Creating User Profiles from Feelee Database

Adi Amitai (2738786)

Jose Chacon (2699643)

Elze Satunaite (2726648)

Said Cetinkaya (2723453)

Supervisor: Robert van der Mei

Person of Contact: Levi van Dam

Host Organization: Garage 2020

Vrije Universiteit Amsterdam

Faculty of Science

June, 2024

Contents

1 Introduction

Recently, there has been growing interest in measuring emotional activity to create accessible and personalized mental health support for young adults. Due to the unattractiveness of the traditional mental health care system, many people lack the proper guidance or proper medical treatment. As a result, health issues like anxiety and depression have rapidly increased to an all-time high, causing long-term problems for everyone, especially young adults.

To address the global scope of mental health issues, many organizations and companies are working on providing more and better mental health-related services. One such initiative is Garage2020's mobile application, Feelee, developed with financial and technical support from the network provider, Odido. Feelee allows people to track their emotional states by completing a five-question questionnaire using emojis. Furthermore, it tracks the user's physical activity by extracting data from their mobile phone.

The main objective of Feelee is to provide adaptive advice and interventions for its users based on their historical data and their latest emotional and physical trends. The team aims to employ machine learning techniques such as time series clustering and k-means clustering to translate the patterns in emotional states and other features into specific user groups. These groups will share a distinctive characteristic that differentiates them from others, enabling psychologists to provide more personalized and effective mental health support.

A key challenge for Feelee is correctly measuring the mental health of young users. To effectively address the project's objectives, it is divided into two subproblems. The first subproblem focuses on transforming the responses from the questionnaire into meaningful and usable insights. The second subproblem involves developing a machine-learning model capable of grouping users with similar characteristics. Dasboroug et al. (2008) explored methods of measuring emotions and the associated challenges [1]. Furthermore, the research conducted by Alslaity and Orji (2022) explored machine learning techniques for emotion detection, providing some insights that are relevant for both subproblems [2]. These studies offer a starting point for accurately measuring and understanding the mental health of the users.

While these studies are successful in understanding individual elements of mental health problems, they fail to solve the main problem of how to measure emotional states and provide personalized interventions effectively. This gap is particularly relevant due to the complex nature of mental health problems in young adults. Therefore, the main question addressed in this paper is how to create an algorithm capable of translating both the responses from the 5-question survey and the activity patterns into a comprehensible user profile. The output of the algorithm would be several groups of users that share the same mental health characteristics. Ultimately, the goal of Feelee is to utilize these user profiles to offer users valuable guidance.

The remainder of this paper is divided into five sections. The first section involves an in-depth exploratory data analysis to understand patterns in the user data. The second part employs statistical tests and visualization tools in the categorical data to differentiate between important groups. The third section covers the design and implementation of our machine learning algorithm, which is designed to create user profiles for Feelee Users. In the fourth section, these profiles are examined to search for key characteristics that give the clusters meaning. Lastly, the report concludes with a discussion and suggestions for future development.

1.1 The Feelee App

This report delves deep into the Feelee application user data. This section will clarify the specifics of the app itself and will lay out the basic terms regarding the Feelee application. The initial step after

downloading the app to a mobile device is registering into the system, in doing so, the user provides private information such as their first name, gender, birth year, email address, and password, after which the user can start using the app daily. Once users log into Feelee, they are presented with five screens, each containing a question about their well-being and emojis representing the possible answers. All questions are consecutive in a way that the answer to one question determines the possible emoji answers to the next question. The first question is "How are you feeling?" when the possible answers are limited to "Good", "Okay", and "Not Okay". The following question is a further description of how one is feeling, where some of the possible answers are: "happy", "joyful", "tired", and "anxious", next, the three remaining questions are: "How come you feel X?", "What are you doing?", and "Who are you with?". At the end of this process, the user gets an overview of their past answers, as well as a count of the steps they walked in the past hours (this requires approval of the retrieval of steps from the user's device).

2 Exploratory Data Analysis

2.1 Dataset

This dataset, sourced from Feelee, contains 59299 entries representing individual responses of users of the Feelee app. Due to privacy reasons, The user's name and email address were omitted from the dataset. Each entry has the following attributes: user ID, OS (whether the user's device is Android or apple operated), gender, birth year, date, time, a column combining the date and the time (when the response was registered within the app), answers to questions one to five including, and twenty-four columns representing steps taken in each hour of the day. Since every entry represents one login to the app, user IDs can be repeated in the dataset as many times as a user has used the app.

