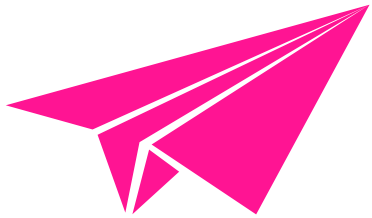
HOW MUCH DO YOU *MATCH*?

process book

OUR PATH



What will be the topic of our website? That was the first question we needed to answer to embark on this journey. We were certain that we wanted to work on something funny yet interesting to us. So, we delved into Kaggle's datasets in search of the perfect gem, and that's when the idea of working on a speed dating data set struck us.

We chose the **speed dating** data set primarily because there were very few visualizations available for it. While there were numerous visualizations and analyses for dating application data sets, speed dating remained largely unexplored.

Once we had settled on the main topic, we conducted data analysis to determine the available information and identify the most interesting aspects to display.

Dataset

Initially, the data set consisted of a series of dates between two individuals, along with their demographic information, characteristics, interests, and whether the date was successful or not. We decided to focus on extracting information about the participants and created a new data set where each row represented a participant, including their demographic information, characteristics, interests, as well as the number of dates and matches they had. This clean data set served as the foundation for our subsequent visualizations.

After completing the cleaning and analysis process, we had to decide on the structure of our website and the types of visualizations we wanted to incorporate.

Initial idea

Our initial idea was to develop a **matching predictor** that would estimate the percentage of compatibility between two individuals based on their demographic and characteristic information, as well as their interests. This idea was inspired by the games we played as children, when we had a crush and wanted to know if they were our soulmate.

From there, it became clear which other types of visualizations we wanted to focus on. Three questions arose:

1. What are people looking for?

2. What are they interested in?

3. Who matched?

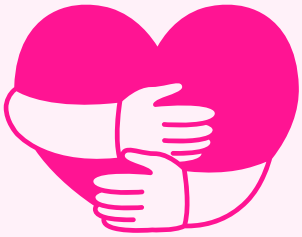
Visualizations

We then had to determine the most suitable type of visualization for each question. Additionally, we wanted these visualizations to be interactive and visually appealing, seamlessly integrating with our website. This is where the coding aspect came into play. We utilized the knowledge we had acquired from our course, employing the D3 library and coding in JavaScript, which, for most of us, was a first-time experience.

From this point onward, we had all the components of our website, and our next focus was on how to structure and organize the visualizations effectively.

Naturally, we planned to place our matching predictor as the final visualization. We aimed to establish a cohesive flow among all our visualizations.

Since the other three visualizations mainly involved visualizing existing data from our data set, we decided to keep them together. Among the three, the visualization pertaining to matches would logically be positioned at the end.



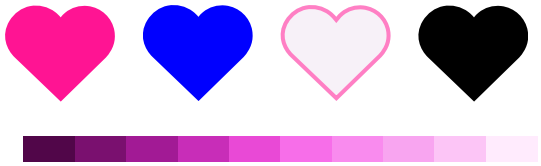
Website design

In terms of visual layout, we envisioned a main page consisting primarily of text, with a title and a brief introduction. The subsequent page would introduce our data set and provide general information. The interactive visualizations would follow, and the journey would culminate with the matching predictor.

Designing a website entailed choosing a visual identity. We focused on selecting a color palette and a consistent typography to be used throughout the website. In this regard, we drew inspiration from the vintage games we enjoyed during our childhood. The visual design predominantly incorporated various shades of pink, complemented by hints of blue. To create contrast with these vibrant colors, black played a dominant role in our website's overall aesthetic.

ROBOTO

Georgia



Feedback

As a final step, we gathered feedback from our friends and family to enhance our design and visualizations. This feedback primarily assisted us in refining the minute details that ultimately made a significant difference in the overall user experience.

The challenges

Compared to other datasets, ours is not the most diverse or largest. We needed to examine the different pieces of information and determine if there was anything interesting or relevant we could do with them.

During our first exploratory analyses, we found that the dataset seemed interesting, but not without its problems. The main issues we encountered were missing or inconsistent values, and identification of date participants.

Missing values

Concerning missing values, we imputed those that were important, but had to drop certain columns, such as the number of matches a person was expecting to get. As for the inconsistent values, they primarily manifested in the participants' field of study. Since participants had the freedom to write down their field of study using any phrasing they preferred, each field appeared in the dataset with multiple variations. To address this challenge, we constructed a map by hand that associated each phrasing with it's corresponding field, and used it to transform the values.

