Project Report – OkCupid Matching Algorithm
Group 10 - Adarsh Prajapat, Gunjan Sharma, Jasmine Gohil, Parita Patel

1) Data Preparation

Data Overview: - Profiles (rows): 59,946

- Features (columns): 31 (3 numerical - age, height, income and 28 categorical)

1. The dataset had a lot of missing values and a detailed report of the cleaning steps undertaken for preprocessing in each column is provided in Table1.

2. Focusing on EDA to uncover insights, we started by generating key graphs to understand the distribution and relationships between different variables, providing insights into user demographics, preferences, and behaviors which included:

- Gender wise orientation and smoking and drinking preferences distribution graph – 3 major orientations: straight, gay and bisexual and big gender gap between male and females.
- Understand the location spread of the users (geo map). California has the highest number of users of OkCupid.
- Most spoken languages and the distribution of education level – English is the most spoken followed by Spanish. The distribution of the number of languages spoken against different education levels does not reveal any insights.
- Basic Sentiment Analysis without NLP by creating categories for preferences and importance for religion, zodiac signs and likeness towards pets. Only small portion of the users feel very strongly towards religion and signs.
- Plotting the word cloud to see most frequently used words/topics etc. The most used words are love, like, prefer, say, indicating that a detailed text and sentiment analysis will help in figuring out what people are looking for on the app.

2) Analysis Plan

1. After doing the EDA we plan to further dive into identifying inherent groupings or patterns in the data that might not be immediately apparent, for which we used KMeans Clustering. We One-Hot Encoded the 29 categorical features.
2. To solve the issue of high cardinality, in our case, we thought of doing dimensionality reduction through PCA, making the dataset easier to work with and visualize through KMeans clustering.
3. We then plan to apply sentiment analysis on subsets of data based on the orientation, with the goal of understanding user sentiments, the emotional tone behind user profiles, whether positive, negative or neutral and thereby enhancing the user experience based on it. This will also help in understanding the user's social behavior and their preferences.

3) Data Anlaysis:

- There is a gender disparity in user distribution across age groups on OkCupid. Notably, males aged 19-30 and 31-50 appear to be more prevalent compared to females in these age categories. In the age bracket 51-80 there is equal distibution between the number of females and males. By acknowledging the gender imbalance and potential age-related preferences, the algorithm can be tailored to account for these biases and potentially adjust matching criteria to ensure fairness and inclusivity.

- A glimpse into the geographic distribution of OkCupid users, with a notable concentration appearing in the United States. California seems to have a higher density of users compared to other regions displayed. We can also observe a users in New York, Hawai, United Kingdom and Europe.
- Using these results, we subsetted the main dataset for San Fransico users and utilized KMeans Clustering(6 clusters) to divide users into distinct groups by interests, preferences, and behaviors, setting the stage for tailored recommendations that boost user satisfaction and engagement.
- Refined Matching Algorithms: Insights into user clusters, informed by a range of features, suggest pathways for enhancing matching algorithms. This strategic refinement is poised to elevate compatibility, success rates, and user retention.
- Leveraging User Sentiments: Analyzed user sentiments to identify marketing opportunities and areas for platform improvement, utilizing positive feedback and addressing negative sentiments to foster a more engaging user environment.
- After clustering, we tried our model by matching for a particular profile. We look for a match for the profile 476, in its cluster based on the orientation and sex.
- Targeted Analysis and Clarity in Clustering: By narrowing our focus to sexual orientation and applying PCA, we achieved more defined clustering. This approach, coupled with text analysis on user summaries, underscored the importance of selecting significant features for meaningful segmentation and laid the foundation for advanced, personalized matchmaking strategies.
- Optimal Clustering and Profiling: Determined that six clusters represented the optimal K for our dataset. We then engaged in profiling using TF-IDF and Cosine Similarity for text analysis and Word2Vec for deeper semantic understanding, enhancing our capability to match profiles with high precision.

4) GitHub Link:
https://github.com/Parita2442/OkCupid-NLP-and-Recommendation-Algorithm

5) Limitations:

- **Large Dataset Handling:** Managed a substantial dataset of 60,000 profiles, each with combined essay entries of 5-6 sentences, necessitating a subset due to the prolonged processing time of TF-IDF and spell checking algorithms.

- **Strategic Subsetting:** Initially segmented the dataset based on user location, focusing on profiles primarily from San Francisco, followed by clustering via K-Means to manage and analyze the data efficiently.

- **Targeted Profile Recommendations:** Implemented a model that selects a specific profile (X) as a reference point, isolates the cluster to which X belongs based on orientation, and employs text analysis with cosine similarity to recommend compatible profiles.

