Final Project

Hobbies and Desirable Attributes in Speed Dating

upf. BARCELONA SCHOOL OF MANAGEMENT

M.Sc. in Management

Advanced Statistical Methods

Professor

Michael Greenacre

Handed in by

Fabian Pfeffer

25.06.2022

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The following report describes the methodology and results of a data analysis project conducted throughout the course *Advanced Statistical Methods* taught by Michael Greenacre at UPF Barcelona in 2022. In particular, this report focuses on the analysis of a dataset that captures the results of a speed dating experiment conducted in 2004 at Columbia Business School.

The analysis which I conduct has mostly been driven by an explorative approach. However, from the very beginning I was interested in a particular subset of the variables of the dataset. Apart from the satisfaction that the participants expressed towards their matched partners in the speed dating setting, the dataset in question contains various attributes through which individuals could express their general interests and desires when finding a partner. Most importantly for this report, the participants rated their interests in a set of hobbies such as sports, movies, or shopping, which became the object of interest for a major part of the following analysis.

Besides that, to capture some variables that are particularly related to the dating setting, another set of variables which describes the preferences of each individual regarding six attributes of potential partners has attracted my attention. Namely, these six attributes are sincerity, fun, ambition, attractiveness, shared interests, and intelligence.

Concerning the methodology, I first prepared the dataset for analysis. More detail on this will be given during the report. For the hobby attributes, as well as for the desirable feature attributes, I conducted a principal component analysis (PCA) to find associated and dissociated hobbies and attributes respectively. I also investigated potential differences in the PCAs with respect to gender. Lastly, I conduct a K-means clustering analysis for the desirable attributes. This allows the identification of three groups of individuals that expose different preferences when it comes to choosing a partner.

After a description of the dataset and the methodology, the results are presented and analyzed. Finally, a conclusion is drawn and some ideas for further analysis are given.

2. Data

The dataset that is being used in this study was collected by Ray Fisman and Sheena Iyengar from the Columbia Business School in the period of 2002 to 2004 for their paper *Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment* (Fisman & Iyengar, 2006). Essentially, the researchers subjected 551 participants in 21 waves to a speed dating setting, in which each participant would be paired with each other participant of the opposite gender in the same wave for four minutes. If both participants of one pairing were interested in each other after the time elapsed, they were provided with each other's contact information.

This experiment setting was intended to imitate the process of private speed dating programs, which were popular in the region of New York at the time (Fisman & Iyengar, 2006). The subjects were students and employees of Columbia Business School who were then invited to partake in the actual speed dating at a bar/restaurant. Apart from the created matches, the researchers recorded several demographic variables as well as interests of the participants (e.g., music, movies, sports). The core of the dataset is a set of attributes which participants were asked to rate according to the importance of each attribute in a potential partner. In addition, the same attributes were used to ask the participants in how far they think these attributes matter to the opposite sex, in how far they measure up, and how others perceive them. Table 1 displays an extract of the available variables and their respective labels and scales. Due to the multitude of variables, an extract of the dataset will not be displayed in this paper.

Label	**Question**	**Scale**
sports, tv, movies, etc.	How interested are you in the following activities? (17 activities such as sports, movies, shopping etc)	Scale of 1-10
attr1_1, sinc1_1, intel1_1, fun1_1, amb1_1, shar1_1	What do you look for in the opposite sex?	distribute 100 points among 6 attributes: sincerity, intelligence, attractiveness etc.
attr2_1, sinc2_1, intel2_1, fun2_1, amb2_1, shar2_1	What do you think the opposite sex looks for in a date?	distribute 100 points among 6 attributes: sincerity, intelligence, attractiveness etc.

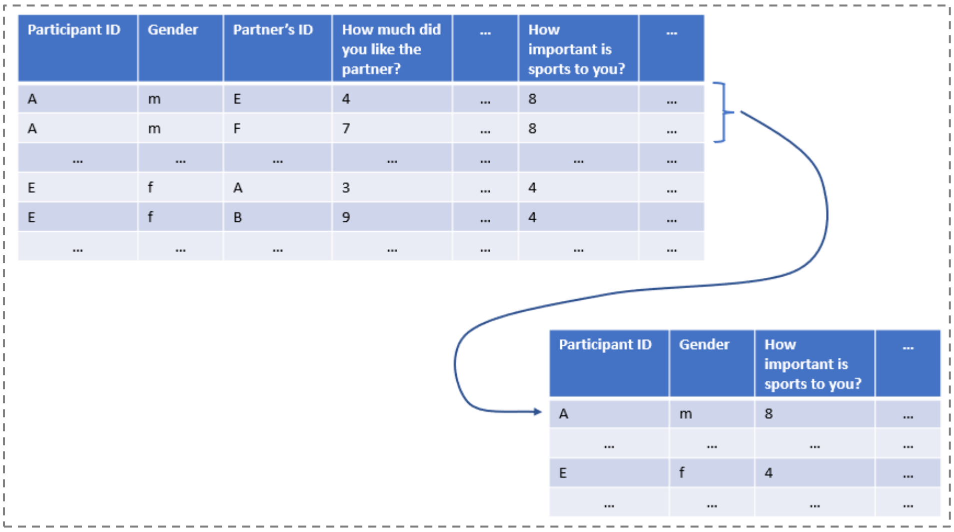
attr3_1, sinc3_1, intel3_1, fun3_1, amb3_1, shar3_1	How do you think you measure up?	scale of 1-10 for the 6 attributes: sincerity, intelligence, attractiveness etc.
attr5_1, sinc5_1, intel5_1, fun5_1, amb5_1, shar5_1	How do you think others perceive you?	scale of 1-10 for the 6 attributes: sincerity, intelligence, attractiveness etc.

