Recommendation Systems

Team Name - Black Coffee

Team Members
IMT2020059 - Manas Agrawal
IMT2020083 - Mukul Gupta
IMT2020131 - Amar Pratap Singh
IMT2020129 - Deep Shashank Pani

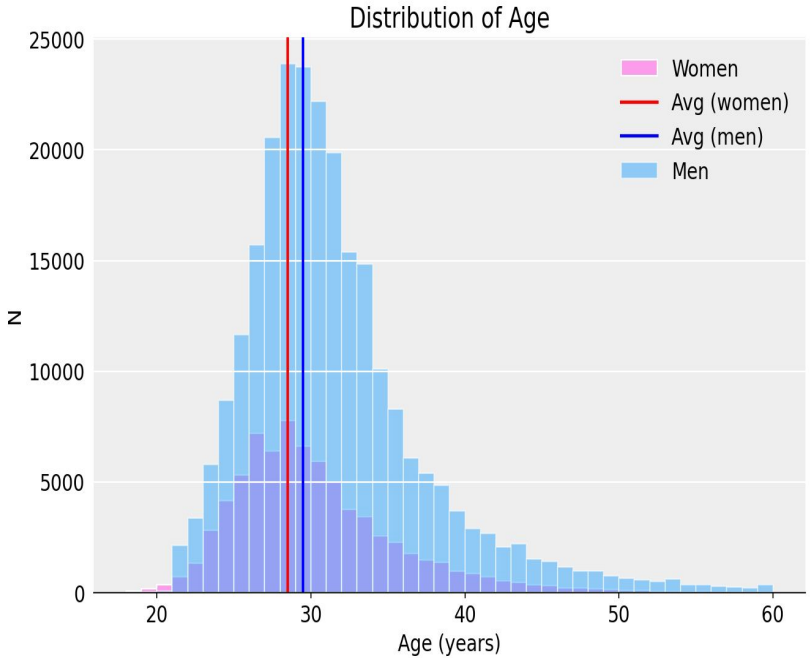
Dataset

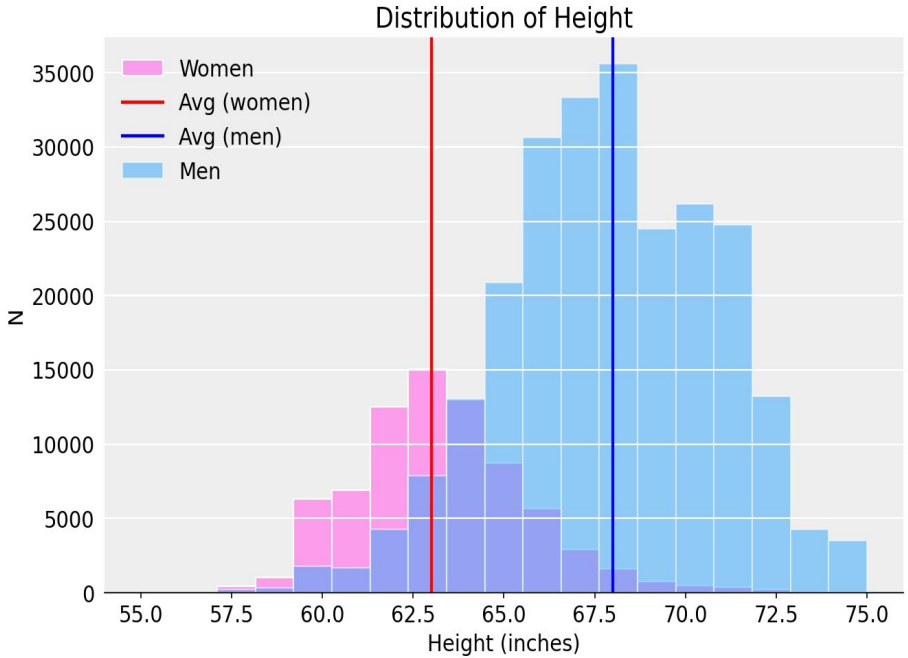
- Our dataset was sourced from SimplyMarry.com, a matrimonial website.
- We employed Puppeteer, a python library, to automate the process of web scraping.
- Our methodology involved systematically navigating through profiles on SimplyMarry.com and extracting relevant data fields.
- The dataset included columns like about1, about2, age1, age2, caste1, caste2, etc.
  - Suffix “1” denote that the data is about the user.
  - Suffix “2” denote the requirements which the user wants in his/her partner.

