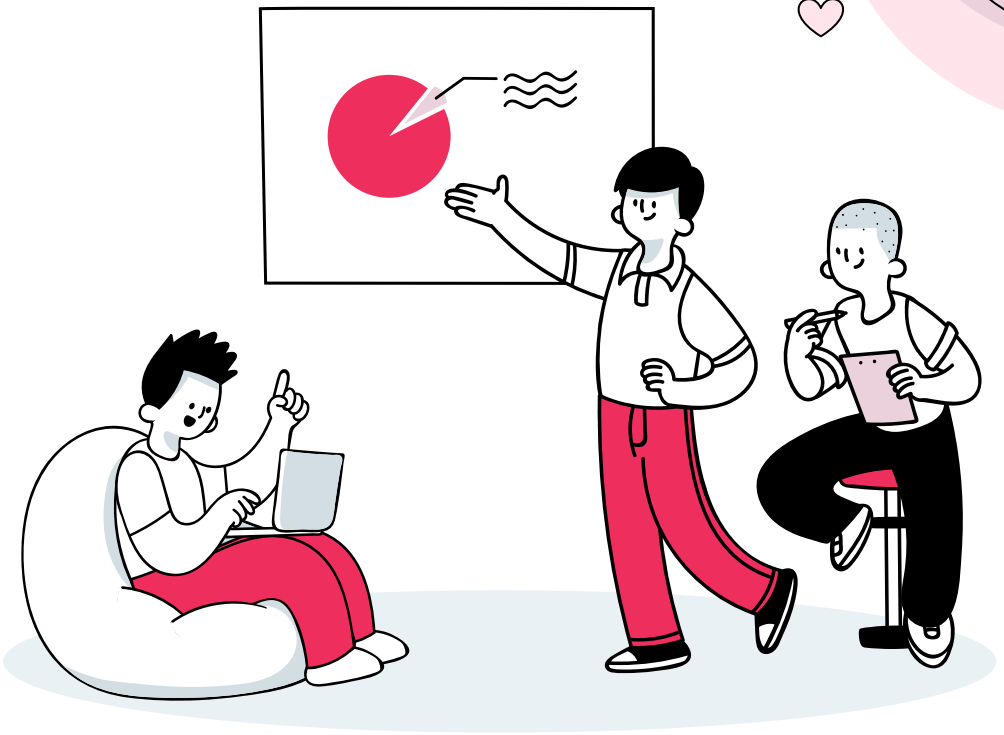
Speed Dating

Here is where you can meet your true love

— ki, Sophia, Rainfield



Agenda

01
Dataset &
Business Value

02
Data Collection

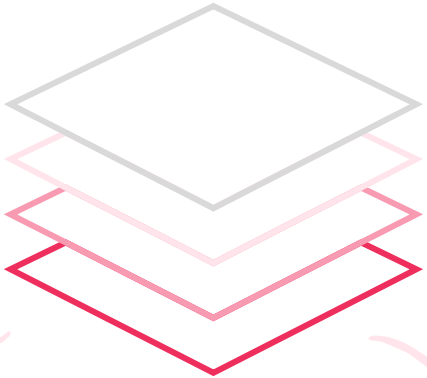
03
Data Preprocessing

04
Model Creation

05
Model Evaluation

06
Conclusion &
Improvements

Our Dataset



Data Source

"Speed Dating Experiment" from Kaggle.com

Study Goal

Gender Differences in Mate Selection

Experiment Design

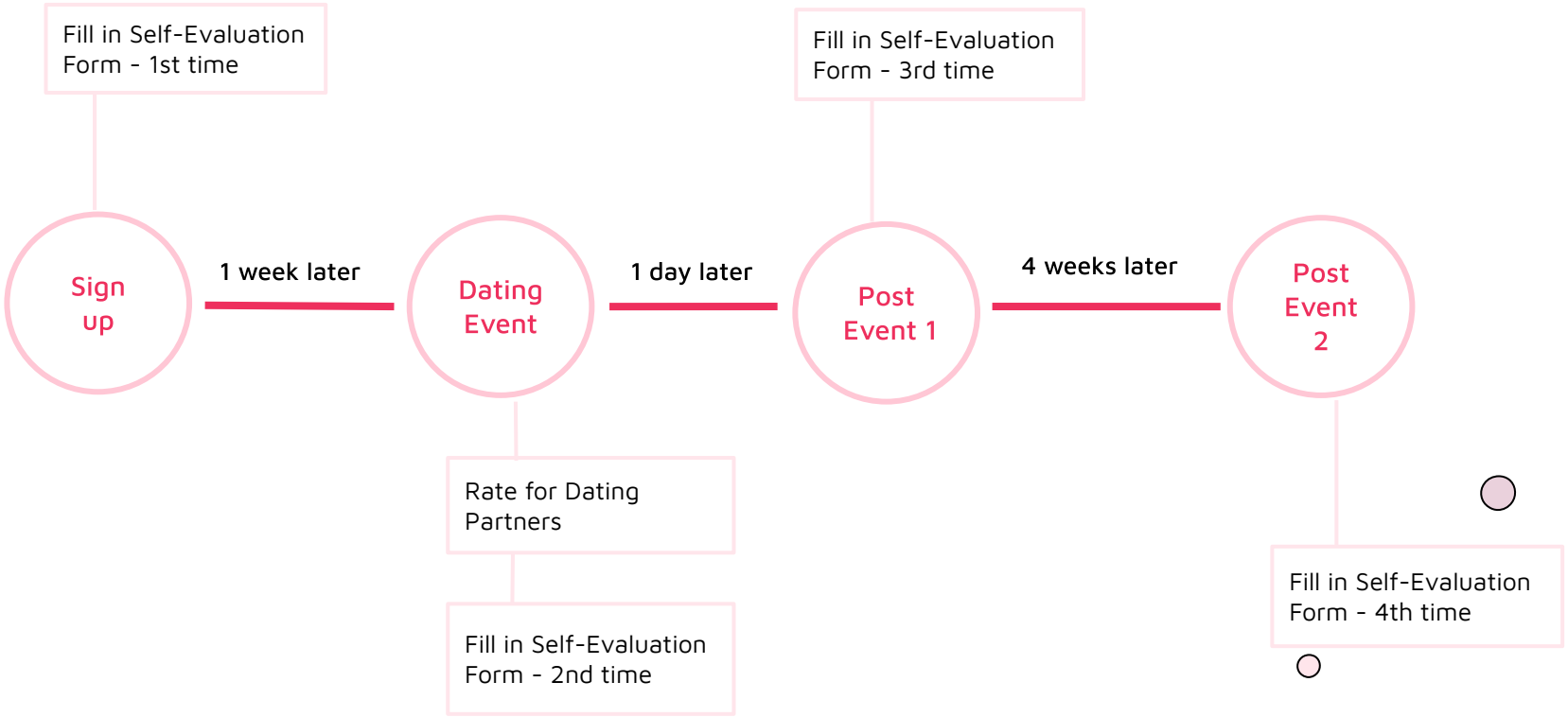
Serial Experimental Speed Dating Events from 2002-2004

Data Attributes

Demographics, Dating Habits, Lifestyle, Self Perceptions, Ratings on the Dates

...

Experimental Flow



Business Questions

Clusters

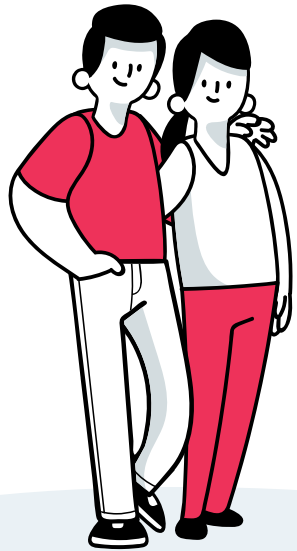
What are the Clusters with Highest Match Rate ?

Attributes

Which attributes are the most desirable in potential mate selection ?

Business Value

Offer solutions for speed-dating agencies and dating apps to improve match success rates



Preprocessing

Data Extraction

Extract Useful Attributes

Feature Selection - Part 1

Select 30 out of 200 Features which best describe participants' personalities (Demographics, Interests, Self Perceptions)

Data Cleaning

Impute Missing Values with Mean

Feature Scaling

Apply Standard Scaler

Feature Selection - Part 2

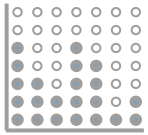
Keep 5 Self Perceptions Features & Use K-Means to pick extra 3 Features (C^{25}_3) with Highest Silhouette Score

