H&M Personalized Fashion Recommendations

H&M

Group 7

Team member:

Chien-Hsin Lee, Jia-Jia Yu, Raymond Su, Xinbo Lu

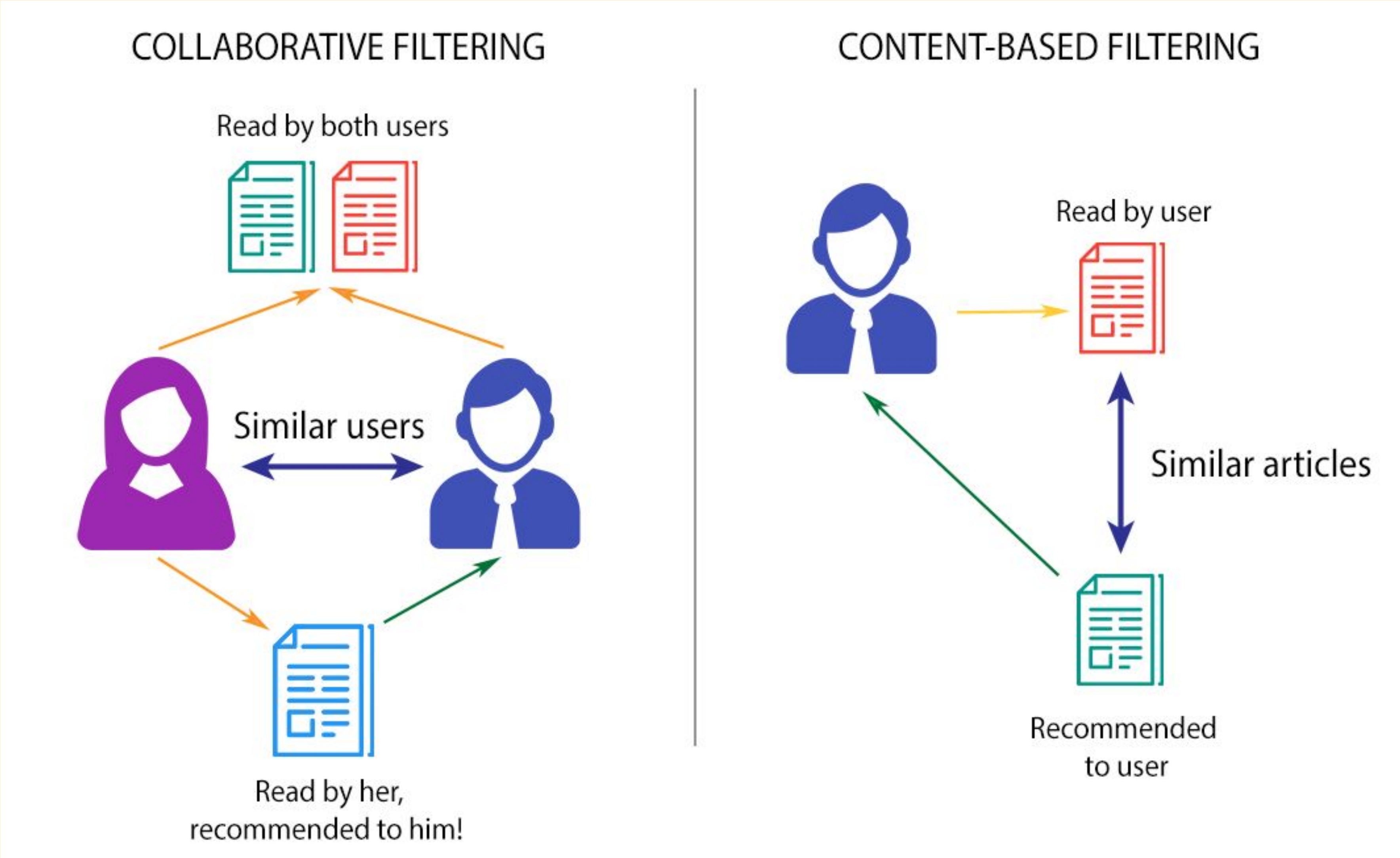
Agenda

1. Previous work

The Existing Recommendation System Construction

User-based Collaborative filtering

Content-based filtering



1. Previous work

K-Menas, PCA, Logistic regression

To create our own recommendation result, we combine PCA, K-means, logistic regression to build our recommendation system.

k-means in fashion: https://www.sciencedirect.com/science/article/pii/S2210832717300315
logistic regression in fashion: https://www.computer.org/csdl/magazine/mu/2014/02/mmu2014020072/13rRUILtJw9
https://blog.jovian.ai/logistic-regression-on-fashion-mnist-e3473ca496f0

2. Motivation

Difficult to find what we are looking for on the home website

Sustainability & minimizes emissions from transportation

Complete & available dataset on Kaggle

Enhance customers’ shopping experience on H&M

2. Research Question

- How can we implement **a recommendation system** differently among loyal customers and less frequent/new customers ?

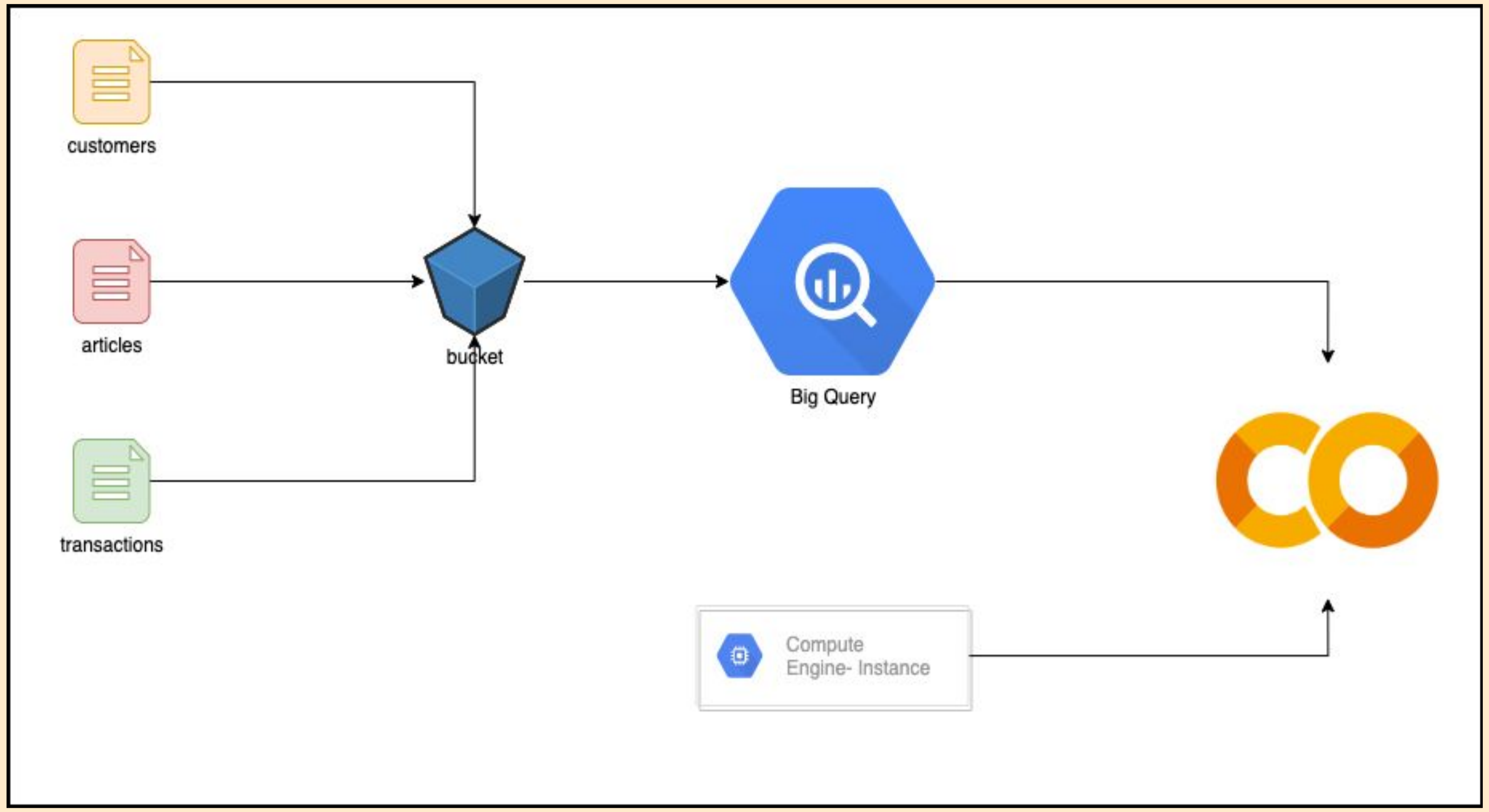
3. Data, Analysis, Results

Data Collection

Our H&M Personalized Fashion Recommendations data source is from kaggle website.

