How to Find a Potential Destined Lover – OkCupid

Final Report – Team 3

Members: Hanyu Chen, Yipeng Guo, Luke Hong, Ting-Ann Lu, Yuyan Ma

A) – The business problem that motivated us to conduct this project is launching improvements to dating apps. We propose to explore this in the following ways -

- Fitted match/lover recommendation with unsupervised learning clustering and text analysis
- Exploration of dating profiles and preferences

The rationale behind this project is that each person has their own specific status and preference when choosing their date. For example, either user can be vegetarian hence wants to date a vegetarian as well, or one can be gay or lesbian who craves a same-sexual couple. Even further, maybe someone who suffers from depression cannot appreciate their match more if their match was someone more outgoing. Sometimes the pairing & recommendation system will not do this kind of specific filter for people, leading to a higher possibility of misunderstanding and waste of time.

However, realistically speaking, the pairing system can be updated. For instance, we can launch clustering algorithm to distinguish gay or lesbian user with straight users, drug users with non-drug ones, so on and so forth.

With tons of dating apps in the market nowadays, the following data set has become outstanding for containing extremely specific characteristics of users; how they would define themselves and what kind of date they are looking for in detail.

B) - Description of our Dataset and Data Cleaning

The dataset was explicitly granted and supported by **OkCupid** (the dating app itself). With thousands of new users joining in and completing their personal status together with their preferences, this dataset allows us to dig deeper than we would imagine. The raw dataset includes 31 attributes & characteristics of around 60k users through years.

Regarding the attributes, some specific ones (columns in the dataset) triggered our interest as listed:

- Basic information: Age, Relationship status, Height, Sex, Sex orientation etc.
- Living habit: Diet, Drink or not, Drugs or not, smokes or not
- Education, Ethnicity and Language
- Income and Job
- Description: One sentence about yourself, one about your hobby, what can bring a smile on your face, about your future etc.

This dataset is indeed sophisticated, but there are still a lot of missing values which may mess with our further analysis. So, we applied the **data cleaning** as follows.

- Set an ID for each user and use it as an index, this will help us with matching.
- Drop Columns: last online/ speaks/ job/ sign/ ethnicity/ location, which do not contribute to our recognition of characteristics.
- Replace reasonable NaH values to reduce the null value proportion.
- Drop the rest of entries containing null values.
- Combine all the essays into one as 'Full_essay' and keep it as the main text components we are going to analyze.
- Rest of them we either make them into dummy or give them a scale.

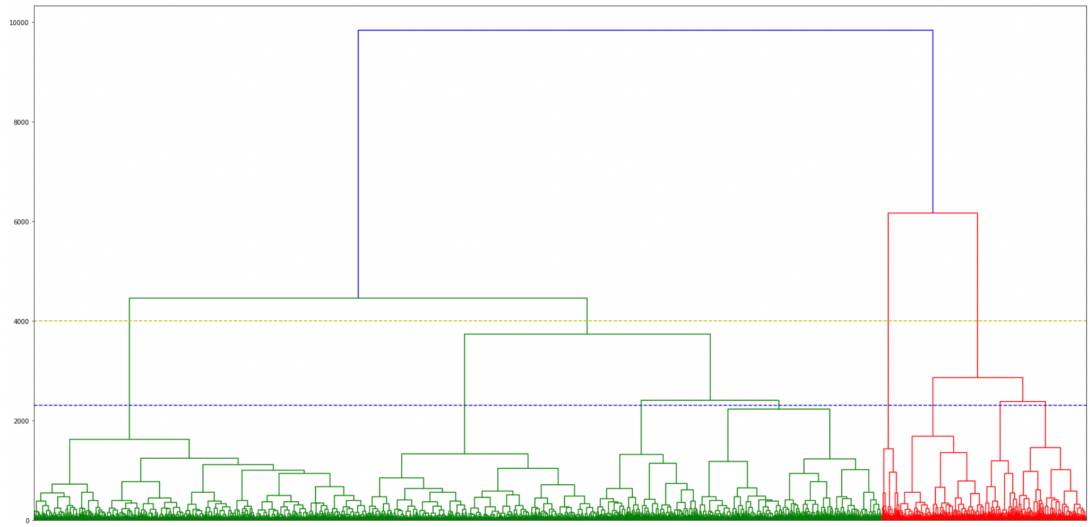
After the data cleaning, we have a dataset shaped as (9055,16). It contains a column of user id, 14 columns of numerical variables and 1 column of text variable (user's essay towards themselves). Except for 'id' and 'Full_essay', those numerical variables are listed as follows.

