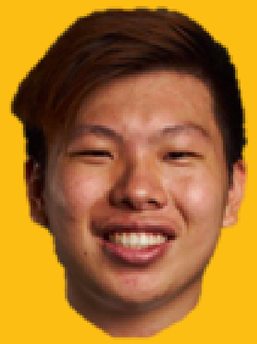
DSA3101 ASSIGNMENT 2

6 October 2020

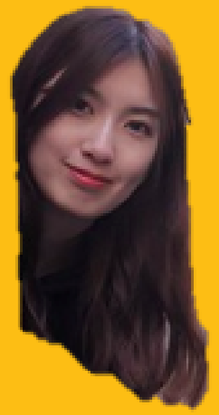
Meet the team!

Chan Yu Hang



Chua Cheng Ling

Chua Huaren



Cai An Qi

Joash Chin

Clarence Ong

ASSIGNMENT 2

- Customer Segmentation
- Principal component analysis
- Collaborative filtering

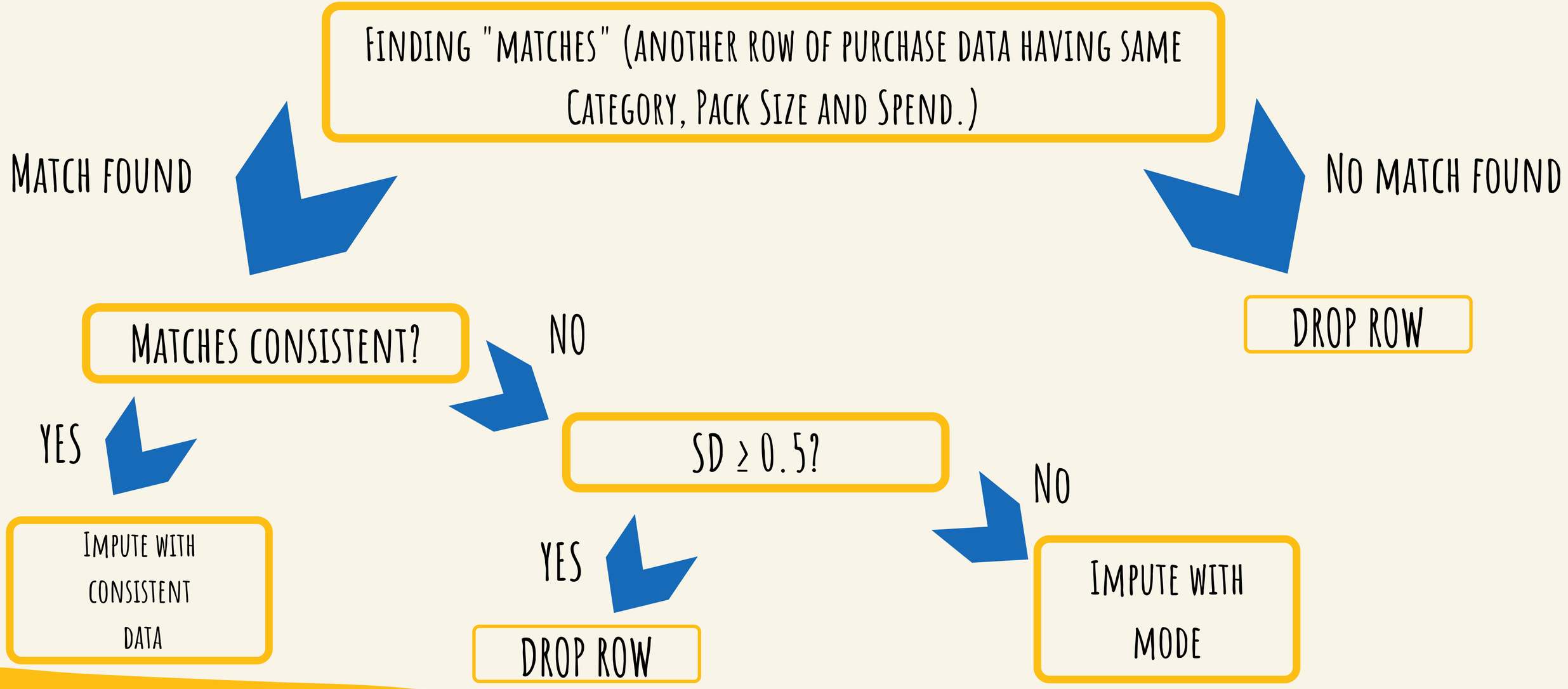
DATA IMPUTATION

Previously in Assignment 1, there were 49,539 instances of missing/mis-recorded data. Our team tried to use kNN imputation, however we were unable to do so and chose to remove the data instead as it was only 3.8% of the dataset.