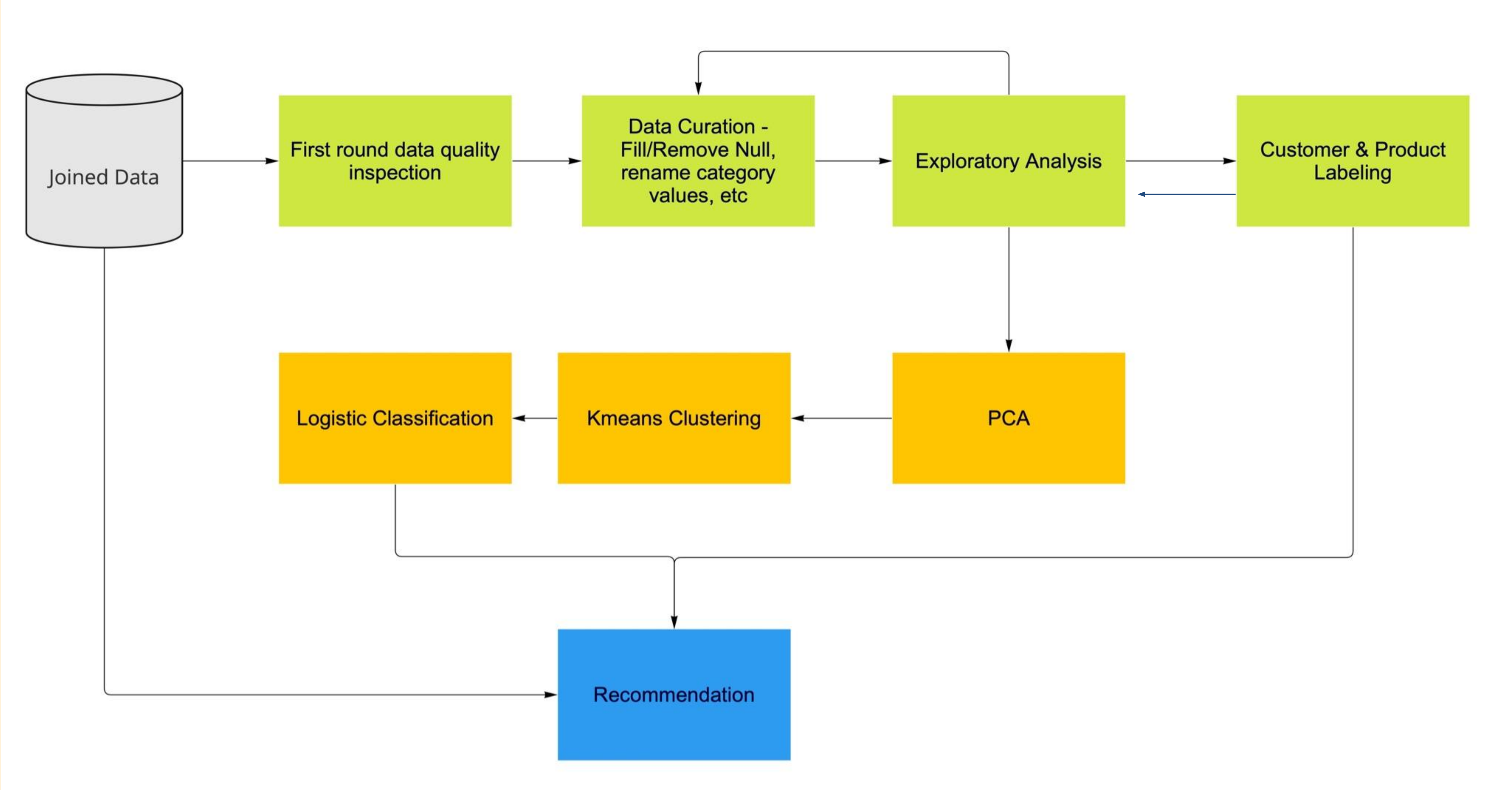
Our data contains three major themes which are customer, article, and transaction. The transaction contained millions of records, so the data curation phase is difficult for us.

Build Data Collection Pipeline

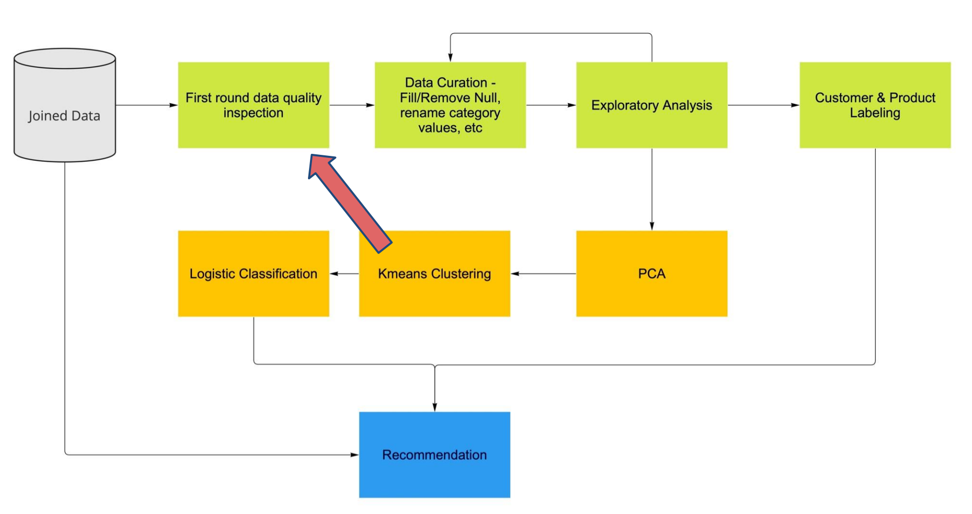


3. Data, Analysis & Result

Workflow overview



3. Data, Analysis & Result



Data Overview

Articles

article_id	product_c...	prod_name	product_ty...	product_ty...	product_gr...	graphical_...	graphical_...	colour_gro...	colour_gro...
0108775015	0108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	09	Black
0108775044	0108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	10	White

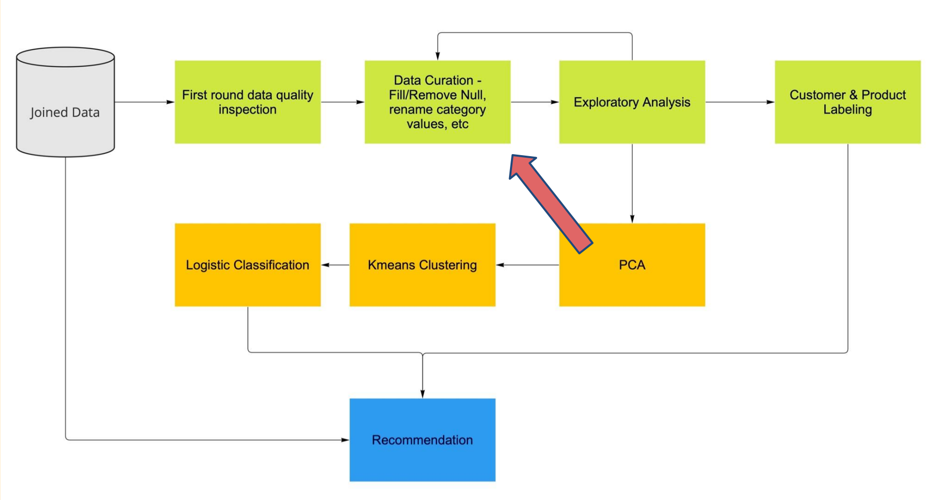
Customers

customer_id	FN	Active	club_mem...	fashion_ne...	age	postal_code
00000dbacae5abe5e23885899a1fa44253a17956c6d1c3d25f88aa139fdfc657			ACTIVE	NONE	49	52043ee2162cf5aa7ee79974281641c6f11a68d276429a91f8ca0d4b6efa8100
0000423b00ade91418cceaf3b26c6af3dd342b51fd051eec9c12fb36984420fa			ACTIVE	NONE	25	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93f4c830291c32bc3057

Transactions

t_dat	customer_id	article_id	price	sales_channel...
2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318	0663713001	0.050830508474576 4	2
2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318	0541518023	0.030491525423728 81	2

- **Duration:** Sep 20th, 2018 -> Sep 22nd, 2020
- **Observation:** 105,542/ 31,788,324/ 1,371,980
- We sample **300K observations.**



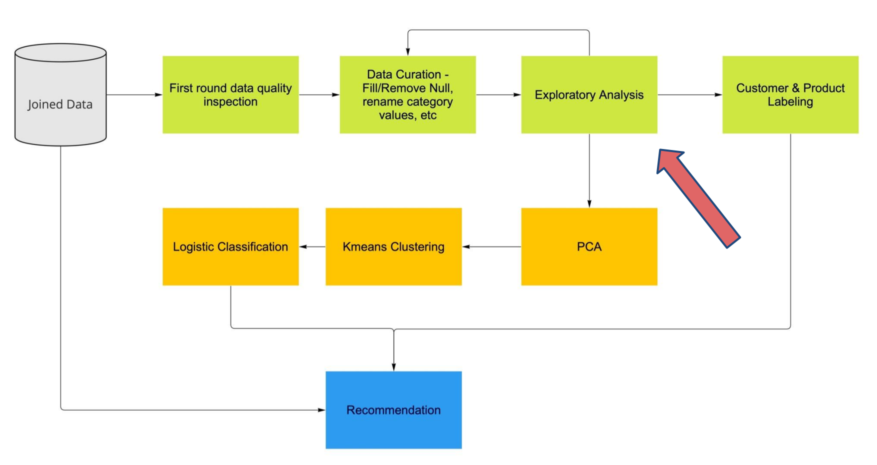
3. Data, Analysis & Result

Data Curation

Tidy and organize our datasets

1. Removed and filled null values
2. To ensure the consistency, we correct spelling (“None” and “none”) and labeled categorical values using LabelEncoder
3. Normalized numeric data by using standard scaler
4. Extracted validation data based on each customer’s last purchase

3. Data, Analysis & Result



Exploratory Analysis

Past EDA result- no significant trend:

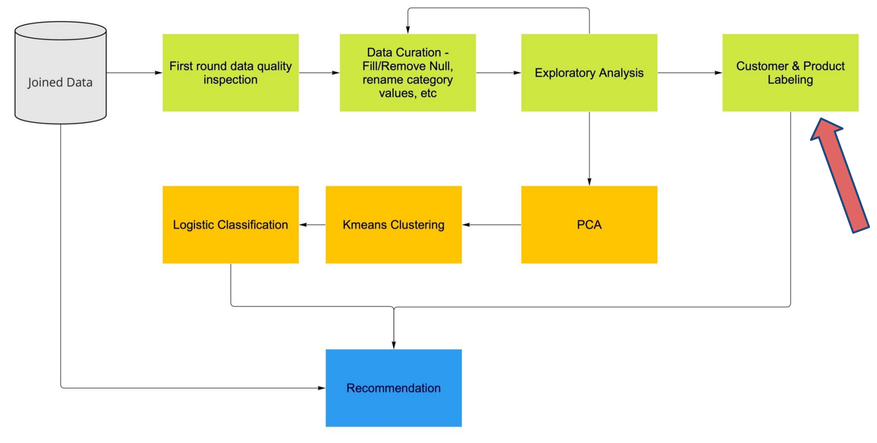
1. Purchase frequency & top 10 favorite product type
2. Purchase frequency & Fashion news familiarity
3. Age group & top 10 favorite product type
4. Postal_code & top 10 favorite product type

frequency	fashion_news_frequency	orders_revenue	order_sum(%)
most frequent	Monthly	98.14	0.02%
most frequent	NONE	347571.12	53.76%
most frequent	Regularly	298822.72	46.22%
frequent	Monthly	128.93	0.07%
frequent	NONE	127022.44	65.87%
frequent	Regularly	65700.47	34.07%
one-time	Monthly	31.12	0.08%
one-time	NONE	30800.68	74.52%
one-time	Regularly	10502.94	25.41%