Table 1 - Extract of features in the dataset and the respective scales

The variables of the dataset are on different types of scales. The most important ones are either on a scale of 1 to 10 with 10 being the highest importance rating, or on a 100-points distribution scale which required participants to spread points across several attributes. When individuals did not fill in the required data completely, they were removed from the dataset. If an individual did not fill in zeros for the attributes they wanted to allocate zero points, but the remaining attributes summed to 100 points, the record was considered to be complete and zeros were filled for the missing values.

3. Methods

To prepare the dataset for analysis, some transformations were necessary. The original dataset is structured in a way that each row corresponds to one participant A meeting and evaluating the experience with participant E. Or, to put it into other words, one row represents one encounter of two participants in the speed dating. This encounter would then be evaluated by participant A in terms of several other variables such as in how far they liked the other person. However, as this report focuses on general survey information of each participant it was necessary to aggregate the data to individual-level in order to remove the data on encounters and get a tidy dataset of the survey responses per participant. A simplified illustration of the original dataset and its aggregated form can be seen in Figure 1. Through this process and the selection of relevant variables, the dataset was reduced to 26 relevant columns of which 23 contained ratings made by each subject. The number of complete observations after removing incomplete records amounts to 313 records.



Participant ID	Gender	Partner’s ID	How much did you like the partner?	...	How important is sports to you?	...
A	m	E	4	...	8	...
A	m	F	7	...	8	...
...	...	...	...	...	...	...
E	f	A	3	...	4	...
E	f	B	9	...	4	...
...	...	...	...	...	...	...

Participant ID	Gender	How important is sports to you?	...
A	m	8	...
...	...	...	...
E	f	4	...
...	...	...	...

Figure 1 - Aggregation of data to individual level

Concerning the methodology for analyzing the dataset, I first employ a PCA of the hobby columns. A PCA allows to identify variables (in this case hobbies) that are related to each other. I.e., a PCA of the hobby attributes as presented in section 2 will hopefully reveal which activities receive interest from the same individuals and which activities are unrelated or even dissociated (opposed) to each other. To enrich this analysis, I first conduct a PCA for the whole sample containing males and females. In a second step, I subset the sample based on gender to investigate potential differences between the two demographic groups. This type of analysis can reveal hobbies that e.g., for males are closely associated with each other, while being dissociated for females.

For the analysis of desirable attributes, I conduct another PCA. The idea behind this PCA is essentially the same as for the hobby variables. In this particular case however, the principal components can demonstrate which set of attributes are related to each other in a sense that individuals who care about one of the attributes in a set of related attributes also have some likelihood of caring about the remaining attributes of that group. In addition, the PCA of desirable features also reveals in how far individuals who care about one attribute do not care about other attributes. This means that in a speed dating setting someone who is e.g., very sincere may not have to worry extensively about the opposed attributes of sincerity because those attracted to sincerity will not value the opposed attributes. Since the speed dating experiment was set up in a way where females were only matched with males and vice versa, I conducted the PCA directly for each gender separately.

Lastly, I conduct a K-means clustering analysis on the desirable attributes. This approach is employed with the goal of identifying a number of groups in the dataset that can be differentiated and described according to their preferences in a potential partner. The clustering allows to find salient attributes of each cluster and characterize the clusters according to their preferences which ultimately reveals the differences in preferences of the groups present in the dataset. In the following section, I present the results of the employed methods and analyze them.

4. Results

a) Analysis of Hobby Interests

I first want to analyze the results of the PCA concerning the hobbies of the participants. The PCA plot of the hobbies can be seen in Figure 2. Males and females are plotted as “m” and “f” respectively. The arrows represent the hobby vectors of the dataset. When interpreting the PCA, it is important to keep the original question of the survey in mind. Participants were asked in how far they are interested in the listed activities, on a scale of 1-10. The question therefore aims at inquiring about a general interest rather than actual practices (although these two might be strongly correlated).