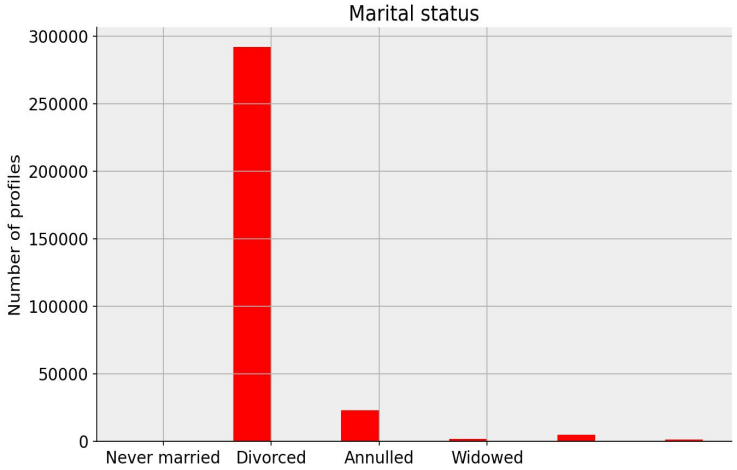
Data Preprocessing

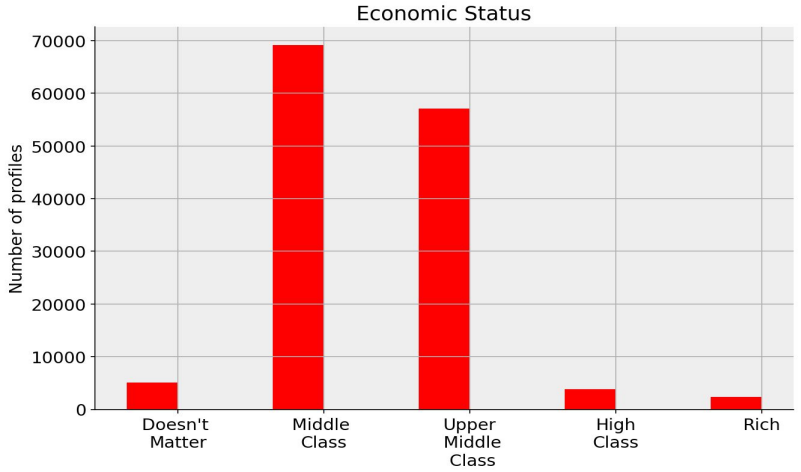
- There were 325688 rows and 67 columns in the dataset.
- There were many columns with categorical data which were handled as follows:
  - Converted into numerical data
  - Nan Values were handled by analysing the distribution of data along each column and then either nan values were replaced by some value (as per distribution) or the column was removed.
  - Again the numerical values were converted into categorical data for further preprocessing.
- List of columns with suffix “1” and suffix “2” were merged and a proper sentence was formed which was then replaced with the data present in about1 and about2 respectively.
- Many redundant columns were removed.

