

Multivariate Statistics for
Machine Learning
MGSC-661-075 Final Project

The logo for OkCupid, featuring the lowercase letters "okc" in a bold, blue, sans-serif font. The letters are set against a white, cloud-like or bubble-like background that has a soft, irregular outline. This entire graphic is centered on a solid, vibrant pink rectangular background.

Table of Contents

Introduction and Objectives	3
Data Description	4
Data Exploration with Principal Component Analysis	6
Model Selection – K Means Clustering	9
Clustering Results	11
Alternative Model – PCA + K Means	12
Conclusions and Recommendations	13
Appendix	14
Appendix Figures	14
Appendix Tables	18

Introduction and Objectives

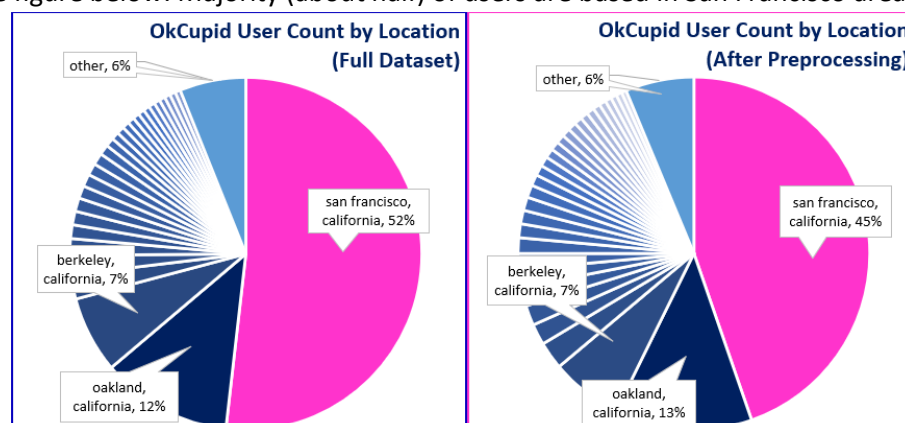
For our project, we decided to dive into the online dating industry. The online dating scene has been heralded as an increasingly popular and socially normalized way to find dates, a romantic partner, even true love. With the rise of online dating empire, there are now an app or site for almost every niche out there from your vanilla match.com to the most specific of flavors like Ashley Madison.

On the surface it may appear to be innocuous fun and games but over the years, numerous online dating sites have been evolving in a dark direction due to the increasing pressure for monetization and growth. This has been pervasive from the top apps like Tinder to older services like OkCupid (which our dataset is based on). Techniques and systems mechanics like gates and microtransactions designed to limit functions and can be removed with a fee. Take eHarmony's subscription, which without it blocks you from being able to send a message back to a potential suitor (who's subscribed) or Tinder's 100 daily swipe limit (you only get more if you sign up for tinder+), or even better, their [Super Boost](#), which boasts to help a user "cut to the front [of the line] and be seen by up to 100x more potential matches", suggesting that there's a line to begin with and that getting matches might not always work for some, which we won't get into.

The more desperate the user (for love, or for their desire not to die alone), the more money online dating services make. Clearly, these mechanics are working as these enterprises are becoming ever so popular and profitable, to give some perspective, top apps like [Tinder now rival the likes of Netflix](#) in terms of profitability. The gamification of it all has made the online dating so convoluted nowadays and knowing how the industry effectively monetizes the desperation of its users, that we felt like going back to the basics a bit. Specifically, our project will seek to develop efficient user clusters that could be used to suggest to its users only potential dating partners that are aligned to the users' unique preferences and dispositions.

To test our model, we selected a dataset from OkCupid which contains various information on its users that are located around the California area from year 2012:

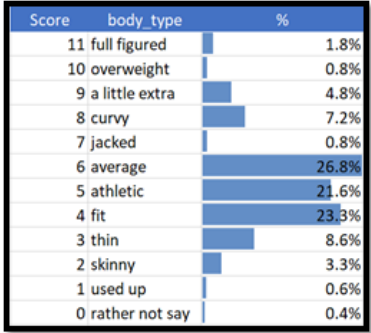
- The full dataset contains nearly 60K observations with 31 features (71 features after preprocessing).
- The dataset contains numerous fields not mandatory for the users, thus for simplicity, this analysis only included users that had full (non-missing) data.
- Refer to the figure below: Majority (about half) of users are based in San Francisco area



- The raw dataset can be found at: <https://www.kaggle.com/andrewmvd/okcupid-profiles>

Data Description

Rows and Features Not Used or Dropped - Observations that have a missing value were dropped. Features which were unused are **last online**, **location** and also the **10 Essay columns**.

Feature	Description / Preprocessing Step	Other Info																																							
Age	Represents the age of the user. This feature ranges from 18-69.	Most of the datapoints are in the 20-30 range.																																							
Status	Represents the relationship status of a user, there are 4 categories: single, available, seeing someone and married. Each been dummified into a binary variable of 0 and 1s.	The vast majority of users are single.																																							
Orientation	represents the sexual orientation of a user, there are 3 categories: straight, gay and bisexual. Each been dummified into a binary variable of 0 and 1s.	The vast majority of users are straight.																																							
Body Type	Users can select from a list of body types. We turned this variable into a scale according to research online for the order of OkCupid's body types .	 <table border="1"> <thead> <tr> <th>Score</th> <th>body_type</th> <th>%</th> </tr> </thead> <tbody> <tr><td>11</td><td>full figured</td><td>1.8%</td></tr> <tr><td>10</td><td>overweight</td><td>0.8%</td></tr> <tr><td>9</td><td>a little extra</td><td>4.8%</td></tr> <tr><td>8</td><td>curvy</td><td>7.2%</td></tr> <tr><td>7</td><td>jacked</td><td>0.8%</td></tr> <tr><td>6</td><td>average</td><td>26.8%</td></tr> <tr><td>5</td><td>athletic</td><td>21.6%</td></tr> <tr><td>4</td><td>fit</td><td>23.3%</td></tr> <tr><td>3</td><td>thin</td><td>8.6%</td></tr> <tr><td>2</td><td>skinny</td><td>3.3%</td></tr> <tr><td>1</td><td>used up</td><td>0.6%</td></tr> <tr><td>0</td><td>rather not say</td><td>0.4%</td></tr> </tbody> </table>	Score	body_type	%	11	full figured	1.8%	10	overweight	0.8%	9	a little extra	4.8%	8	curvy	7.2%	7	jacked	0.8%	6	average	26.8%	5	athletic	21.6%	4	fit	23.3%	3	thin	8.6%	2	skinny	3.3%	1	used up	0.6%	0	rather not say	0.4%
Score	body_type	%																																							
11	full figured	1.8%																																							
10	overweight	0.8%																																							
9	a little extra	4.8%																																							
8	curvy	7.2%																																							
7	jacked	0.8%																																							
6	average	26.8%																																							
5	athletic	21.6%																																							
4	fit	23.3%																																							
3	thin	8.6%																																							
2	skinny	3.3%																																							
1	used up	0.6%																																							
0	rather not say	0.4%																																							
Diet	Users can select their diet preference. The different diets include: anything, halal, kosher, vegan, vegetarian, and other. Each been dummified into a binary variable of 0 and 1s.	The most prominent diet is anything followed by vegetarian.																																							
Drinks	A user's self-defined drinking habits. We simplified this to a scale of 0-5 where the original values were: not at all (0), rarely, socially, often, very often, desperately (5).	Most users drink socially.																																							
Drugs	A user's self-defined drug use habits. We simplified this to a scale of 0-2 where the original values were: never (0), sometimes, often (2).	Most users never do drugs.																																							
Height	A user's self-specified height (in inches), the range goes from 43 to 95.	Height follows a normal distribution with a mean of around 68.12 inches.																																							
Income	A user's self-disclosed income level. The range goes from 20k to 1M. The -1 values are likely non-disclosed values.	Most users that disclosed a value make below 100k.																																							
Smokes	A user's self-defined smoking habits. We scale these habits from 0-4 where the original values were: no (0), trying to quit, sometimes, when drinking, yes (4).	Most users do not smoke.																																							