2.2 Data Cleaning

Cleaning the data is an essential part of an analysis, as it ensures the accuracy and reliability of the results obtained by the research done using the data. The cleaning process of the Feelee data involved dealing with inconsistencies in the data, as well as the removal and completion of missing values. First, the data were scanned for discrepancies, to do so, all unique user IDs were examined, and for each user, the algorithm checked for multiple unique values for "gender" and "birth year". Next, for users that showed discrepancies, the algorithm created new user IDs, such that every user ID was linked to a specific gender and birth year. This step resulted in an increase in the total number of unique IDs from 2525 to 3043. For users who had null values for gender or age, the information was retrieved from a different entry that had the same user ID.

Secondly, in the cleaning process, the removal of null values and unnecessary columns was performed. The column "os" was determined unnecessary for the rest of the analysis and decision-making process and was removed. Rows in the data that had null values for question one (how are you feeling?) and question two (a further description of how you are feeling) were removed as well (the application allows the user to skip questions three, four, and five and thus null values for those were kept in the dataset). All rows which consisted of the answer "weet ik niet" (English: I do not know) as an answer to question four were omitted from the data set as well since this is not a valid answer to this question according to the Feelee app. In addition, approximately 35 rows where daily steps were seen in individual columns that represented different hours of the day were omitted from the dataset.

Furthermore, the 24 columns that contained hourly steps were merged and a "daily steps" column was created and replaced the separate "hourly steps" columns. Additionally, an additional

column representing the user's age at the time of the response in the app was created.

2.3 Data Exploration

The objective of this passage of the report is to uncover patterns and characteristics of the cleaned dataset, to understand its structure, and provide preliminary insights to inspire further analysis. The data set comprises numerical columns such as user ID, birth year, age, and the time and date the user responded in the Feelee app, as well as categorical values which are the user's gender and responses to the five questions and their daily steps when provided.

The dataset consists of 59273 rows which correspond to entries made by 3043 unique Feelee app users, and 13 columns. It spans a time frame of almost two years, the earliest entry is dated 12/04/2022 and the latest to 28/02/2024. With 4049 reports, 2007 users identify as female, and they account for the majority of the dataset. males come in second place with 16937 entries and 954 users, followed by users who identify as they/them with 1110 entries and 41 users and lastly, 41 users who preferred not to state their gender and reported 322 times. Furthermore, access to daily steps was provided in 21140 of the entries. The majority of the users are between the ages 9 and 30, while the average user age is 26.5, the median age is 26, the youngest user in the dataset is 9 years old and the oldest is 81 years old. Table 5 in the Appendix displays the different age groups and number of entries per age group.

The following bar chart (figure 1) illustrates the average number of daily steps taken by users across various age groups, segmented by four gender categories. The x-axis represents the age group of the users, while the y-axis indicates the average daily steps. It provides insights into the physical activity patterns across the different age groups and genders.