Identification

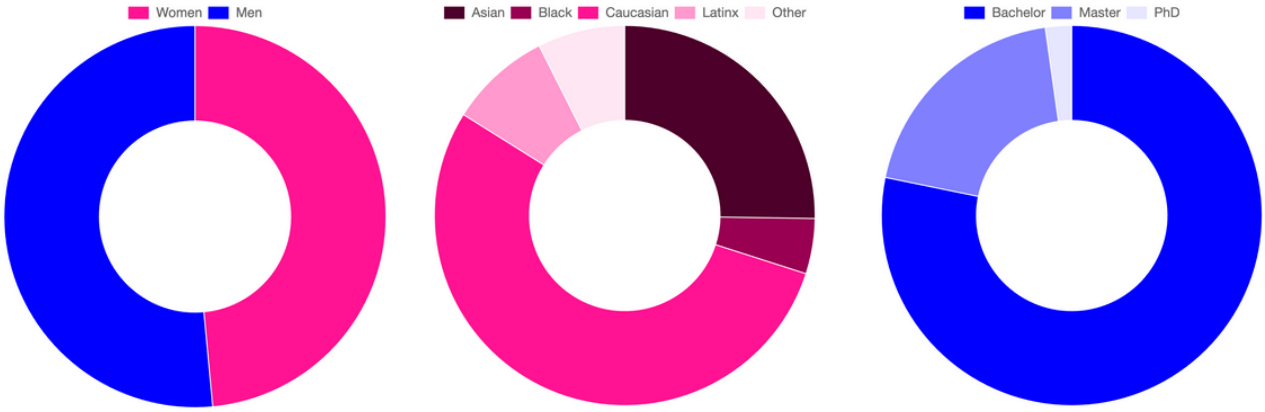
The final and most significant dataset challenge was the fact that the participants were not given an identifier in the dataset - specifically, the dates were logged from the perspective of one participant, with information about the "main" participant, and the "other" participant. Enough information about the "main" participant was included to allow us to construct an identifier for them, but there was not enough information to identify the "other" participant. Consequently, we had to abandon one of our initial visualization ideas - to display a bipartite graph of the participants, connecting them when there was a match. However, since we couldn't clearly distinguish the pair of individuals participating in the date, we dropped this idea and had to rethink to another type of visualizations that would display the number of successful dates.

Lastly, we faced the challenge of structuring our website. Initially, we created each visualization separately, and later we had to develop a coherent structure and narrative to follow. Initially, we did not have an introduction page that provided essential basic information about our dataset. It is crucial to specify the study's context and the specific group of people that were studied since different contexts would yield different results. Without this introductory page, we were concerned that users would dive too deeply into the dataset without grasping important information.

The sketches

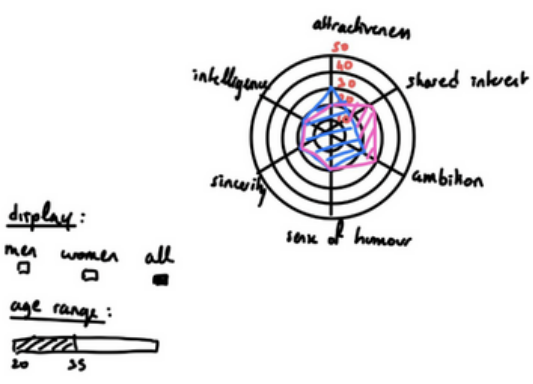
Introduction chart

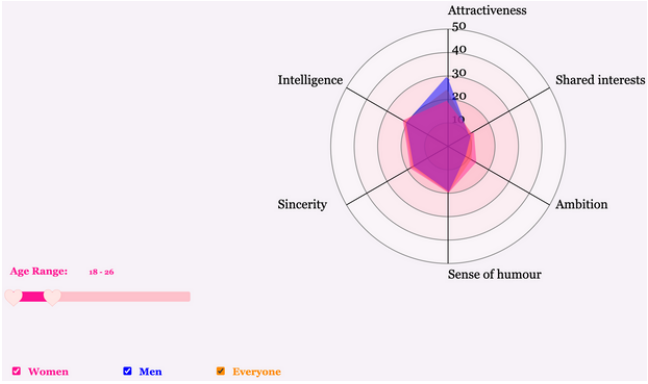
To introduce our analyses and make sure the concept of the events we analysed was clear, before delving into our more complex, interactive visualizations, we decided to include an introduction section, where we explain how the speed dating events worked, and show some basic information about the students who attended them. The goal of these visualizations was to be simple, clean, and informative. Because of this, we decided to include information about only a single attribute per chart - either gender, field of study, or ethnicity. We presented the information in a donut chart, using the Chart.js library.