Feature	Description / Preprocessing Step	Other Info
Speaks	It represents the number of languages a user speaks with missing values defaulted to 1.	Most users speak 1-2 languages.
Education	This is a simplified feature representing a user's education level. We scaled it from 0-5 where the (0) includes categories like high school, space camp and 2-year college, (1) college/university, (3) law/med school, (4) masters, (5) PhD.	Most users belong to category 2 (college/university).
Dropped Out	This is a binary variable there 1 is an indication of a user having dropped out of whatever education they last did.	Only a small portion of users has been flagged as a dropout.
Job	Originally there were 21 categories which we combined into 9: Artistic / musical / writer, Business and management, Education (includes students), Entertainment / media, Law and government, Medicine / health, Non-Technical, Technology, and Other. Each dummified into binary variables of 0/1s.	The largest categories are: business and management, education, technology and other. The lowest number of users were in entertainment / media and law and government.
Religion	A user's self-disclosed region, there are 9 categories: - Agnosticism, Atheism, Buddhism, Catholicism, Christianity, Hinduism, Islam, Judaism, Other Each religion was dummified into binary variables of 0/1s.	The largest categories are: Agnosticism, Atheism, Catholicism, Christianity and Other while lowest categories are Hinduism and Islam.
Relig_Imprt	An additional feature separated from religion. We simplified this to a scale of 0-3 where the original values were: laughing about it (0), not too serious about it, very serious about it (3).	Most users selected are not very serious about their religion.
Sign	A user's self-disclosed astrological sign, there are 12 categories, one for each sign. Each dummified into a binary variable of 0 and 1s.	The distribution across each sign is more or less even.
Sign_Imprt	An additional feature separated from sign. We simplified this to a scale of 0-2 where the original values were: it doesn't matter (0), it's fun to think about, it matters a lot (2).	Most users picked that it doesn't matter or it's fun to think about, the latter being higher.
Ethnicity	A user's self-disclosed ethnicity, they can have more than one and the 9 choices are: Asian, White, Black, Hispanic/Latin, Pacific Islander, Native American, Middle Eastern, Indian, and Other. Each dummified into a binary variable of 0 and 1s.	Most users are White (60%), with Asian being the next largest group (13%) and the rest below 10%.
Have Kids	Whether or not a user has a child or children: 1 = yes, 0 = no, -1 = omitted.	Most users do not have kids.
Want Kids	Whether or not a user wants a child or children: 1 = yes, 0 = no, -1 = omitted.	Most users did not disclose this information. Of those that did, users that want are about even.
Likes Dog	Whether or not a user like dogs: 1 = yes, 0 = no.	Most users like dogs.
Has Dog	Whether or not a user has a dog: 1 = yes, 0 = no.	Most users do not have a dog.
Likes Cat	Whether or not a user like cats: 1 = yes, 0 = no.	Most users like cats.
Has Dog	Whether or not a user has a cat: 1 = yes, 0 = no.	Most users do not have a cat.

- Hispanic / Latin ethnicity is highly correlated with the other ethnicity, if we were to consider reducing the number of ethnicity and adding them into the other category, Hispanic and Latin would be a very good candidate
- There does not seem to be any discernible patterns with relationship status as confirmed by the previous plot.

On the third subset analysis³ we investigated user's diets as well as their preferences for drinking, smoking and drugs. Some other variables such as sex, height and body type are included into the plot. We observed that in terms of diet, there is a major distinction between users that eat anything vs having a specific diet, the most popular specific diet is vegetarianism.

Interestingly, we see that there is not a relationship between diet and body type. Preferences for drinking, smoking and drugs are correlated together, but not in conjunction with a diet, but rather males instead. While on the topic of male, we see that they are negatively correlated with body type, dictating that females tend to have larger body types compared to males.

On the fourth subset analysis⁴ we investigated the type of jobs field and education level of our users. This plot also includes other competency related features such as the number of languages they speak, their income and whether they have dropped out of school or not. It is colored by the level of education, high being pink and low being blue. Some interesting observations include:

- People that work in a technology field tends to speak more languages
- Users that are older tend to fall into business job categories whereas younger users tend to work in the education field (which includes students as well)
- Level of education (defined as low = 1 which is high school and high = 5 which encompasses PhD and specialized education) is not correlated with a quite a few numbers of job categories but seems to be highly negatively correlated with non-technical jobs and being a drop out but interestingly to a small extent income as well

On the fifth subset analysis⁵ we investigated users' preferences for cats, dogs and kids. We see that users that has cats do not like dogs. We see that preferences for pets are not correlated to preferences for kids. On the topic of kids, we see that people that are older tend to have kids while younger users tend to want kids but there is a distinction that people who want kids do not have kids and vice versa.

On the fifth subset analysis⁶ we investigated the different religion our users are affiliated with. Some additional variables included in here for interest are preferences for kids, preferences for drugs, drinking and smoking.