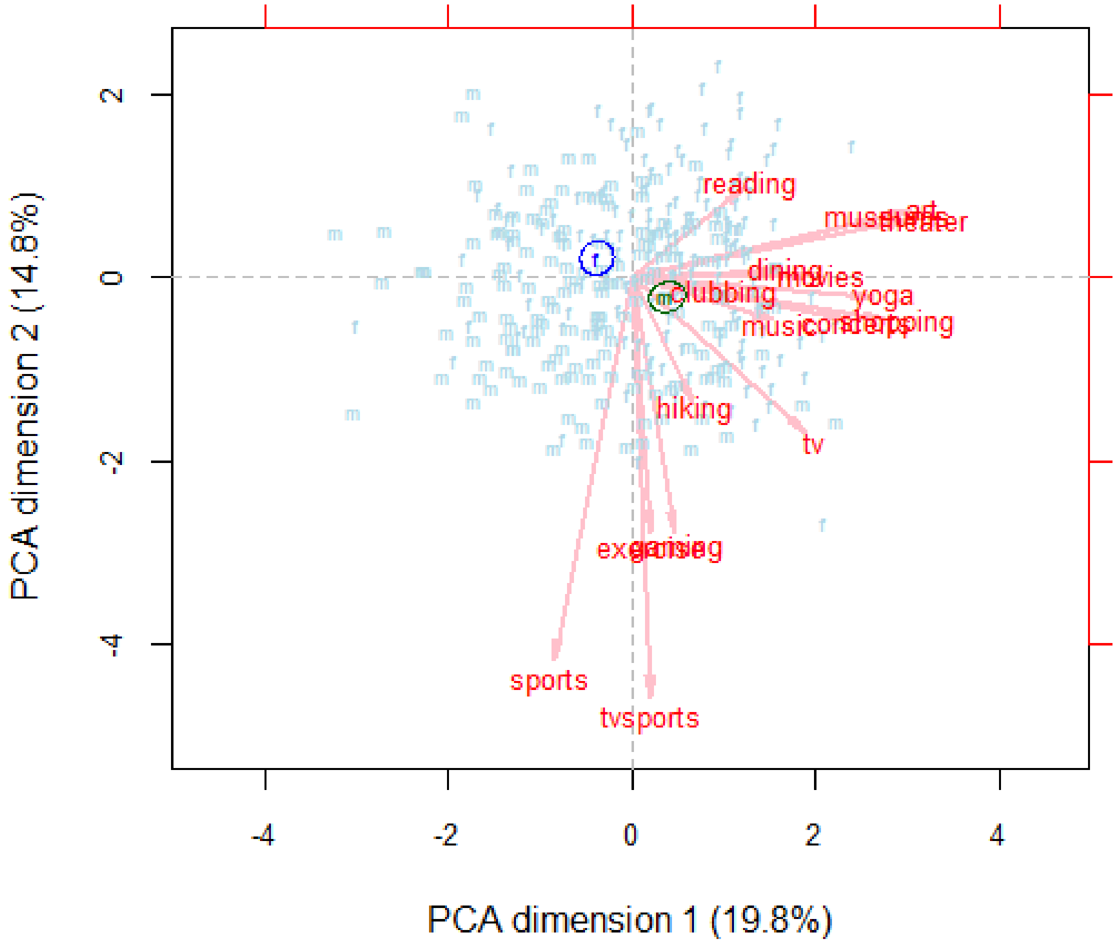


Figure 2 - PCA of hobbies and confidence ellipses for males and females

When looking at the plot, it appears that some hobbies are closely related to each other. Individuals that are interested in one hobby in a group of related hobbies are likely to also be interested in the other hobbies of that group. A very intuitive association can be confirmed by the PCA when investigating the hobbies of sports, tvsports, and exercising. All of these hobby vectors have the same orientation and account for much of the data dispersion on the second PCA dimension. Individuals interested in these activities also seem to have an interest in gaming, as this hobby vector is almost identical to the one of exercise.

Another group of closely related hobbies can be seen in the top right corner. The vectors of art, museum, and theater are almost identical, indicating that most individuals interested in one of these activities will likely have some interest in the other activities as well. The group of artistic hobbies is more or less orthogonally plotted to the above-described group of sports/exercise/gaming. This orthogonality can be interpreted as having no relation to each other. Thus, I deduce that when observing the hobby interests of the whole population, there is no clear tendency that someone who is interested in artistic activities is very interested or not at all interested in sports/gaming activities (and vice versa).

However, there are other hobby vectors which are to some extent opposed to each other. Reading exhibits roughly the same orientation as the artistic hobbies. Ultimately, however, it is at least somewhat opposed to the sports/gaming group of activities. Consequently, individuals tend to be either interested in sports/gaming or reading. Being interested in both is possible, but to some extent unlikely.

Between the already described activities relating to sports/gaming, reading or arts, there is a group of other hobbies containing hobbies such as clubbing, shopping, yoga, or concerts. Some of these vectors such as clubbing are very short, thus only accounting for a small proportion of the dispersion of points in the plot. All of the hobbies in that group have some touch of consumerism or hedonism to them. Some are related to consuming music (concerts, clubbing, music), others are related to simple consumption of goods and services (dining, shopping, movies). Yoga seems to be the black sheep in this group of hobbies as it is a specific form of exercise and hence could be expected to be plotted together with the exercise/gaming group. However, it seems as if people interested in yoga are rather interested in hedonistic hobbies while having no clear correlation (neither negative nor positive) with other sports and gaming.