Spider chart - The characteristics

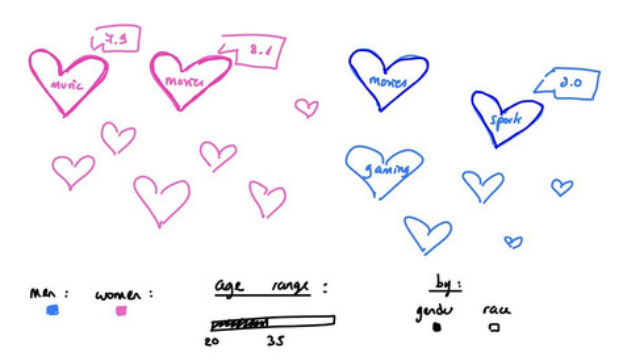
Our first more complex visualization comes in the form of a spider chart representing the participants' preferences - what they look for in a partner. To keep the chart readable and clean, we only include the possibility of comparing the preferences by gender and changing the age group analysed. To implement the chart itself, we modified a spider chart implementation from the D3 Graph Gallery, while for the range slider, we used the jQuery UI JS library.

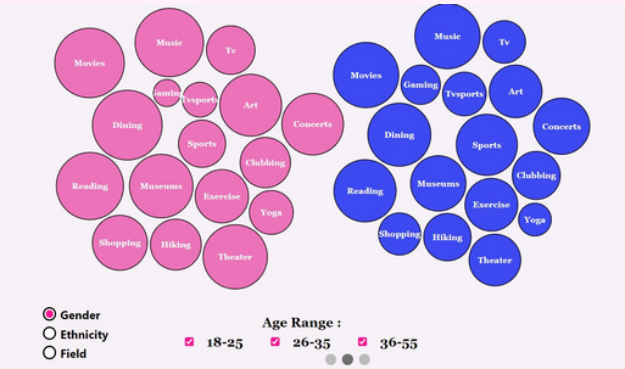




Bubble chart - The interests

Regarding the interests section, we mostly settled with our original design idea. However, we made improvements by incorporating new buttons to display the average interests based on gender, ethnicity, and fields of study/work.
In addition, we replaced the age range slider with check boxes. This decision helped simplify the coding process for the bubble chart. By using check boxes, users can easily select multiple options or choose none at all, providing them with flexibility and the ability to observe any variations.
Unfortunately, we encountered coding issues that prevented us from implementing the original idea of using hearts instead of circles. As a result, we had to stick with circles for the visualization. To implement the bubble chart, we modified an implementation from D3 Graph Gallery.