Interesting observations from the plot include:

- Catholicism and Christianity are highly correlated with users that think religion matters a lot
- There does not seem to be a specific religion that is correlated preferences for kids, but we see that people that are atheist tend to not have kids

³ Refer to Appendix - Figure 3: Principal Component Graph – Diet and Preferences

⁴ Refer to Appendix - Figure 4: Principal Component Graph – Job and Education

⁵ Refer to Appendix - Figure 5: Principal Component Graph – Kids and Pets

⁶ Refer to Appendix - Figure 6: Principal Component Graph – Religion

- Users that are agnostic seem to be the party people that are correlated with high drug use, smoking and drinking compared to users that are Catholic and Christian which can be said to be more conservative

On the final subset analysis⁷ we investigated users' specific astrological signs. While a user's sign is something that is mutually exclusive, we can see for fun that there is not really any specific sign that thinks religion or signs are particularly that important. We see that Scorpios tend to be older whereas Pisces tend to be younger. Also, it seems that males are negatively correlated with sign importance, meaning that it does not really matter to them, interestingly signs matter to females a lot.

⁷ Refer to Appendix - Figure 7: Principal Component Graph – Astrological Signs

Model Selection – K Means Clustering

After completing our data exploration, a handful of features were identified as the most relevant due how they could help differentiate users, giving us a total of 24 pre-selected features we would use to explore clustering. However, it was determined through our data exploration that some of these features were highly correlated. Hence, through some additional feature engineering (refer to Table 2) the clustering dataset dimension were reduced to a final 21 variables.

Table 2: Feature Engineering Summary

Engineered Featured	Feature Description
diet.vegan.vegetarian	The user's diet is vegan or vegetarian
job.business.and.health	The user works at job related either to business or health
religion.christian.catholic	The user is either Christian or catholic

In Figure 2, we can analyze that our final selected features are consistent enough to differentiate our user dataset. This shows that even though we will differentiate users through only 29.57% of their characteristics, users will be different enough to generate logical clusters and a more scalable model.

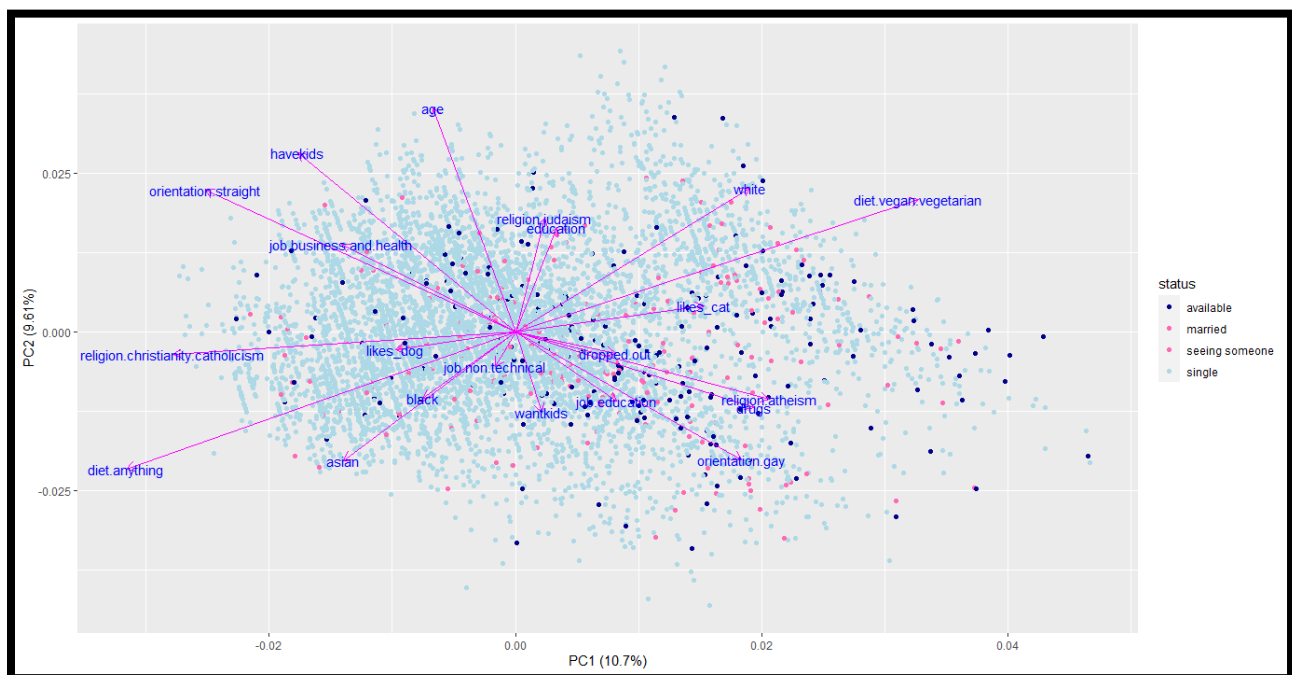


Figure 2: Principal Component Graph of final selected features

The model selected for clustering our users is K Means Clustering. This is a machine learning algorithm that uses Euclidean distance to determine how to subset the data into similar groups based on the features provided. The model will assign each user to a cluster ensuring that they will be very similar to the other users within its cluster and very different to users in other clusters. We will thus be able to

leverage this algorithm to get closer to our initial objective of matching users to other similar users, maximizing matching by user similarity, preferences, and dispositions.

It must be mentioned that in K Means the number of clusters created is determined by the data scientist. Hence, we must optimize the tradeoff between minimizing within cluster variance (represented below by SSE) as we increase the number of clusters used. We see that at a certain point, the marginal gain in lower model inertia begins to wane as we increase the numbers of clusters we use. That cutoff point is identified using the “Elbow Method” which is depicted on the plot below and is used to determine the “optimal” number of clusters to use. In the end 7 clusters was selected as our K Means hyperparameter.