The two-dimensional PCA plot shown in the graph can explain an accumulated 34.6% of the variance in the data. The above derived insights thus need to be taken with a grain of salt, as some of the generated conclusions might differ when accounting for the third and the following dimensions as they represent a considerable amount of the variance. The first five principal components can be seen in Table 2.

Dim	Value	%	Cum%	Scree plot
1	1.105604	19.8	19.8	*****
2	0.823513	14.8	34.6	****
3	0.625262	11.2	45.8	***
4	0.448911	8.0	53.8	**
5	0.412373	7.4	61.2	**

Table 2 - First five principal components of hobbies PCA

The 95% confidence ellipses of males and females are plotted close to the zero point which means that the two groups do not deviate extensively from the total population's average. However, the ellipses are not overlapping, meaning there is a significant difference between hobby interests of males and females. That is why it might be worth to take a look at the PCA for each gender individually.

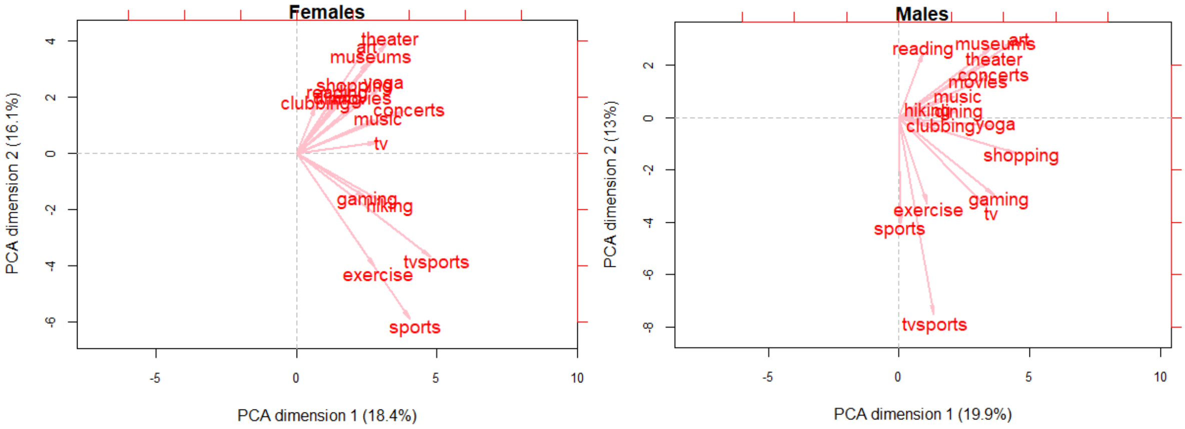


Figure 3 - PCA of hobbies for demographic subgroups of females and males

For the females, we see that along with the other hobbies, the sports/gaming activity group has gotten a new member, which is hiking. For the males, hiking seems to be less strongly related to the other activities of that group. The group, in fact, has dissolved itself a little bit and is now less clearly plotted as one dense set of hobbies. For the males, reading is even more opposed to the sports/gaming groups than observed in Figure 1 whereas for females, reading has a strong association with the artistic and hedonistic hobbies. Overall, there are some minor differences, but most groups is still intact. Females and males therefore differ in associated hobbies to some extent, but largely exhibit similar patterns of related activities of interest.

b) PCA of Desirable Attributes

In the following section, I want to analyze the survey data on desirable attributes when dating. Participants in the experiment had several questions to answer on this topic. The first question of interest was the following one: Please rate the importance of the following attributes in a potential date: sincerity, fun, ambition, intelligence, attractiveness, and shared interests. I computed another PCA for males and females individually and I briefly want to discuss the results.

The two PCAs for males and females are similar on first sight. They have a similar amount of variance explained with 66.7% and 63.9% for males and females respectively. For both plots, attractivity explains the majority of the dispersion along the first principal component. Attractivity and sincerity are for both demographic groups rather unrelated. A difference is that males who care about attractivity tend to care less about shared interests, intelligence, or ambition. For females the negative correlation between attractiveness and the two attributes shared interests and ambition can still be observed but is less strong. In comparison to males, the negative correlation of attractiveness and intelligence is less strong for females, with the two vectors rather gravitating towards orthogonality. Fun seems to be less important to females who care a lot about attractiveness (and vice versa) whereas males on average rank the importance of fun almost independently of attractiveness (indicated by a very short fun vector).