pets	pets contains: like both – 0, dog lover – 1, cat lover – 2, cat hater – 3, dog hater – 4, both hater - 5
age	age of every user currently
status	marriage status contains: single – 0, available – 1, seeing someone – 3 and married – 4
sex	sex contains: male – 0, female – 1
orientation	orientation contains: straight – 0, gay – 1, bisexual – 2
body_type	body_type contains: skinny – 0, average – 1, curvy – 2, overweight – 3, athletic – 4, rather not say – 5
diet	diet contains: anything – 0, vegetarian – 1, halal – 2, kosher – 3, other – 4
drinks	drinks contains: not at all – 0, rarely – 1, socially – 2, often – 3, very often – 4, desperately – 5
drugs	drugs contains: never – 0, sometimes – 1, often – 2
education	educations contains: high school – 0, undergrads – 1, masters – 2, PhD – 3, law school – 4, med school – 5, space camp – 6
height	height of every user currently
income	income of every user in dollar
offspring	offspring contains: no or don't want – 0, no and want – 1, yes or want more – 2, yes and don't want more – 3
smokes	smokes contains: no – 0, when drinking – 1, trying to quit – 2, sometimes – 3, yes – 4

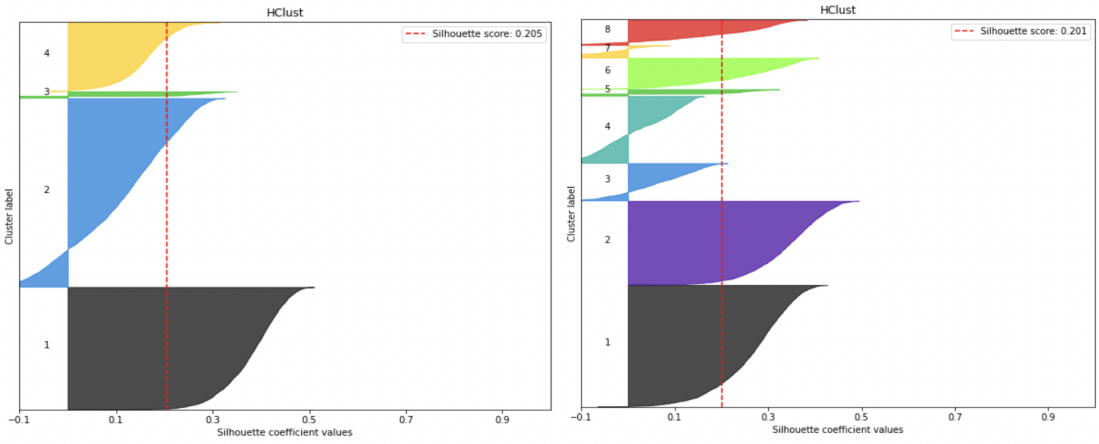
C) - The methodologies we applied & basic findings based on them

- **PCA**
  - PCA for numerical variables.
  - Variables have low correlation.
    Principal components did not contribute to reducing noise much.

- **H-Clustering**
  - Re-do PCA to check whether we can shrink a little bit - **9 variables** are used in the H-Clustering model.
  - Linkage method of "ward" since it is more balanced and faster.



  - Initially select 4 and 8 groups for our clustering numbers based on the dendrogram.
    4 groups - Average silhouette score of 0.205; The distribution is uneven and cluster 2 and 3 have many negative values.
    8 groups - Average silhouette score of 0.201; The distribution is getting better, but there are still some negative values.



