Project - Advance Business Analytics with R

Abhishek Jhunjhunwala

11/23/2020

PROBLEM STATEMENT

Develop strategies or features that can increase the user engagement and traffic on the app.

DATASET

The Dataset contains information about the profiles of around 60K users on the online dating app – OK Cupid.

The Dataset contains information entered by the user on the app while creating their profile.

It contains information such as demographics, likes, dislikes, interests and self description of the users.

INTRODUCTION

The idea is to develop an algorithm that can show relevant or similar profiles to users based on the profile data entered by them on the app.

We believe that this can increase the interest level of the users as they see relevant profiles on the app.

This will help increase the user engagement and traffic on the app.

FEATURES

The following information is available for each user in the dataset:

Body_type - rather not say, thin, overweight, skinny, average, fit, athletic, jacked.

Diet - mostly/strictly; anything, vegetarian, vegan, kosher, halal, other.

Drinking habit - very often, often, socially, rarely, desperately, not at all.

Drug abuse - never, sometimes, often.

Education - graduated from, working on, dropped out of; high school, two-year college, university, masters program, law school, med school, Ph.D program, space camp.

Height - inches

Income - (US $, -1 means rather not say) -1, 20000, 30000, 40000, 50000, 60000 70000

Job - student, art/music/writing, banking/finance, administration, technology, construction, education, entertainment/media, management, hospitality, law, medicine, military.

Offspring - has a kid, has kids, doesnt have a kid, doesn't want kids; ,and/,but might want them.

Orientation - straight, gay, bisexual.

Pets - has dogs, likes dogs, dislikes dogs; and has cats, likes cats, dislikes cats.

Religion - agnosticism, atheism, Christianity, Judaism, Catholicism, Islam, Hinduism, Buddhism, Other.

Data Cleaning and Transformation

The dataset was treated for missing values in the following ways:

The rows or data points that had a lot of missing values (more that half the columns) were dropped.

The columns which had a lot of missing values were dropped.

The columns with a few missing values were treated using mode if the data was categorical and median if the data was numeric.

DUMMY VARIABLES

All categorical variables were converted to dummy variables which resulted in a total of around 500 variables.

Top 35 variables with the most variance were selected out of the 500 variables.

These 35 variables were used for creating clusters and further models.

CLUSTERING

We start by finding groups or categories among users that can be used to develop more meaningful strategies for increasing user engagement.

The idea is to show profiles of users to each other among the same group or category.

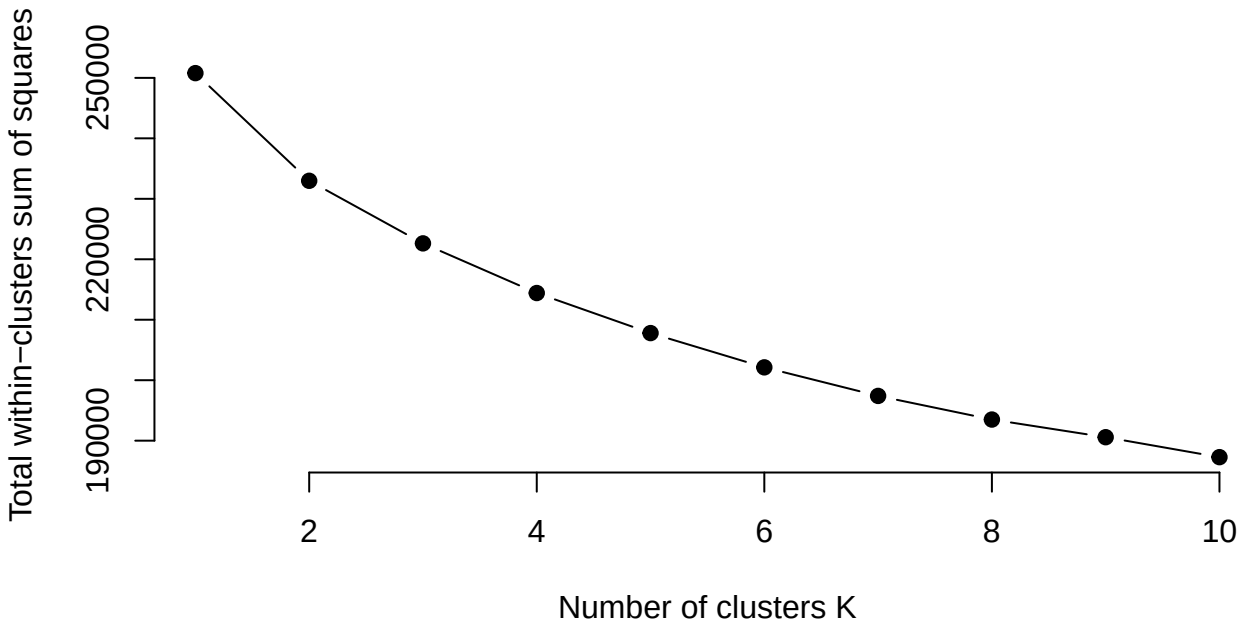
For example, if a user belongs to Group A, he/she will be able to see only the profiles of other users in Group A.

