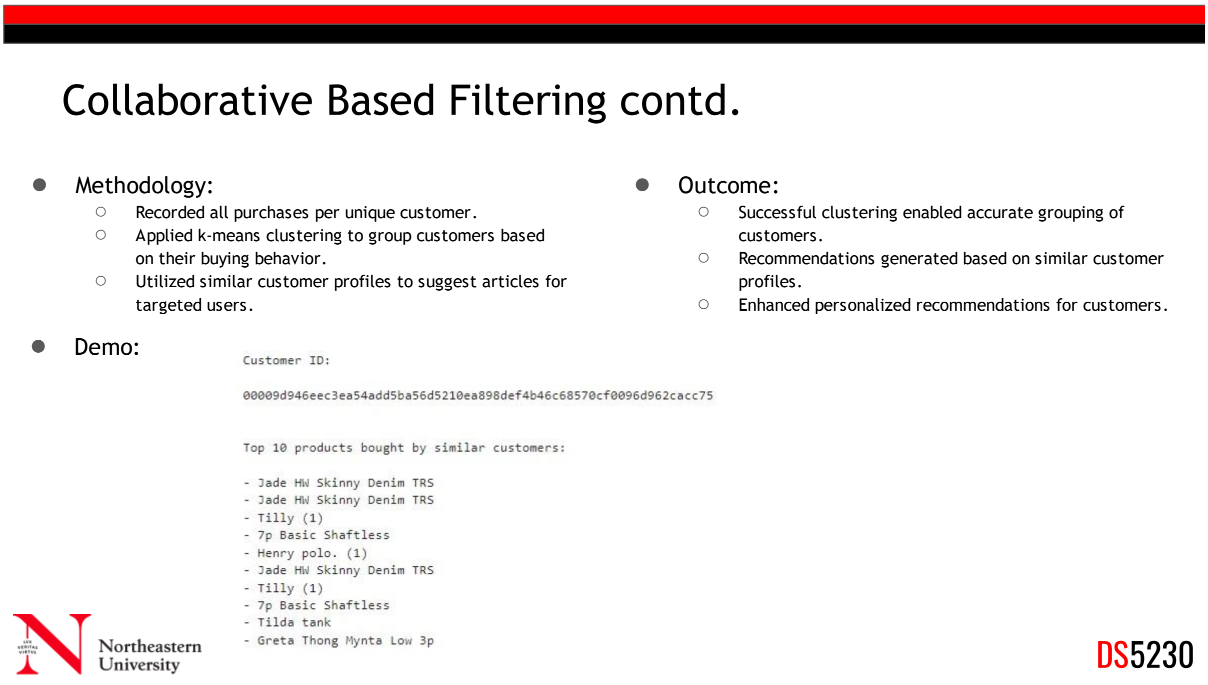
DS5230

Collaborative filtering hinges on two fundamental concepts:

i. User-User Similarity:Identifies similar users

ii. User-Item Similarity: Pinpoints items bought or interacted with by similar users.

This method relies on recorded past interactions between users and items to generate fresh recommendations. Its core premise revolves around leveraging historical user-item interactions within the system to detect resemblances among users or items. Through this process, the system extracts insights and estimations to make predictive recommendations based on these identified similarities.



Collaborative Based Filtering contd.

- Methodology:
  - Recorded all purchases per unique customer.
  - Applied k-means clustering to group customers based on their buying behavior.
  - Utilized similar customer profiles to suggest articles for targeted users.
- Demo:

```
Customer ID:

00009d946eec3ea54add5ba56d5210ea898def4b46c68570cf0096d962cacc75


Top 10 products bought by similar customers:

- Jade HW Skinny Denim TRS
- Jade HW Skinny Denim TRS
- Tilly (1)
- 7p Basic Shaftless
- Henry polo. (1)
- Jade HW Skinny Denim TRS
- Tilly (1)
- 7p Basic Shaftless
- Tilda tank
- Greta Thong Mynta Low 3p
```