Feature Selection - Part 3

Adopt PCA to Reduce Dimensions in Remaining 22 Features

Merge Datasets

Modelling



Confusion Matrix

Display Match Rate
between Clusters

Clustering

Feature Selection

30 Features

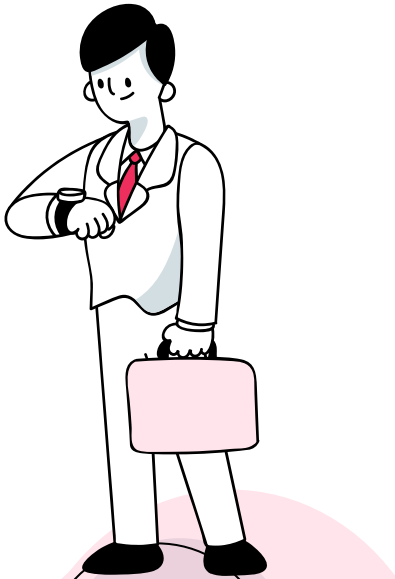
10 Features

with Higher Silhouette Score

$30C_{10} =$

30,045,015

Too many Combinations



Feature Selection

01

Manual Selection

5 Self-Perception Features:
"Attractiveness", "Sincerity", "Fun"

02

Selection with K-Means

Pick extra 3 Features (C^{25}_3) with Highest Silhouette Score in K-Means

03

PCA

Adopt PCA to Reduce Dimensions in remaining 22 Features into 2 Features

04

Merge Datasets

Total 10 Features

Final Dataset

	iid	id	gender	idg	condtn	wave	round	position	positin1	order	...	sinc3_3	intel3_3	fun3_3	amb3_3	attr5_3	sinc5_3	intel5_3	fun5_3	amb5_3	class
0	1	1.0	0	1	1	1	10	7	NaN	4	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
1	1	1.0	0	1	1	1	10	7	NaN	3	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
2	1	1.0	0	1	1	1	10	7	NaN	10	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
3	1	1.0	0	1	1	1	10	7	NaN	5	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
4	1	1.0	0	1	1	1	10	7	NaN	7	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
5	1	1.0	0	1	1	1	10	7	NaN	6	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
6	1	1.0	0	1	1	1	10	7	NaN	1	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
7	1	1.0	0	1	1	1	10	7	NaN	2	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
8	1	1.0	0	1	1	1	10	7	NaN	8	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
9	1	1.0	0	1	1	1	10	7	NaN	9	...	7.0	7.0	7.0	7.0	NaN	NaN	NaN	NaN	NaN	f0
10	2	2.0	0	3	1	1	10	3	NaN	10	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
11	2	2.0	0	3	1	1	10	3	NaN	9	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
12	2	2.0	0	3	1	1	10	3	NaN	6	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
13	2	2.0	0	3	1	1	10	3	NaN	1	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
14	2	2.0	0	3	1	1	10	3	NaN	3	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
15	2	2.0	0	3	1	1	10	3	NaN	2	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
16	2	2.0	0	3	1	1	10	3	NaN	7	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
17	2	2.0	0	3	1	1	10	3	NaN	8	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
18	2	2.0	0	3	1	1	10	3	NaN	4	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0
19	2	2.0	0	3	1	1	10	3	NaN	5	...	6.0	9.0	9.0	4.0	NaN	NaN	NaN	NaN	NaN	f0

20 rows × 196 columns

K-Means

Which Algorithm we choose?