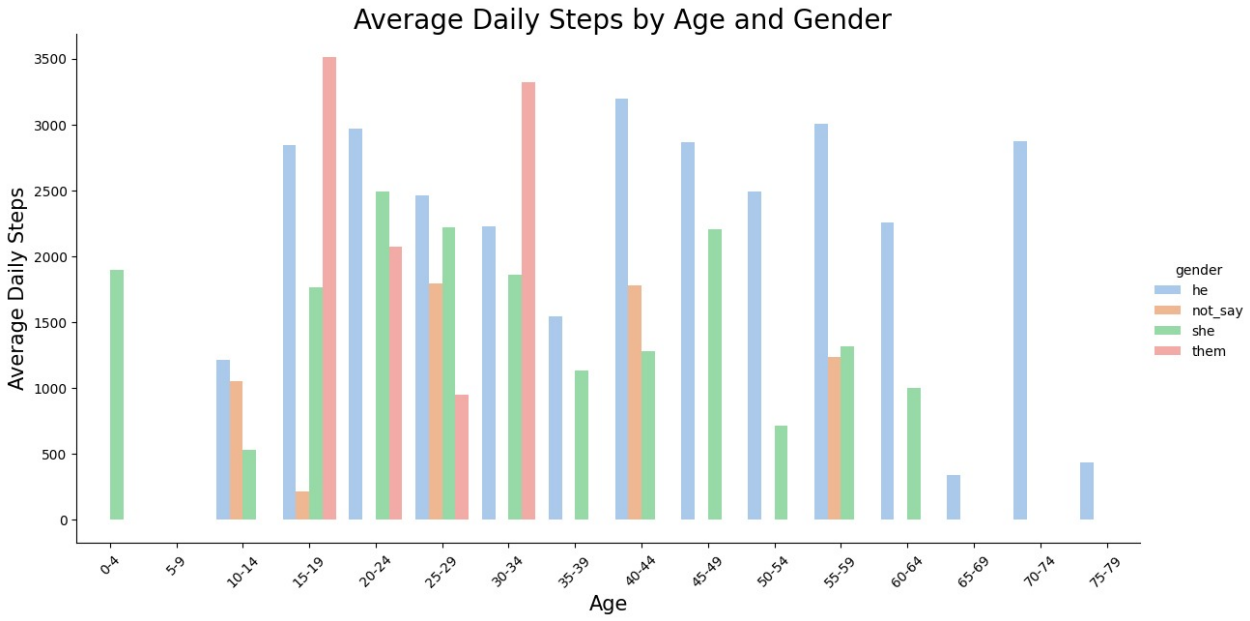

Figure 1: Total Steps Regarding Age and Gender

The graph above shows a clear concentration of a higher count of daily steps in younger female age groups, particularly between the ages of 20 to 34. Regarding males, it is seen that they tend to walk more on average in older age groups such as 40 to 33, 55 to 59, and 70 to 74. This can suggest younger female and older male users tend to be more active. In addition, there is a clear high number of average steps for users who identify as "them" between the ages 15 to 19 and 30 to

34. While Feelee's focus is on young adults, there can be seen data about all age groups, even an outlier in the age group of 0 to 4 with female users who walked almost 2000 daily steps on average. Moreover, throughout all age groups, males tend to have a higher activity level and more steps per day than other genders.

Moving on to the exploration of answers of Feelee users to the five questions regarding their well-being, it can be seen that out of the three possible answers to the first question, how are you feeling, the two leading ones are "Good", with 24460 responses, and "Okay" with 23696. The four most popular answers to questions 2, 3, 4, and 5 are shown in table 1.

How are you feeling? (Q2)		**Why are you feeling X? (Q3)**	
Tired	11880	Something else	10714
Calm	9726	Do not know	5202
Satisfied	6741	Conversation	4802
Happy	6129	Ok night	4277
What are you doing? (Q4)		**Who are you with? (Q5)**	
Relaxing	14569	Alone	26208
Something else	8079	Close Family	6199
Work	7947	Relatives	5500
Nothing	6373	Family	4847

Table 1: Top Answers to Questions 2-5

Considering the two common answers to question 1 are "Good" and "Okay", it is reasonable that the four answers that are repeatedly provided as a further description of how a user is feeling reflect positive and medium feelings. Following, users are asked to determine the reason for feeling X, and it is seen that most of them do not know or can not find the most appropriate answer within the options presented to them. In addition, It is observed that when filling in the information in the Feelee app, most users are either at work or are not particularly occupied and are either by themselves or with a close or distant relative. Moreover, it was found that every possible response for every question was answered at least a hundred times, thus proving that all possibilities are relevant to at least a certain number of users.

To further explore and later use the data to cluster and profile users of the Feelee app, the consistency of the users was examined, meaning the number of times each user contributed to the application overall. On average, one uses the Feelee app 19 times in the span of two years in the dataset. The minimum number of entries per person was found to be one entry, which indicates only one use, and the maximum was a user who filled in the questions in the app 1618 times in total. In addition, the 50 percentile is 4 times and the 75 percentile is 15 times. Additionally, it was found that the majority of the consistent users are youth born between the years 2000 and 2010 (ages 14-24), this can be seen in the following figure 2.

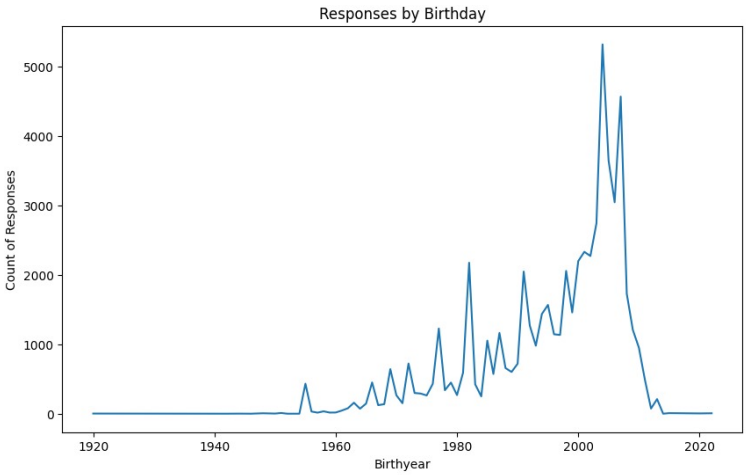

Figure 2: Consistent Users by Birth Year

3 Statistical Data Analysis

In this section, a deeper exploration of the findings from the preliminary exploratory data analysis is conducted through statistical data analysis. Insights into the dataset are gained by employing statistical tests, with a focus on identifying statistical differences between certain groups based on specific features. The findings of this analysis are then used to improve the machine-learning model.