6) Conclusion:

The pair of profiles from San Francisco shows a highly promising compatibility, sharing a passion for movies, comedy, and a vibrant social life that includes dining out and attending live concerts. Their mutual love for the city's vibrant culture and enthusiasm for outdoor activities suggests numerous opportunities for shared experiences that could nurture a deep bond. This congruence in their profiles highlights our model's success in matching individuals based on their similarities, underscoring our platform's strength in facilitating fruitful connections.

7) Coding Contribution

Everyone was motivated and contributed pretty much equally for the project execution. The data preprocessing and cleaning took up a lot of time and everyone gave ideas for how to do the same. We had different ideas for clustering as well – based on demographics, lifestyles, orientations and sentiments – and everyone took up one way to cluster to visualize.

Member Name	**Contribution**	**Number of Commits**	**Status**
Adarsh Prajapat	Grouping a few categorical columns, analyzing top 5 education levels by age and sex, lifestyle and behavior clustering, subsetting data based on location.	14	Complete
Gunjan Sharma	Initial cleaning, checking for null values, dealing with outliers, creating correlation graphs, analyzing average income by job category, clustering on sample dataset.	14	Complete
Jasmine Gohil	Grouping a few categorical columns, analyzing number of individuals by age bracket and gender, gender-wise drinking status, gender-wise smoking status, users across the globe, demographic-based clustering.	14	Complete
Parita Patel	Analyzing gender orientation distribution, grouping a few categorical columns, creating sentiment columns, visualization of sentiment from the dataframe, clustering based on orientation.	14	Complete

APPENDIX

Table 1: Data Cleaning and Preprocessing

Column Names	Description	Task Performed
age	Contains age of the users	Range 18-110; adjusted any anomalously high ages to a maximum of 100; removed a row with unconvertible age value.
status	Shows the relationship status of the users.	NA
sex	Contains User reported sex	NA
orientation	categorizes the sexual orientation of the users.	NA
body_type	categorizes the self-described body type of the users.	Performed mapping of the categorical values making it more uniform and easier to analyze. Example grouping athletic and fit to one category as fit.
diet	categorizes dietary preferences of the users.	Transformed the data to reduce the complexity of diet information by consolidating similar diet types into broader, more manageable categories, handling any missing diet information.
drinks	Categories drinking habits/preferences of the users	Filled the missing values with ‘prefer not to say’.
drugs	categorizes the frequency of drug use by the users	Filled the missing values with ‘prefer not to say’.
education	Categorizes highest level of formal education attained by the users	Performed mapping of the categorical values making it more uniform and easier to analyze and filled the missing values with ‘Prefer not to say’.
ethnicity	Categorizes user’s self-identification with a particular ethnic group	Filled the missing values with ‘prefer not to say’.
height	Contains height of individuals within the dating profile in inches.	Filled missing 'height' values with median value as central tendency measure.
income	Contains annual income of the users	Replaced the -1 values with 0 income.
job	categorizes the primary occupation or professional interest of the users	Filled the missing values with ‘rather not say’.
last_online	Contains timestamp of the users when last logged in	Changed to datetime format

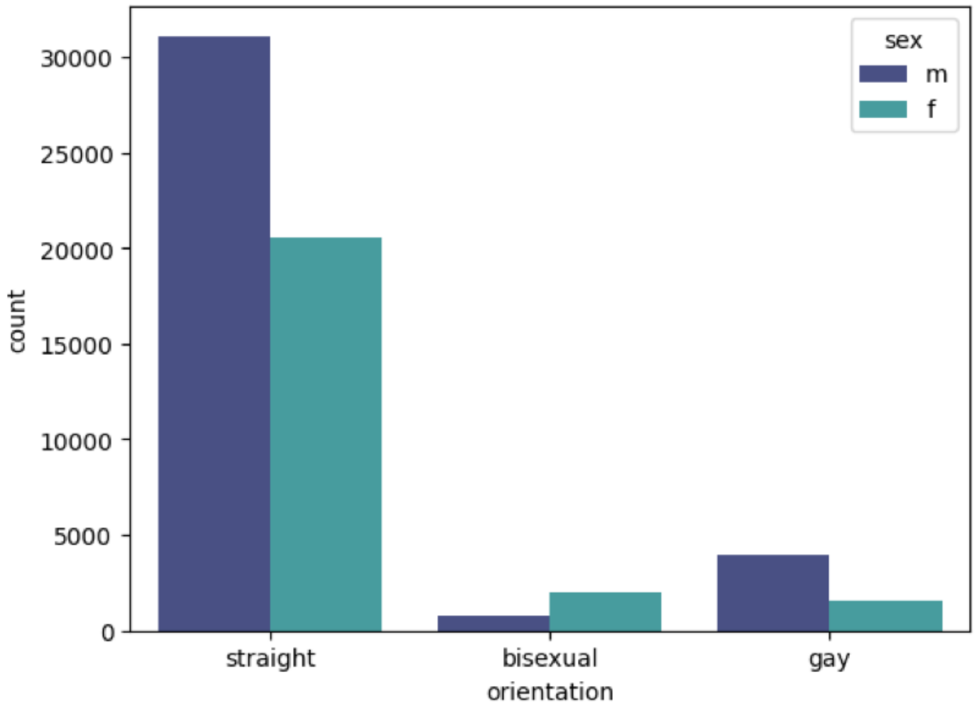
location	Contains the geographical location associated with the users	Mapped to find coordinates to plot a geographical map.
offspring	Categorizes user preferences regarding children	Performed mapping of the categorical values making it more uniform and easier to analyze. Handled the missing values with ‘prefer not to say’.
pets	Categorizes user preferences regarding pet	Filled the missing values with ‘prefer not to say’.
religion	categorizes religious beliefs into several distinct categories	Filled the missing values with ‘prefer not to say’.
sign	Categorizes user zodiac sign and qualitative measure of importance	Filled the missing values with ‘prefer not to say’.
smokes	Contains smoking habits of individuals	Filled the missing values with ‘prefer not to say’.
speaks	Categorizes languages spoken by the users, along with their fluency level	Filled the missing values with ‘prefer not to say’.

