CS532: Final Project Report

OK CUPID PROFILES ANALYSIS

Dishant Bhatnagar
Computer Science Dept.
Thomas J. Watson College of Engineering and Applied Science
NY, USA
dbhatnagar@binghamton.edu

Sagar Naresh Sidhwa
Computer Science Dept.
Thomas J. Watson College of Engineering and Applied Science
NY, USA
ssidhwa@binghamton.edu

Saloni Manoj Wandile
Computer Science Dept.
Thomas J. Watson College of Engineering and Applied Science
NY, USA
swandile@binghamton.edu

I. NOSQL QUERIES

Query 1: Matching Preferences Exploration

1.1 Patterns and Trends in Matching Preferences: Age, Height, and Body Type

In our exploration of OkCupid matching preferences, we delved into both structured and unstructured attributes of user profiles. The initial focus was on structured attributes like age, height, and body type, aiming to discern patterns across different age groups and body types. This preliminary investigation laid the groundwork for a more nuanced understanding of how these attributes influence match interactions.

Transitioning to unstructured attributes, we analyzed the language used in users' essays to gain insights into how self-descriptions impact interactions on the platform. Our goal was to uncover whether specific age groups, heights, or body types would attract distinct types of matches based on these attributes.

The MongoDB aggregation pipeline was crafted to filter and process the relevant data for this analysis. Starting with a match stage that filtered documents based on the presence of age, body type, and essay fields, we then projected these fields for further analysis. We also added a categorization field for age groups and grouped the data by age group and body type to calculate average height and essay language.

The resulting dataset was then used to generate a bar plot visualizing the average height across different age groups and body types. The percentages represented in the legend provide a clear depiction of the distribution of body types within each age group.

1.2 Essay Content Analysis across Age Groups

Building on our previous exploration of OkCupid user preferences, we shifted our attention towards analyzing the textual content of users' essays. This analysis aims to identify common themes and words used by different age groups, thereby providing insights into the language patterns and interests that may resonate within these demographics.

The MongoDB aggregation pipeline was tailored to filter and process documents based on the presence of essential fields, such as age, body type, and essay0. The essays were then split into individual words, and we focused on filtering and counting these words to determine their frequency.

To organize our findings effectively, we categorized the results into three age groups: '19-35', '36-45', and 'Other'. Each category was further analyzed to identify the most frequent words used in the essays. The top 40 words with counts exceeding a specified threshold were selected for each age group to provide a focused view of the prevalent language patterns.

Visualizations were crafted to display these findings, showcasing the top 40 most common words for each age group. The bar plots highlight the frequency of these words, offering a clear representation of the linguistic preferences within each demographic.

Query 2: Profile Popularity Analysis

2.1 Decoding Popularity: Income, Lifestyle Choices Insights

In our exploration of "Profile Popularity," we aim to unravel the factors that contribute to a user's popularity on the OkCupid dating platform. Our investigation spans both structured attributes—such as income, dietary habits, drinking preferences, smoking habits, and gender—and unstructured attributes, notably the essays users write about themselves.

By analyzing structured attributes, we seek to understand if specific factors like income levels, dietary choices, and lifestyle habits significantly influence a user's attractiveness to potential matches. On the other hand, examining unstructured attributes, like the content of users' essays, offers deeper insights into their personalities, interests, and communication styles.

The MongoDB pipeline we employ starts by filtering users based on the presence of essential structured and unstructured attributes. Then, we project relevant fields for deeper analysis, like income, dietary habits, drinks, smokes, and essays. We further categorize income levels into groups to facilitate our analysis.

2.1.1 Income and Smoking Preferences

We explore the relationship between average income and smoking habits, differentiating between genders. Our findings might shed light on whether income correlates with smoking habits and if this relationship varies between male and female users.

2.1.2 Income and Drinking Preferences

Similarly, we delve into the connection between average income and drinking preferences, segmented by gender. This analysis aims to discern whether there's an association between income levels and drinking habits and if this connection is gender dependent.

2.2 Extending Decoding Popularity by Unraveling Preferences: Smoking, Drinking, Gender, and Essay-Based Analysis

In this analysis, we deepen our understanding of profile popularity by exploring the interplay between smoking and drinking habits across genders. We aim to identify patterns in

income distribution and correlate them with lifestyle choices to uncover potential preferences or trends.