Purchase frequency

&

Monetary



3. Data, Analysis & Result

Customer Labeling

purchase frequency

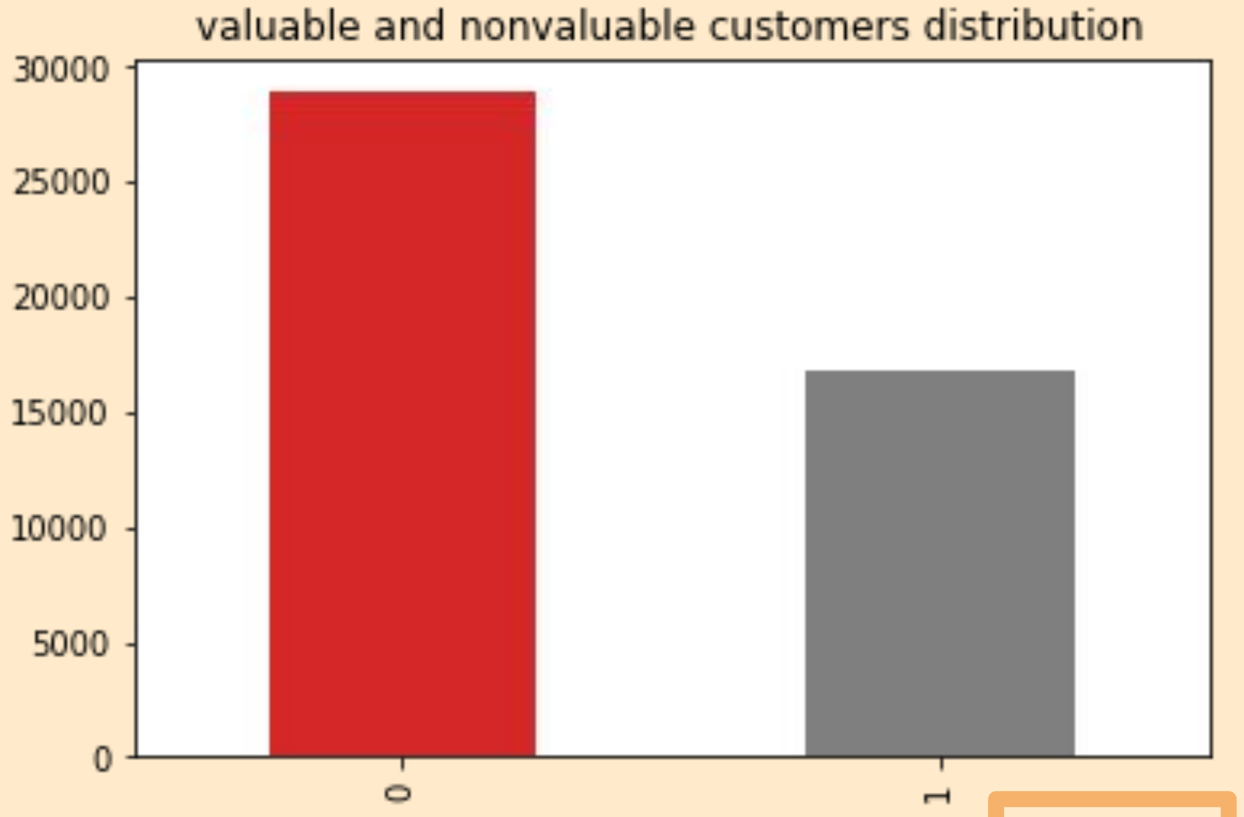
purchase revenue

Select 2 representative variables

frequency score

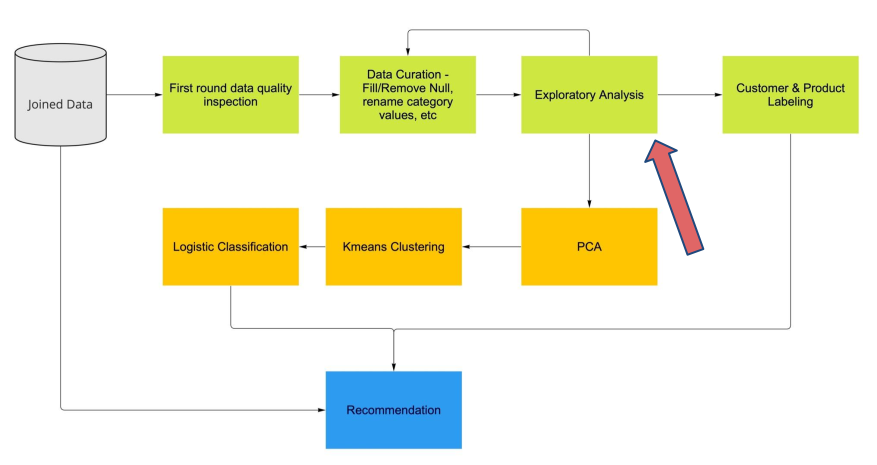
monetary score

Both Score >= 3 as valuable customers 0

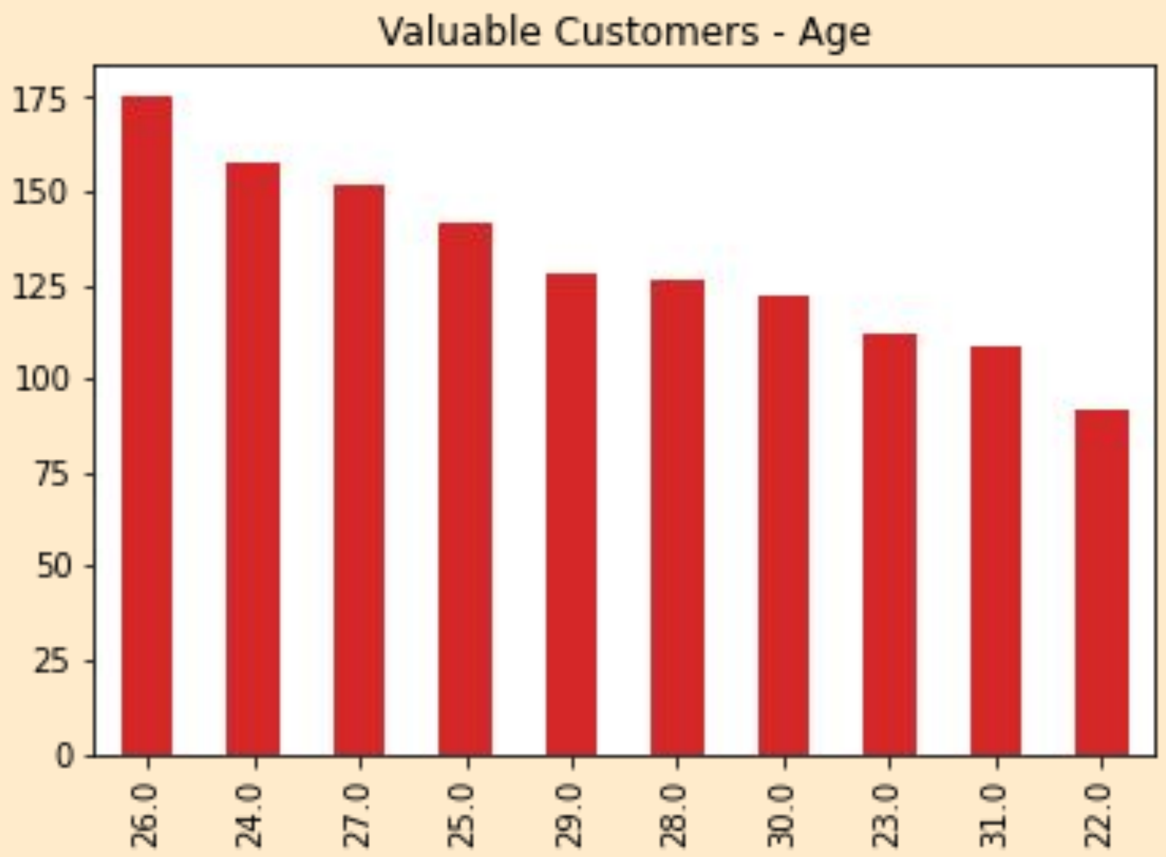


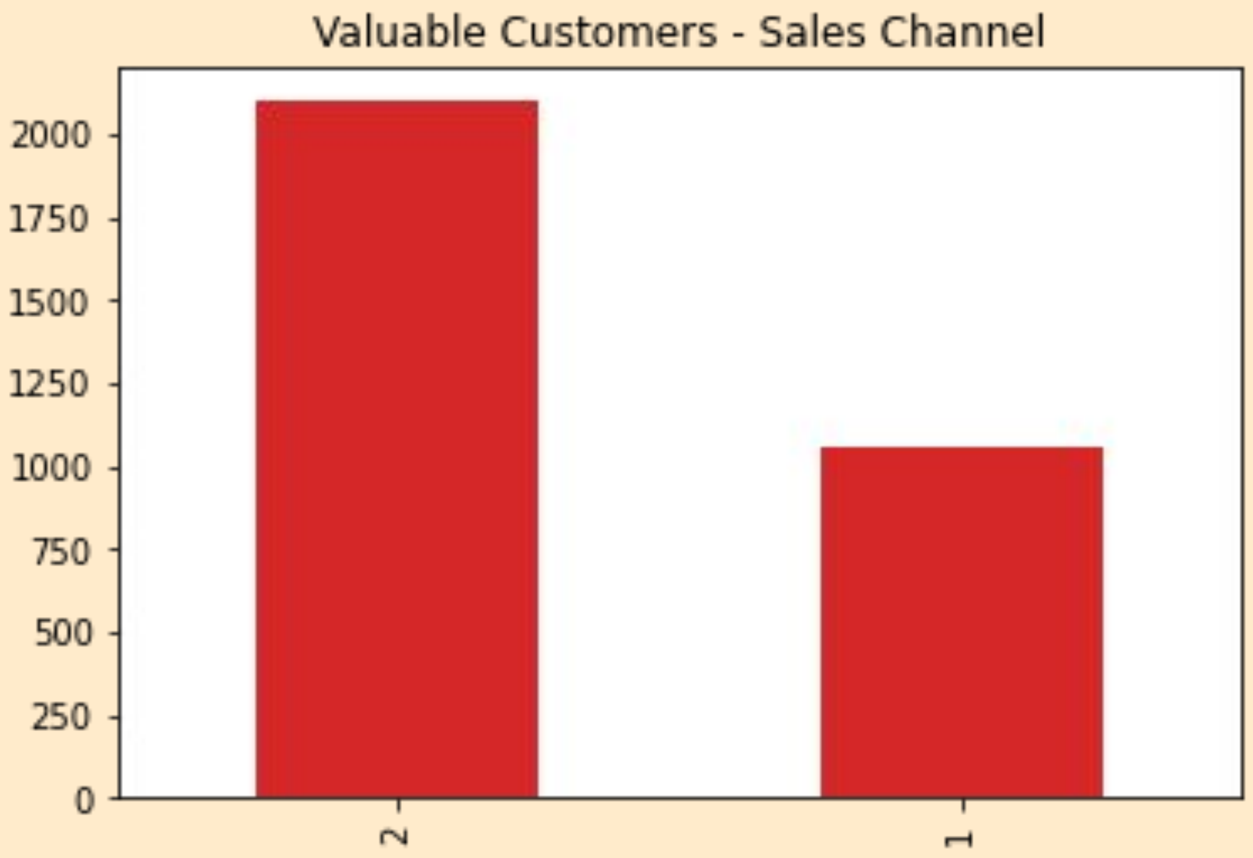
customer_id	Frequency	Monetary	frequency_score	monetary_score	valuable
7edda079a53460373db073a0f975de158239811448ac6428aea6ddda9a4ef423	1	0.001677966101694920	3	1	0
7ee17b7552d3977bb0e69d84b21cf27f3b256b7b19381d460ecf21b8427fe5b6	1	0.004220338983050850	3	1	0
7ee1a4abb01ff588be0c15dad34d9de577b1c307b52b30351123d8a065c7fa32	1	0.033881355932203400	3	4	1
7ee1ae23a2996d1e282dc53123d99bd05e8934afe98befb7a1af99c9f3dd93bb	2	0.0948813559322034	5	5	1
7ee3ca19df5fd981d61c480763e271dd280541b379cd5d67a51abcba1842313a	1	0.025406779661016900	3	3	1

3. Data, Analysis & Result



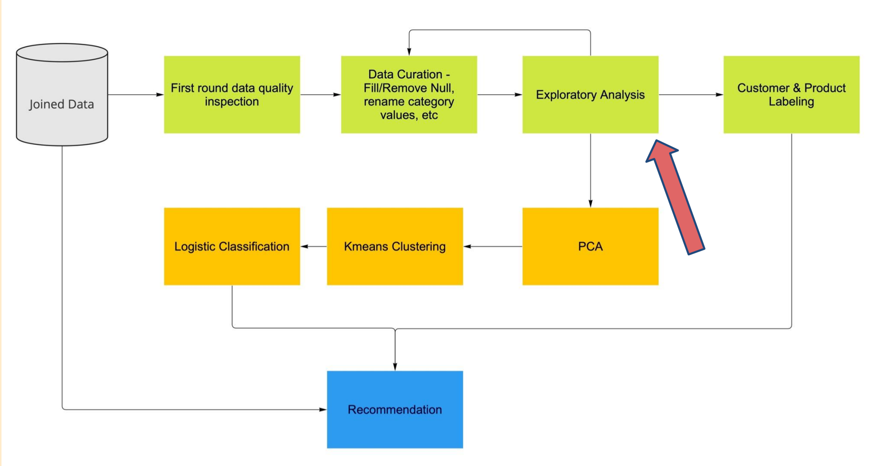
Exploratory Analysis - Valuable Customer