OPTICS

DBSCAN

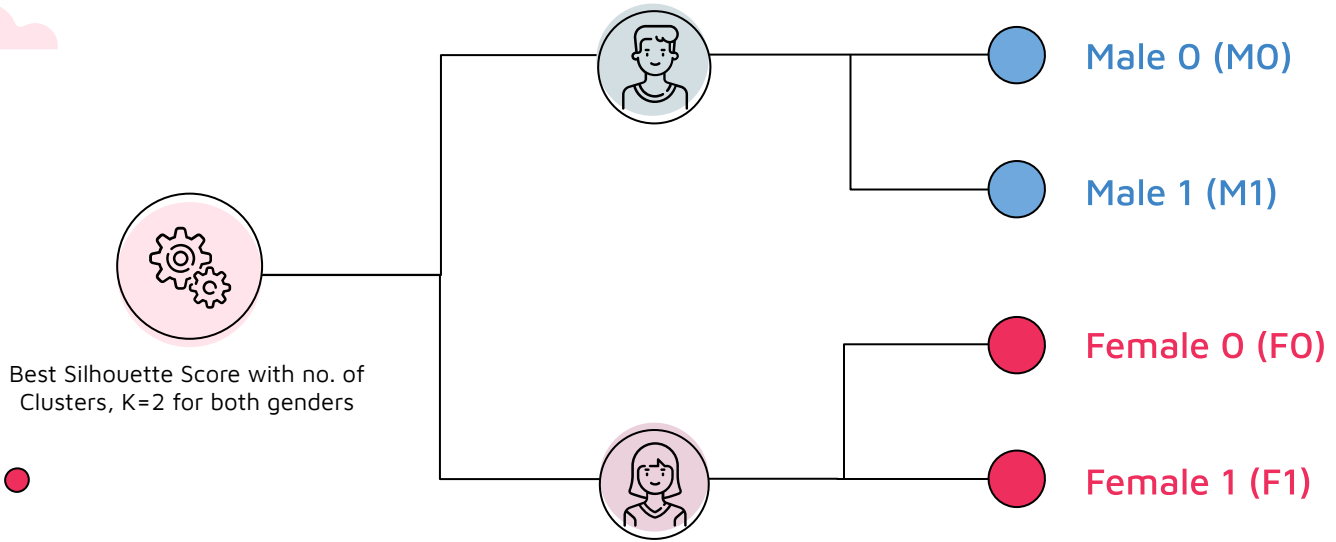
HDBSCAN

Model Evaluation

$$-1 < \text{Silhouette Coefficient} < 1$$

Analysis

Distribution



Match Result

	F0	F1
M0	17.3% 161 / 929	**19.2%** 332 / 1731
M1	11.1% 66 / 595	14.1% 131 / 929

F1M0 Combination has the Highest Match Rate

Dominant Class - By Male

	F0	F1
M0	17.3% 161 / 929	19.2% 332 / 1731
M1	11.1% 66 / 595	14.1% 131 / 929

F1 is the Dominant Class in mate selection of Males

Dominant Class - By Male

	F0	F1
M0	17.3% 161 / 929	19.2% 332 / 1731
M1	**11.1%** **66 / 595**	**14.1%** **131 / 929**

F1 is the Dominant Class
in mate selection of Males

Dominant Class - By Female

	F0	F1
M0	17.3% 161 / 929	19.2% 332 / 1731
M1	11.1% 66 / 595	14.1% 131 / 929

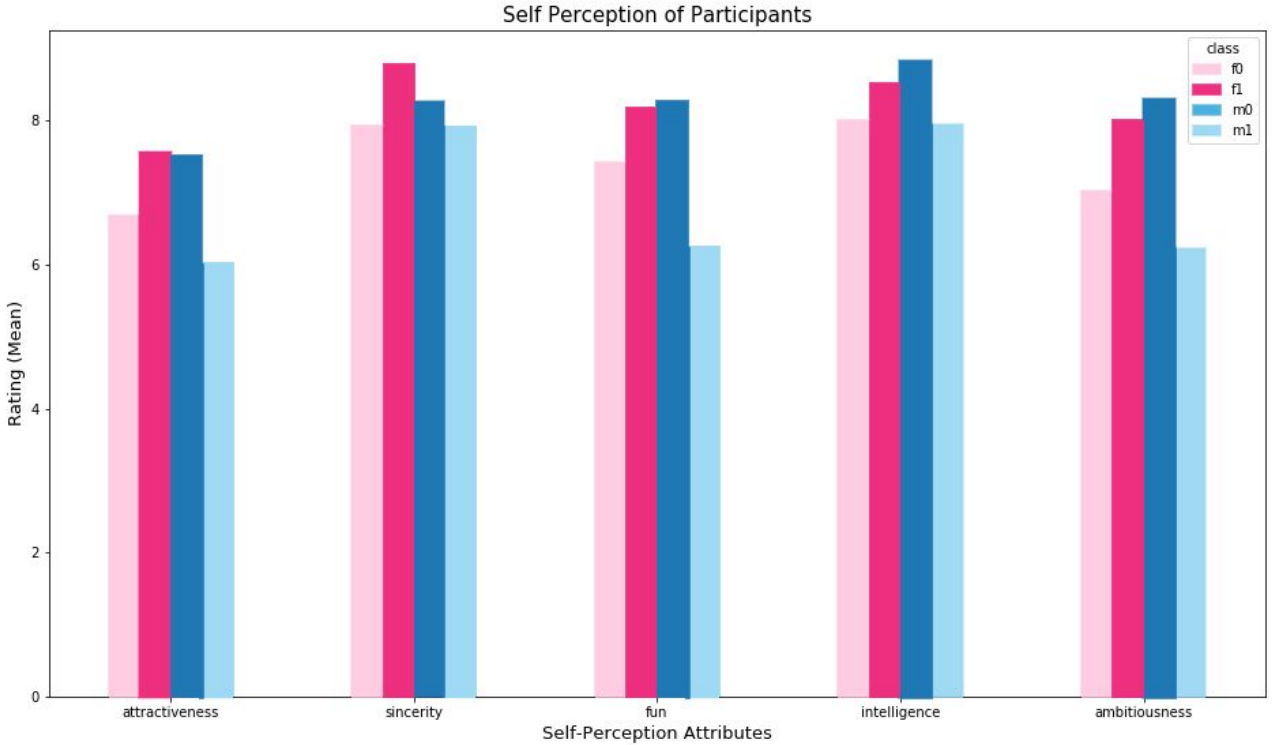
M0 is the Dominant Class in mate selection of Females