Unsupervised learning techniques - k-means clustering

The k-means clustering algorithm was used to find groups or categories among users.

The elbow diagram based on the within sum of squares distances was used to decide the number of clusters.

From the elbow diagram, it is clear that there should be 5 clusters or groups of users.

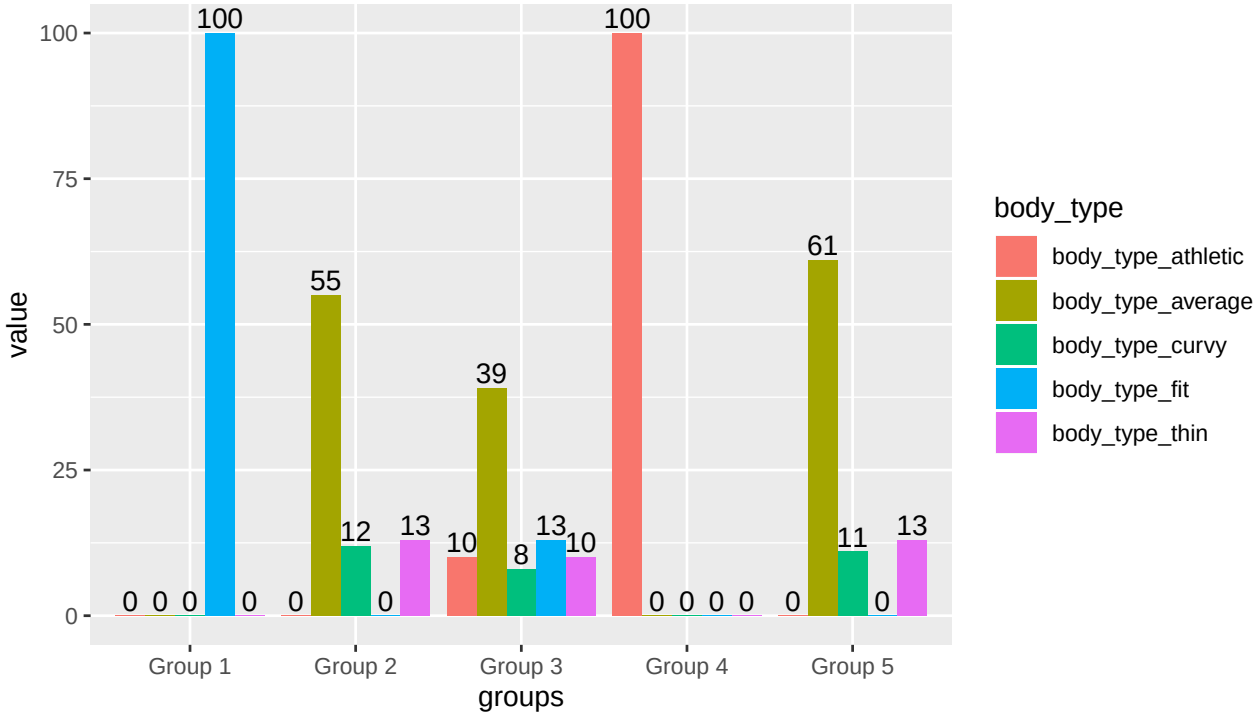


From the elbow diagram, it is clear that there should be 5 clusters or groups of users.