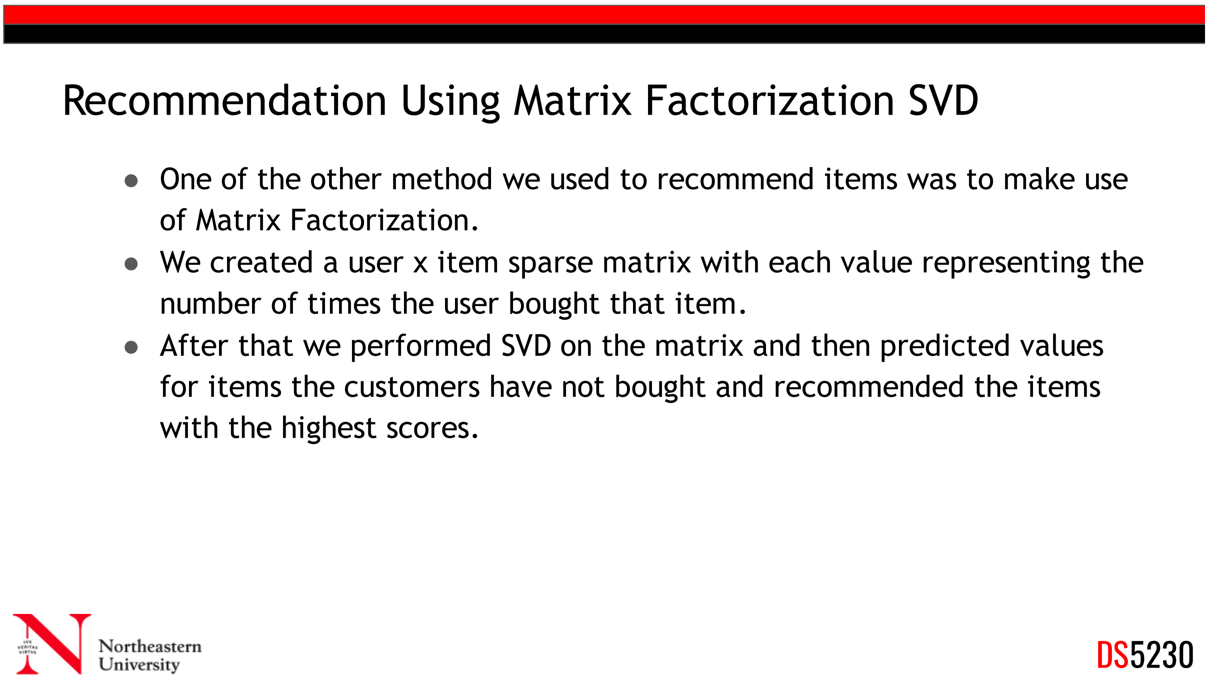
- Outcome:
  - Successful clustering enabled accurate grouping of customers.
  - Recommendations generated based on similar customer profiles.
  - Enhanced personalized recommendations for customers.

Northeastern University

DS5230

One of the first approaches of Collaborative Filtering techniques that we implemented aimed to identify **similar user groups**, grouping them into clusters based on their preferences to offer recommendations aligned with each cluster's preferences.

Based on past interaction given in transaction data(i.e., articles bought against each customer id), we find clusters of similar users through embedding. Once we identify the cluster most similar to targeted customer(**user-user interaction**), we recommend products , that have been most frequently bought(**user-item interaction**).



Recommendation Using Matrix Factorization SVD

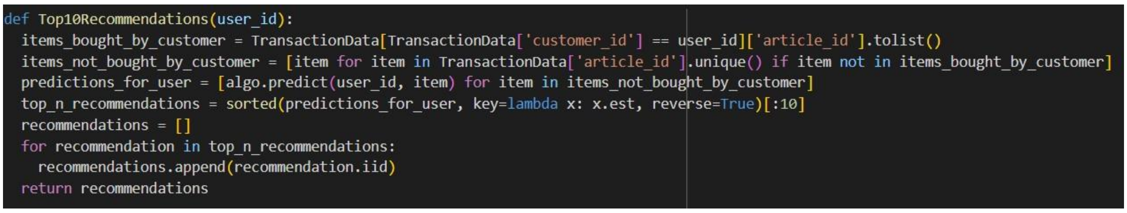
- One of the other method we used to recommend items was to make use of Matrix Factorization.
- We created a user x item sparse matrix with each value representing the number of times the user bought that item.
- After that we performed SVD on the matrix and then predicted values for items the customers have not bought and recommended the items with the highest scores.

