BU Boston University Questrom School of Business

BA 820: Unsupervised Machine Learning Spring 2021

BA820 Team Project – OkCupid User Profile Analysis

Team 11: Peter Mankiewich, Chenzhi Pan, Phoenix Wang

Problem Statement

OkCupid is a dating application that allows users to make a profile by answering a series of questions (your gender, ethnicity, smoking habits, etc.). Then, the user can swipe on other profiles, and view the percent match, which is a metric automatically calculated based on the contents of both profiles.

Amid a global pandemic, it can be difficult if not impossible to engage in traditional dating while still staying safe, and adhering to public health guidelines. Dating applications, such as the popular OkCupid, become a go-to solution, allowing users to meet remotely, and even plan a date over Zoom without coming into contact with another individual. However, understanding the virtual dating landscape ahead of time can help get the best results from the application. The dataset, which includes features about a user's profile, can help us **understand the types of users that are on the platform**. By creating distinct clusters of users, we can help someone make better decisions while swiping, leading to better results for match seekers. Gaining these additional insights is particularly useful given the fact that users only have a limited number of swipes they can make per day, and of course, our time is valuable.

About the Dataset

Our dataset was taken from Kaggle (https://www.kaggle.com/andrewmvd/okcupid-profiles). It contains 59,946 rows and 31 columns. Each row represents an anonymous user profile found on OkCupid, with structured information as well as open-ended descriptions. The available information about a user includes:

Age	Relationship status	Sex	Sexual orientation
Body type	Diet	Education level	Ethnicity
Height	Income	Job	Location
Kids	Pets	Religion	Zodiac sign
Drinking habit	Drug use frequency	Smoking habit	Language skills

There are 3 numeric columns in the dataset – age, height, and income – while the rest of them contain categorical or textual data. Each user's answers to several open-ended questions were also recorded. The questions were "About me", "Current goals", "My golden rule", "I could probably beat you at…", "The last show I binged", "A perfect day", "I value…", "The most private thing I'm willing to admit", and "What I'm looking for (in dating)". The answers were stored in 10 "essay" columns in the dataset, yet not necessarily in the same order as the questions listed above. Last but not least, only 7 columns in the dataset do not contain missing values.

Data Cleaning

One of the challenges we faced while working with data from a dating app is that many users choose to not fill out certain questions or fields when creating their profiles. This translates into missing values in the dataset. Instead of dropping the missing values, we took several logical approaches to fill in most of them.

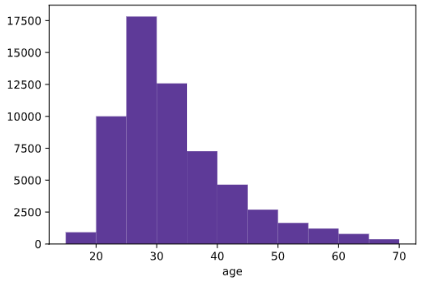
- Imputation – We replaced missing values and outliers in a column with the mean value. For example, in the "age" column, there are a 109-year-old and a 110-year-old. We replaced these two with the average age in the dataset. Another example is the "height" column. We assumed a person's height should be somewhere between 3 and 8 feet. Therefore, we replaced all heights that fall out of this range with the average height in the dataset.

- Reasonable assumptions – For example, in the "drinks" feature, we filled in all missing values with drinking "socially". We made this decision based on the fact that drinking "socially" is the most abundant category, and seems the most reasonable for the general population. Another example would be filling in all missing values in the "languages" feature with "English". Given 99% of the users in the dataset are based in the U.S., it is safe to assume that English is the primary language one speaks.
- "Unknown" – There are features where we could not make reasonable assumptions to fill in missing values. In those cases, we decided to replace missing values with "unknown". We think an "unknown" answer would still provide insight into the clusters we will create later.

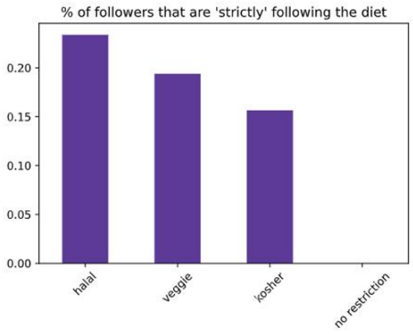
Exploratory Data Analysis

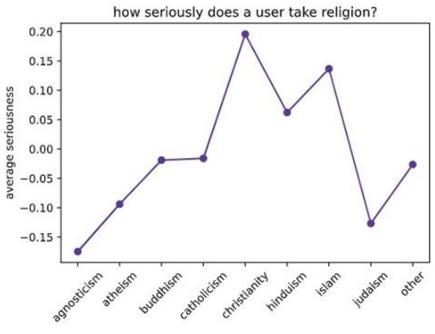
Through the exploratory analysis, we gained insight into the characteristics of the individuals in the dataset. Our initial findings will help to inform models that we will use to understand the types of people who use OkCupid.