This dataset contains 10 columns for essays, which are sections where users can write freely in the app. We merged these essay columns into one called 'combined_essays' and used 'prefer not to say' as a default text for any empty spots.

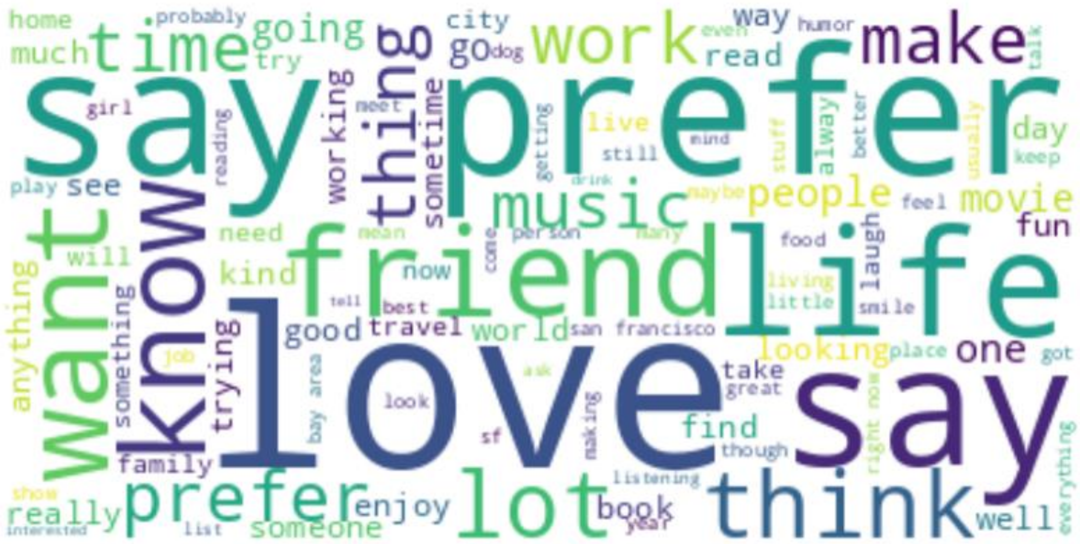
Col. Name	**Renamed Columns As**	**Description**
essay0	self_summary	My self-summary
essay1	current_activities	What I’m doing with my life
essay2	skills_talents	I’m really good at
essay3	first_noticeable	The first thing people usually notice about me
essay4	favorites	Favorite books, movies, show, music, and food
essay5	essentials	The six things I could never do without
essay6	thoughts	I spend a lot of time thinking about
essay7	friday_night	On a typical Friday night, I am
essay8	private_admission	The most private thing I am willing to admit
essay9	message_reasons	You should message me if...