Northeastern University

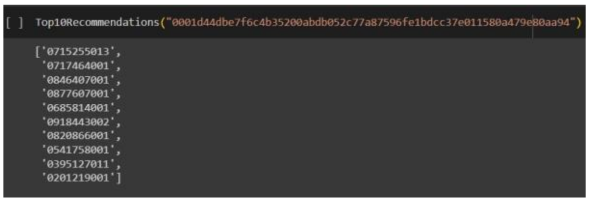
DS5230

Another collaborative technique we used to recommend items was to create a sparse matrix of user x items. The values represent the times the customer bought the item which will be used as a form of a rating and then we performed matrix factorization using SVD. After performing SVD our model was ready and then we used that model to predict values for items the customer has not bought. Once we predict scores for the items the customer has not bought, we recommend the items for which we got the best scores to the customer that is how we are using matrix factorization method to recommend items to the customers. In total our matrix had 189510x26252 values in the sparse matrix.

Demo for Matrix Factorization Method



```python
def Top10Recommendations(user_id):
  items_bought_by_customer = TransactionData[TransactionData['customer_id'] == user_id]['article_id'].tolist()
  items_not_bought_by_customer = [item for item in TransactionData['article_id'].unique() if item not in items_bought_by_customer]
  predictions_for_user = [algo.predict(user_id, item) for item in items_not_bought_by_customer]
  top_n_recommendations = sorted(predictions_for_user, key=lambda x: x.est, reverse=True)[:10]
  recommendations = []
  for recommendation in top_n_recommendations:
    recommendations.append(recommendation.iid)
  return recommendations
```

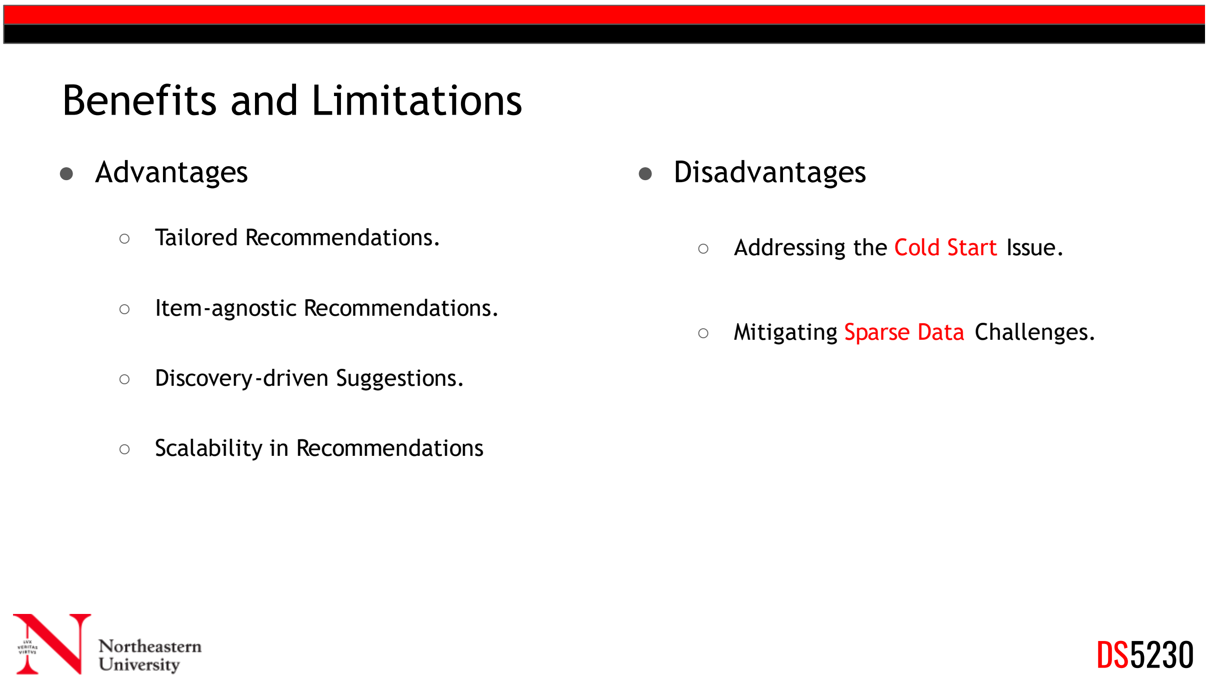


```
[ ]  Top10Recommendations("0001d44dbe7f6c4b35200abdb052c77a87596fe1bdcc37e011580a479e80aa94")

   ['0715255013',
    '0717464001',
    '0846407001',
    '0877607001',
    '0685814001',
    '0918443002',
    '0820866001',
    '0541758001',
    '0395127011',
    '0201219001']
```