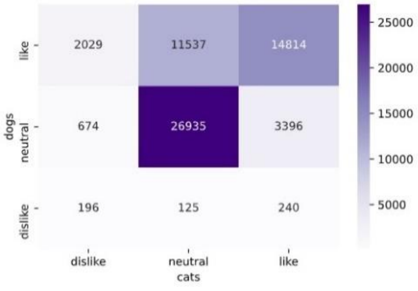
Most of the anonymous users in our dataset are in their 20s or early 30s, while the youngest is 18 years old and the oldest is 69 years old. 60% of the users are male and 40% are female, which is similar to the gender ratio in each age group. Since OkCupid is a dating app, 96% of the users are available for dating. However, we did find that the other 4% are either in a relationship or married. The majority of the users are straight, while around 10% are gay and around 4% are bisexual.

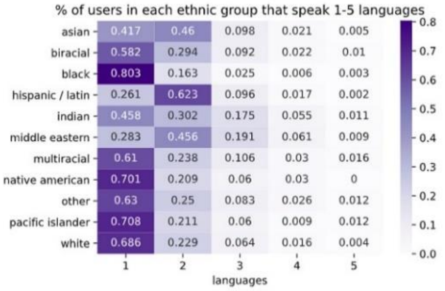


In addition to the basic demographic information, we were able to do some further analysis using the users' answers to more personal questions. For example, we calculated how strictly one follows a diet and found that the halal diet has the highest percentage of followers who are "strictly" following it. Coincidently, when we measured how seriously a religious user takes his/her religion, we discovered that Muslims take their religion the second most seriously.







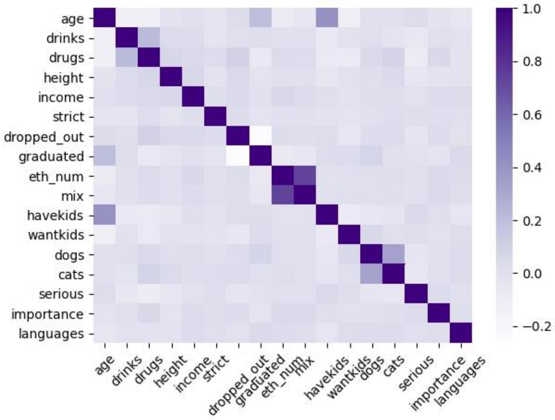


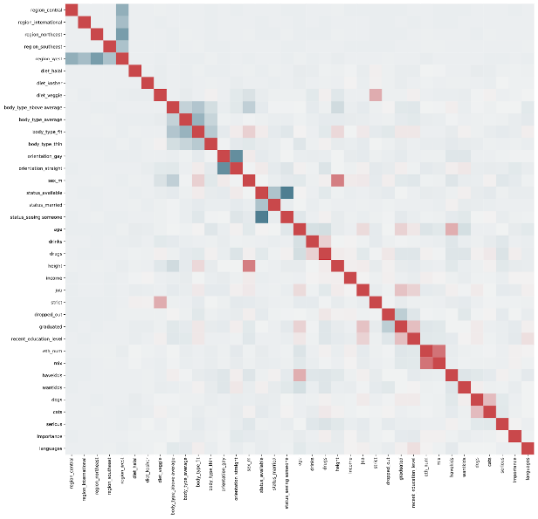
We also found that when it comes to pets, most people like both dogs and cats. Among those who only like one of them, there are almost 4 times more people who prefer dogs over cats.

The number of languages a user in the dataset can speak ranges from 1 to 5. After some calculations, we discovered that the majority of the Asian, Hispanic, and Middle Eastern populations speak 2 languages, while most people in the other ethnic groups speak only 1 language. Interestingly, we noticed that C++ stood out from a pool of human languages. It ranked the sixth most common language that the users "speak" among all 77 languages in the dataset.