Dominant Class - By Female

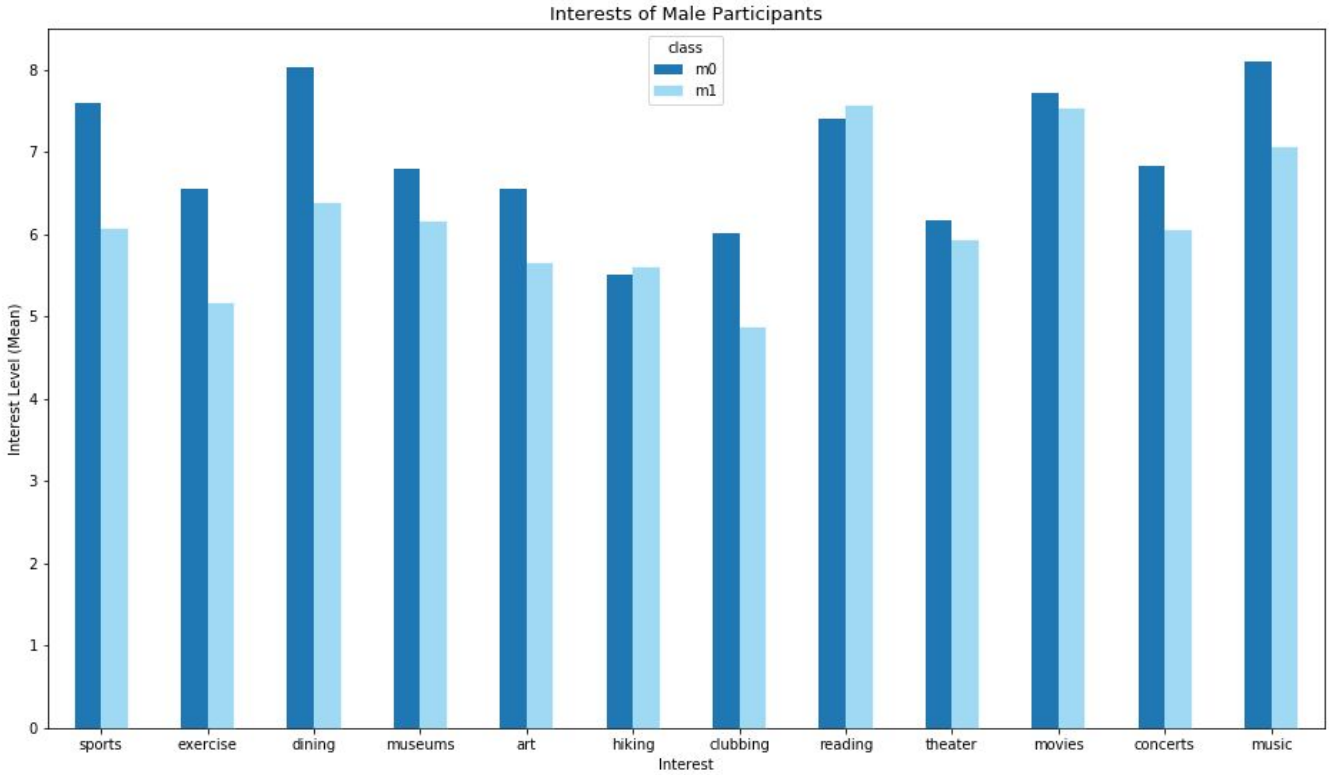
	F0	F1
M0	17.3% 161 / 929	**19.2%** **332 / 1731**
M1	11.1% 66 / 595	**14.1%** **131 / 929**

