Project Proposal - BA 820 A1

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 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 social behavior and their preferences.

3) Preliminary Results

We began by conducting preliminary clustering on a sample to understand potential user groups within the dataset. Encouraged by our initial findings, we sought to explore the full dataset but encountered the challenge of identifying the most relevant features for segmentation. To address this, we took a multifaceted approach, examining demographic factors such as age, location, and gender, as well as lifestyle behaviors like smoking and drinking habits. Additionally, we delved into sentiment analysis to understand user attitudes towards various topics. However, our initial clustering efforts produced unclear results, prompting us to refine our approach.

Undeterred by the initial setbacks, we focused on a targeted analysis by segmenting users based on their sexual orientation and analyzing sentiment within these groups. Employing Principal Component Analysis (PCA), we streamlined the data to focus on the most significant features, resulting in clearer clustering. Through various evaluation methods, we determined the optimal number of clusters, indicating effective segmentation into distinct groups. This laid the groundwork for our next phase: tokenizing user summaries to conduct in-depth text analysis.

4) Next Steps

Given our dataset's categorical nature, we'll explore KModes clustering to improve cluster formation and enhance our matching algorithm. Given the challenges of visualizing higher-dimensional clusters, we plan to use TensorFlow for visualization. Once similarity clusters are formed, our aim is to analyze and offer matching recommendations based on user preferences. To accomplish this, we'll employ Natural Language Processing (NLP) and sentiment analysis on profile essay questions. This approach will help us extract insights and patterns from textual data, enabling better matchmaking decisions.

5) Coding Contribution

Everyone was motivated and contributed pretty much equally for the project execution. The data preprocessing and clean 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.
Adarsh Prajapat - grouping few categorical columns, top 5 education level by age and sex, lifestyle and behavior clustering.
Gunjan Sharma - Initial cleaning, checking for null values, dealing with outliers, correlation graphs, average income by job category, clustering on sample dataset.
Jasmine Gohil - grouping few categorical columns, number of individuals by age bracket and gender, gender wise drinking status, gender wise smoking status, users across the globe, demographic based clustering.
Parita Patel - gender orientation distribution, grouping few categorical columns, creating the sentiment columns, visualization of sentiment from the dataframe, clustering based on orientation.

6) GitHub Link:

https://github.com/S-Gunjan/BA820-A1-OkCupid/blob/main/BA820_ProjectPhaseTwo.ipynb

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’.

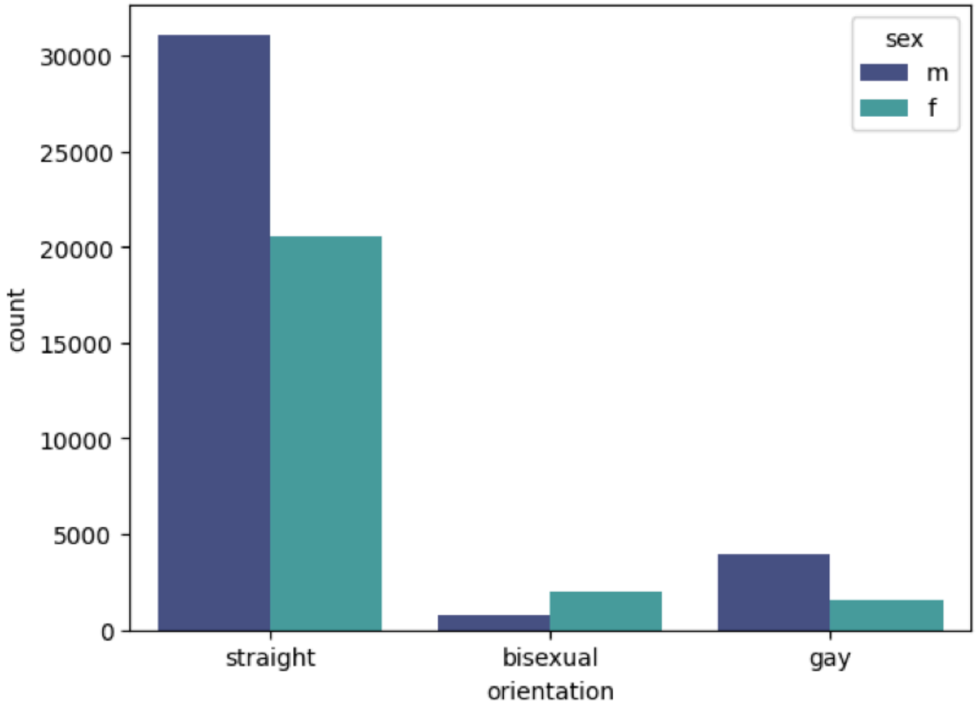
This dataset contains 10 columns for essays, which are sections where users can write freely in the app. We merged these essay columns into one called 'combined_essays' and used 'prefer not to say' as a default text for any empty spots.