DATA IMPUTATION

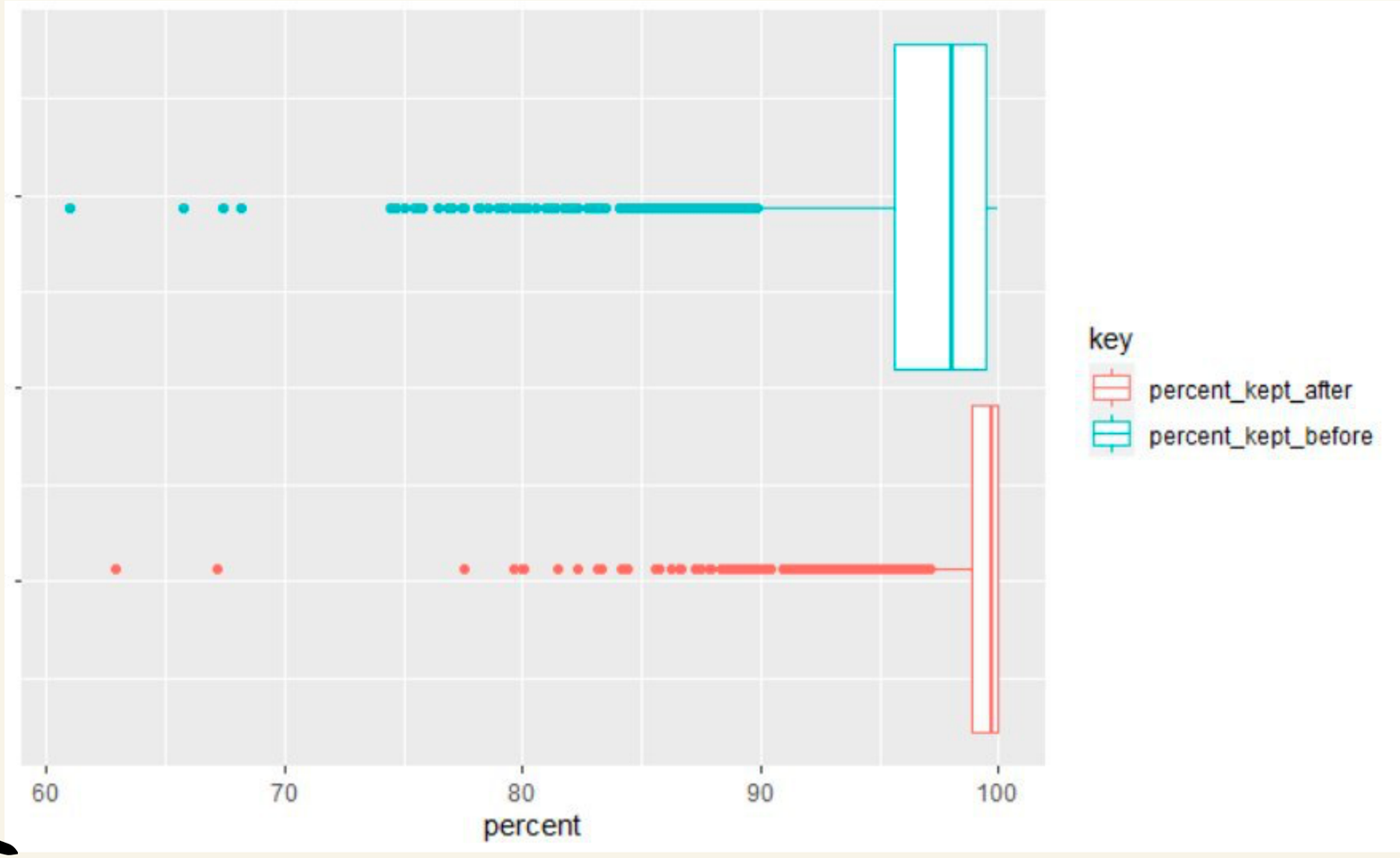
For assignment 2, we decided to do data imputation because removing all incorrect data would drop around <u>207 customers</u> (losing more than 10% of their purchase data), which could affect the accuracy of suggestions especially if there is some significance to the missing data

Conservative approach towards data imputation



EFFECTIVENESS?

THE MEAN INCREASE IN USABLE ROWS IS 9.1%!