2.2.1 Smoking and Drinking Preferences by Gender

We begin by grouping users based on their smoking and drinking habits, segmented by gender. By calculating the average income within these groups, we intend to discern if there's a relationship between lifestyle choices and income levels, and whether this relationship differs between males and females.

2.2.2 Word Cloud Analysis of Essay Content

To gain deeper insights into the personalities and interests associated with different smoking and drinking habits, we generate word clouds for male and female users based on their essays. Word clouds provide a visual representation of frequently used words, offering a glimpse into the topics and themes that resonate most with each group.

Through this multifaceted analysis, we aim to uncover correlations and patterns that illuminate the nuanced dynamics affecting user popularity on OkCupid. By examining structured attributes like income, dietary habits, drinking preferences, smoking habits, and gender, alongside unstructured attributes such as the content of users' essays, we seek to understand the deeper layers influencing user attractiveness and compatibility. Understanding these intricate relationships can provide valuable insights to refine OkCupid's recommendation algorithms, enhancing user experience and fostering more meaningful connections.

Additionally, we present word clouds generated for male and female users, capturing the essence of their essays. These visualizations can potentially reveal underlying themes or common interests, offering further depth to our exploration and understanding of user preferences and personalities.

***Query 3*: Analysis of User Characteristics**

The aim of this analysis is to gain insights into user characteristics by examining their essays from columns essay5 and essay7, in conjunction with their sex and age. Specifically, we seek to determine the user's relationship preferences (love, casual, or mixed) and their preferences on how to spend their weekends.

3.1 Understanding User's Relationship Preferences

The analysis employs three distinct MongoDB aggregation pipelines to discern users' relationship preferences based on their essays contained in column 'essay5'. To ensure that there are no invalid or empty values, the query uses match and group stages. These pipelines effectively categorize users into three distinct groups: 'love', 'casual', and 'mixed' relationship preferences. Through the extraction of keywords indicative of each preference category from the essays, these pipelines utilize regular expression patterns to match synonyms associated with both 'love' and 'casual' preferences. Following this categorization process, the pipelines further group the identified documents by gender, subsequently computing the count of users for each preference category.

3.2 Understanding User's Ideal Weekend Plan

The query aims to uncover user preferences for weekend activities by leveraging MongoDB aggregation pipelines. Through the analysis of user essay 7 and demographic data, the goal is to extract keywords indicative of preferred activities, including socializing, dining out, staying in, and engaging in work-related tasks. By aggregating and processing this information, the query seeks to reveal patterns in user preferences across various gender and age groups.

***Query 4:* Analysis of Habits and Relationships**

The aim is to analyze preferences in habits from essay 2 and essay 6 to understand how they influence relationship statuses. Specifically, explore preferences regarding drinking, smoking, diet, etc., and correlate these with the individual's relationship status.

The queries for diet, drinks, and smokes use a combination of match and group stages to aggregate the data, ensuring that only valid and non-empty values are considered.

The query for extracting keywords from essay columns combines various stages to filter, project, and split text data from 'essay2' and 'essay6' columns, aiming to create a comprehensive list of keywords for each individual.

Overall, these NoSQL queries serve as the foundation for our subsequent analysis, helping to uncover patterns and preferences that can offer valuable insights into how habits may influence relationship statuses.

II. NoSQL Database and Dataset

We utilized MongoDB database to analyze the OKCupid dataset. The dataset contained information regarding dating profiles of users, in which columns essay 0 to essay 9, were unstructured and in these columns, users explained about their habits, dating preferences, work, daily habits, etc.