EXPLORATORY DATA ANALYSIS BASED ON CLUSTERS

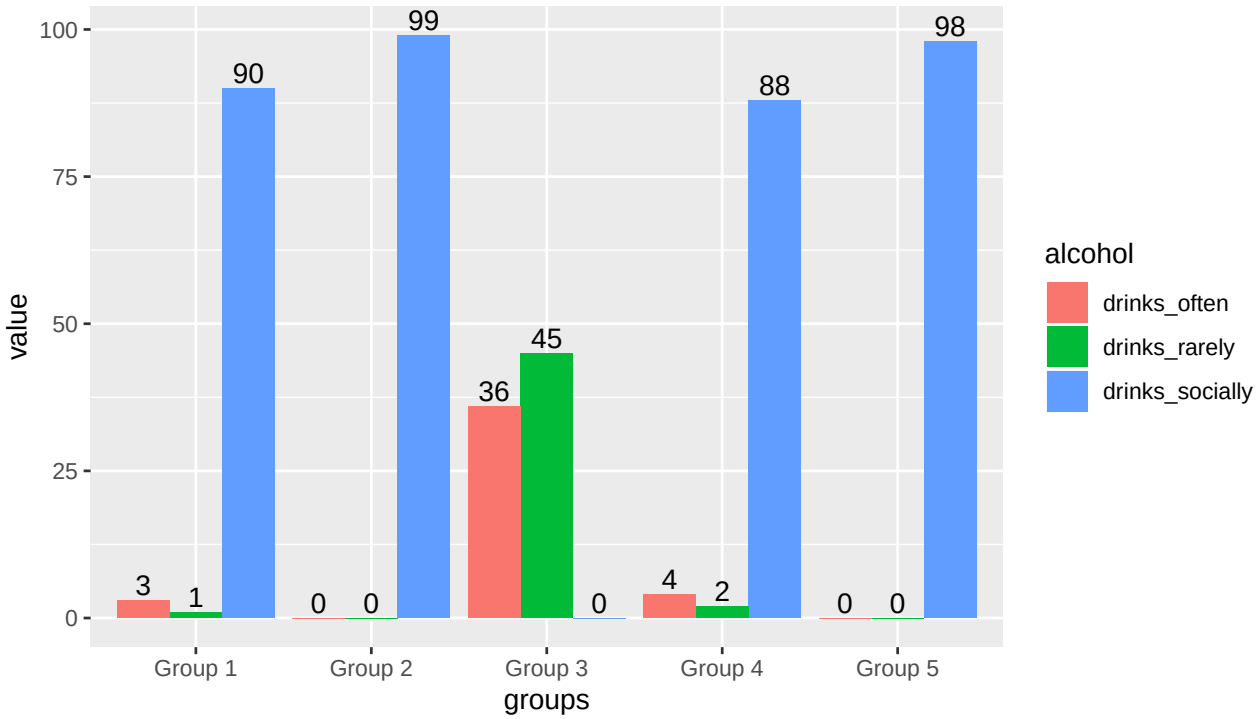
Trying to read the similarities within users of the same cluster.

And trying to figure out the difference between users of different clusters.