Feature Correlations

Post-cleaning, the dataset now contains the original features and over 30 variables we created. We then generated a correlation heatmap, excluding the dummy variables, to learn more about the relationships between features. For example, whether a user has kids or not seems to be positively correlated to their age. In other words, if later one of the clusters we create consists of older users, we might observe that there is a higher proportion of users who have kids in this group as well.





By creating a correlation heatmap that includes all dummy variables, we noticed that if a user indicates he/she has a preference toward a vegetarian or vegan diet, there is a higher chance that this user is "strictly" following the diet. Also, it's not surprising for us to find a positive relationship between gender and height. The correlation coefficient is around 0.67. Hence, if the average height of users in a cluster is significantly higher, we might suspect that there are more male users in this group. We were curious to find out if the trends and correlations we just discovered in this initial analysis will appear again when we tried to understand the clusters

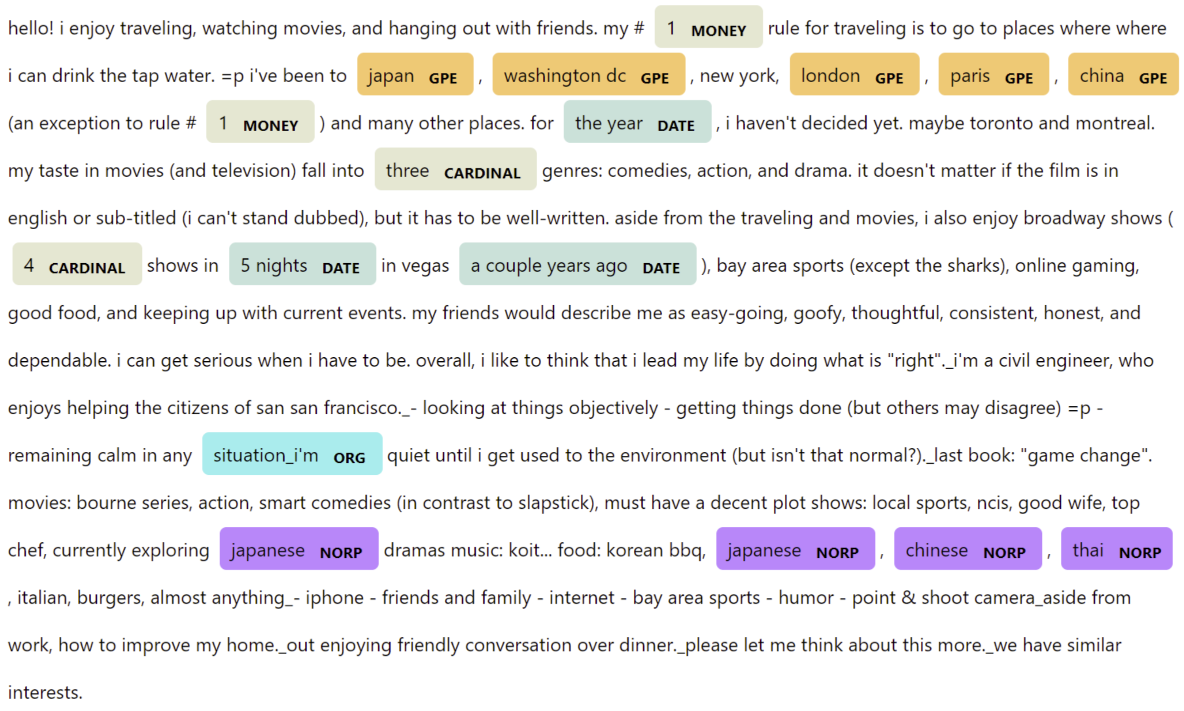
Text Analysis

The dataset contains 10 different text columns each containing user-generated answers to various questions presented in the app when users create their OkCupid profiles. In addition to using the categorical and numeric variables available to us, we would also like to leverage this text data in our models.

The first step involved cleaning these columns to ensure that all of the text data could be used in the clustering. Using the Python package *langdetect*, we iterated through all 10 essay columns for all of our users and determined if the text was in English. While we found that all of the text was English, the process did uncover some strange values, including links and special characters that we removed in order to not affect the later text analysis steps. As mentioned previously, each column corresponds to a different question, however, there were discrepancies in which columns represented which question, and so we decided to concatenate the text for each question together and look at our analysis on an aggregated basis for each user.

We then decided to calculate the sentiment and subjectivity for each piece of text. These features could be meaningful in the later clustering steps, allowing us to classify users by how positive they are, and how opinionated their answers are to the questions in the app.