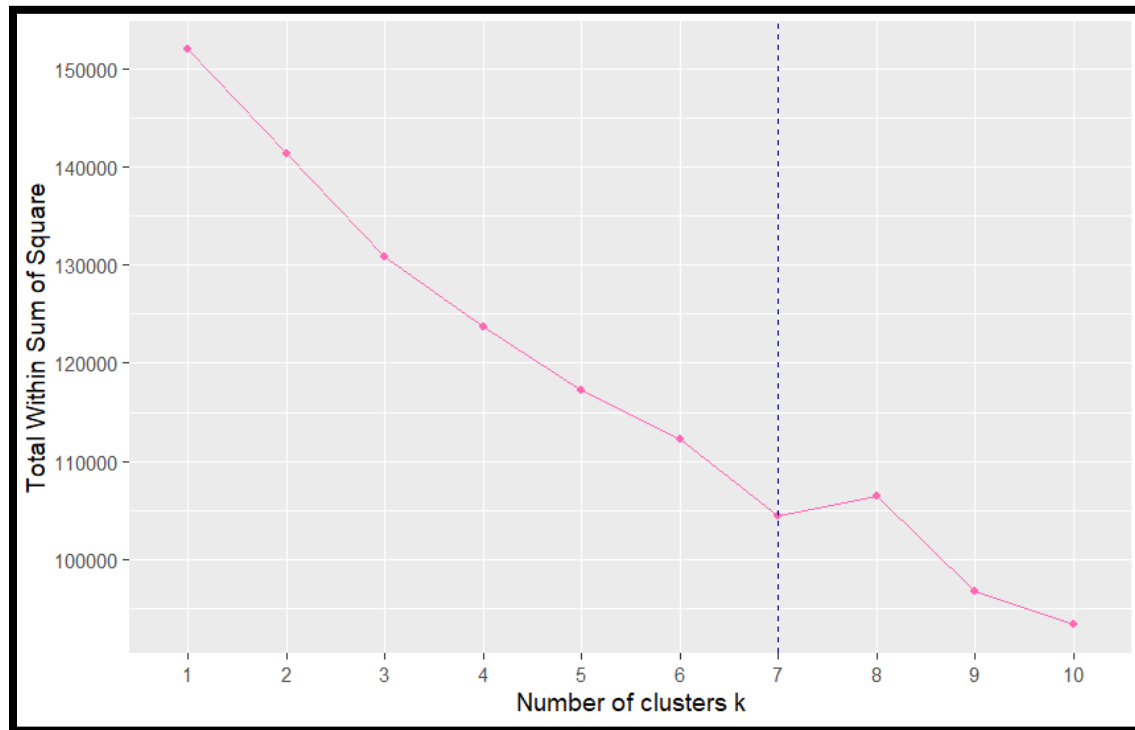


Figure 3: Elbow Method Graph for selected features K Means Clustering

Note that the dataset’s feature scales have been standardized, and K Means was applied to separate the user dataset into 7 groups / clusters.

Clustering Results

As our 7 clusters are affected by 21 features a plot is will not be effective to show how each cluster is different from the others. Therefore, we have decided to identify the main differences between the clusters by analyzing each clusters' centroids which will give us information on the average attributes of the users in each group by interpreting the average features values.⁸

Table 3: Cluster Features Trend

Feature	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6	CLUSTER 7
Sexual Orientation	Straight	Straight	Mostly Straight	Straight	Straight	Mostly Straight	Straight
Diet	Non-Vegetarian	Mixed	Non-Vegetarian	Non-Vegetarian	Non-Vegetarian	Vegetarian	Mixed
Drug usage	Average	Average	High	Low	Low	Average	Average
Ethnicity	Mostly White	White	White	Mostly White	Mostly Asian	Mostly White	Mostly Black
Age	29.70	38.44	29.89	42.23	30.09	34.94	30.91
Education	College /Univ	College /Univ	College /Univ	College /Univ	College /Univ	College /Univ	College /Univ
Job	Educators or non-technical	Business, health, or education	Mixed	Mostly business and health	Business, health, or education	Education or business and health	Education or business and health
Likes dogs	Yes	Yes	Yes	Yes	Yes	Mostly yes	Yes
Likes cats	Mostly yes	Mixed	Mostly yes	Mostly yes	Mixed	Mostly yes	Mostly yes
Have kids	Not reported	Mostly no	Not reported	Mostly yes	No	No	Mostly no
Wants kids	No opinion	No opinion	No opinion	No opinion	No opinion	No opinion	No opinion
Religion	Others	Jewish	Atheist	Christian/Catholics or others	Mixed	Mostly non-Christian/Catholics	Mixed (Mostly Christian/Catholics)

Based on the results from our K Means analysis we can interpret each cluster at a high level as:

- Cluster 1: Straight late 20s non- vegetarian users with mostly white ethnicity
- Cluster 2: Straight late 30s Jewish users with a mixed diet, white ethnicity, and no kids
- Cluster 3: Late 20s non-vegetarian atheist users with white ethnicity and high drug usage
- Cluster 4: Straight early 40s non-vegetarian users with white ethnicity, kids, and low drug usage
- Cluster 5: Straight early 30s non-vegetarian users with Asian ethnicity, no kids, and low drug usage
- Cluster 6: Mid 30s vegetarian users with mostly with ethnicity, no kids, and average drug usage
- Cluster 7: Straight early 30s users with a mixed diet, Black ethnicity, no kids, and average drug usage

⁸ Refer to Appendix - Table 1: Cluster centroids by feature

Alternative Model – PCA + K Means

Although our previous selected model is good for clustering, we might be omitting some features that could be very important when matching users with their other half. However, having so many features in our model might lead to problems as some features might be repetitive, or information overload. To address some of our previous model's limitations, we decided to explore Principal Component Analysis to create independent features (principal components) that can be later used for clustering.

To develop this model, PCA was applied to all features creating 71 principal components that are independent linear combinations of the original 71 features. However, our model will not use all the created components, so we used the Percentage of Variance Explain plot to determine our optimal number of components. For our PVE threshold of 80% the number of components was 45.

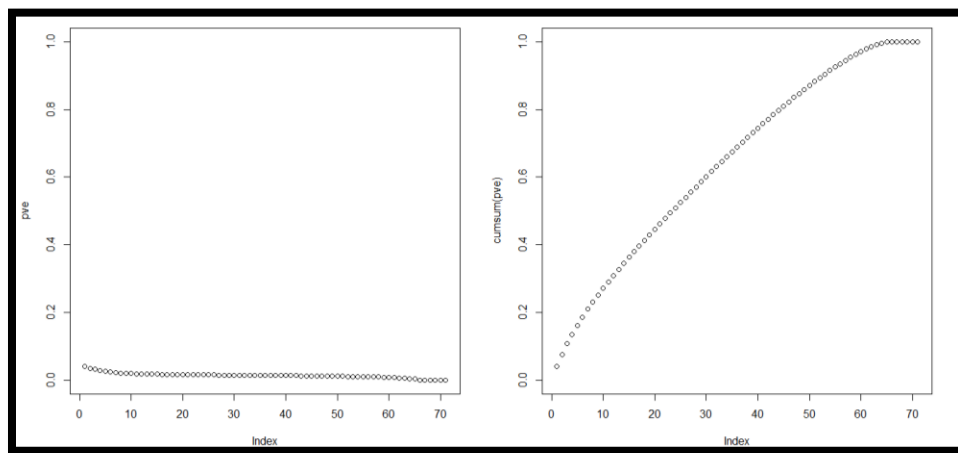


Figure 4: Percentage of Variance Explain plot

Next, the original dataset was transformed to its PCA form and standardized to apply K Means clustering, resulting in 7 clusters. Note that while we have 7 clusters of similar users, we lose the ability to interpret clusters as clearly as we did in our previous model using manually selected features.