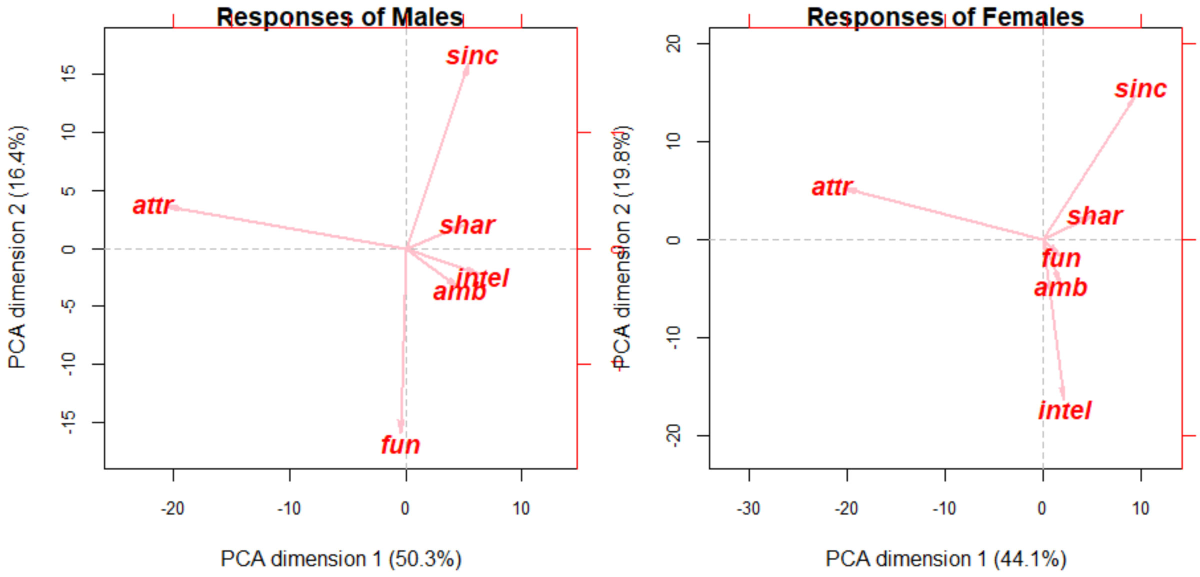


Figure 4 - PCA of desirable attributes according to males and females

Besides that, both genders expose some attributes that are correlated, which means that the individuals who care about one attribute are likely to also care about the other correlated attribute. Males and females share the correlation between ambition and intelligence. As seen in both plots, the vectors for these attributes roughly point in the same direction. Although less obvious, shared interests and sincerity have a similar orientation for females. For males, these two attributes are also somewhat correlated but less strongly.

Overall, the PCA for the males and females mainly reveals similarities in related and opposed attributes with some differences in the composition of the associated and contradictory features that people care about when dating.

c) K-Means Clustering of Desirable Attributes

Next, I will present and analyze the results of a K-means clustering approach. To conduct the K-means clustering, I will first determine which number of cluster means gives the best clustering results. No single statistic should determine the number of clusters. However, as seen in class, I will use the ratio of between sum of squares (BSS) to total sum of squares (TSS) as a first indicator of the optimal cluster number.

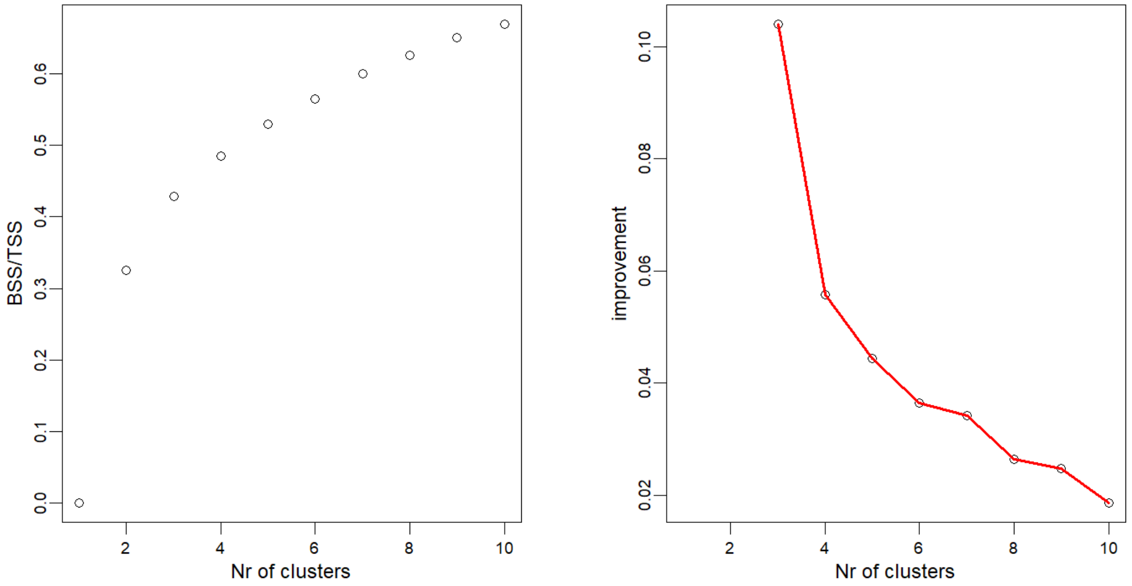


Figure 5 - Different cluster counts with the respective BSS/TSS ratios (left) and the improvement of adding clusters (right)

Figure 4 shows the plots for the BSS/TSS ratios and the improvements on that ratio attained by adding an additional cluster mean. The common way of choosing a good number of cluster means is by looking at a bend (knee) in the curves. Considering the right graph of improvements, we can observe that from two to three clusters (the first point to the left of the plot) the improvement is still significant at a value above 10%. Afterwards, the improvements decrease rather smoothly. This smooth decrease makes it difficult to identify a clear point of the graph for which further improvements are negligible. The right graph of figure 4 therefore does not provide one clear number of clusters.