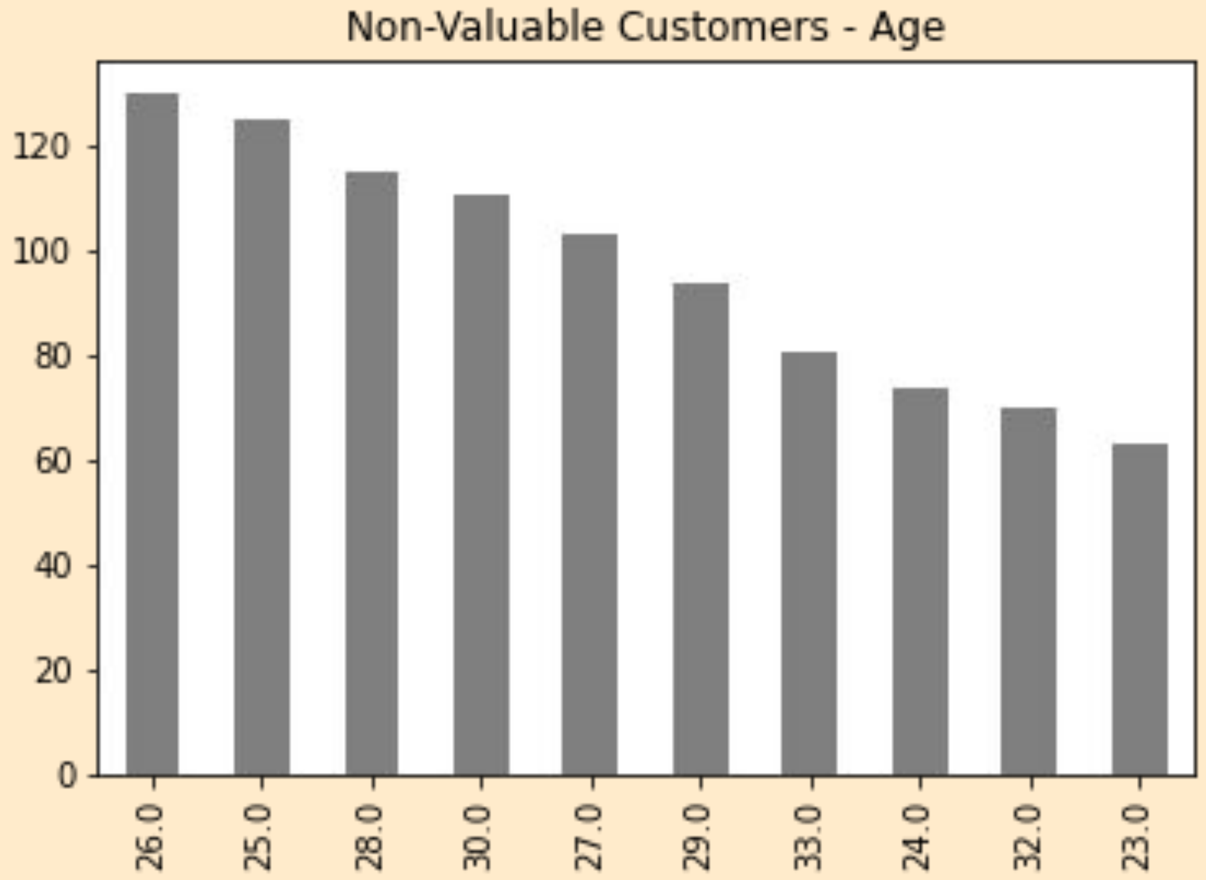


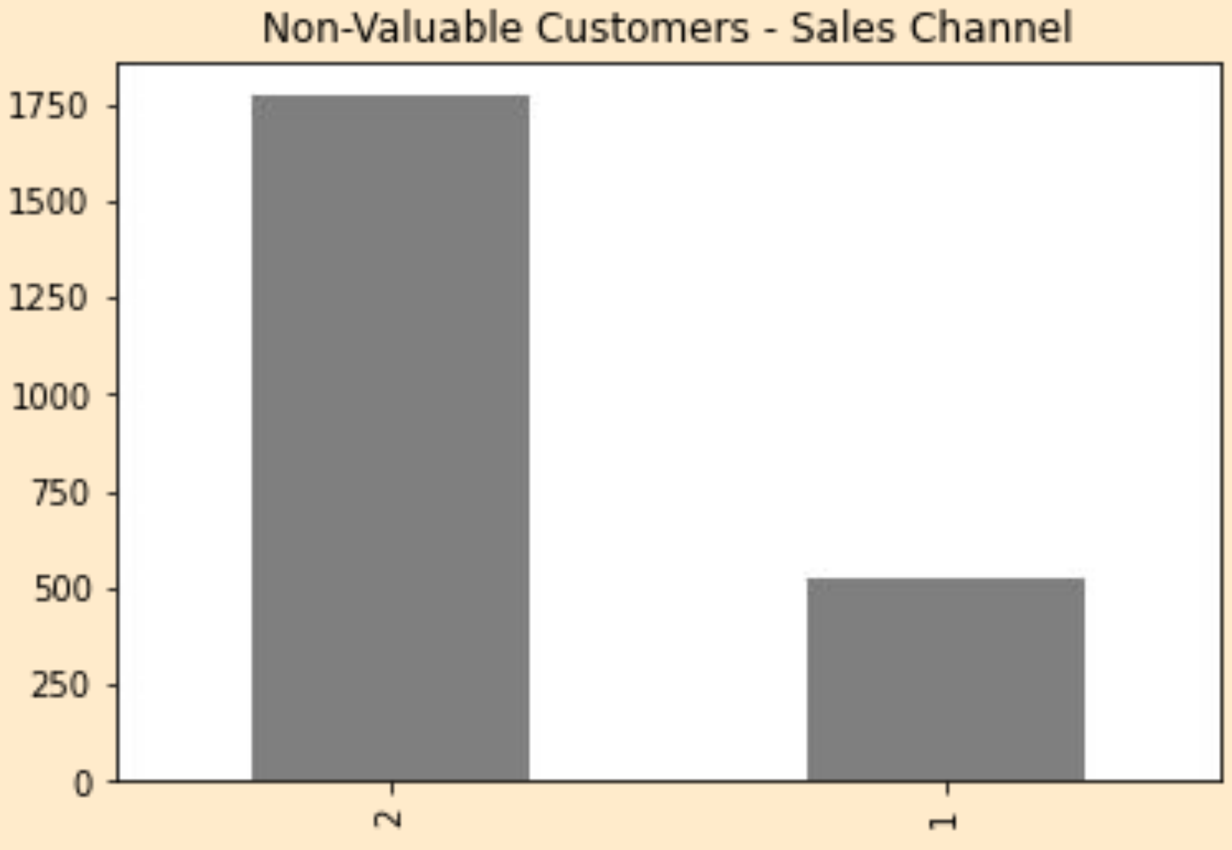
For the valuable customer, its top 10 age range is between 22 to 30.
Its mainly sales channel is 2. Its top 1 product code is 372860, Basic 7p Shaftless.

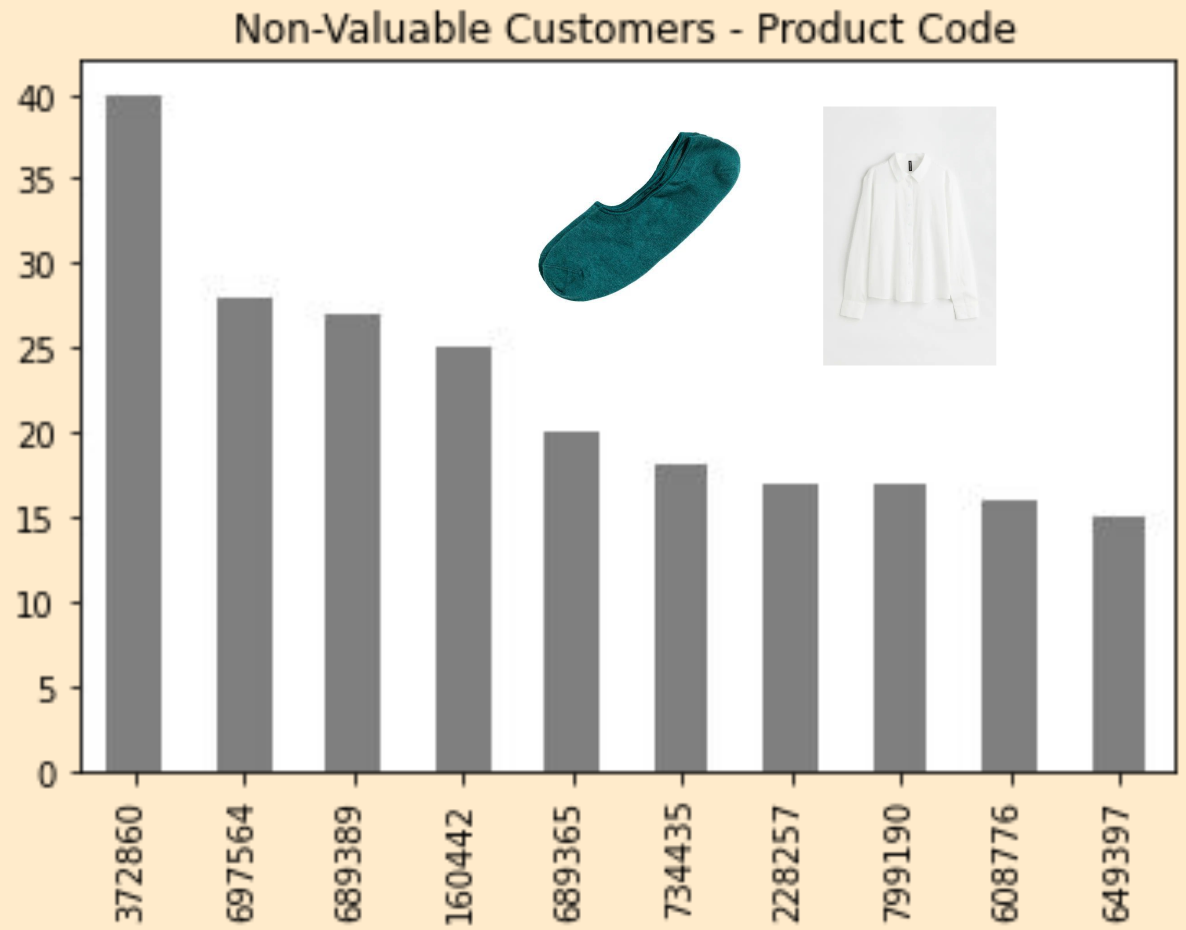
3. Data, Analysis & Result



Exploratory Analysis - Non-Valuable Customer

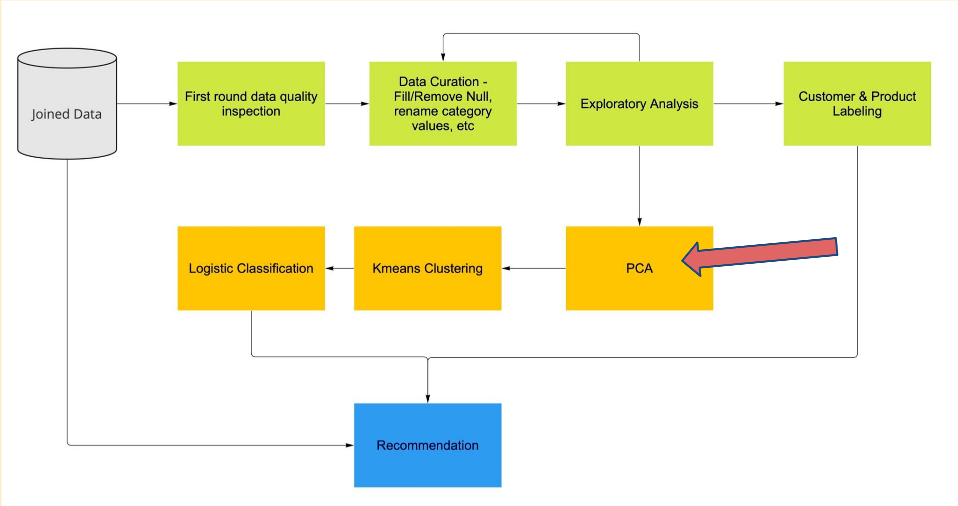






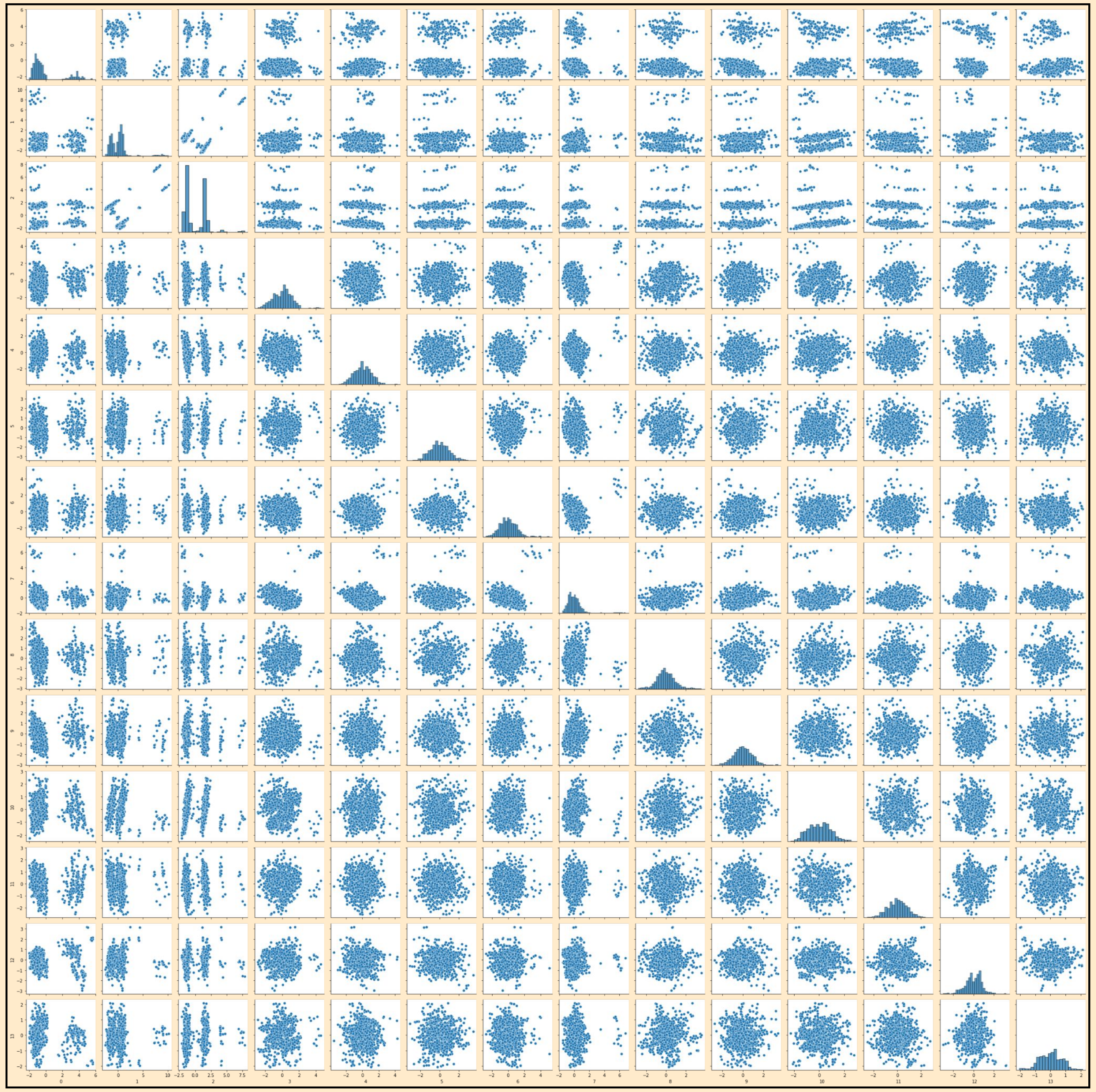
For the non valuable customer, its top 10 age range is between 23 to 33.
Its mainly sales channel is 2. Its top 1 product code is 372860, Basic 7p Shaftless.