While we were interested in the general characteristics of the text, we also wanted to understand what sort of topics users were talking about in their responses. To do this, we used *spacy* to find different entities or specific subjects in the text. For example, if a user travels a lot, then *spacy* would automatically recognize words like “Paris”, or “Japan”, as locations (GPE). This data can help us understand users who talk about similar topics and have similar interests. Below is a screenshot of one of the responses written by a user.



hello! i enjoy traveling, watching movies, and hanging out with friends. my # 1 MONEY rule for traveling is to go to places where where i can drink the tap water. =p i've been to japan GPE , washington dc GPE , new york, london GPE , paris GPE , china GPE (an exception to rule # 1 MONEY) and many other places. for the year DATE , i haven't decided yet. maybe toronto and montreal. my taste in movies (and television) fall into three CARDINAL genres: comedies, action, and drama. it doesn't matter if the film is in english or sub-titled (i can't stand dubbed), but it has to be well-written. aside from the traveling and movies, i also enjoy broadway shows (4 CARDINAL shows in 5 nights DATE in vegas a couple years ago DATE), bay area sports (except the sharks), online gaming, good food, and keeping up with current events. my friends would describe me as easy-going, goofy, thoughtful, consistent, honest, and dependable. i can get serious when i have to be. overall, i like to think that i lead my life by doing what is "right"._i'm a civil engineer, who enjoys helping the citizens of san san francisco._- looking at things objectively - getting things done (but others may disagree) =p - remaining calm in any situation_i'm ORG quiet until i get used to the environment (but isn't that normal?)._last book: "game change". movies: bourne series, action, smart comedies (in contrast to slapstick), must have a decent plot shows: local sports, ncis, good wife, top chef, currently exploring japanese NORP dramas music: koit... food: korean bbq, japanese NORP , chinese NORP , thai NORP , italian, burgers, almost anything_- iphone - friends and family - internet - bay area sports - humor - point & shoot camera_aside from work, how to improve my home._out enjoying friendly conversation over dinner._please let me think about this more._we have similar interests.

There are many useful entities identified, including some locations that might give us the idea that this person enjoys traveling. It is also important to note that some of the entities are not correctly labeled. For example, the number 1 is labeled as money when given the context, it is not money. Also, *spacy* was unable to recognize New York as a location.

While we feel that these entities are useful in understanding a user’s interests, and their level of seriousness when it comes to writing meaningful responses, we felt like the model would benefit from more domain-specific entities. We decided that we would identify hobbies and interests in the responses not only to match users with similar interests together, but to understand if a user is serious about using the app, and sharing detailed information about themselves.

We pulled a dataset off of Kaggle with a list of 666 different hobbies and then checked this list against each of the written responses. For the example given above, we found the below hobbies/interests. In addition to searching the string for the given hobbies, we also used *spacy* lemmatization; if the hobby was “acting”, then we would also search the documents for “act”.

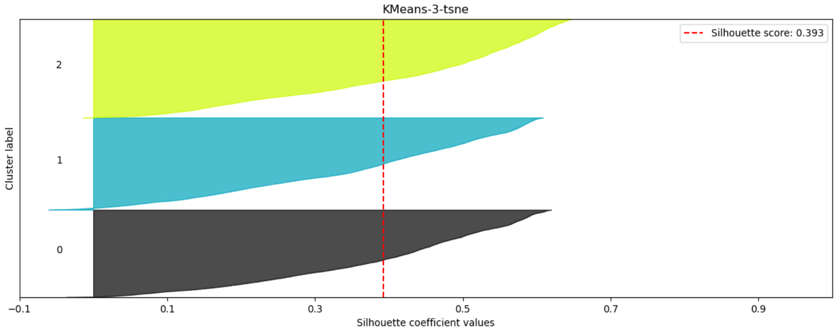
> *['acting', 'sports', 'travel', 'traveling', 'traveling', 'acting', 'arts', 'drama', 'games', 'gaming', 'internet', 'shooting', 'sports', 'traveling']*

It is important to note that while this method does give some useful results, it does have its limitations, and could be improved upon in the future. For example, the word “act” was compared to the response, and since the beginning of some of the words in the document starts with “act”, this user was marked as talking about acting even though those words have nothing to do with acting. On the other hand, the word “travel” was picked up, which is very important for this user since they talked a lot about this in their responses. When it comes to creating the clusters, we generated a column that is the count of the number of hobbies for each user. We felt that this would be useful because it indicates how detailed the user was in their response, and if they have a lot of interests.

