Milestone 1

Hongyi Shi, Ajkuna Seipi, Nataša Krčo

5th April 2023

1 Dataset

In this project, we will delve into the Kaggle Speed Dating dataset, which captures data from twelve speed dating events held between 2002 and 2004 on the campus of Columbia University in the City of New York. During each event, participants would engage in a series of four-minute dates with partners of the opposite sex. After each date, they would record their impressions of the partner and whether they would be interested in meeting again. Participants also provided basic demographic information, their interests, and their preferences in a partner. All of this information is compiled in the dataset, where each row represents a four-minute date, including the participants' final decision and whether it led to a match.

Although the dataset is rich in information and well-structured, some preprocessing is required. Firstly, we need to create an identifier for each participant based on their demographic data, interests, and preferences. Secondly, the current recording of each date is from the perspective of each participant separately, resulting in two rows for one date. These rows contain information on the interests of only one participant in addition to demographic information about both. Therefore, we need to take this into account when analyzing the data. Section 3 provides more information on the exact preprocessing steps and basic statistics.

2 Problematic

Until recently, meeting potential partners was not such an easy task. People met each other through relatives, friends and acquaintances, at parties, bars, and also through speed dating. Nowadays, with the multiplication of dating apps, such as Tinder, Bumble or Grindr, meeting someone has never been so easy. But ironically, finding a life long partner has never been as hard.

The digitalisation of intimacy raises a lot of questions among scientists. Is there a link between the multiplication of dating applications and the low rate of life long partnership and marriages? Some research states that the possibility of meeting a lot of people easily, provided by dating apps, has changed the view of having a life long partner in the young population. There is no more this fear to stay alone and to not find your other half, or simply someone with whom to share your life. Dating apps provide continually a way to meet easily someone for multiple reasons, whether to find a life long relationship or to share a one-night stand.**love**

On the other hand, we have countries with more traditional cultures that still provide more traditional setups, such as speed dating. Also, the culture of hook up applications is way less present and, furthermore the rate of marriages and relationships is way higher than in some countries such as the US.**marriage**

Finally, we have recently seen a rise of speed dating in Switzerland with companies that provide such services. One of the reasons for this rise, according to those companies, is that people try to meet more

people organically after spending those last years with the Covid pandemic which restricted physical contact with other people.

Does speed dating really work? So, do we see a good matching rate at the end of a speed dating session? Given some criteria, what is the probability of matching between two people?

To extend our visualization, we can also analyze if the initial criteria of the person corresponds to the description of its actual match. This is an interesting point, because by meeting someone organically, we can have a different feeling than messaging them through dating applications. Another interesting aspect is to consider the race, the age and the background of the people that matched. In fact, it has been shown that most people tend to have a relationship with someone from the same background as them and, in some cases, with the same race. As said above, having the chance to meet different people from different backgrounds organically can alter our initial viewpoint, and grow feelings toward someone we never expected.

Therefore, our visualization aims to answer all those questions

3 Exploratory Data Analysis

After preprocessing, the resulting dataset includes information from 16 speed dating events with a total of 404 participants and 5895 dates. The participants are evenly split between genders, with 49% being women and 51% being men. They represent 198 different fields of study, with Law and Business being the most common fields overall. Women tend to study Social Work and Law, while men are more likely to be in Law or Business. However, since the number of people from each field is generally low, no field is significantly overrepresented. Out of all the dates, 17.2% resulted in a match. Interestingly, men were more likely to express interest in another date, with 47% of them being willing to go on a second date, while only 39% of women were interested in doing the same. Regarding partner preferences, participants were instructed to rate certain characteristics on a scale of [0-100], in a way such that the ratings sum up to 100. The following characteristics were given: attractiveness, intelligence, how funny a potential match is, sincerity, shared interests, and ambition. Physical attraction was rated the most important factor with an average score of 24.2, followed by intelligence with a score of 20.4, and a sense of humor with 17.4. Lastly, we take a look at the participants' interests. Music, movies, and dining were the most popular across all participants, and this trend remained consistent across gender. It's worth noting that women, on average, rated interests higher than men, and there was a slight difference in the most common interests, which are depicted in Figure 1.

4 Related Work

The dataset was originally shared as the source of an example from the book "Data Analysis Using Regression and Multilevel/Hierarchical Models" by Andrew Gelman and Jennifer, professors at the University of Columbia in the City of New York, published in 2006. We did not get access to the book and thus can not give more details about the example for which it was used. Then, the same dataset was shared recently, on February 2023, by Ulrik Thyge Pedersen on Kaggle, which is the version we are using for this project. A few codes using this dataset has been shared on Kaggle, most of them being classification prediction models, but none of the code shared contains major visualizations, except the most recent one "Dimensionality Reduction & Feature Selection" shared by Harini Sundar on the 2nd April 2023. As its name suggests, this last code displays some graphics related to PCA, UMAP or t-SNE, and a heatmap containing the top 20 features selected under various conditions. In summary, nothing is quite close to what we are planning.

It was common activity to compute love compatibility between two of your classmates in the school yard,

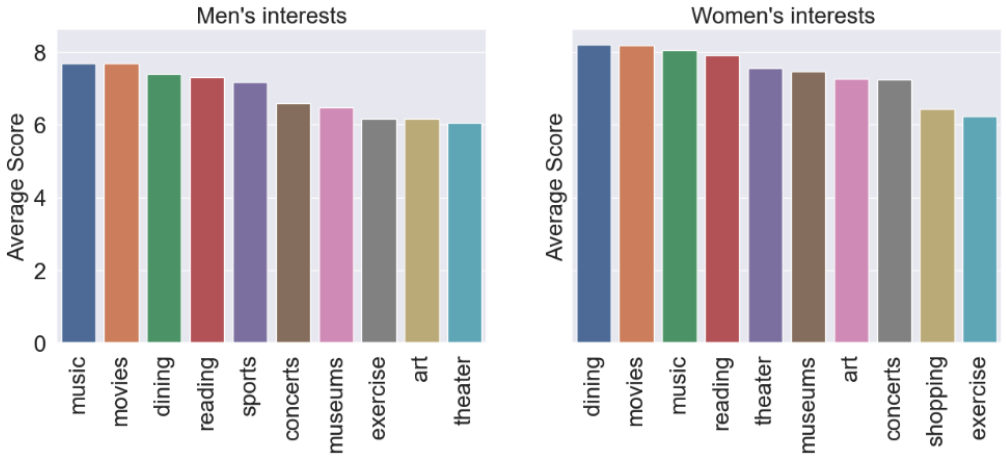

FIGURE 1
Top interests by gender

and there were plenty of different way of doing this: by name, birth date, astrological sign, and more. Kids used to play with letters and numbers to make up some crazy results which, now looking back, probably did not make a lot of sense. Most of the time, those new techniques were picked up from some random teenager magazines, and it used to be popular enough so it was also a thing on various online websites. Although those games did not show any accuracy, there were a lot of fun. So now, what if those results were actually (kinda) reliable? Would it not be even funnier? That is our starting point, getting inspired from our childhood games and display our data in a playful manner in a retro vibe.

To make the speed dating theme show up in the design, we plan to take inspiration from otome games, dating simulation games, making use of diagrams to show compatibility and so on, and from various love calculators that can be found online.