3. Data, Analysis & Result

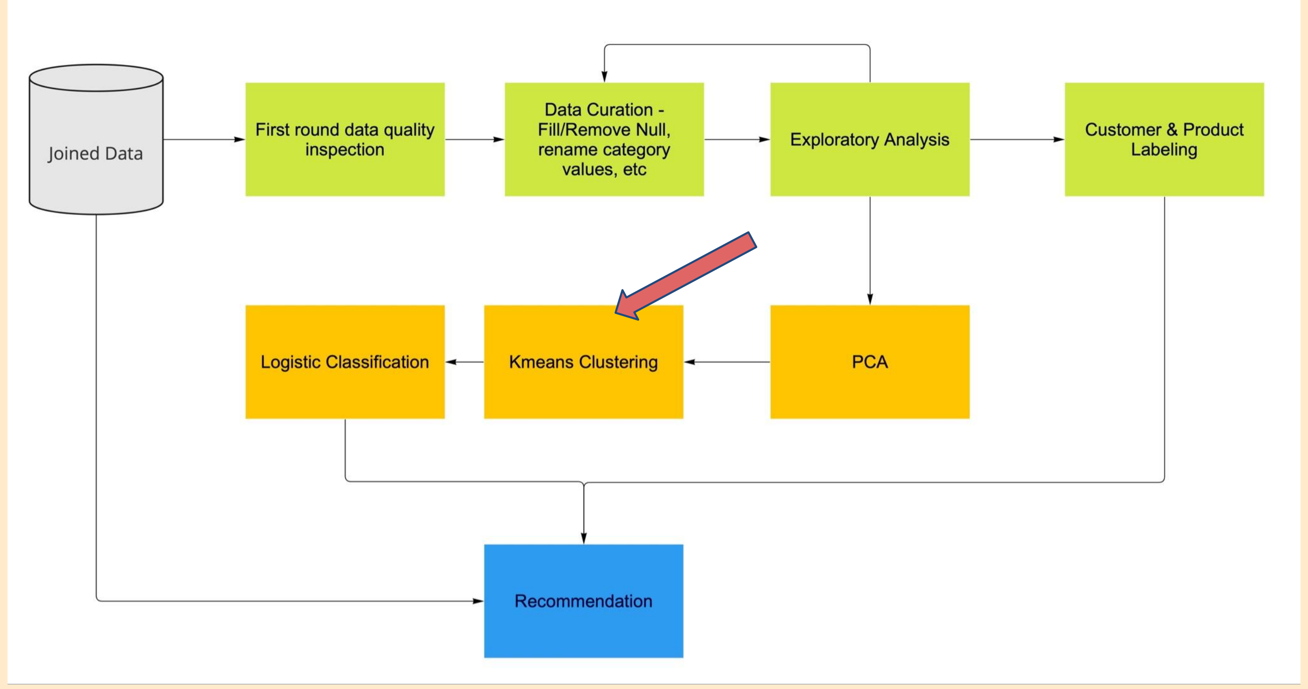


Principal Component Analysis

1. The features included in the joined data is overwhelming
2. PCA has several benefits
   a. Removes Correlated Features
   b. Improves Algorithm Performance
   c. Reduces Overfitting
   d. **Reduce the "Curse of Dimensionality"**
3. PCA components covered 95% of the explained variance (15 features) out of the original data(37 features)