Dimension Reduction and K-Means Clustering

Post-cleaning, our dataset contains 36 numeric features including all the dummy variables we created. The last step to take before we conduct cluster analysis was to standardize the data. We standardized the numeric features so that no feature has more influence on the clustering algorithm than the others. We first used two methods to reduce the dimension of our dataset and then used the newly-generated features to create clusters. Due to computational constraints, we decided to perform K-Means clustering only and leave out hierarchical clustering.

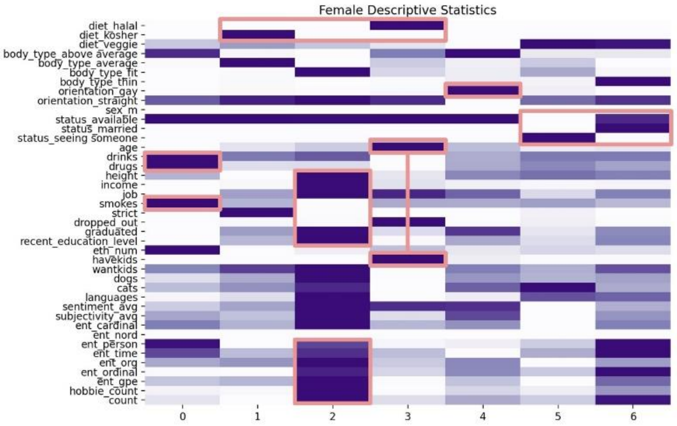
PCA is the first method we used to reduce the dimension of our dataset. However, because we created a large number of dummy variables, most of these variables have specific meanings and we couldn't further reduce the number of variables. For example, if we want to explain 90% (or 85%) of the variance, we will need 26 (or 24) principal components. If we further decrease the number of principal components in PCA, we face losing meaning in our cluster analysis. With 25 principal components, the optimial number of clusters needed in terms of maximizing the average silhouette score and minimizing inertia is over 10. Therefore, PCA and K-Means might not be the best combination to choose for cluster analysis for this dataset.



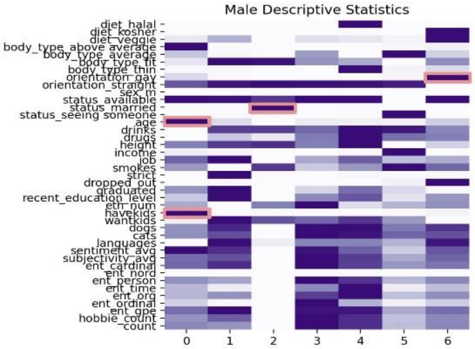
t-SNE is another way to reduce dimension. We used the same standardized dataset in the t-SNE model and got 2 output variables that can be used to create clusters. Then, we examined the change in inertia and the average silhouette score for different numbers of clusters. By looking at the descriptive statistics and the plots, we found that we could choose to generate 3 clusters for the whole dataset and gain sufficient insights into the characteristics of each cluster. In short, t-SNE did a better job reducing dimension than PCA while still providing us with useful information in the clustering process that helped us better understand the users in our dataset.

Clusters Examination and Summaries

We built several different clustering models for various splits of the data. We first split the data by gender and created clusters for women and then for men. We also split the data by people who are above and below the age of 30 and then created clusters for these two subsets. To determine the optimal number of clusters, we examined both the inertia and the silhouette score, and combined this knowledge with what we felt would be the most useful number of clusters given the context of the problem.