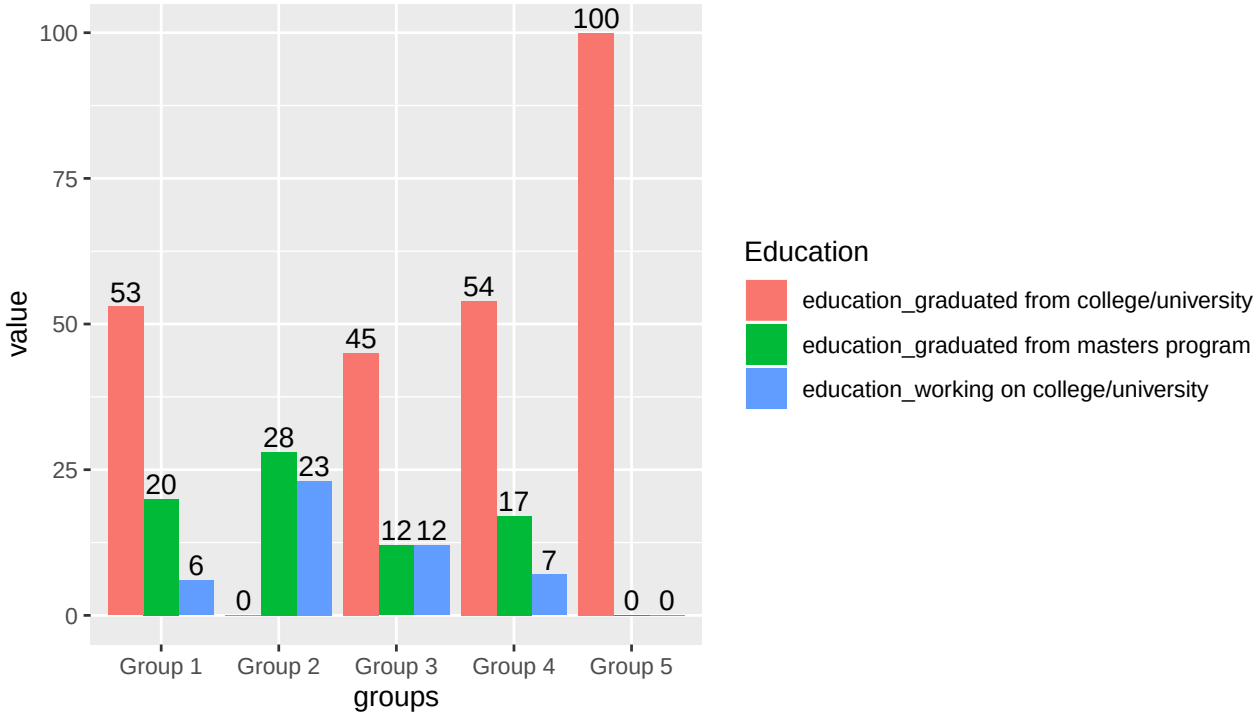
We can see that group 1 has 100% fit users. Group 4 has 100% athletic users.

Group 2 and Group 5 do not have any athletic or fit users.