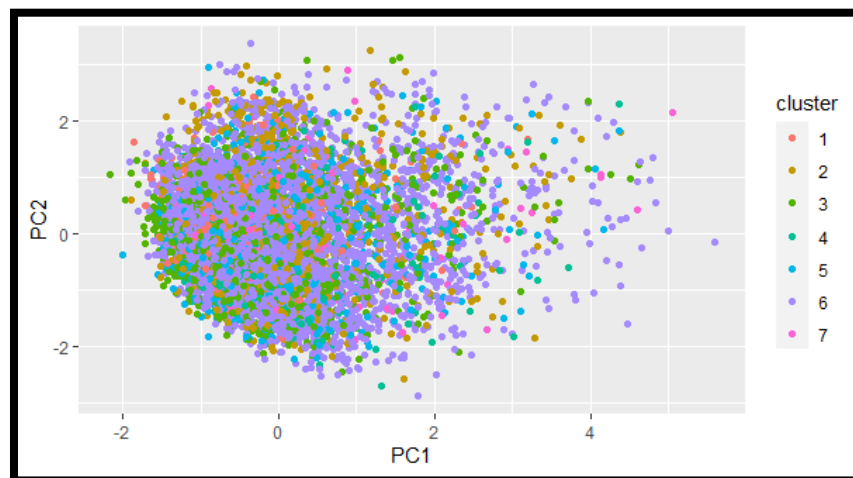


Figure 5: PCA K Means Clusters in 2D plot

Conclusions and Recommendations

To summarize, we first decided to explore our dataset with its sizeable number of features with Principal Component Analysis. We were able to discern which features were related (or had no relations) and by how much in terms of correlational strength. By using PCA, we were able to perform exploratory data analysis in a methodological way even with upwards towards 70+ features.

With the knowledge gleaned from our data exploration, we were then able set out and manually select a list of around 20 features that were the most relevant in representing the of the users in the dataset. Armed with these selected features, we were able to perform K Means clustering in order to segment users into 7 groups where within each group's users had the most similar characteristics, tastes, and dispositions. We were then able to do a deep dive into each of these 7 clusters and look at the characteristics and interpret what each segment looks like.

We also explored clustering with K means but instead of manual features, we used the principal components from our PCA. This approach theoretically produces even more representative clusters, but we lose a bit of interpretability in exchange. However, in our scenario / objective, if we did not care of cluster interpretation and just wanted to group our users in the best way possible, this is useful. Our next steps would be to use our generated clusters and do matching / recommendations of users to each other within their assigned cluster. The roadmap or extension will likely entail filtering each group by the appropriate sex and orientation and then you would have the matches which are compatible to each other.

One thing to know is that the nature of our data, being from an online dating platform, is that it's self-reported at the end of the day. Users that fall into the online dating scene are probably already of a specific kind of people in terms of age and demographic. People are people and may not always be honest and therefore there are limitations to how effective the recommendations will be due to all these factors mentioned above. Perhaps that is nature of the game after all and that whether or not a date, relationship, or love actually results, there will always be an element of chemistry, kismet or serendipity no matter how data driven our approach turns out to be, as it is only one (but arguably a large) part of the equation.