III. Project Outcome

Analysis 1: Matching Preferences Exploration

1.1 Patterns and Trends in Matching Preferences: Age, Height, and Body Type

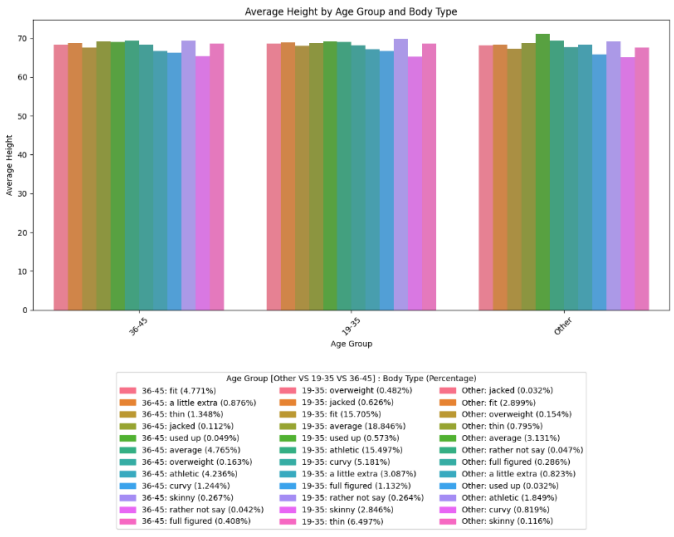

Fig 1.1 : Average Height by Age Group and Body Type

This analysis uncovers intriguing patterns in matching preferences based on age, height, and body type on the OkCupid platform. While the average height shows

consistency across various age groups, the distribution of body types presents notable differences. Specifically, the "fit" body type is predominant among younger users aged 19-35. In contrast, "average" and "a little extra" body types are more evenly represented across all age groups, including those aged 36-45.

The introduction of an "other" body type category adds depth to our understanding, though its distribution across age groups warrants further investigation due to potential data limitations. Despite these considerations, these findings offer a preliminary insight into potential matching preferences based on age and body type. This understanding can aid in tailoring OkCupid's recommendation algorithms to better align with user preferences, ultimately enhancing user experience and fostering more compatible connections.

1.2 Essay Content Analysis across Age Groups

By integrating the findings from both sections, we gain a holistic understanding of how age influences both the physical and textual dimensions of user profiles on OkCupid. The initial analysis (Fig 1.1) highlighted a predominant preference for "fit" body types among younger users (19-35), laying the groundwork for exploring how these physical preferences intersect with language use in user essays (Fig 1.2.1, Fig 1.2.2, Fig 1.2.3).

The essay analysis reveals distinct variations in word usage across different age groups. For instance, visualizations, such as bar charts, depict the most frequently used words in essays for the 36-45 age group, including terms related to career, hobbies, travel, and interests. When juxtaposed with body type preferences for this age group, which tend to favor "average" or "a little extra," these linguistic patterns suggest a potential alignment between self-presentation and partner preferences.

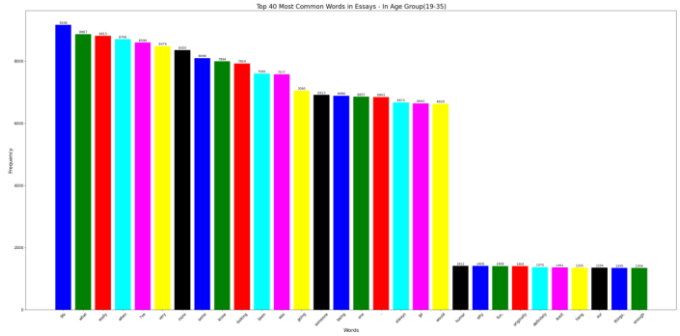

Fig 1.2.1: Top 40 Most Common Words in Essays - In Age Group(19-35)

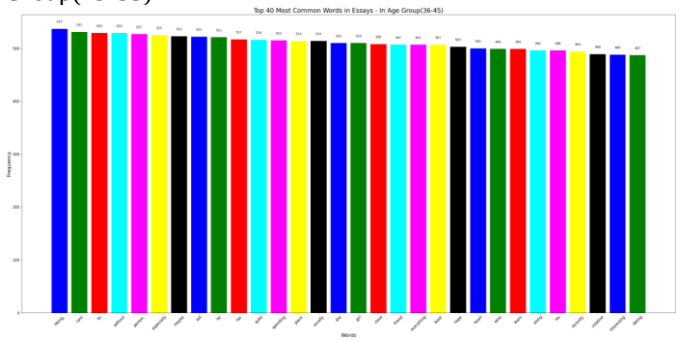

Fig 1.2.2: Top 40 Most Common Words in Essays - In Age Group(36-45)