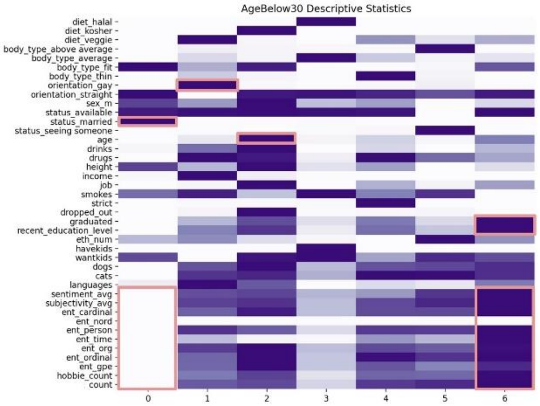


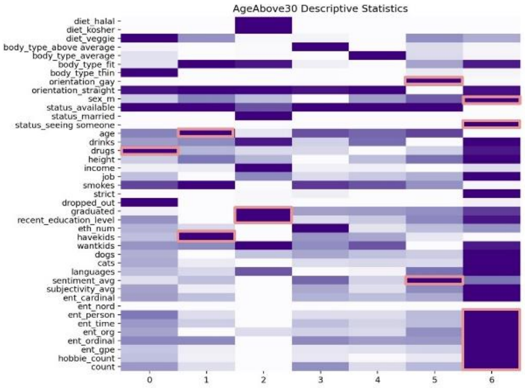
When examining the 7 clusters that we created for just women, we observed some interesting trends and distinct groups that would be useful for OkCupid. We can see that overall, the different diets are distinct between the groups. For example, group one is distinctly people who eat Kosher. Also, we have a group of older people with children (this was true for a few other subsets as well). We can observe that people who put more effort into their essay (those with lots of entities and hobbies mentioned) on average tend to have a higher income and a good job. Also, relationship status and orientation were very distinct across groups.



When it comes to the male group, we see some distinct differences in how the seven groups are arranged. While one cluster tends to have older people with kids, we also see that filling out the essays completely is not as strongly associated with high income and a good job (compared to the analysis of females). Again, being married does not seem to be strongly grouped with putting a lot of effort into the essays. Unlike for females, the diets are less distinct, and Halal and Kosher seem to be more grouped in the same cluster.

Next, we created clusters for only the subset of people who are above 30 and the subset that are below 30. For those younger than 30, we can again see that there is a cluster who are married and don't put a lot of effort into their essays. In this group, people who do drugs and drink are also in a cluster who spend a lot of time on their essays and are potentially more serious. Also, we see that people with a higher education level are more serious about their profiles (which is an observation similar to some made in previous clusters). There is also a cluster of people who are gay and put an average amount of effort into their essays.





Finally, we performed cluster analysis on people who are over the age of 30. In this case, people with a high income are less likely to be grouped with people who put effort into their essays. Again, there is a group of slightly older people with children. There is also a group that is gay and has a positive sentiment.

We developed some descriptive phrases to describe different groups in the analysis, such as "self-made entrepreneurs", "healthy lifestyle followers", or "single parent". These groups that appear in many of the cluster models could be used by OkCupid to provide people with good match recommendations and improve the overall user experience.

Business Values and Recommendations

Building on top of our models and anlaysis, we believe OkCupid can benefit from doing further investigation, analysis, and optimization in the following areas.

1. **OkCupid can utilize the essay columns to rate each of its users for internal and external use.** Since written answers are not required when creating a profile, we think it's safe to assume that users who do provide them are more serious match seekers. With 10 essay columns in the dataset, we were able to perform textual analysis, count the number of hobbies, and identify named-entities each user mentioned in their written answers. Based on the quality of their answers, OkCupid can create a score for each user. Internal-wise, this score can be used as a way to detect scam accounts or to measure how invested a user is in the app. By identifying scam accounts and recommending serious match seekers to users, OkCupid can improve user experience.

2. **OkCupid can recommend profiles to users based on their previous swiping behavior.** Once a user sets the search filters based on his/her preference, OkCupid can gather a pool of candidates that fit the requirements. Then, as we did in this project, OkCupid can create clusters and put candidates into like-groups. From this user's previous swiping behavior, OkCupid can identify which cluster fits the characteristics of this user's "type", i.e. has a higher chance of having candidates that will be a match for this user. On the other hand, instead of always recommending candidates that have similar characteristics, OkCupid can also try recommending candidates from other clusters and test out if there are undiscovered characteristics that would also be attractive to this user.
3. **OkCupid can create more intelligent search filters and generate prompts for users when they're updating their profiles.** Looking at the clusters generated for each user, OkCupid can combine written answers and structured information to create more flexible filters for a user to choose from. For example, instead of filtering age, religion, income, etc. one by one, users will have a chance to choose from a set of descriptions OkCupid generated, such as "young professionals who enjoy traveling to Europe", "dog-lovers who like medieval literature", or "humorous college students who enjoy outdoor activities". Additionally, OkCupid can also consider the characteristics of people that a user usually likes and use association rules to inspire users to update their profiles. For example, if a user tends to like vegan candidates, OkCupid can send prompts such as "Would you describe yourself as an animal lover?" to this user and suggest that he/she mention it in his/her profile.