However, when looking at the BSS/TSS ratios plotted in the left graph, we see that from three clusters to ten clusters, the BSS/TSS curve is almost linear if we compare it to the BSS/TSS curve from one cluster to three clusters. It seems as if there is a bend at three clusters. It would also be possible to argue that the bend can rather be observed for four clusters. But taking the limited sample size of 313 observations into account, I decided to conduct the cluster analysis for three cluster means as to reduce the probability of producing very small clusters.

Having decided on computing three clusters, it needs to be noted that the input data for the clustering was not standardized or transformed in any way. As a reminder, the participants of the experiment were asked to distribute 100 points among the six attributes. Therefore, the scores of the participants can be treated as ratios, as they would allocate a certain amount of the 100 available points to each attribute. That is why I decided against transforming the data to a 0-to-1 scale as the points can already be understood as ratios. This means, that the scales of the boxplots differ as the participants limitation to 100 distributable points caused them to not give the same range of importance to all attributes. To illustrate this, refer to Figure 6 which displays the histograms for the six attributes. As can be seen, participants used almost the whole scale for attractiveness, whereas shared interests only received a maximum rating of 30 points.

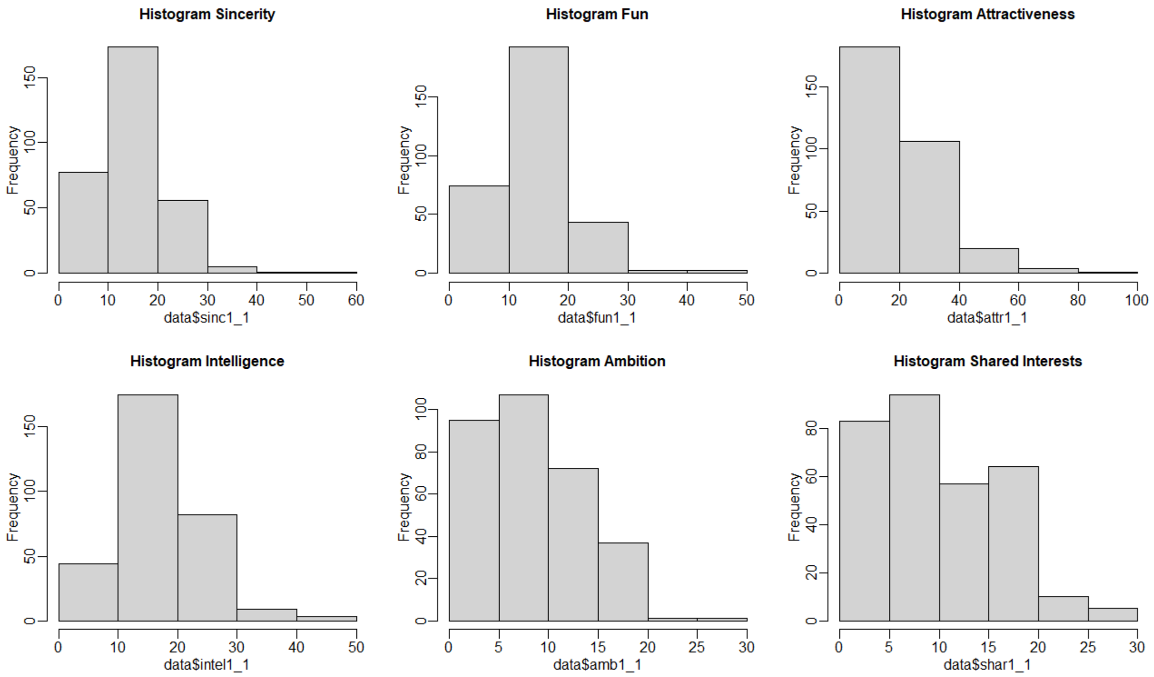


Figure 6 - Histograms of the importance of six attributes

Figure 7 displays the importance of the six attributes for the three clusters in the form of box plots. In the following, I will explain the characteristics of each cluster while building on these box plots.