DS5230

This is for the demo of the matrix factorization method where we take in the value of the customer ID. For each article or product that the customer has not bought we predict the values for those items for our customer using the model and then recommend the top highest scoring products to the customer that our SVD model has predicted using the Matrix factorization method.



Benefits and Limitations

- Advantages
  - Tailored Recommendations.
  - Item-agnostic Recommendations.
  - Discovery-driven Suggestions.
  - Scalability in Recommendations
- Disadvantages
  - Addressing the Cold Start Issue.
  - Mitigating Sparse Data Challenges.

Northeastern University

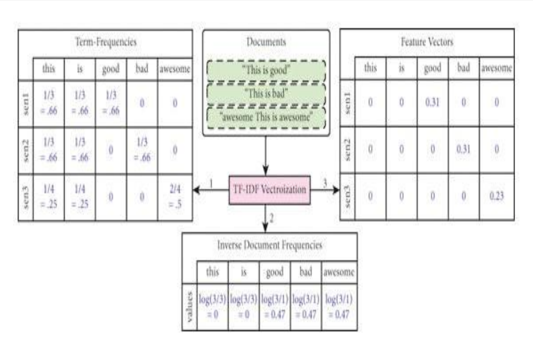
DS5230

In the realm of Collaborative based recommendations, the advantages are evident: tailored suggestions based on user behavior, a reliance on user interactions rather than exhaustive item details, and the introduction of serendipitous discoveries through recommendations based on user similarities. Moreover, the **scalability** of this approach with larger datasets is a notable benefit.

However, when delving into the drawbacks of collaborative filtering, the spotlight falls on the challenging "**Cold Start Problem**," posing difficulties in recommending for new users or items with limited historical data. Additionally, the **sparsity of data** can impact accuracy when records of user interactions are limited, and there's a potential for bias towards popular items, potentially overshadowing niche, or less-known products.

Content Based Filtering

Vector Representation



Term-Frequencies

	this	is	good	bad	awesome
sen1	1/3 = .66	1/3 = .66	1/3 = .66	0	0
sen2	1/3 = .66	1/3 = .66	0	1/3 = .66	0
sen3	1/4 = .25	1/4 = .25	0	0	2/4 = .5

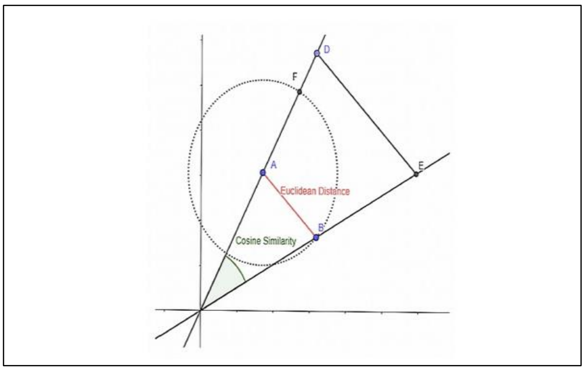
Inverse Document Frequencies

	this	is	good	bad	awesome
values	log(3/3) = 0	log(3/3) = 0	log(3/1) = 0.47	log(3/1) = 0.47	log(3/1) = 0.47