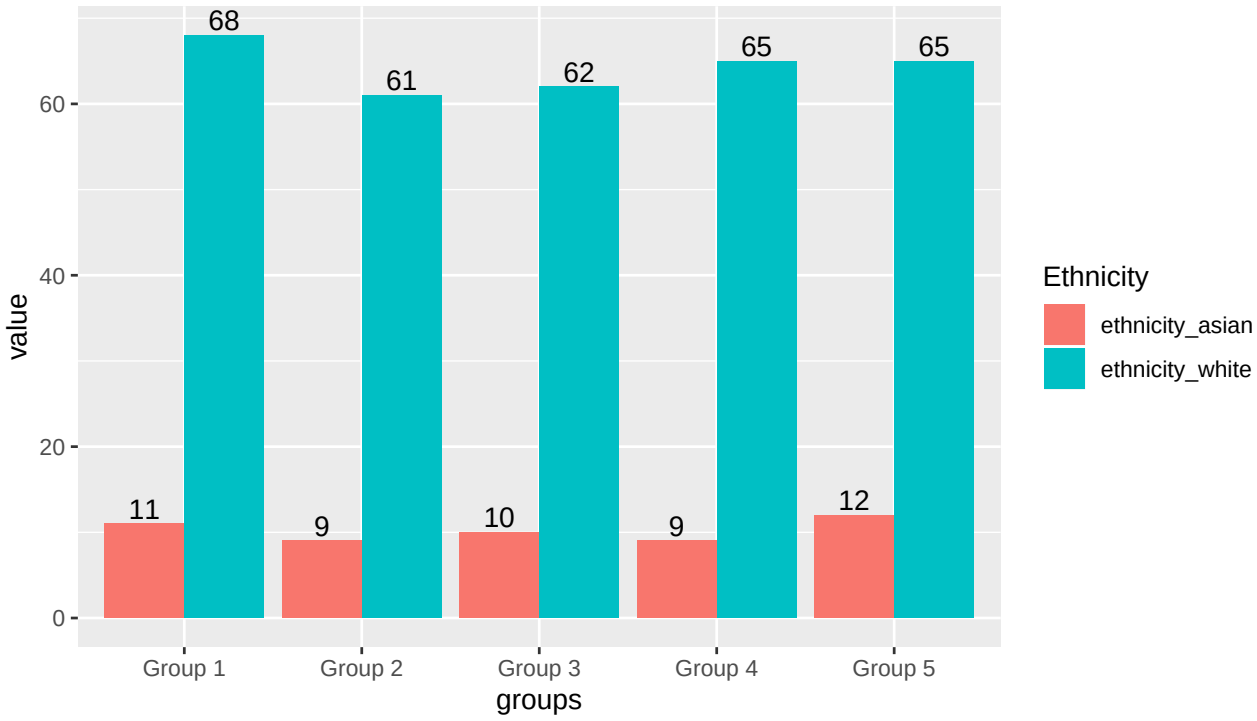
Group 2 and Group 5 have no users who drink often or drink rarely.

Group 3 have no users who drink socially.