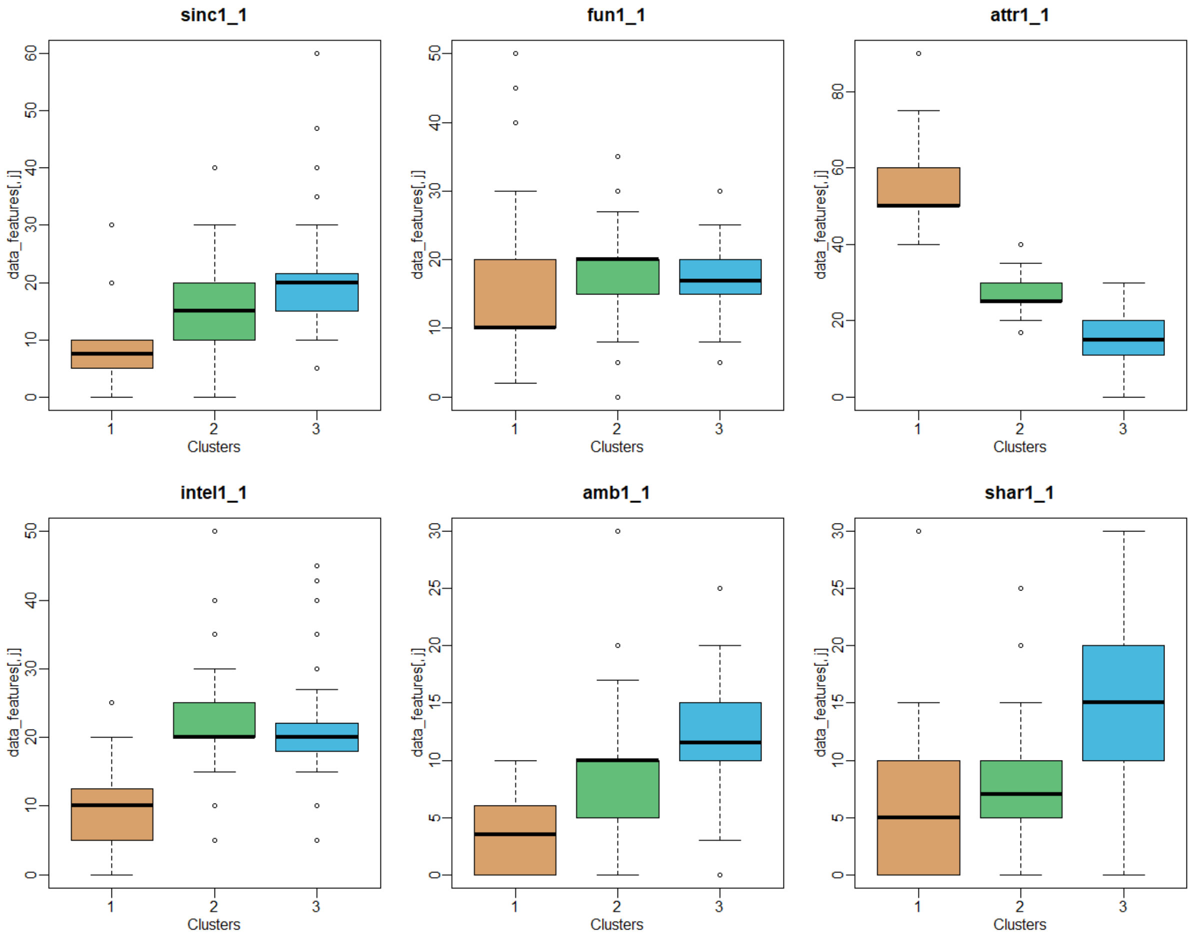


Figure 7 - Box plots of the six attributes for each of the three clusters

Attractivity is the attribute in which participants had the largest dispersion between the minimum and maximum of allocated points. It seems to be an attribute that sets itself apart from the other ones. Concerning the clusters, cluster 1 has rated attractiveness with the highest importance. The participants of that group have given a median of 50 to attractiveness which corresponds to half of the available 100 points. Consequently, members of cluster 1 have few points left to distribute among the other dimensions. Compared to cluster 2 and 3, they care less about all remaining attributes, with ambition and shared interests scoring especially low with medians of 3.5 and 5 respectively. Only the fun attribute seems to draw at least some attention from cluster 1 with the interquartile range between quartile 1 and quartile 3 (i.e., the box in the box plot) ranging from 10 to 20. Thus, after attractivity, members of cluster 1 mostly care about fun and disregard the other attributes. Based on this analysis, I argue that cluster 1 cares about short-term pleasure and wants to be entertained. Members of that group prioritize fun and attractiveness over features which might correspond more to long-term orientation such as ambition or shared interests.

When looking at Cluster 2 and 3, it can broadly be stated that these groups share some commonalities. For various attributes these two clusters expose similar importance scores. That is why I will compare the two clusters apart from cluster 1. One example for similar attribute ratings is intelligence, in which both clusters exhibit the same median at 20 points and the interquartile range is also similar. Resembling to that, members of cluster 2 and 3 have an almost identical box plot for the fun attribute with the interquartile range being the same and the median only differing by three points (20 for cluster 2 and 17 for cluster 3). Cluster 3 has more interest in sincerity as an attribute in a potential dating partner. It also contains some outliers at the upper end of the scale with the highest outlier allocating 60 points to sincerity. The boxes of the two clusters on that dimension have a large overlap though, so they are still somewhat in agreement on the importance of the sincerity feature.

Where cluster 2 and 3 differ visibly is on the remaining attributes. Cluster 2 prioritizes attractiveness compared to cluster 3. The medians differ by 10 points with medians of 25 (cluster 2) and 15 (cluster 3). In fact, cluster 2 rates attractiveness with the highest importance among all attributes. Cluster 3 in comparison sets itself apart from cluster 2 by allocating more points to ambition and shared interests. On both of these features, there is no overlap between the interquartile ranges of the two groups.