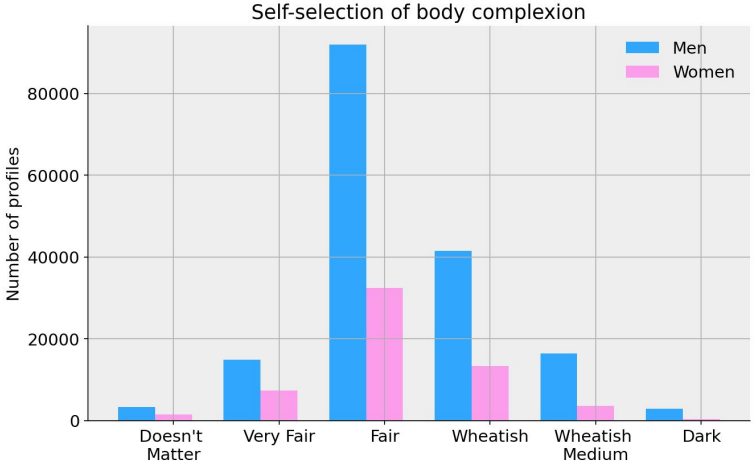
Data Distribution

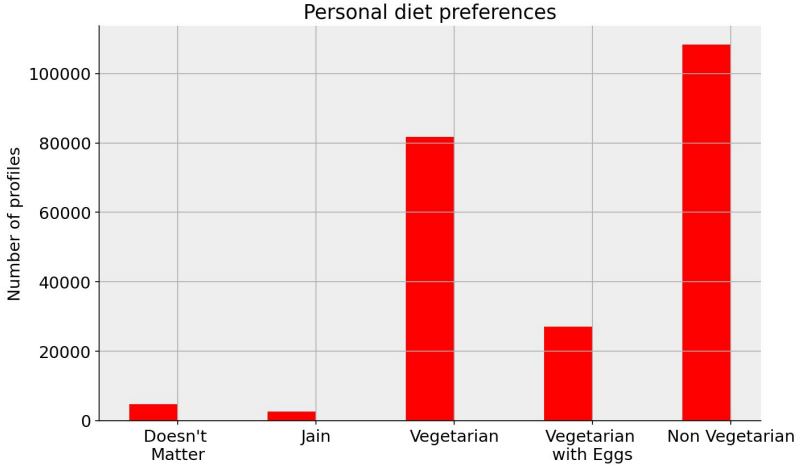


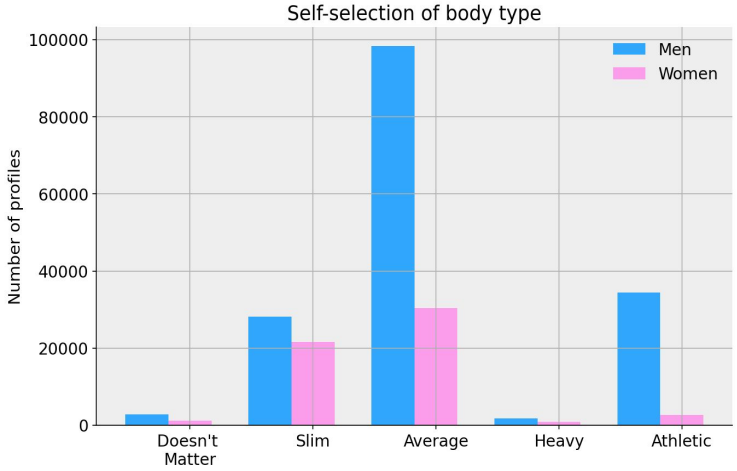


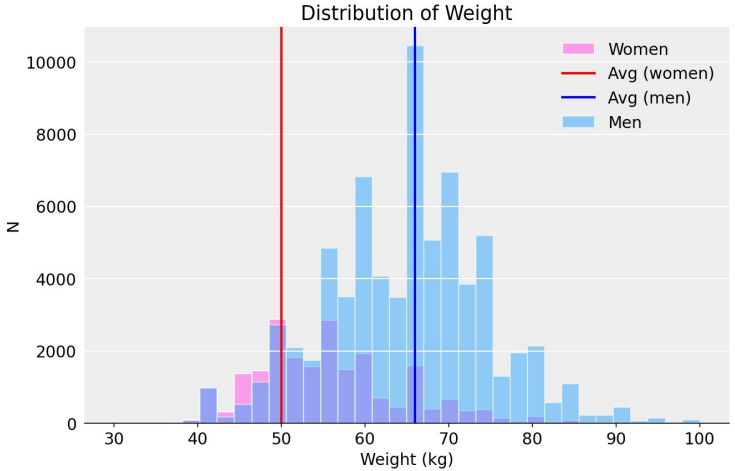

Marital status
Number of profiles
300000
250000
200000
150000
100000
50000
0
Never married
Divorced
Annulled
Widowed

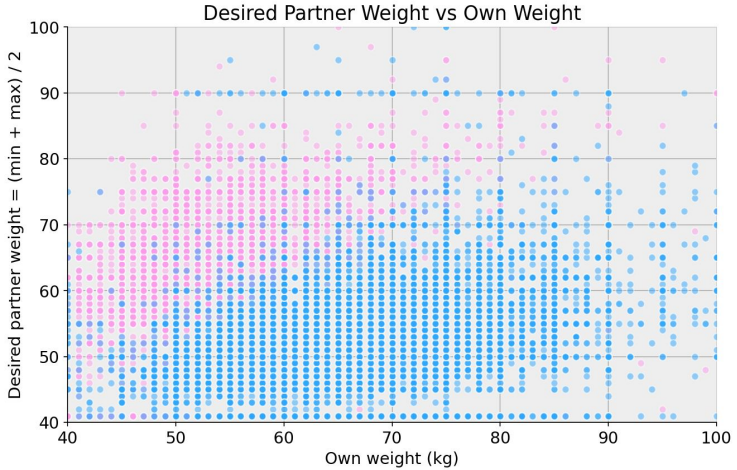

Economic Status
Number of profiles
70000
60000
50000
40000
30000
20000
10000
0
Doesn't Matter
Middle Class
Upper Middle Class
High Class
Rich