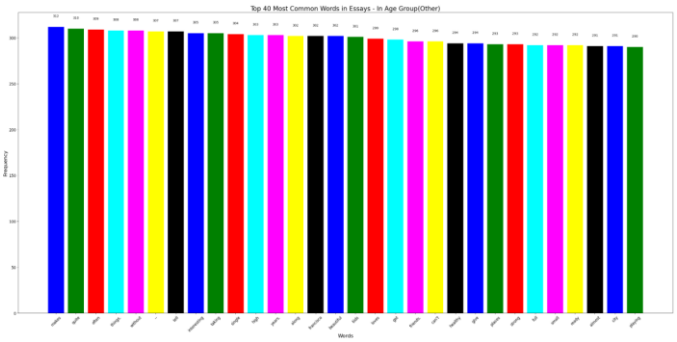

Fig 1.2.3: Top 40 Most Common Words in Essays - In Age Group(Other)

Focus on Interests and Activities: The bar chart (Fig 1.2.2) illustrates frequent words like "love," "time," "life," "work," "family," "kids," "good," "people," "want," and "make." These linguistic choices indicate that users in the 36-45 age group prioritize expressing their interests, values, and relationship goals in their essays. This focus on shared activities or interests may influence their matching preferences, possibly favoring connections based on mutual interests rather than specific body types.

Potential for Diverse Preferences: While the limited view of the chart does not directly link essay content to body type preferences, the emphasis on general interests and life aspects suggests a potential openness to a wider range of body types. This contrasts with younger age groups (fig 1.2.1), where a preference for "fit" body types was more pronounced. Thus, the essay analysis hints at a more diverse set of preferences among older users, possibly reflecting a broader range of interests and values.