M0 is the Dominant Class in mate selection of Females

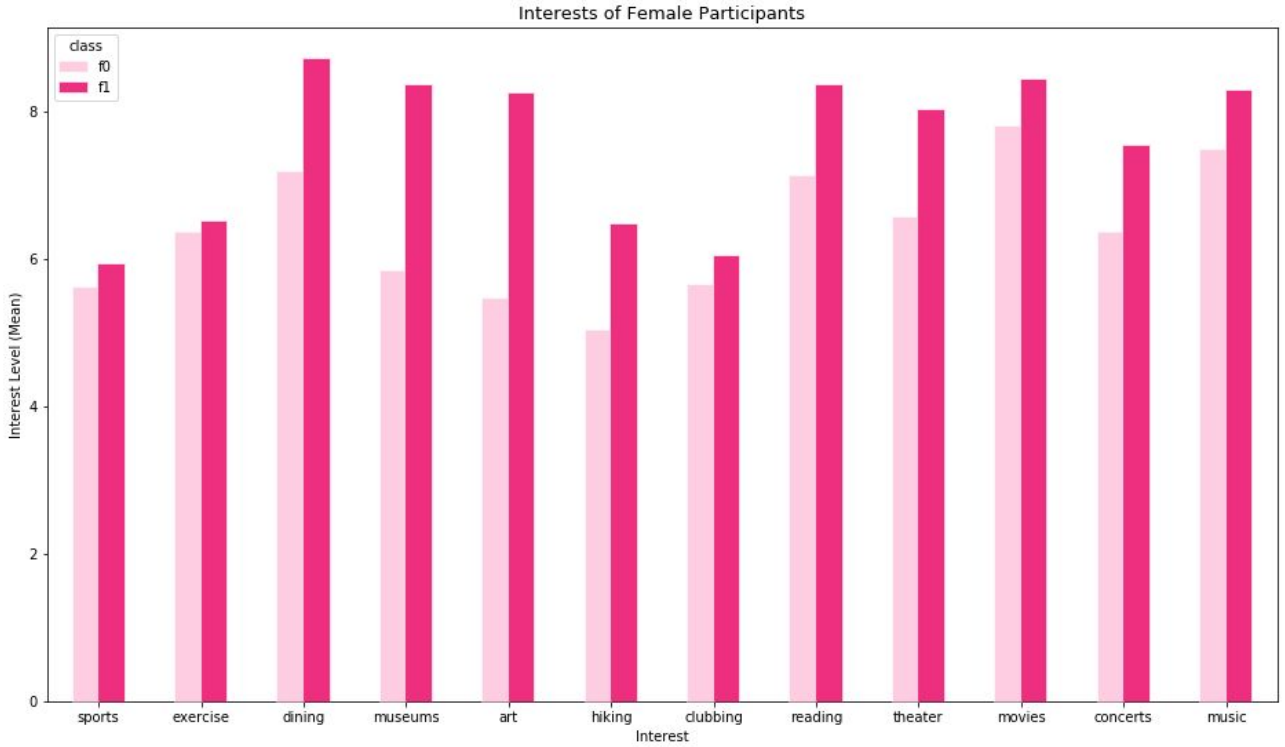
Self Perception Analysis



Interest Analysis - by Male Clusters



Interest Analysis - by Female Clusters



Secrets to Increase your Match Success Rate

Self Confidence

x

Diverse Interests

Recommendations to Dating App

	F0	F1
M0	17.3% 161 / 929	19.2% 332 / 1731
M1	11.1% 66 / 595	14.1% 131 / 929

Avoid 'F0M1' Combinations!

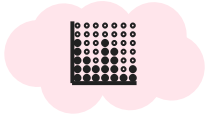
Limitations & Improvements

Model Limitation

Try out more Clustering Models (Centroid/ Density-based)

Limited Samples

Small Sampling Size & Sample Variety



Computing Power

Limited time to attempt 30 million Feature Combinations

Questions?