Customers' Location

- North
- "Central"
- East Coast
- South

CENTRAL

	CSD	Rice	.....	Belacan
CUSTOMER A	$	$	.....	$
CUSTOMER B	$	$	.....	$

Customer segmentation - Principal Component Analysis

- <u>Low correlation</u> between the various components in the input vectors, making it difficult to use PCA to significantly reduce the number of dimensions used to represent each customer effectively

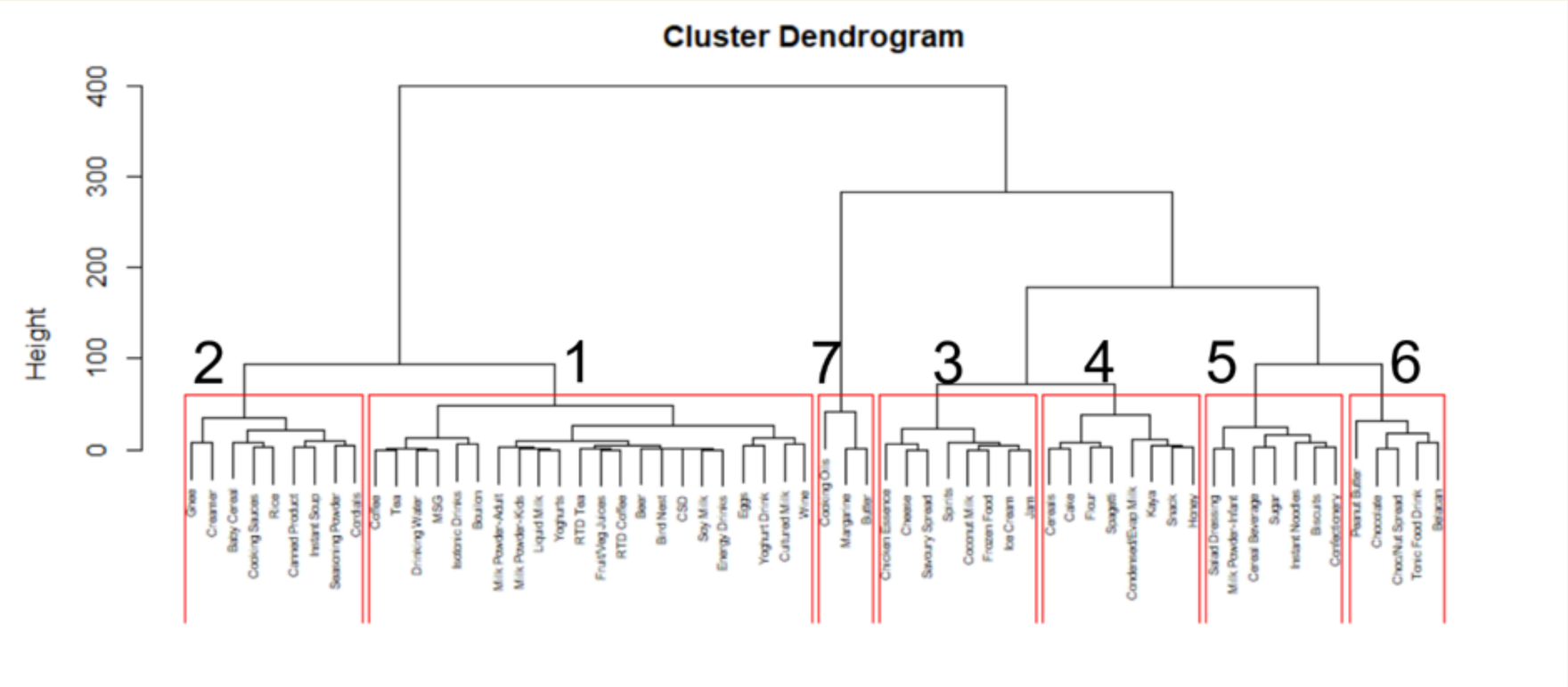
Top Absolute Correlations

Category	Category	
Flour	Sugar	0.406735
Beer	Isotonic Drinks	0.404076
Condensed/Evap Milk	Sugar	0.399785
Cooking Oils	Rice	0.391161
Isotonic Drinks	RTD Tea	0.[illegible]7523
Biscuits	Liquid Milk	0.3[illegible]821
CSD	Isotonic Drinks	0.36[illegible]97
Biscuits	Tonic Food Drink	0.352[illegible]5
	Cooking Oils	0.35[illegible]4
	Instant Soups	[illegible]548424
Beer	RTD Tea	0.345429
Cheese	Yoghurts	0.344703
Cooking Oils	Tonic Food Drink	0.343524
Yoghurt Drink	Yoghurts	0.326345
Rice	Sugar	0.325497
Salad Dressing	Spagetti	0.324520
Cheese	Instant Soup	0.323259
Beer	CSD	0.322699
Cooking Oils	Cooking Sauces	0.322407
Eggs	Sugar	0.322243

FAIL

"Hierarchical clustering"

- Idea: segment the different categories into <u>7 larger overarching categories</u> based on the Calories per 100g of each of the 62 original categories
- Agglomerative manner



clusters with a smaller label number value represent categories with lower values of Calories per 100g