Col. Name	Renamed Columns As	Description
essay0	self_summary	My self-summary

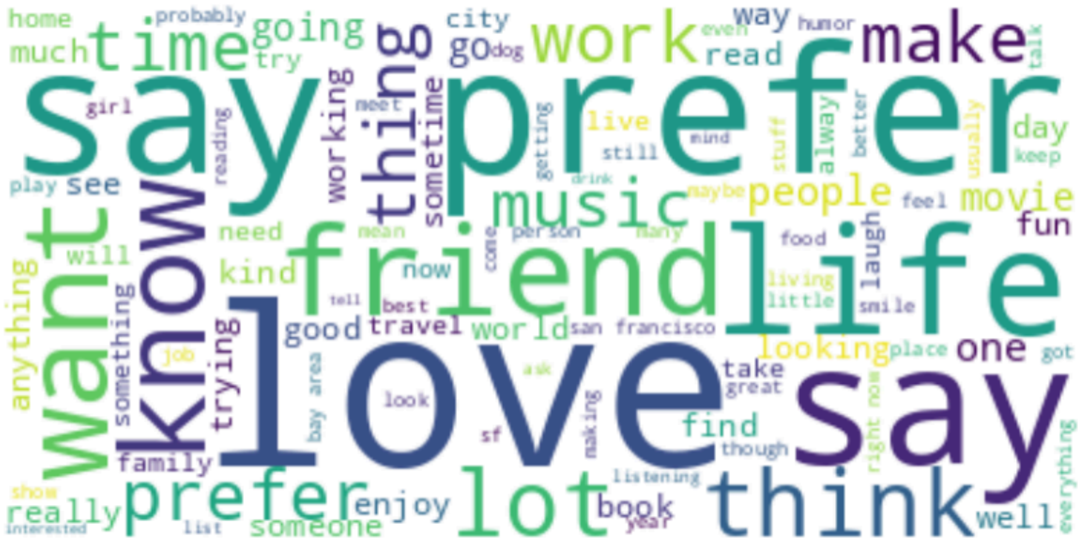
essay1	current_activities	What I’m doing with my life
essay2	skills_talents	I’m really good at
essay3	first_noticeable	The first thing people usually notice about me
essay4	favorites	Favorite books, movies, show, music, and food
essay5	essentials	The six things I could never do without
essay6	thoughts	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...