First, the statistical methods used are described, providing an obvious reason for their selection. Next, the tests are employed, and the results are presented along with the insights gained from this analysis.

The dataset was used in its original form and structure, preserving the categorical variables without transforming them into numerical features. This decision was motivated by the need to maintain the integrity and meaning of the data. Consequently, statistical tests well-suited for categorical data were selected. Specifically, the Chi-Square test was chosen due to its high performance with categorical data [3].

3.1 Statistical Tests

3.1.1 Chi-Square Test

Categorical variables are represented as counts or frequencies and are conveniently arranged in contingency tables, often designated as $r \times c$ tables (where r is the number of rows and c is the number of columns).

- **Purpose:** Tests the independence or association between two categorical variables.
- **Suitable Data:** Categorical data arranged in contingency tables.
- **Use Case:** Testing if there is an association between a specific group (age groups, or gender groups), and the answers to all questions.

3.1.2 One-Way Analysis of Variance (ANOVA)

- **Purpose:** Determine if there are statistically significant differences in the means of a continuous dependent variable across multiple groups defined by a categorical independent variable.

- **Dependent Variable (Numerical):** The number of daily steps (the level of physical activity).
- **Independent Variable (Categorical):** The different groups being compared (age groups, gender groups, overall feeling group/answers to question 1).

3.2 Conclusions on SDA

3.2.1 Chi-Square Test

Gender and answers to all questions: Since the p-value (8.60e-11) is much smaller than the conventional significance level of 0.05, the null hypothesis of independence is rejected. This means a significant association exists between gender and the responses to "question_1". The expected frequencies show what would be expected if there were no association between gender and responses. The observed frequencies would deviate from these expected frequencies. In summary, a significant difference in responses to "question_1" across genders is concluded.

The same procedure was performed for questions 2 through 5. The results of all tests coincide with the findings of Q1. There is a significant difference in responses to all questions across the four registered genders (he/she/them/not specified).

Age groups and answers to all questions:
The test results suggest a significant association between the variables 'youth' and 'question_1', as indicated by the very small p-value, which is 1.55e-26. This means that the distribution of responses for 'question_1' varies significantly across the adult and young adult groups.

Similarly to the gender groups, there is a significant difference in the responses to all questions across the age groups. These results suggest that each group has its own way of responding. These insights will be used later in the creation of the machine learning model.

3.2.2 ANOVA Test

In the research paper "Statistical Analysis of Emotional Response through Physiological Signals," the authors explored the relationship between emotional stimuli and physiological indicators [4]. They utilized the ANOVA method to analyze the effects of different groups of emotional pictures on various physiological signals. This approach was translated to the current project, where the effects of different groups on different physical activity levels measured by daily steps were analyzed.

The ANOVA test was performed twice. First, to measure the effect of gender in physical activity, measured by daily steps. Secondly, the relationship between age groups and physical activity was measured. In the following table 2, the results of the ANOVA test are summarized.

Factor	F-statistic	P-value	Significant Difference
Gender	103.45	1.74e-66	Yes
Age	1.97	0.16	No

Table 2: ANOVA Test Results

This table outlines that there is a significant difference in daily steps based on gender under the 5% significance level, but there is no significant difference between youth and adult groups in terms of daily steps. In other words, young adults and adults perform indistinctively from each other in terms of daily steps. The gender test results conclude that there is a significant difference in daily steps and gender. To summarize, the age of the user is only determinant when measuring

the question-1 answers and not when measuring the steps. The following graphs will illustrate the results of these tests. Furthermore, the descriptive statistics of the main groups against daily steps are included in the Appendix.