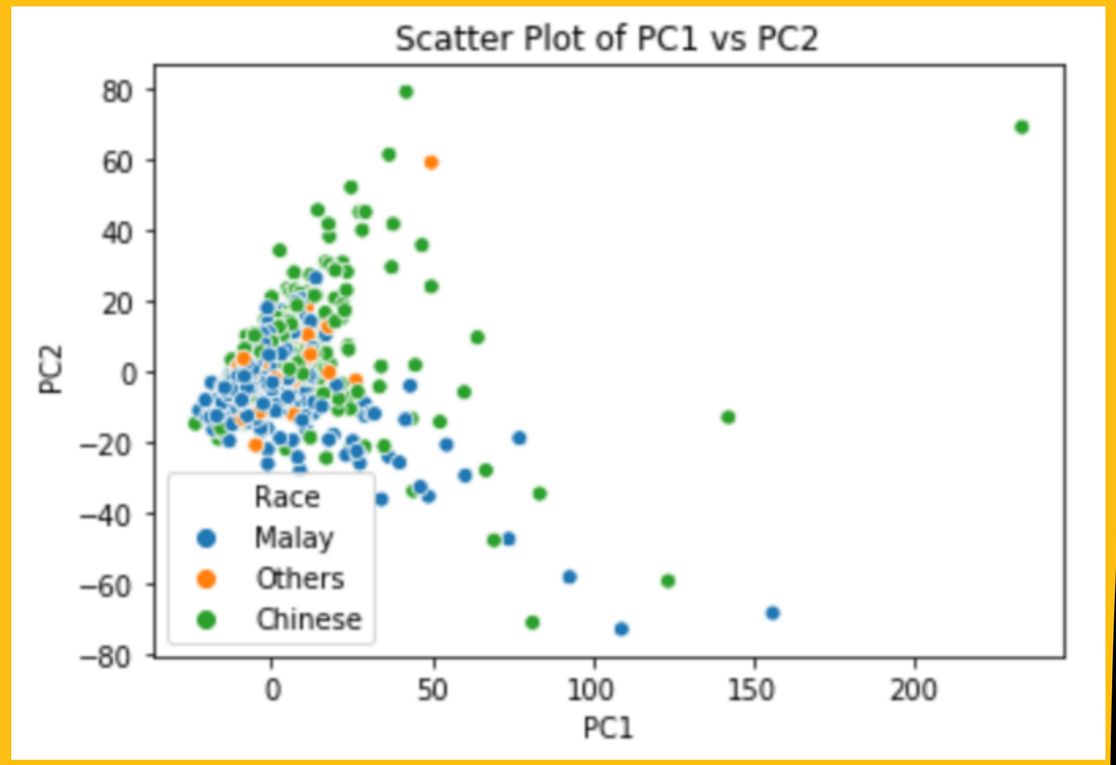
PRINCIPAL COMPONENT ANALYSIS

Using k = 2 (PC1 and PC2), our team has identified that:

1. PC1 is has **items of relatively high calories** like Chocolate, Peanut Butter and icecream which can be considered as comfort food.
2. Pc2 can be represented by **items of low calories** like soy milk, fruits and vegetable, and items that are **rich in carbohydrates**, like biscuits, instant noodles and cereal beverages

rank	PC1	PC2
1		0.827071
2		
3		
4		
5		0.791783
6	0.711918	
7	0.741586	

VERY IMPORTANT FINDING!



THERE SEEMS TO BE A RELATIONSHIP BETWEEN RACE AND HIS/HER PURCHASING HABITS!

FOR MALAYS, THERE IS A NEGATIVE LINEAR RELATIONSHIP BETWEEN THE 2 PC

FOR CHINESE, THERE ARE VARYING PURCHASE PATTERNS!

IDEA!

Since we have established a strong relationship between a customer's race and their purchasing habit, let's apply it to our recommendation engine!

PROBLEM WITH THE DATA

We are only provided with the 62 categories. Using these categories will not provide any meaningful insight.

SOLUTION

Create arbitrarily unique items categorized by the spend/volume ratio. We generated a list of 36,237 items.

WEB APPLICATION

- Historical purchases
- Discounts and promos
- Historical Views
- Trending Purchases
- Collaborative filtering

NAIVE MODELS

Global mean to predict every item's ratings

RMSE = 1.59

Panel id's mean to predict for the corresponding panel id's items' ratings

RMSE = 1.56

Item's mean to predict for the corresponding item's ratings

RMSE = 1.55

→ RMSE for collaborative filtering = 1.52

CONCLUSION

Given slightly more information (either knowing the panel/product mean) can obtain an even lower RMSE

SUMMARY

Data imputation

Since reco engine is mostly about customer satisfaction, we do not want to risk losing the 10% of customers which could be costly

Principal Component Analysis

Race has a strong correlation on the purchasing habits of the customer.

Recommendation Engine

- deployable and provides key insights
- EMA used so model learns with time
- Collab filtering model (matrix factorisation) outperforms naive models

THANK YOU!