Sample plots

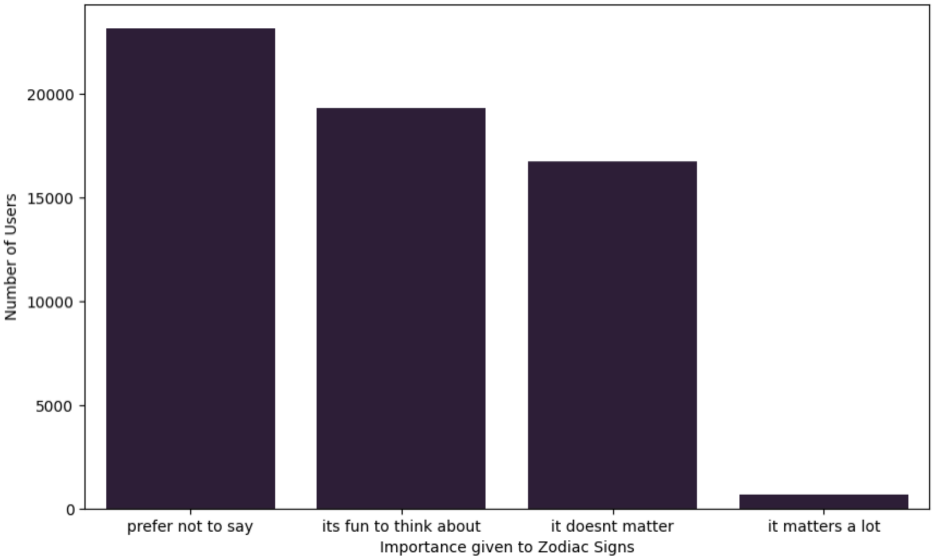
1) Gender-wise orientation distribution



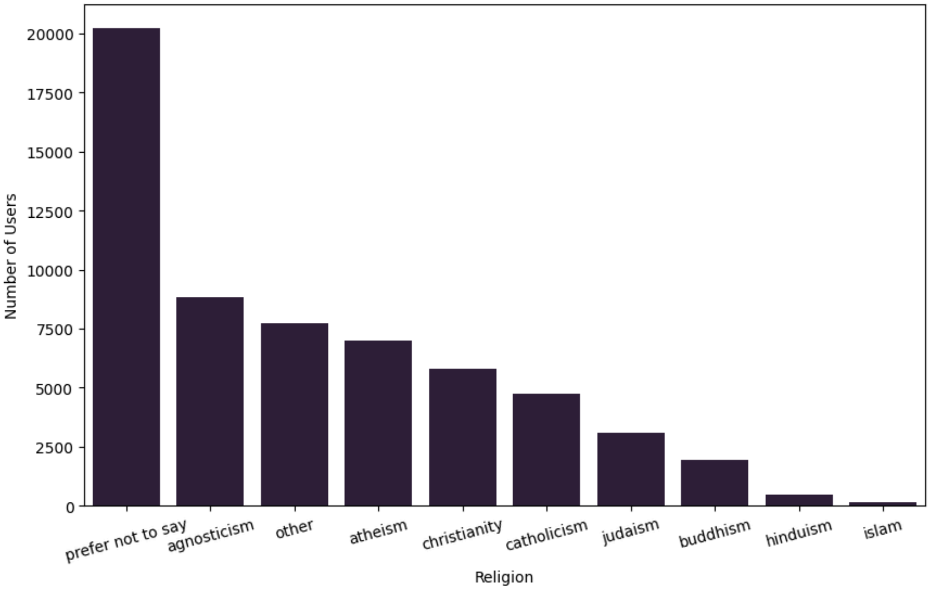
2) Word Cloud



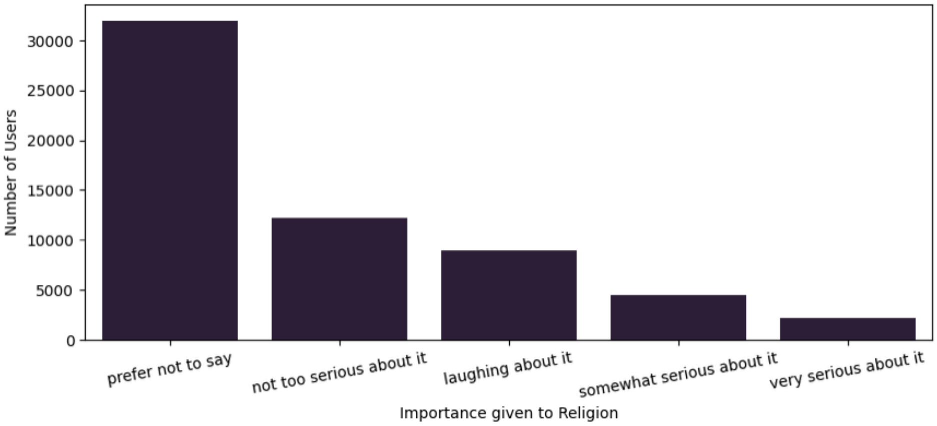
3) Zodiac Sign Importance



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