- **K-Means Clustering**
  - Apply PCA also - **9 variables** are used in the K-Means Clustering model.
  - Inertia plot and Silhouette score plot are used to make the decision of the K

    The inertia with the number of clusters plot illustrates the elbow pattern where the optimal solution for k lays between 6 to 10.
    The peak point of the silhouette plot lies on the k of 8. (figure 1)

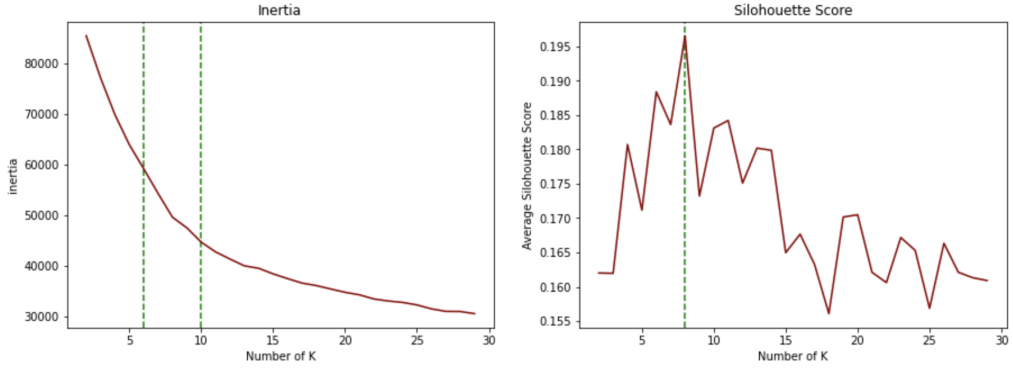


Figure 1

K of 8 is chosen after trading off between the two plots.

  - Fit our numerical data with the K-means model

    By looking at the cluster distribution with a silhouette score plot. We get the average silhouette score of 0.196. (Figure 2)

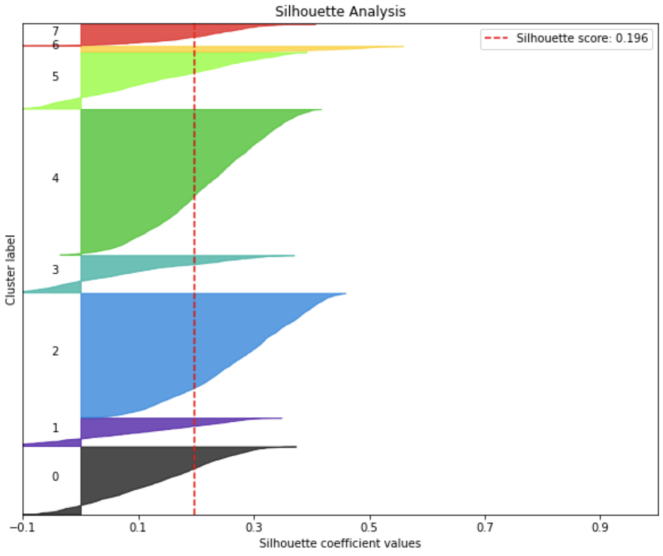


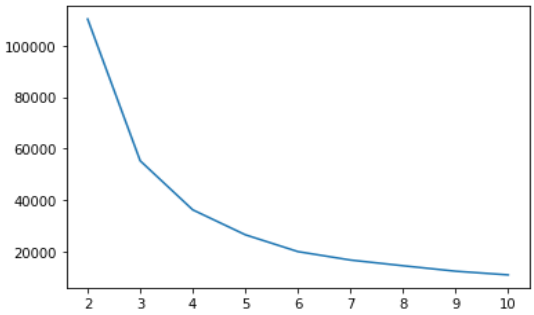
Figure 2

- **The Trade-off between H-clustering and K-means Clustering**
  - **From the results showing above, we decided to choose H-clustering and 8 clusters.**