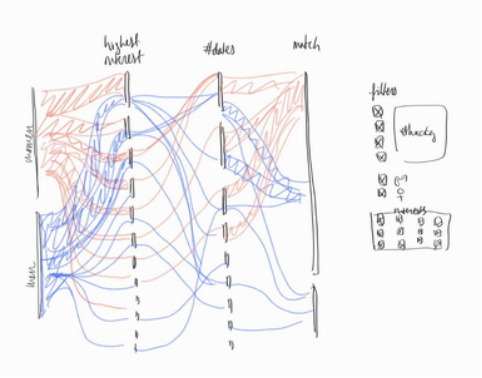


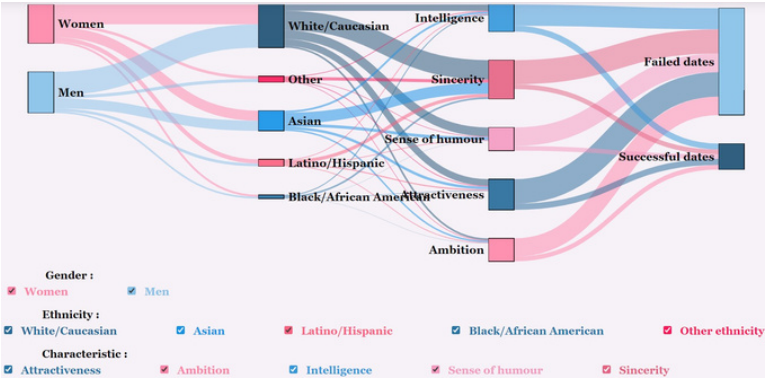


Sankey diagram - The matches

The sankey diagram underwent significant changes during the development process. Initially, our plan was to include the following columns in the diagram: gender, highest interest, number of dates, and number of matches. However, we soon realized that displaying the highest interest would make the diagram unreadable due to the large number of interests (17 in total). To address this issue, we made the decision to replace the highest interest column with the highest characteristic column. This change proved beneficial as we hadn't utilized the individuals' own rating characteristics in the previous visualizations.
Additionally, we encountered limitations in the diagram's implementation that prevented us from having separate columns for the number of dates and number of matches. Since the dataset didn't contain specific fields for the number of dates, only numbers, we were unable to add links and nodes to represent this information in the diagram. As a result, we opted to include a final column that displays the number of successful and failed dates. In this context, successful dates refer to the number of matches, while failed dates represent instances where there was no mutual match.

Lastly, we introduced check boxes for each gender, ethnicity, and characteristic. This feature allows users to select their desired options, providing them with enhanced flexibility and usability when exploring the diagram.
To implement the diagram, we modified an implementation of D3 Graph Gallery.

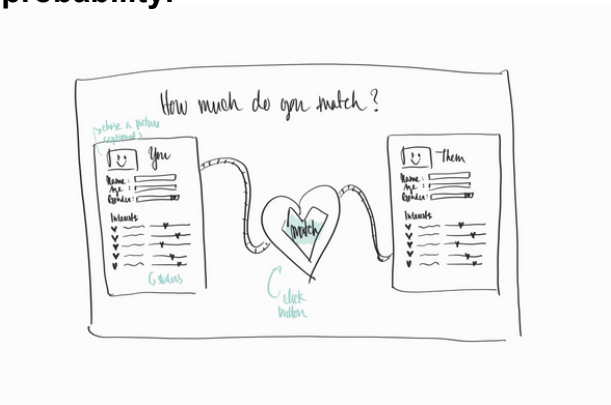


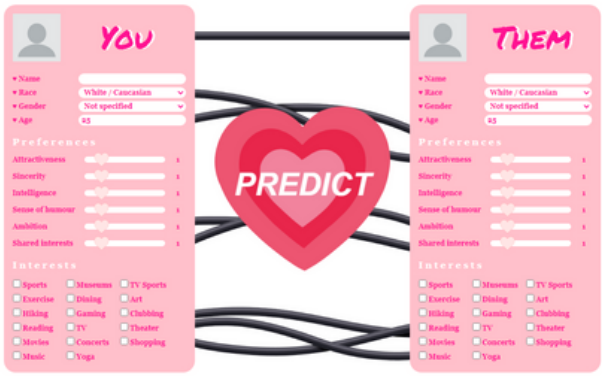


Matching predictor

People love all types of predictions, as evidenced by the widespread popularity of astrology. Taking inspiration from this and quizzes and match prediction games from online magazines from when we were children, we applied this concept to predicting the compatibility of two individuals *before* their date. While accurately determining whether two people will match well before their date is not an easy task, we remained committed to our idea, as we felt it matched the tone of our website and provided a fun and lighthearted final section. The original sketch contains more animations that were a bit too ambitious and thus we dropped and we also added an Interests section to fit the model needs.

To estimate the potential compatibility, we used a Gradient Boosted Tree model. This model was implemented using the Tensorflow library in Python, and we integrated it into our website using the Tensorflow.js library. However, we encountered a challenge during the training process. As most dates did not result in a match, our model was rather pessimistic. Even in cases where matches did occur, the model's confidence was only 70%. To deal with this, we devised a practical approach, comparing the confidence scores to the model's overall behaviour to determine compatibility levels rather than solely relying on the predicted class of 'match' or 'no match', or simply using the raw predicted probability.





Peer assessments

During the whole project, we worked and met every week to discuss about the visualizations, the challenges we faced, and the future work that still needs to be done. We took in consideration each member's ideas and suggestions. Then, we split the implementation work between each one of us. The specific tasks each team member had are detailed below. It is important to note, however, that we worked as a team throughout the project, providing help to each other when it was necessary.

❤ Ajkuna ❤

- Initial website skeleton
- Bubble chart implementation
- Sankey diagram implementation
- Additional data processing

❤ Hongyi ❤

- Initial sketches and website plan
- Website enhancement with interactivity and design refinement
- Match predictor interface implementation

❤ Natasa ❤

- Exploratory data analysis and preprocessing
- Spider chart implementation
- Match predictor model training and website integration