Appendix

Appendix Figures

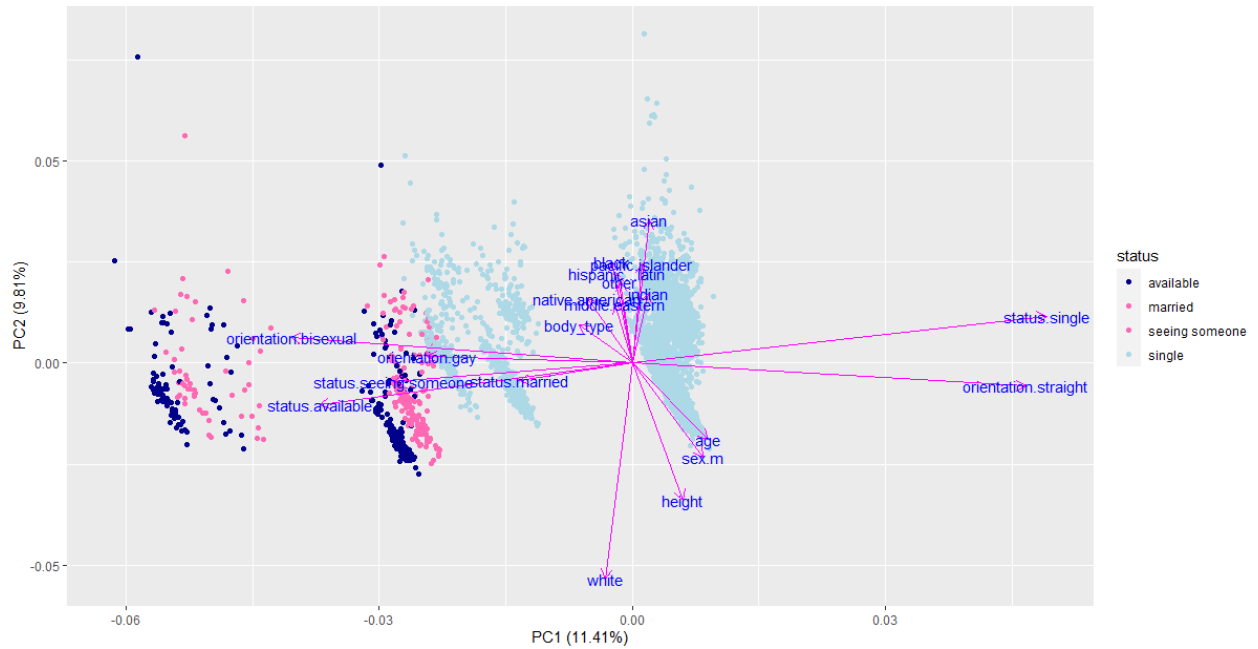


Figure 1: Principal Component Graph – Physical Features

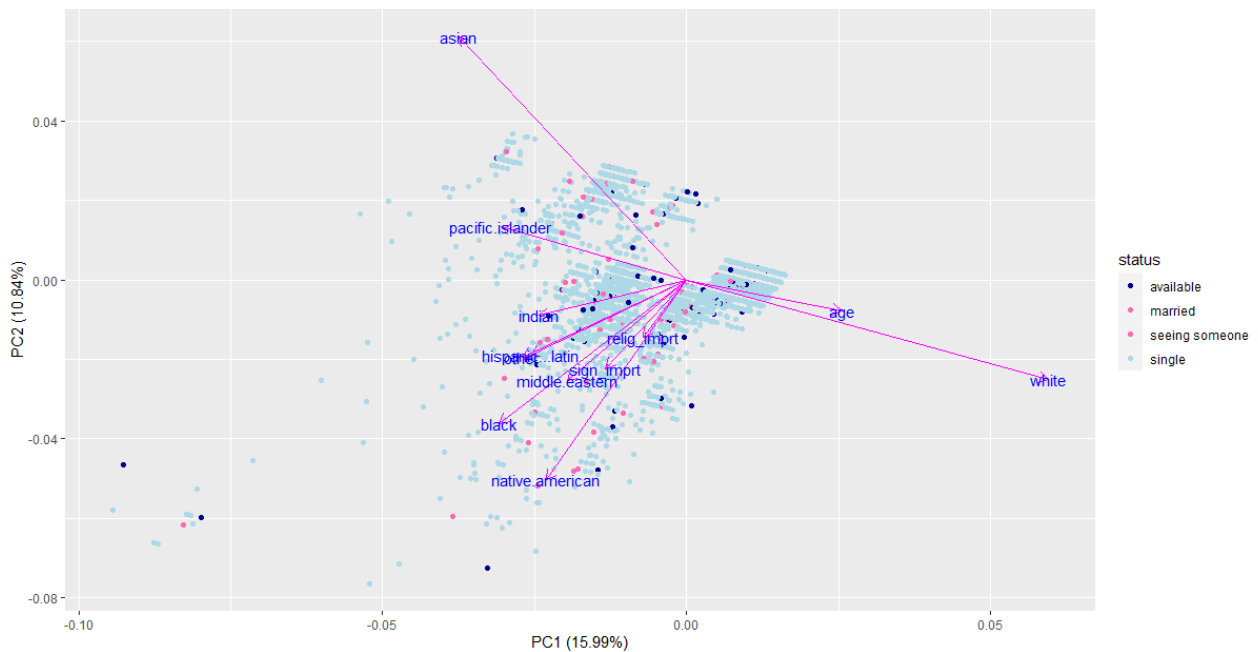


Figure 2: Principal Component Graph – Ethnicities

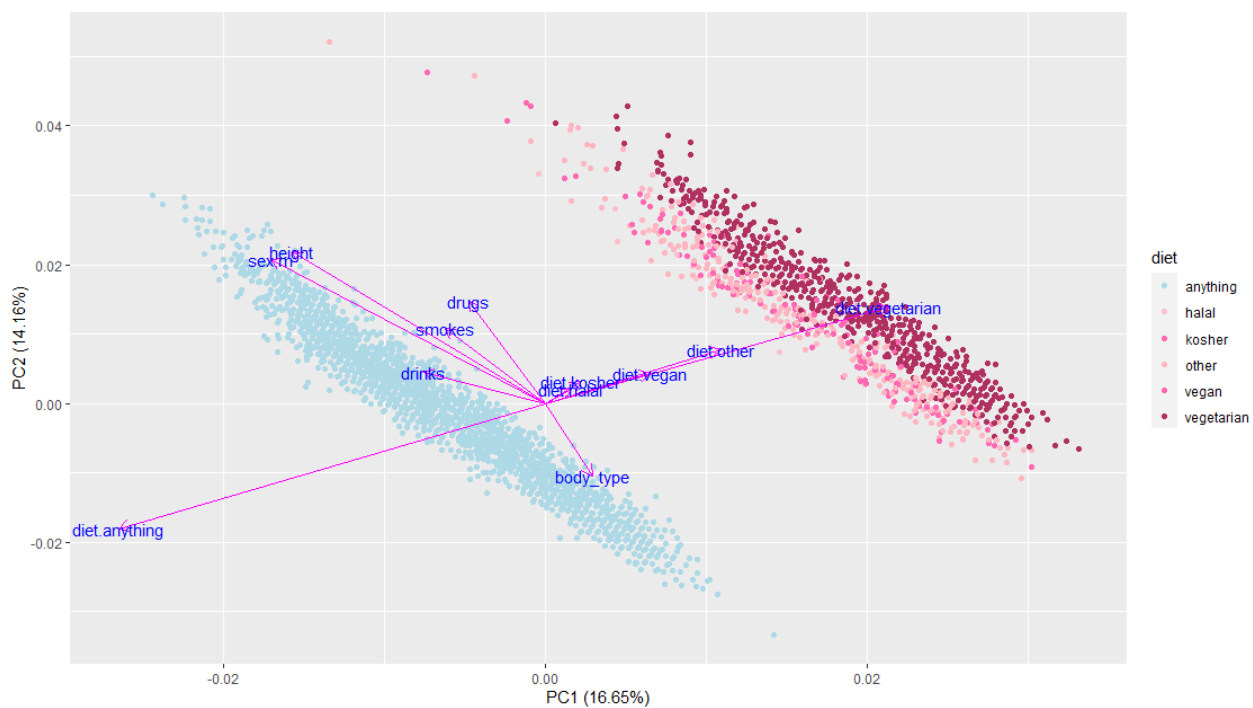


Figure 3: Principal Component Graph – Diet and Preferences

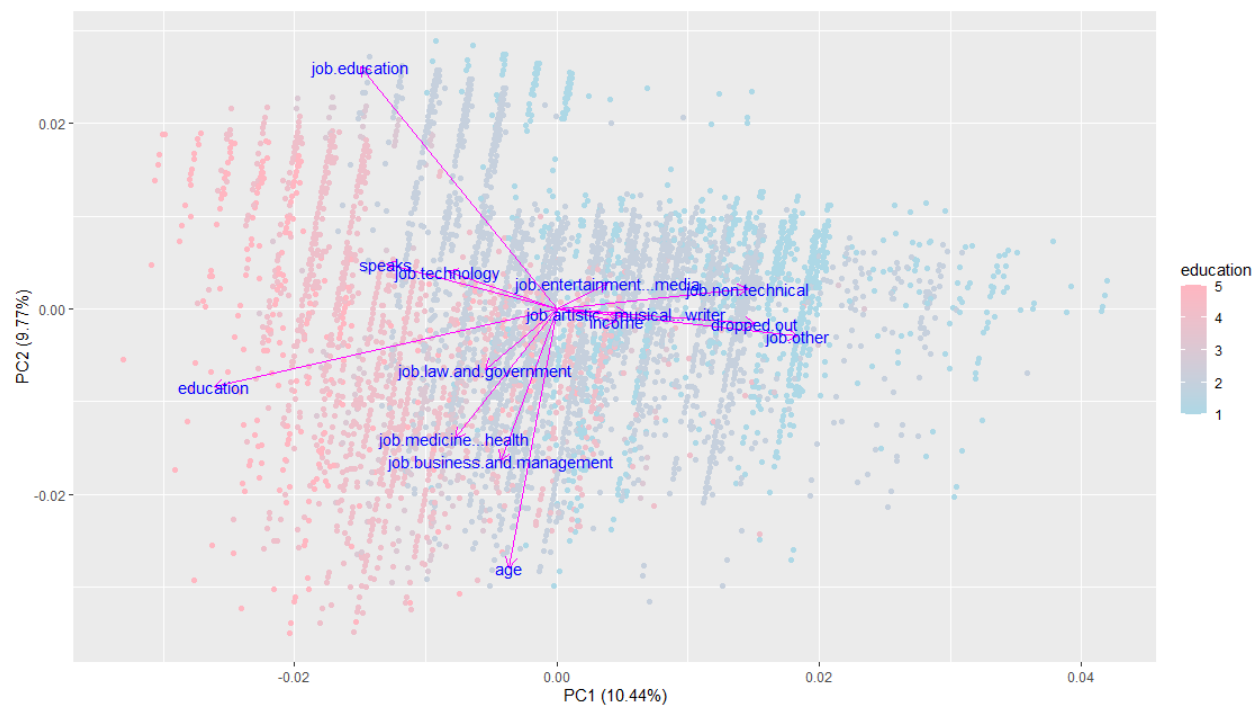


Figure 4: Principal Component Graph – Job and Education

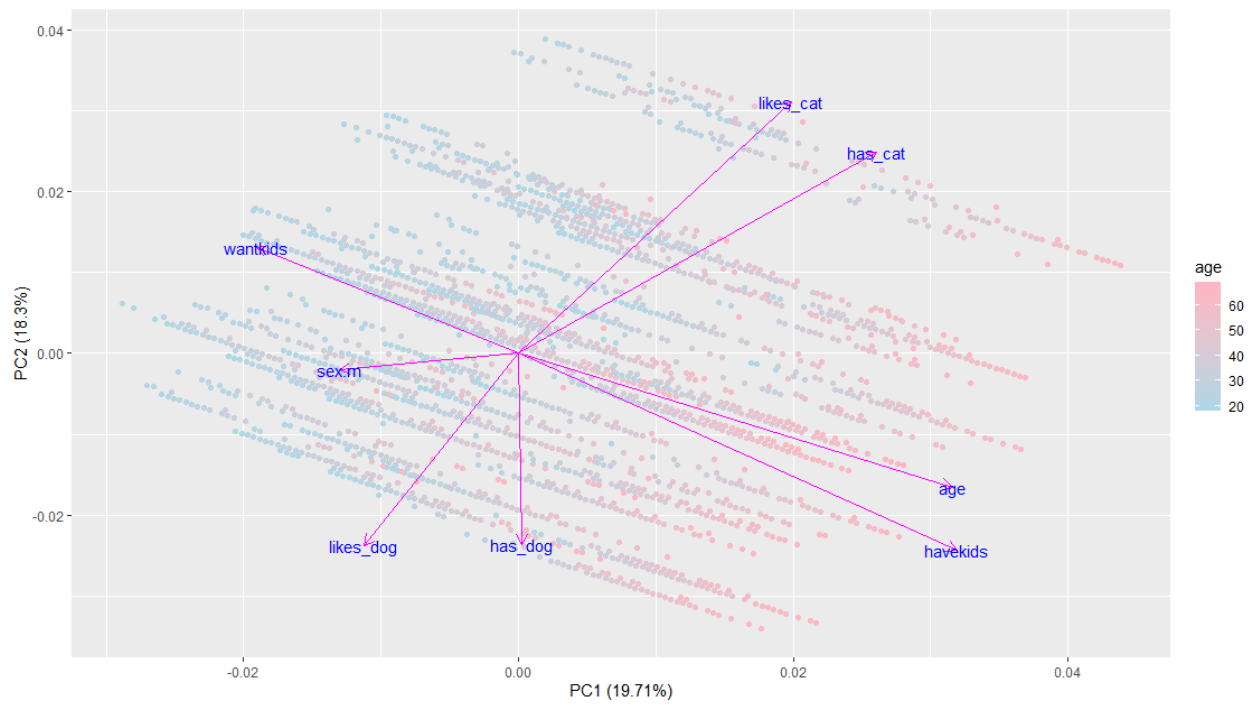


Figure 5: Principal Component Graph – Kids and Pets

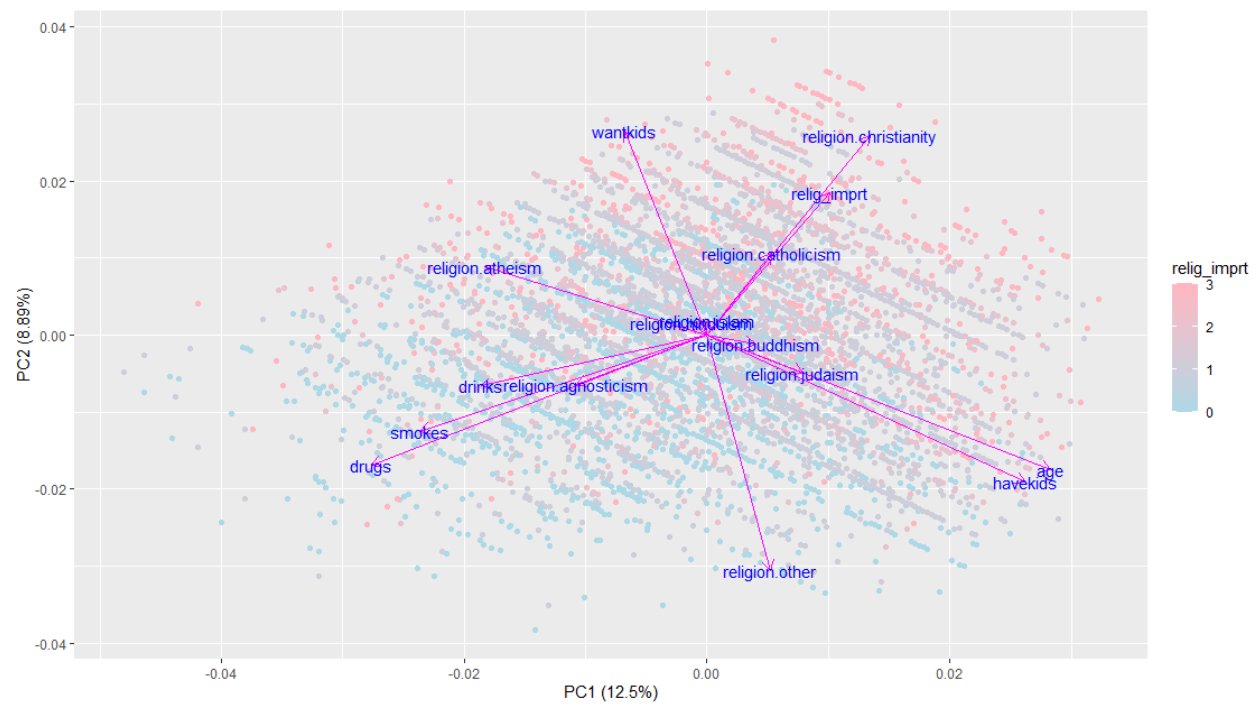


Figure 6: Principal Component Graph – Religion