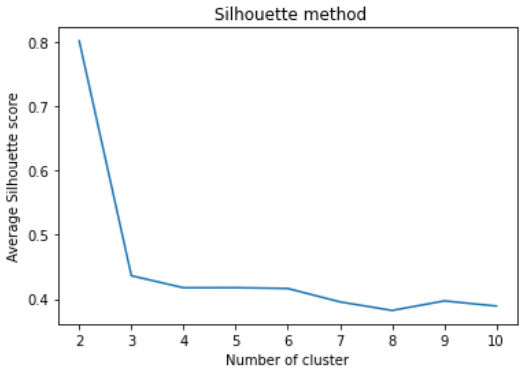
    The Silhouette score of H-clustering is higher than K-means clustering. (0.201 vs. 0.196)
    In H-clustering, although the Silhouette score of 4 clusters is higher than 8, they are extremely close (0.201 vs. 0.205). In addition, we

have to consider the trade-off between H-clustering and K-means. In K-means, the silhouette score of 8 clusters is higher than 4.
We still need to consider the subjective factors of our business issue. Since we have lots of clients, we believe that 8-cluster is more appropriate for the clustering.

- **Text Analysis - Classification**
  - Tokenizer - Spacy
    - Model used " en_core_web_md"
    - Only include 'tok2vec'
  - Reducer - PCA; UMAP
  - Using K mean to do classification
    - Step 1: Determine the number of K using the elbow method and Silhouette score method
    - The First graph is showing the elbow method.



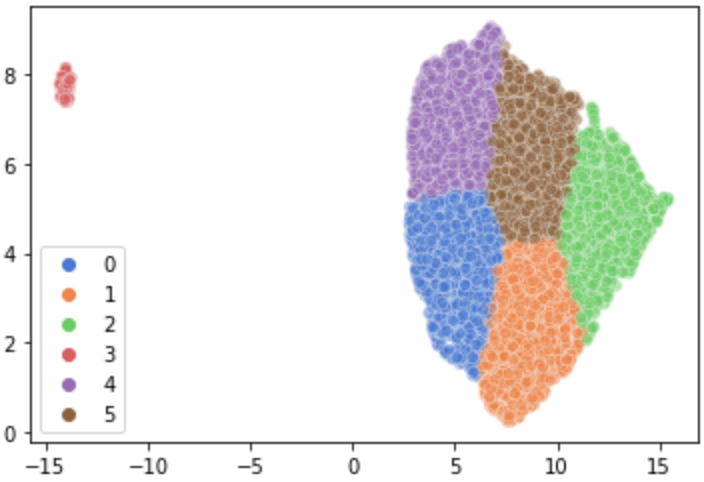
Based on this graph we can see that the best choice for us will be 4 which is on the curve of the elbow, and inertia is about 37000.



This graph is indicating that the silhouette score has a dramatic drooping at n = 2.

And when n = 3 ~ 6 , silhouette scores are around 0.42 which is not that bad.

However, we need to take the business problem into consideration. Only having 2 clusters is not so intuitive for our case. Therefore, We decide to use n = 6, and here is the visualization for K=6 below.



We can easily see that the majority of the data are on the right side, only a tiny bit of data on the left side that might seem to be some outliner that we did not detect. Meanwhile, the five clusters on the right side are pretty evenly distributed which is nice.

D) - Conclusions and recommendations based on our analytical findings

Based on our earlier findings, we are good to see the results

- 8 clusters for our numerical part (basic information of users)
- 6 clusters for our text part (text description about their preference)

The reason that we applied multiple analytical tools on our dataset, is to find a way of promoting the mechanism behind dating apps; to make things easier and more precise for finding a potentially destined lover. After we have these two main clusterings, we will be able to recommend a match for a newbie user. This is how this procedure works:

- Input the required information into the system once a newbie user is in, the system will automatically assign him/her into the basic user pool according to our first clustering algorithm. In this pool, most people will be sharing similar interests, income, living habitats with him/her.
- Once the user has been placed on their ‘planet’, fill in other selective verbal descriptions at least one. That way our second sorting algorithm will be working on it and sort the user into the second pool as well. In that pool, people will be sharing similar preferences and similar characteristics with him/her.
- Generally speaking, we as an analytical team, manage to use the existing data to create clustering. And using that, we are able to use models to train and forecast.

Moreover, we still have some limitations. We currently have no means to deal with the users’ location. But due to our raw material, the main resource of our user is southern California, so we would like to put the precise location into users’ profile after being sorted into two precise pools. That way, people can think for themselves whether they care about it or not.

Appendix - Model designing for the users' profile