Feature Vectors

	this	is	good	bad	awesome
sen1	0	0	0.31	0	0
sen2	0	0	0	0.31	0
sen3	0	0	0	0	0.23

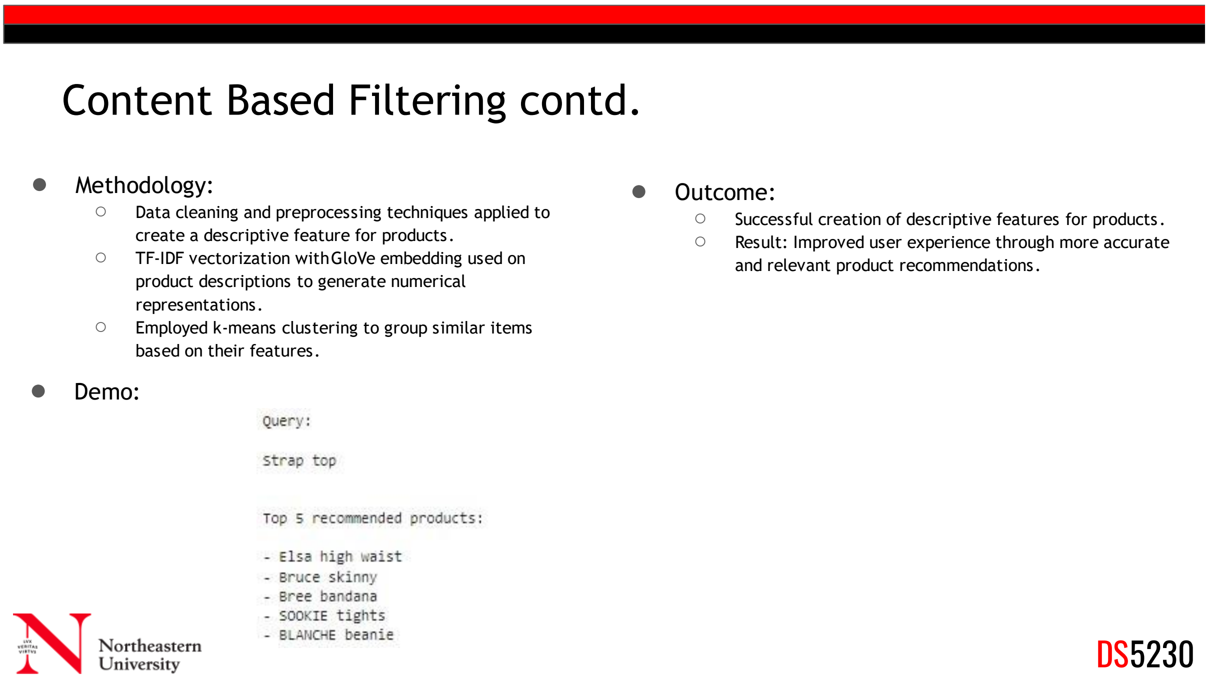
Similarity Score



This is where Content-based Filtering technique comes to play.

The main idea of content-based methods is to try to build a model, based on the **available "features"**, that explain the observed user-item interactions. This approach requires a good amount of information about items' features, rather than using the user's interactions.

It does not need any data about other users since the recommendations are specific to a particular user. This makes it easier to scale down the same to a large number of users, unlike Collaborative Filtering Methods.



Content Based Filtering contd.