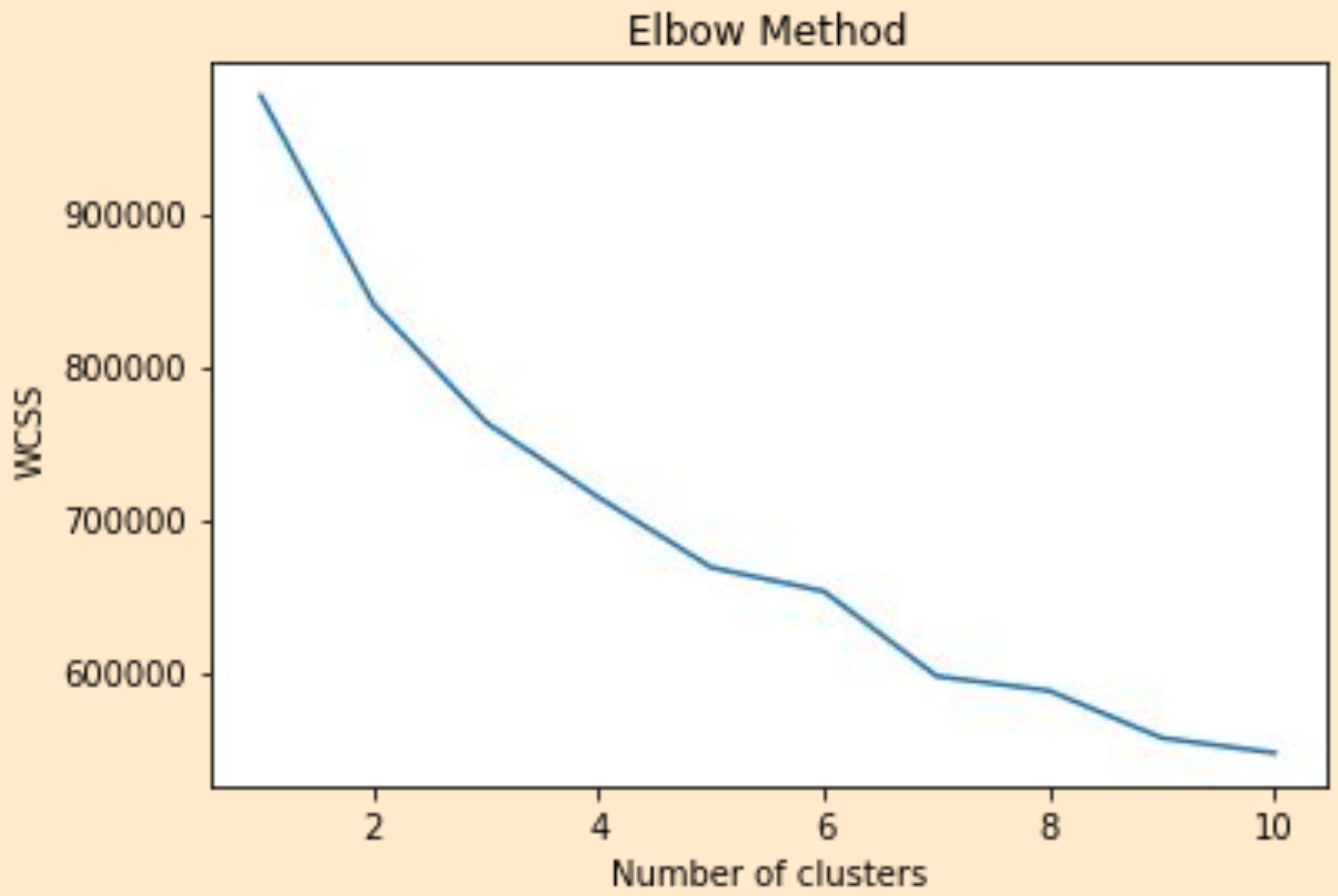


3. Data, Analysis & Result

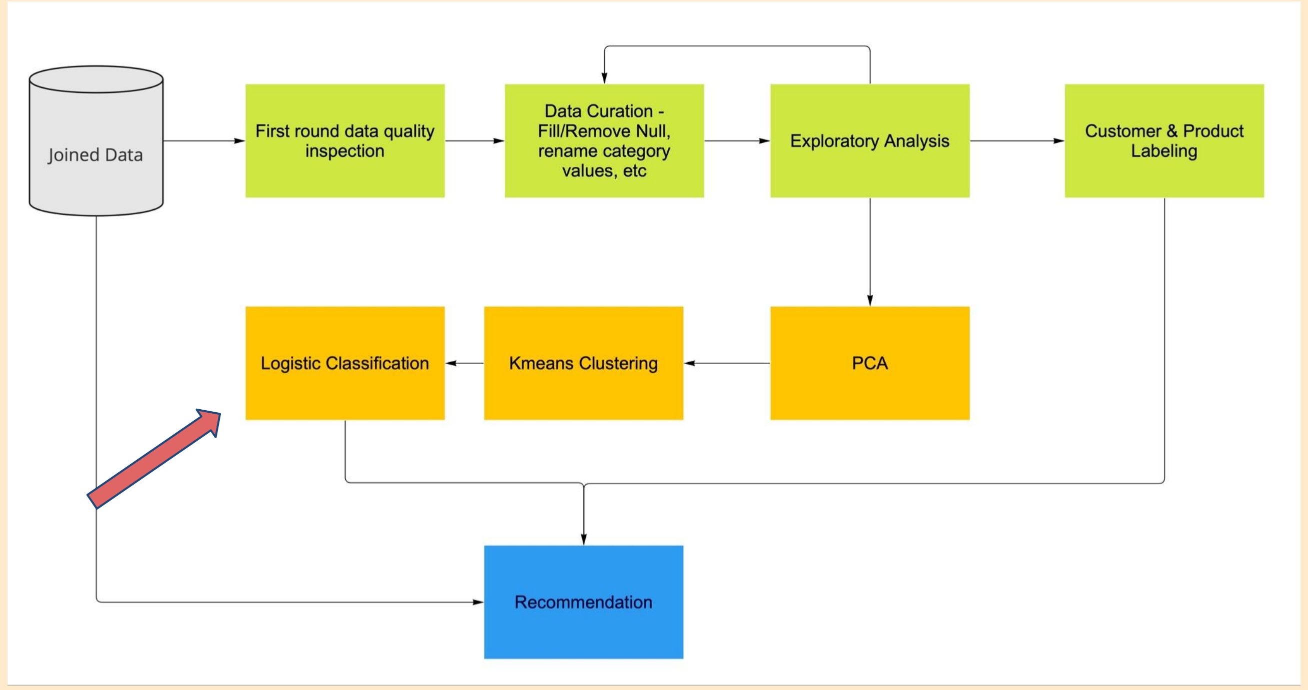


Kmeans clustering

1. Kmeans clustering benefits
   a. Unsupervised
   b. **Scalable to large dataset**
   c. Easy to implement
2. Classified our customer based on the PCA components into 10 clusters (determined by the Elbow method)
3. Serve as one of our X-variables for future prediction

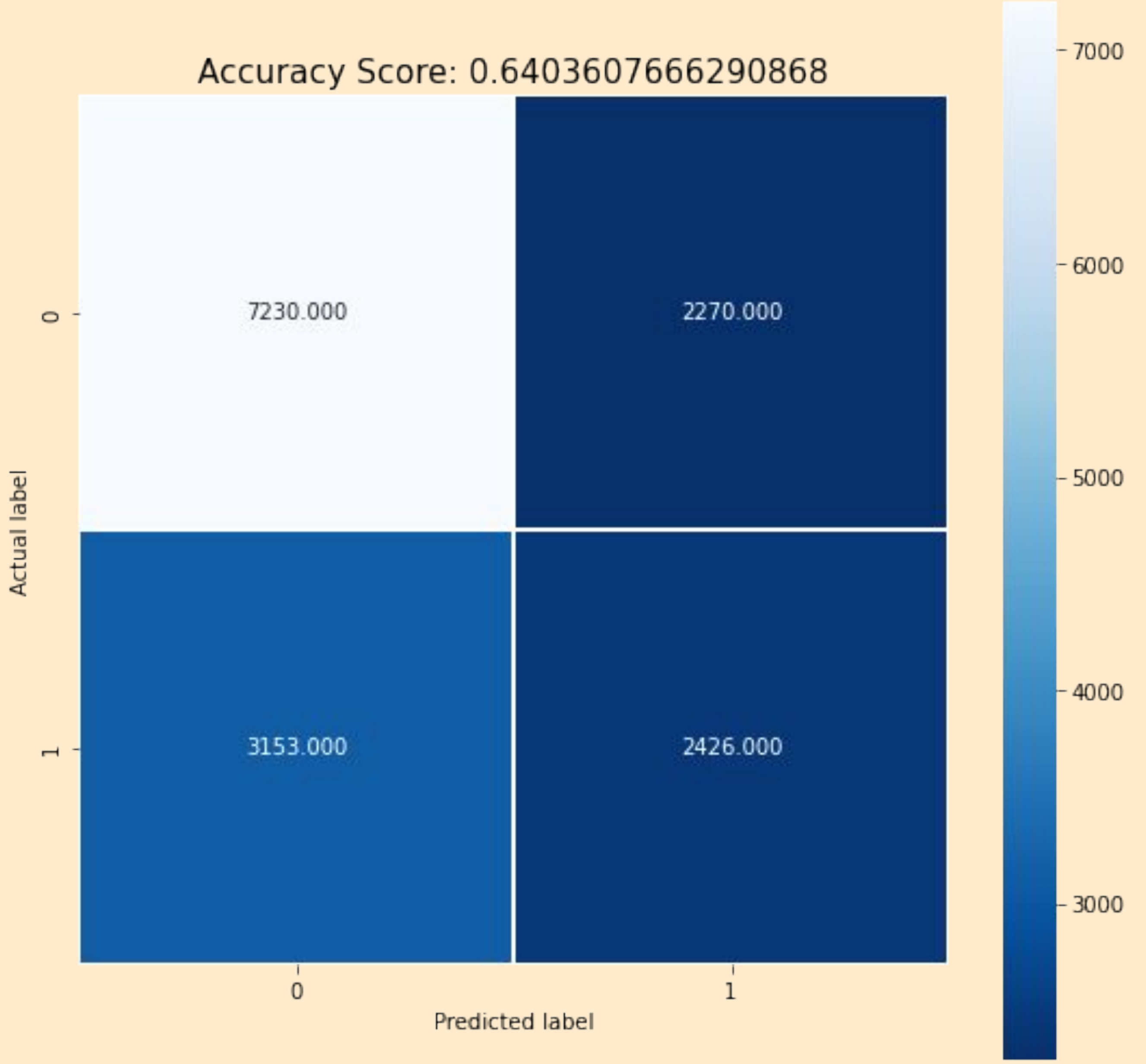


3. Data, Analysis & Result

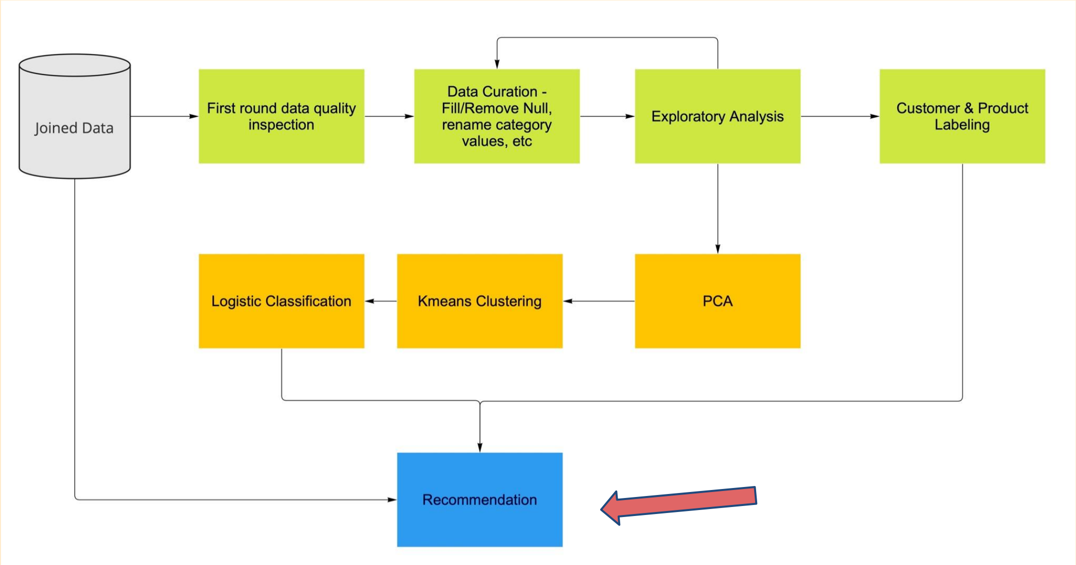


Logistic Regression

1. Goal is to predict whether the customer will purchase in the future (using the valuable customer label, Kmeans cluster, and PCA)
2. The predicted results has a accuracy rate of 0.64
3. From our validation dataset, our model is able to predict 80% of the future purchase