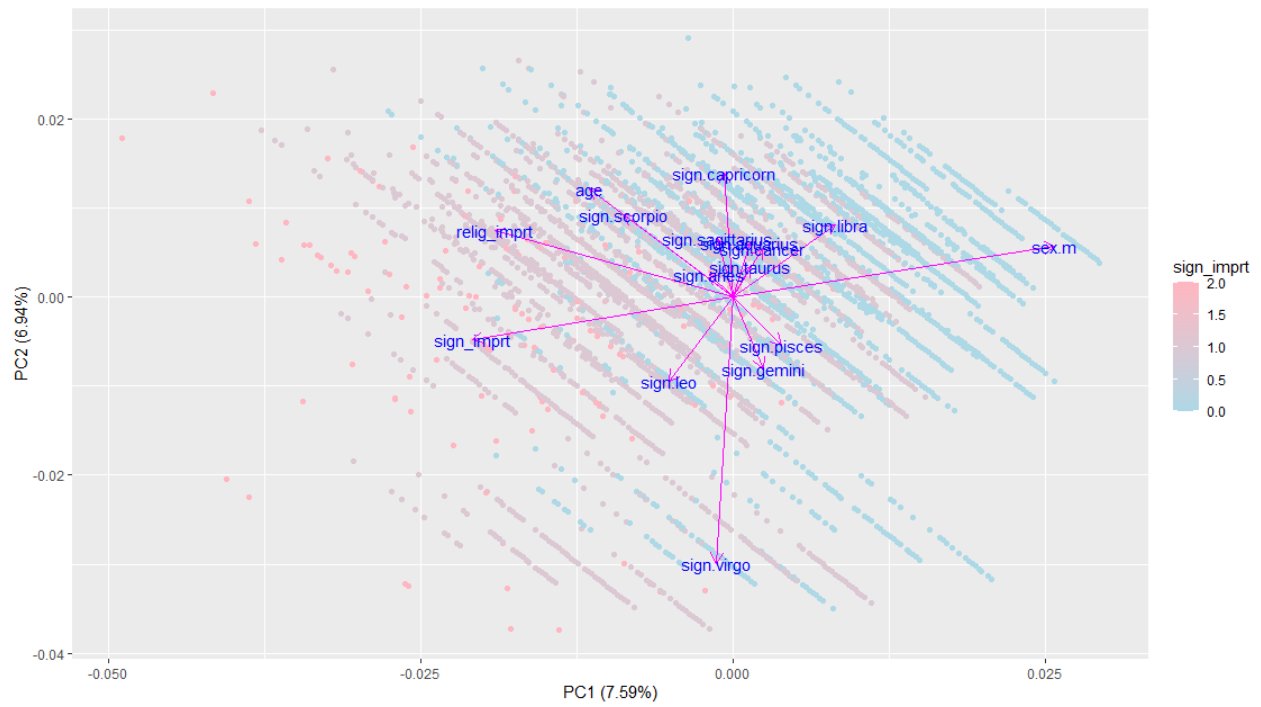


Figure 7: Principal Component Graph – Astrological Signs

Appendix Tables

Table 1: Cluster centroids by feature

Attribute	CLUSTER R 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6	CLUSTER R 7
Orientation.straight	0.87	0.94	0.85	0.94	0.90	0.80	0.87
Orientation.gay	0.07	0.04	0.07	0.04	0.06	0.09	0.07
Diet.anything	0.93	0.72	0.94	0.91	0.93	-	0.77
Diet.vegan. vegetarian	-	0.21	-	0.00	0.03	0.99	0.12
Drugs	0.26	0.17	0.39	0.09	0.13	0.28	0.21
White	0.80	0.94	0.92	0.86	0.12	0.85	0.24
Black	0.00	0.01	-	-	-	0.00	1.00
Asian	0.00	0.01	0.02	0.00	1.00	0.05	0.07
Age	29.70	38.44	29.89	42.23	30.09	34.93	30.91
Education	2.20	2.80	2.32	2.43	2.37	2.63	2.14
Dropped out	0.06	0.03	0.10	0.04	0.03	0.06	0.08
Job.education	0.21	0.18	0.18	0.03	0.15	0.20	0.24
Job.non.technical	0.12	0.04	0.08	0.05	0.07	0.06	0.09
Job.business.and. health	0.03	0.31	0.13	0.64	0.29	0.19	0.19
Likes_cat	0.68	0.59	0.73	0.65	0.56	0.77	0.61
Likes_dog	0.94	0.94	0.90	0.93	0.95	0.87	0.93
Havekids	-0.11	0.21	-0.12	0.47	-0.01	-0.01	0.12
Wantkids	-0.37	-0.37	-0.44	-0.55	-0.34	-0.41	-0.36
Religion.christianity. catholicism	0.32	-	-	0.48	0.34	0.12	0.47
Religion.atheism	-	-	1.00	0.01	0.12	0.25	0.06
Religion.judaism	-	1.00	-	-	-	-	-