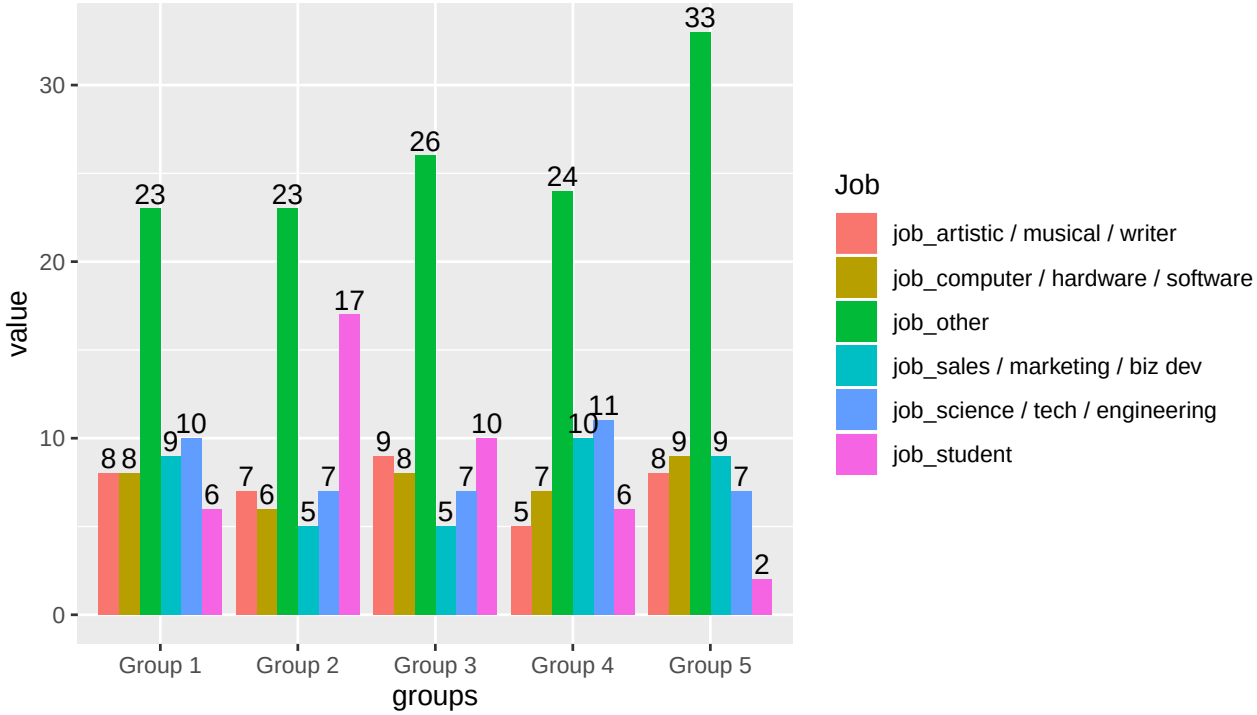


Group 2 have no users who have graduated from college/university.

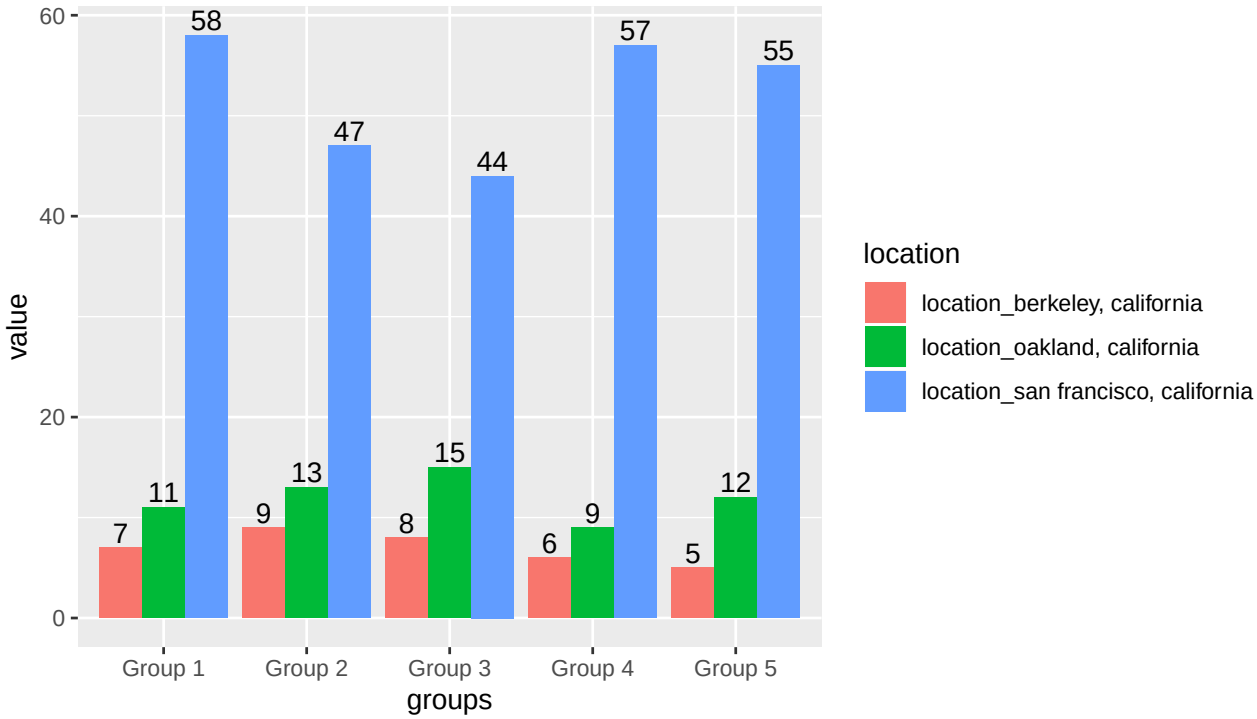
Group 5 have 100% users who have graduated from college/university.



We have a pretty even mix of people from the two races with white being around 65% and asian being around 10%.