Building on the previous assessment of the three clusters, I argue that cluster 1 and cluster 3 having opposing preferences. While cluster 1 prioritizes attractiveness over all the other variables, cluster 3 gives it the lowest importance among the three groups. Instead, members of cluster 3 gave more points to other features. Although cluster 3 still cares about attractiveness to some extent, values like ambition and shared interests were ranked higher in comparison to the other groups. Cluster 2 seems to be somewhere in the middle of the other two. It bears great resemblance with cluster 3 on attributes like sincerity, fun, and intelligence. However, members of cluster 2 decided to allocate the remaining points to attractiveness rather than ambition and shared interests, thus, exposing some common interest in that property along with cluster 1.

Overall, the cluster analysis was able to separate the sample into three sub-samples that expose different preferences. Apart from that, the clusters also differ in size. The number of records as well as the respective proportion of the entire sample in percentage points can be read from Table 3. Cluster 1 is the smallest cluster, followed by cluster 2. Cluster 3 contains more than half of the participants.

	Records in cluster	Percentage	% accumulated
Cluster 1	28	8.95%	8.95%
Cluster 2	113	36.10%	45.05%
Cluster 3	172	54.95%	100%

Table 3 - Cluster sizes and proportions of sample

Finally, I want to report the ratio of between-sum-of-squares to total-sum-of squares. With 0.42924, the cluster analysis using three cluster means captures a significant part of the total sum of squares while maintaining a number of clusters that is low enough to make reasonable interpretations.

5. Discussion and Conclusion

The analysis of the speed dating dataset collected at the Columbia Business School has revealed some interesting insights. After preprocessing and aggregating the information of 313 respondents, an assessment of the hobbies of the participants and the searched-for attributes in potential partners was conducted. For this purpose, principal component analysis and K-means clustering have been employed.

Through the PCA of the interest in hobbies, several groups of related hobbies could be identified. These include for example sports/tvsports/exercise, music/art/theater, or dining/movies/concerts/music/shopping. Individuals who care about one hobby in one of these groups are likely to also show some interest in other hobbies of the same group. In addition, it was possible to identify which hobbies are unrelated to each other. An example is the group of sports-related activities which is uncorrelated with the hedonistic hobbies comprising e.g., dining and concerts, therefore exposing no clear association of these two groups.

The PCA applied to the hobbies of each gender had similar outcomes as the PCA applied to the whole sample. However, some differences arise such as reading being opposed to sports for males whereas for females there is no clear association or contrariety of the two hobbies.

Concerning the desirable attributes in potential partners, the PCA for each gender largely revealed similar related and opposed attributes. Both genders showed the same constellation for attractiveness, sincerity, shared interests with the former two being almost entirely uncorrelated. The comparison of the PCA for the two demographic groups demonstrated that those individuals who care about attractiveness of the opposite gender generally care less about shared interests and ambition. However, some differences for males and females arose. For males, the opposing attributes to attractivity exhibit a stronger opposition than for females.

Especially intelligence does not seem to matter to males who care about attractiveness (and vice versa).

Lastly, the K-means clustering on the desirable attributes variables has yielded three groups of individuals with different preferences. The smallest cluster 1 put great emphasis on attractivity and had little interest in the other attributes with only fun being at least somewhat important. Cluster 3 cared more about shared interests and ambition than the other clusters, all while still allocating some of the available importance points to attractiveness. Cluster 2 seems to take a middle position between the two other clusters with intelligence, fun and sincerity receiving similar scores to cluster 3, while having a stronger tendency to rank attractivity higher in importance, thus, resembling cluster 1 on that dimension. All in all, the three clusters consistently differ on more than one dimensions, thus separating the sample into meaningful groups.

For further analysis, it would have been interesting to analyze in how far hobbies and desirable attributes are related to each other. This could potentially allow for the identification of a set of "dating types", e.g., dating personas that could be characterized by their searched-for attributes and their interests in hobbies. For time and complexity reasons, this analysis has not been conducted in the course of this project.

Besides that, the original dataset offers various other promising variables. One of these is the estimation of participants regarding what the opposite sex looks for in a data. In addition, the information from that question could be paired with the question of in how far individuals think they measure up on the same six attributes. This would allow to analyze what people think others want, comparing that to what they think they provide, and then comparing to what others actually want in reality. This could lead to insights in how far perceptions of others are faulty and how desirable individuals estimate themselves to be compared to evaluation through others.

All in all, the dataset has been proven to contain a considerable amount of interesting information in just a few data columns. The implemented methodology not only improved the understanding of the type of analysis that can be conducted with these models, but also generated some interesting insights into the interests and wishes of individuals when looking for a partner.

6. Appendix

Fisman & Iyengar: *Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment*. The Quarterly Journal of Economics, May, 2006, Vol. 121, No. 2 (May, 2006), pp. 673-697.

Dataset taken from: *https://data.world/annavmontoya/speed-dating-experiment [last confirmed 21.06.22]*.