- Methodology:
  - Data cleaning and preprocessing techniques applied to create a descriptive feature for products.
  - TF-IDF vectorization withGloVe embedding used on product descriptions to generate numerical representations.
  - Employed k-means clustering to group similar items based on their features.
- Demo:

```
Query:

Strap top


Top 5 recommended products:

- Elsa high waist
- Bruce skinny
- Bree bandana
- SOOKIE tights
- BLANCHE beanie
```

- Outcome:
  - Successful creation of descriptive features for products.
  - Result: Improved user experience through more accurate and relevant product recommendations.

DS5230

The given approach to Content-Based method includes:

1. Feature Engineering Item Features:

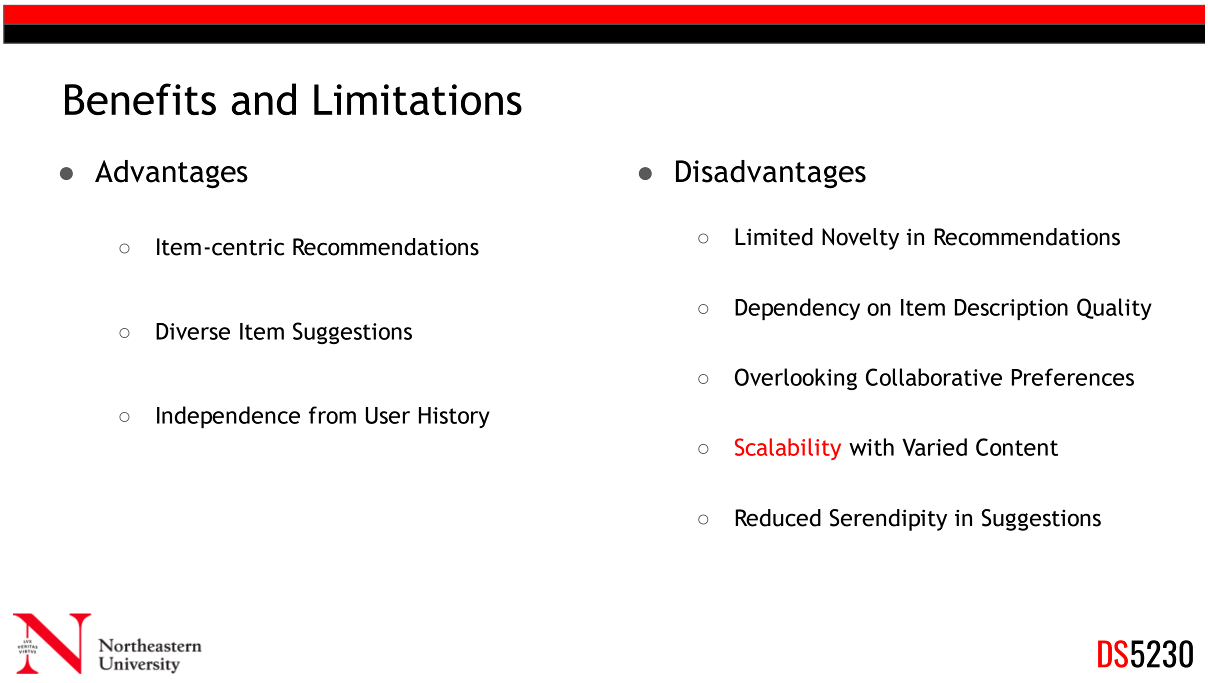
   Each item in the system is described by a set of features or attributes (e.g., genre, descriptions, etc.) which form the basis for comparing and recommending similar items.
2. Vector Representation:

   Implemented techniques like TF-IDF (Term Frequency-Inverse Document Frequency) with GloVe embeddings to convert item features into numerical representations.
3. Item Similarity Calculation:

   The system calculates the similarity between items based on their features.
4. K-Means Clustering Pipeline:

   Performed kmeans clustering based on the similarity score to group siliar items.
5. Item-Item Interactions:

   Based on the user's interaction with a particular item, the system identifies and recommends similar items that share common features or attributes.