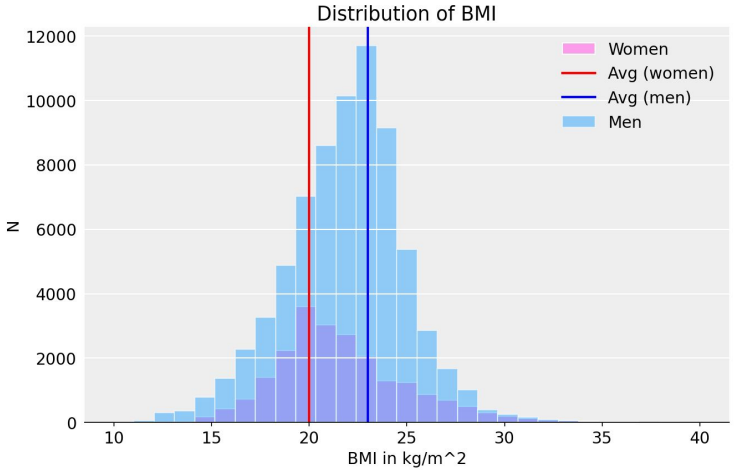

Self-selection of body complexion
Men
Women
Number of profiles
80000
60000
40000
20000
0
Doesn't Matter
Very Fair
Fair
Wheatish
Wheatish Medium
Dark


Personal diet preferences
Number of profiles
100000
80000
60000
40000
20000
0
Doesn't Matter
Jain
Vegetarian
Vegetarian with Eggs
Non Vegetarian


Self-selection of body type
Men
Women
Number of profiles
100000
80000
60000
40000
20000
0
Doesn't Matter
Slim
Average
Heavy
Athletic


Distribution of Weight
Women
Avg (women)
Avg (men)
Men
N
10000
8000
6000
4000
2000
0
30
40
50
60
70
80
90
100
Weight (kg)


Desired Partner Weight vs Own Weight
Desired partner weight = (min + max) / 2
100
90
80
70
60
50
40
40
50
60
70
80
90
100
Own weight (kg)


Distribution of BMI
Women
Avg (women)
Avg (men)
Men
N
12000
10000
8000
6000
4000
2000
0
10
15
20
25
30
35
40
BMI in kg/m^2

Models Used

- Content-Based Filtering
- Item-Item Collaborative Filtering
- Multi-Arm Bandit (ε greedy)
- Multi-Arm Bandit (UCB)

Content Based Filtering

- Cosine Similarity Calculation:
  - Computed cosine similarity between male preferences and female details.
  - Also calculated cosine similarity between female preferences and male details.
- Average Similarity Calculation:
  - Determined the average similarity score by averaging the cosine similarities from both comparisons.
- Sorting:
  - Sorted all profiles based on their average similarity scores in descending order.
- Top Recommendations:
  - Recommended the top 5 profiles with the highest average similarity scores.

Item-Item Collaborative Filtering

For our recommendation system, item is user’s preference.

Found Correlation Matrix of User Details

Based on user’s past experience, we are finding similar users/ dissimilar users.

From that pool of users, we are finding the top k users with highest similarity score.

Multi-Arm Bandit: Epsilon Greedy

Clustering the users based on similarity.

Each cluster is an arm.

- For initial training purpose, we are using a simulation to provide feedback/reward to the recommended cluster.
- At inference stage, we are prompting user to provide feedback.

For reward system, we take into account mutual compatibility and user rating

Reward belongs to the cluster

Novelty

- Incorporation of Family Dynamics:
  - Unlike traditional matchmaking systems, we factor in family background and dynamics such as joint or nuclear family structures, father's occupation, etc, thus, recognizing the significant role they play in marital relationships.
- Customized Recommendation Criteria:
  - Our system allows users to customize their preferences based on cultural, social, and personal values, enabling tailored recommendations that align closely with individual expectations.

Thank You