Analysis 2: Profile Popularity Analysis

2.1 Smoking and Drinking Preferences by Gender

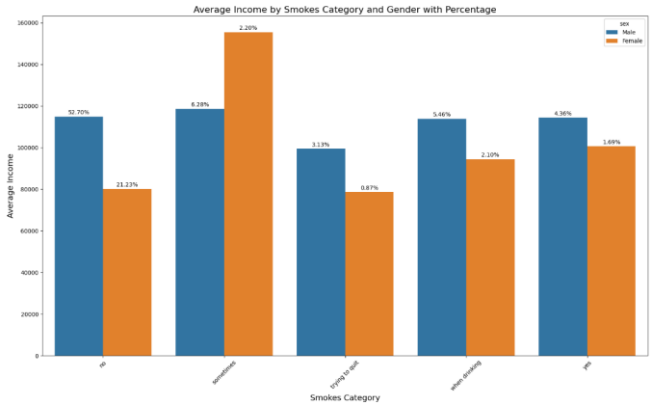

Fig 2.1.1: Average Income by Smokes Category and Gender with Percentage

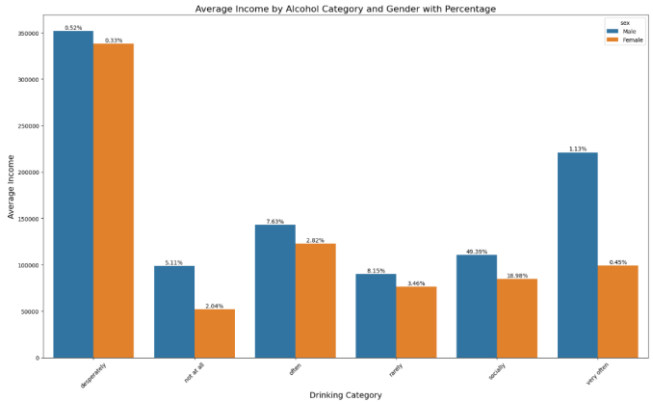

Fig 2.1.2: Average Income by Drinks Category and Gender with Percentage

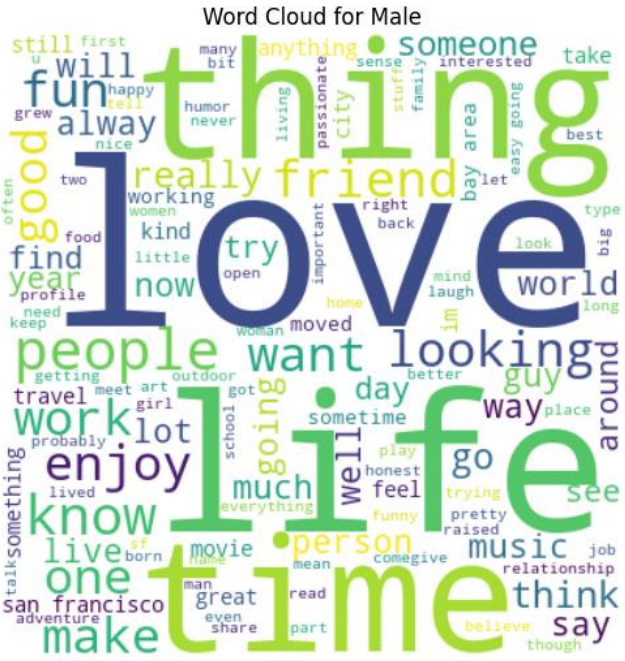

Fig 2.2.1 : Word Cloud for Male based on Smokes and Drinks Category

Fig 2.2.2 : Word Cloud for Female based on Smokes and Drinks Category

In our comprehensive analysis of matching preferences on the OkCupid platform, several intriguing patterns emerged across income, lifestyle choices, and essay content. From (Fig 2.1.1) Non-smokers were found to have a higher average income compared to smokers, suggesting a potential preference for non-smoking partners among certain users, possibly due to health considerations or lifestyle choices. Additionally, (Fig 2.1.1) income variations were observed based on alcohol consumption habits, with social or rare drinkers generally having a higher average income than frequent drinkers. Gender-specific income differences within each alcohol category further highlighted nuanced preferences between male and female users.

On the essay content front, (Fig 2.2.2) the female word cloud predominantly featured terms related to relationships, openness, and positive sentiments. This suggests a strong desire among female users for meaningful connections, with a notable emphasis on work-life balance and personal values. When compared to previous findings on (Fig 2.2.1) male essay content, both genders expressed a keen interest in relationships, but females appeared to prioritize work-life balance more prominently, whereas males leaned towards social activities or career aspirations.

Overall, these insights provide valuable context for understanding the diverse preferences and priorities of OkCupid users, laying a foundation for refining recommendation algorithms to facilitate more compatible and meaningful matches. By considering these multifaceted aspects, OkCupid can enhance the user experience and foster connections that align with users' values and lifestyle choices.

Analysis 3: Analysis of User's Characteristics

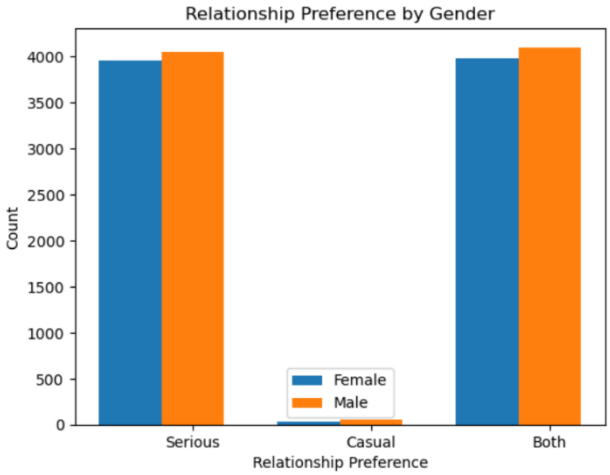

Fig1. Relationship preferences based on gender

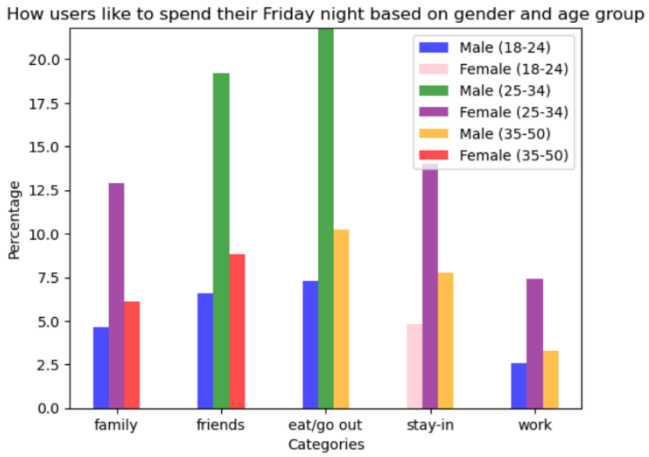

Fig2. Preferences of weekend activity based on gender and age

Attributes used: sex, age, essay5 and essay7

Observations: Observations of the graphs are as follows:

3.1 Fig 1. A higher percentage of users expressed a preference for a romantic relationship compared to those seeking casual encounters. There were also a significant number of users who indicated an interest in both.
Women: Less women users expressed a preference for a romantic relationship on the dating site compared to men. There were also a small number of women seeking casual encounters. A considerable number of women indicated an interest in both romantic and casual relationships.
Men: Contrary to women, a higher percentage of men users preferred a romantic relationship. However, the percentage was lower compared to women. A slightly higher percentage of men sought casual encounters compared to women. Like women, a considerable number of men indicated an interest in both romantic and casual relationships.