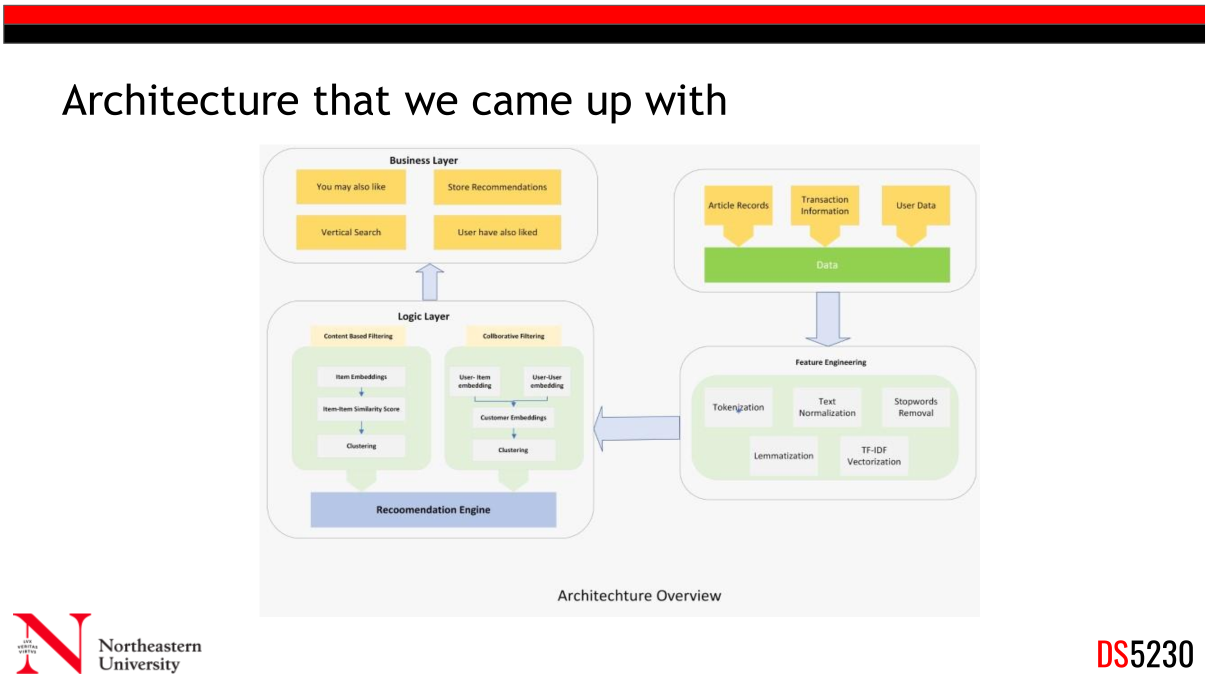
Benefits and Limitations

- Advantages
  - Item-centric Recommendations
  - Diverse Item Suggestions
  - Independence from User History
- Disadvantages
  - Limited Novelty in Recommendations
  - Dependency on Item Description Quality
  - Overlooking Collaborative Preferences
  - Scalability with Varied Content
  - Reduced Serendipity in Suggestions

Northeastern University

DS5230

Content-based technique heavily focuses on item features, enabling a detailed matching process that resonates with specific user tastes, ensuring a **tailored experience**. It promotes diversity by offering varied recommendations encouraging users to explore different facets of their interests. Additionally, its independence from user history makes it advantageous for new users, as it doesn't heavily rely on past interactions. Thus avoiding cold start problem. However, content-based filtering encounters challenges in suggesting entirely new or unrelated items to users. There's also a risk of overlooking collaborative preferences or trending items, potentially missing out on recommendations that align with broader preferences. One of the major concerns apart from these is **scalability issue,** reducing element of surprise.



Architecture that we came up with

Architechture Overview

Northeastern University

DS5230

Considering the advantages and constraints inherent in both filtering methods, we adopted a hybrid approach that amalgamates both techniques emerges as a viable solution.

Content-based recommendations can be displayed categorically, such as '**Store Recommendations**,' showcasing items based on their features.

Simultaneously, collaborative recommendations can be featured in sections like '**Users Have Also Liked**,' drawing from past user interactions to suggest items aligned with similar users' preferences.