We have a pretty even mix of jobs between the different groups.



We have a pretty even mix of location between the different groups.

Compatibility score

Compatibility score is a measure of how suitable/similar two people are to each other.

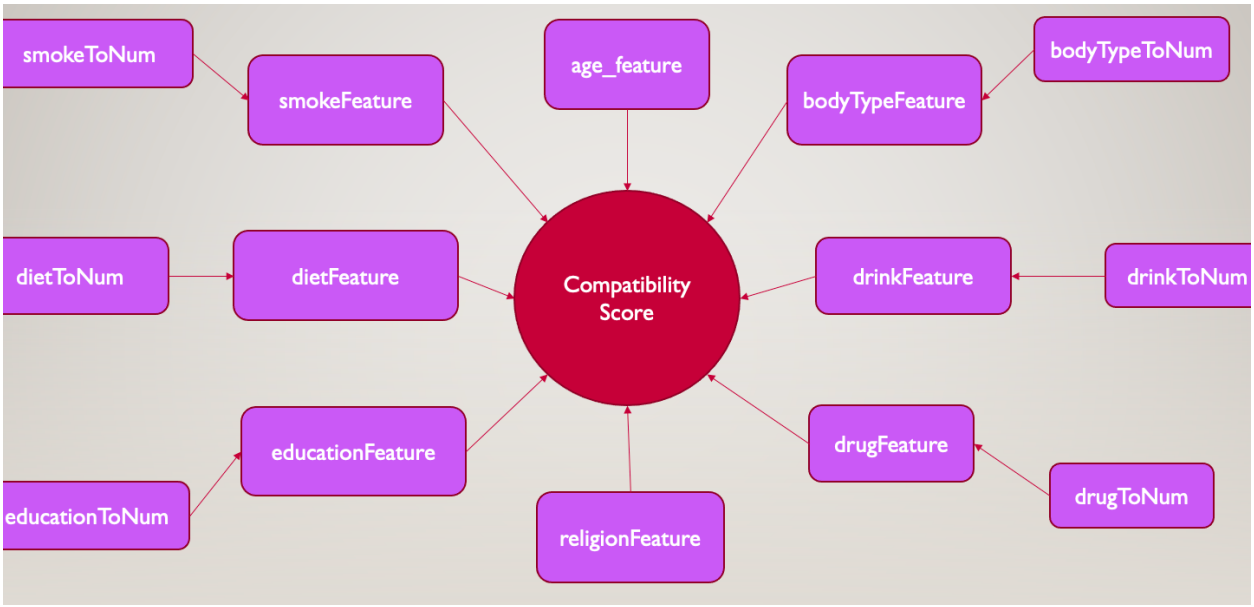
It is calculated using different features available in the dataset like body type, diet, drinks, education, job, ethnicity, etc.

The text based features are first converted to a number implying the significance or order of the users.

Then, a score is calculated for each feature that is considered.

Then we calculate the average of all those scores to obtain the final compatibility score.

The users will then be shown profiles in the decreasing order of the compatibility score with respect to the other users.



Example

Calculating and displaying the top 100 recommendation for a randomly selected user (user id - 8) based on the compatibility score.

```
##   [1]  796 2610 3475  380  517  759  892  995 1331 1345 1478 1554 1787 1838
##  [15] 2115 2124 2191 2245 2575 2918 2958 3098 3427 3430 3772 3822   53   61
##  [29]  259  448  488  531  661  676  684  831  844  866  897 1119 1147 1160
##  [43] 1213 1252 1261 1309 1354 1391 1529 1543 1608 1736 1749 1764 1768 1829
##  [57] 1863 2034 2072 2273 2285 2302 2355 2390 2417 2425 2644 2667 2732 2788
##  [71] 2794 2820 2900 2933 2973 3022 3037 3049 3070 3073 3093 3133 3185 3284
##  [85] 3286 3366 3418 3432 3489 3597 3713 3719 3765 3832 3856 3871 3972   49
##  [99]   74  105
```