3.2. Fig 2. This data suggests that preferences for Friday night activities vary significantly by gender and age. Women are more likely than men to spend Friday nights with family or friends, while men show a higher inclination towards working or staying in. There's a noticeable trend of decreased likelihood of spending Friday nights with family or friends with age, coupled with an increase in dining out or going out. In specific age groups: in the 18-24 age group, both genders prefer spending time with family, while in the 25-34 age group, males tend to eat out or go out, and females still prefer family time. Lastly, in the 35-50 age group, both genders favor dining out or going out on Friday nights.

Analysis 4: Below are the snapshots of the said analysis

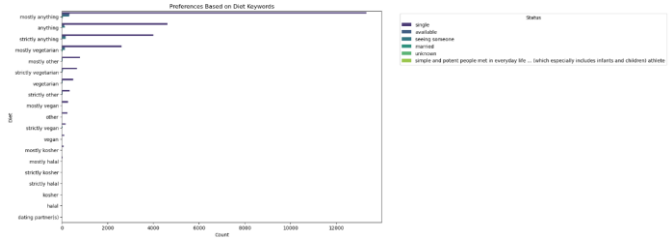

Fig1. Preferences based on Diet Keywords

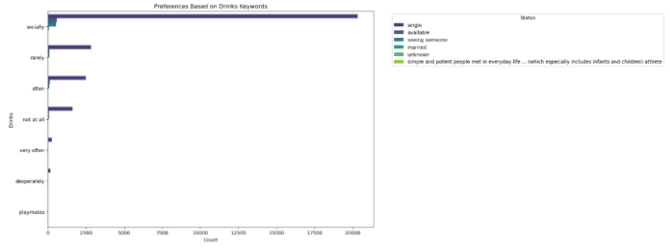

Fig2. Preferences Based on Drinks Keywords

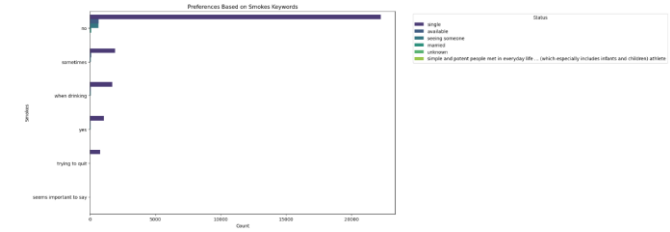

Fig3. Preferences based on Smokes Keywords

Attributes used: Diet, drinks, smokes, essay2 and essay6.

Observations:

1. Dietary Flexibility: Singles tend to gravitate towards less restrictive dietary patterns compared to their married counterparts. This is evident in the higher proportion of singles following "anything" or "mostly anything" diets compared to married individuals.

2. Social Drinking: The data suggests a high prevalence of social drinking among single individuals. This might be linked to factors like socializing in bars or restaurants, or simply enjoying occasional drinks without the influence of a partner's preferences.

3. Health Focus: Despite the prevalence of social drinking, the graph reveals a surprisingly high number of non-smoking singles. This indicates a potential health focus

among this group, with some choosing to avoid smoking while still enjoying occasional social drinks.

What these Traits Tell Us:

These observations paint a picture of singles who prioritize flexibility and social connection. Their dietary choices suggest a willingness to explore different options, while social drinking habits might reflect active social lives. Interestingly, the high number of non-smokers among singles indicates a potential health-conscious side that coexists with their social tendencies.

REFERENCES

[1] R. Elmasri and S. B. Navathe, Algorithms, Fundamentals of Database Systems, Seventh Edition, Pearson, 2017.

[2] Michael Ernst, How to Write a Technical Paper, https://homes.cs.washington.edu/~mernst/advice/write-technical-paper.html .

[3] OKCupid dataset, https://www.kaggle.com/datasets/yashsrivastava51213/okcupid-profiles-dataset.

[4] IEEE paper style, https://www.ieee.org/content/dam/ieee-org/ieee/web/org/conferences/Conference-template-A4.doc