3. Data, Analysis & Result



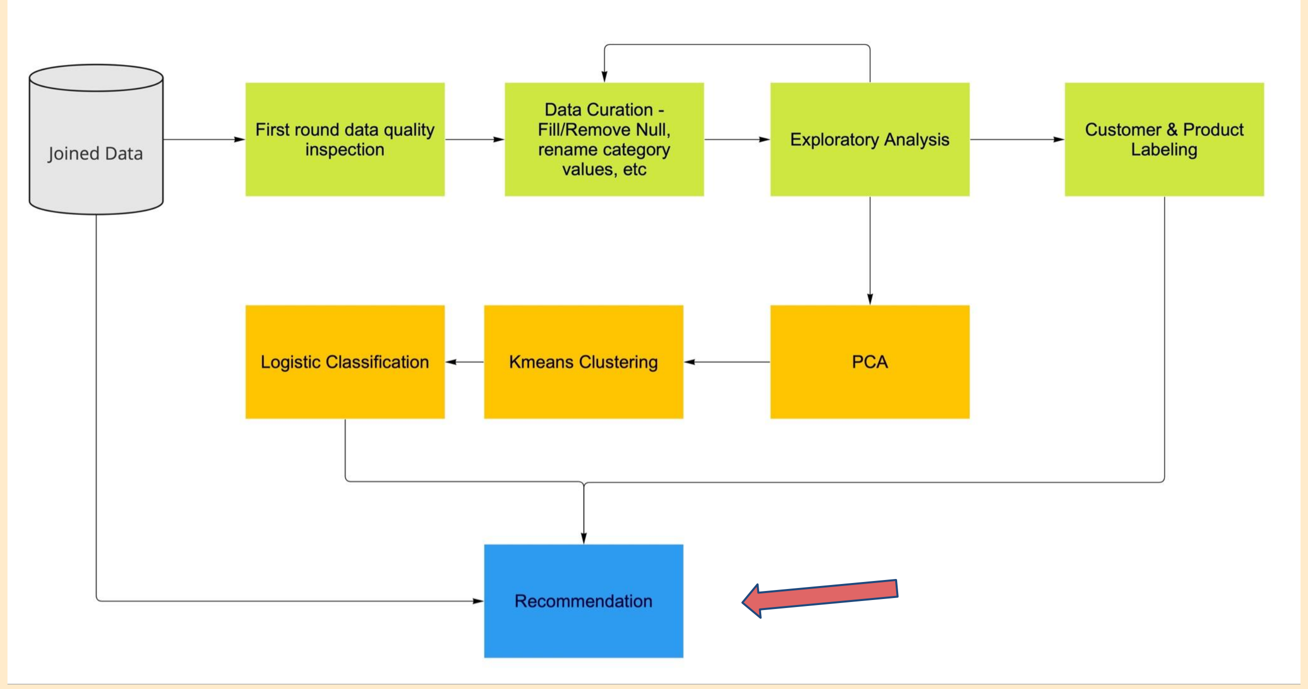
Recommendation- Will purchase

1. Those who will purchase -> **mapping to the cluster**
2. Top 3 **Most frequent** products **in 7 days** in each cluster

kmeans_pred	article_id
0	543035019
0	715024004
0	720504004
1	720504010
1	917300002
1	499334001

	article_id	customer_id	t_dat	postal_code	prod_name	kmeans_pred	val_cus	will_purchase	recommendation
0	917294003	-9158040424663740055	2020-09-19	5522c25dec3bba00a37129b9e5b9ee593f534d40007e44...	BLANKS JACK RELAXED LS TEE	4	0	1	[871517002, 876926001, 871517008]
1	917294005	3413429063454787034	2020-09-07	76071e1ea2a874b5d231fad4d528153760330d1956bc2b...	BLANKS JACK RELAXED LS TEE	0	0	1	[543035019, 715024004, 720504004]
2	917294003	-8845919207528614688	2020-09-06	9cfc6c0e96bb8b7441576581506a88573816ea0a3eddf0...	BLANKS JACK RELAXED LS TEE	4	0	1	[871517002, 876926001, 871517008]

3. Data, Analysis & Result



Recommendation- Extreme Cases

What if the customer is not registered on H&M or the customer is not ready to purchase anything (will_purchase = 0)

- Use postal code and Kmean clusters to rank the top purchases
- Retrieve top three items purchased as our recommendation
- Didn't consider time range because quickly catch the new customer's attentions to checkout is important

```
recommend = get_rcmd_extreme(customer_id = 4965599760277118247 )
```

```
print(recommend)
```

```
[prod_name    CAMILLA OL OFFER
Name: 21155, dtype: object, prod_name    EDC LAURA LACE TOP
Name: 22741, dtype: object, prod_name    CAMILLA OL OFFER
Name: 21163, dtype: object]
```

```
recommend_notRgstr = get_rcmd_extreme(customer_id = None, postal_code= '2c29ae653a9282cce4151bd87643c907644e09541abc28ae87dea0d1f6603b1c')
```

```
print(recommend_notRgstr)
```

```
[prod_name    7p Basic Shaftless
Name: 38344, dtype: object, prod_name    FANTASITC LOW PRICE TEE VP3
Name: 18499, dtype: object, prod_name    Scallop 5p Socks
Name: 37488, dtype: object]
```

4. Issues & Difficulties

Data is enormous

There is 20Gb of data.
VM could not handle.

?% accuracy

We can only recommend item, failed to provide or validate the accuracy.

Cold Start

No click history, cookie, or previous data for reference.

5. Future Directions

Limitation

- Disregard the time sensitivity of the dataset.
- Cannot validate whether the customer will purchase the recommended item.
- Failed to provide recommendations for new by account.

Future Work

- Combine time-series forecasting methods with the ML models to make predictions with time effect.
- Instead of using PCA, use model forward selection.
- Provide top 10 most popular item that was sold past month.

Thank you!

Q&A


70 %

Teammates

Chien-Hsin Lee

Briefly elaborate on what you want to discuss.

Jia-Jia Yu

Briefly elaborate on what you want to discuss.

Raymond Su

Briefly elaborate on what you want to discuss.

Xinbo Lu

Briefly elaborate on what you want to discuss.

Write your topic or idea

Add a main point

Elaborate on what you want to discuss.

Add a main point

Elaborate on what you want to discuss.

Add a main point

Elaborate on what you want to discuss.

Add a main point

Elaborate on what you want to discuss.

Go Back to Agenda Page

Write your topic or idea

Write a column name	Write a column name	Write a column name	Write a column name	Write a column name
Double click to add text				

Write your topic or idea

Add a main point

Briefly elaborate on what you want to discuss.

Add a main point

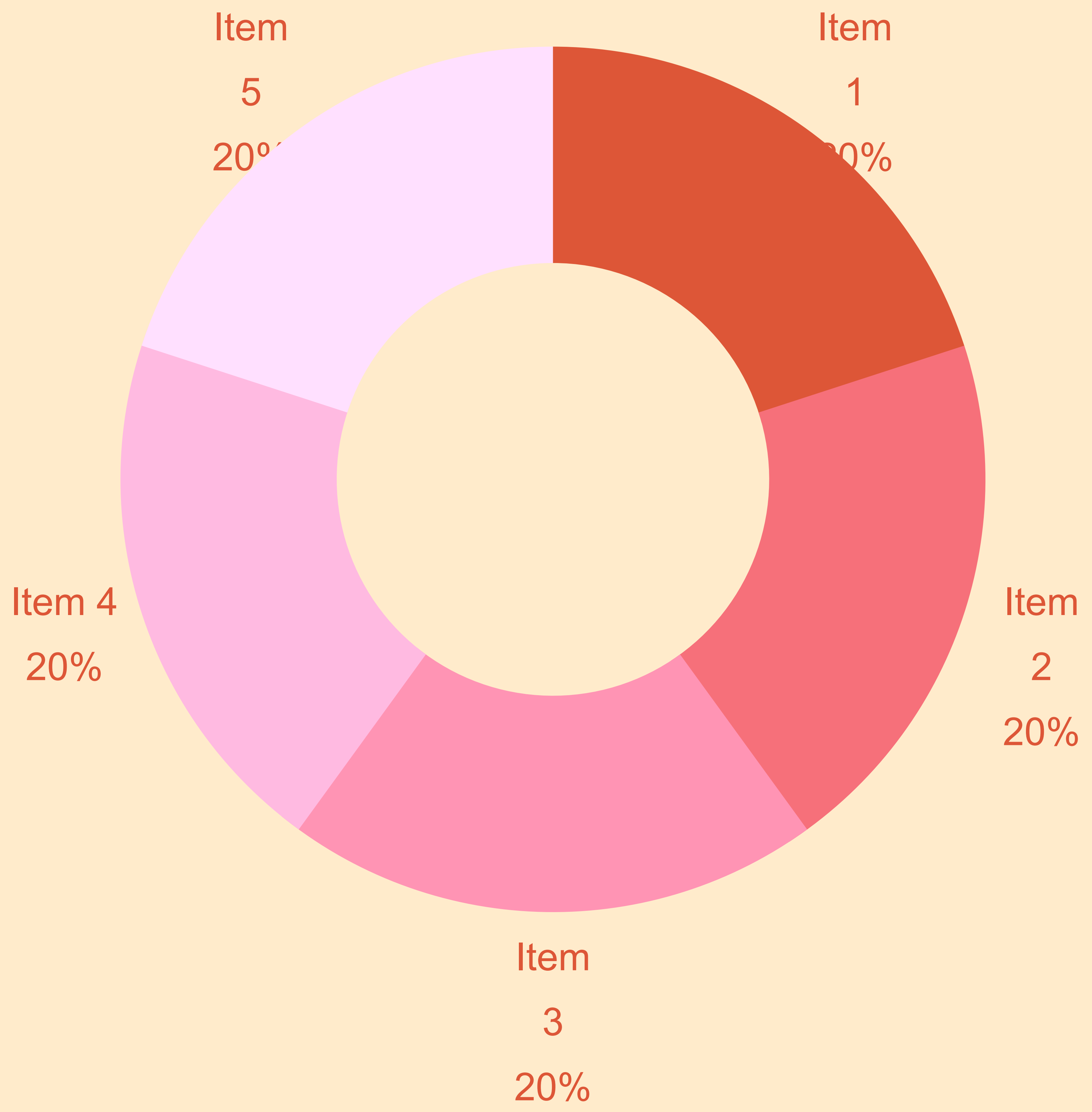
Briefly elaborate on what you want to discuss.

Add a main point

Briefly elaborate on what you want to discuss.

Add a main point

Briefly elaborate on what you want to discuss.



Write your topic or idea

Briefly elaborate on what you want to discuss.

Resource Page

Use these design resources in your Canva Presentation. Happy designing!

Don't forget to delete this page before presenting.

Write your topic or idea

Briefly elaborate on what you want to discuss.

Add a section header

Go Back to Agenda Page

Resource Page

Find the magic and fun in presenting with Canva Presentations. Press the following keys while on Present mode!

Don't forget to delete this page before presenting.

B for blur

C for confetti

D for a drumroll

O for bubbles

Q for quiet

X to close

Any number from 0-9 for a timer

Write an original statement or inspiring quote

— Include a credit or citation