<u>Sample plots</u>

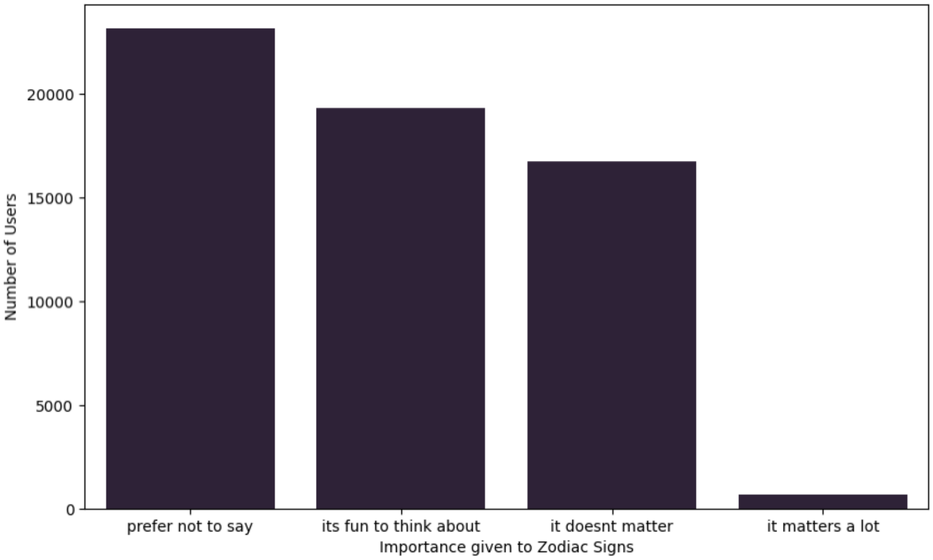
1) Gender-wise orientation distribution



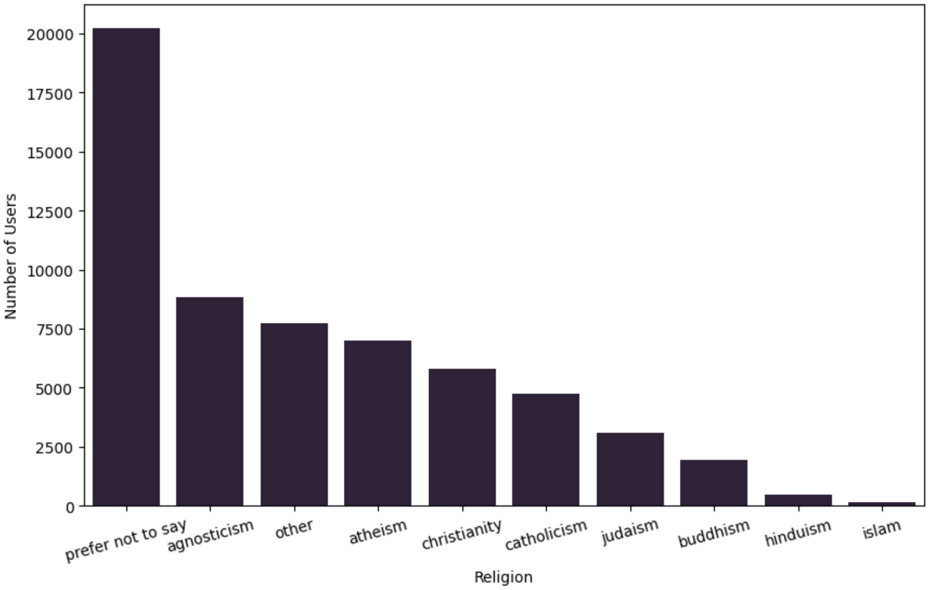
2) Word Cloud



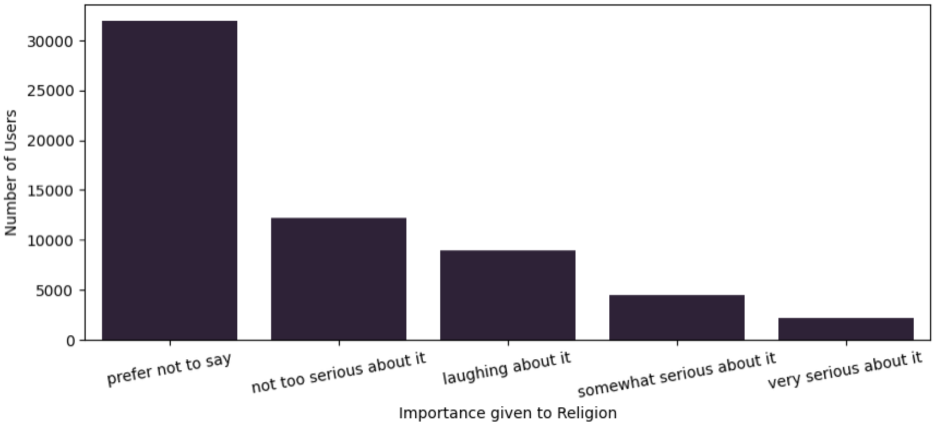
3) Zodiac Sign Importance



4) Religion Distribution



5) Religion Importance


Number of Users
prefer not to say
not too serious about it
laughing about it
somewhat serious about it
very serious about it
Importance given to Religion