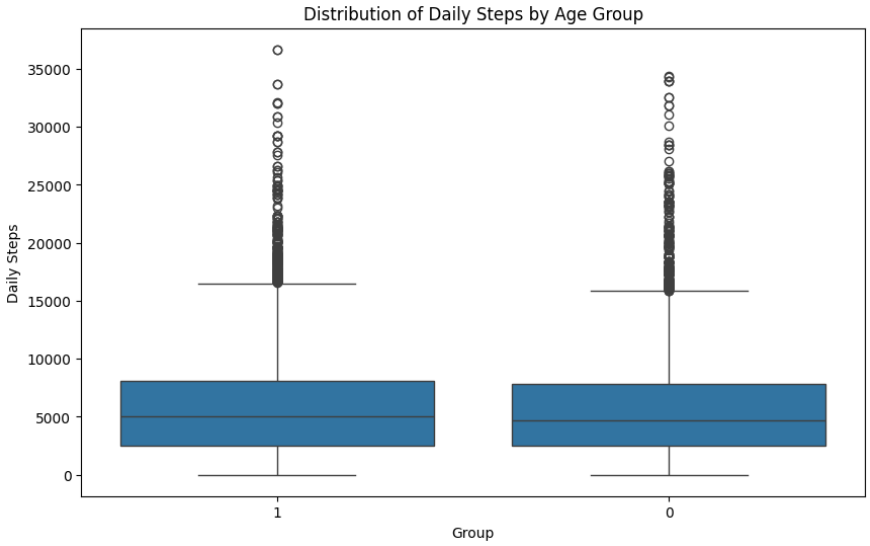

Figure 3: Boxplot of Daily Steps for Age Groups

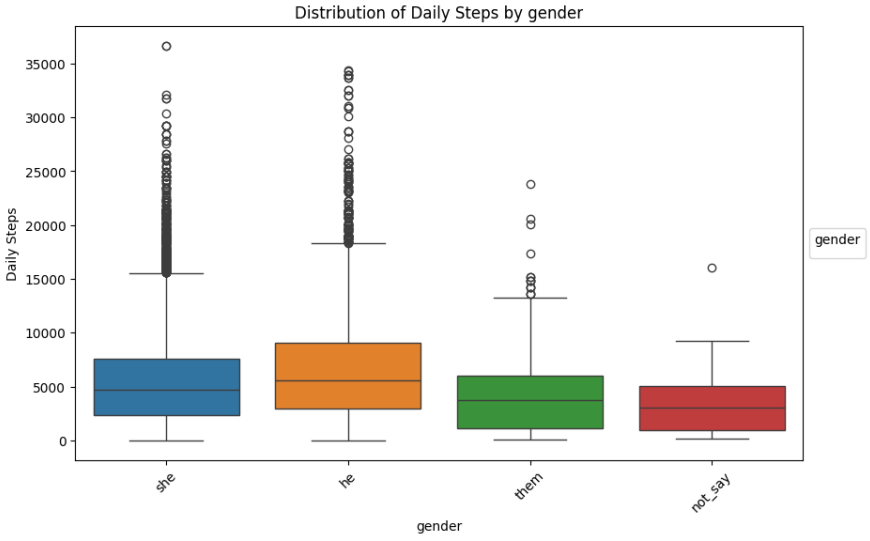

Figure 4: Boxplot of Daily Steps for Gender

4 Methodology

This section outlines the approach used to analyze and process data from the Feelee app to develop user profiles based on their activity and emotional patterns. The primary goal is to prepare the data, filter it appropriately, and transform it into a format suitable for statistical analysis and machine learning. The machine learning models are discussed here as well. The methodology involves several stages: data wrangling, pivoting, quantifying mean and variance, and sentiment analysis.

4.1 Data Wrangling

First, the data is filtered to ensure it is clean, consistent, and ready for analysis. The data is grouped by user ID to handle each user's data separately, and for each group, it is sorted by date in descending order to focus on recent entries. The filtering logic considers various scenarios to manage gaps between days and ensures a consistent dataset for analysis, focusing either on daily patterns or individual entries. This step ensures that the dataset is tailored to the specific needs of the analysis, whether it is concentrating on consecutive days or individual entries.

To handle cases where data is grouped by user ID, user IDs are expanded to ensure each user ID is unique for every set of instances, preventing issues in subsequent grouping operations. This modification is crucial for maintaining the integrity of the data during analysis.

Next, the dataset is transformed into a pivot table format, making it suitable for statistical analysis. The data is reorganized so that each user's responses over seven days are aligned into a structured format. The index columns include user ID, gender, and birthday, while the values columns include responses to questions and other relevant metrics. This transformation spreads each day's responses into separate columns, allowing for easy analysis. Columns with repeated measures, such as response age and daily steps, are processed to retain only the relevant measures.

Next, qualitative responses are transformed into quantitative sentiment scores. Predefined dictionaries map qualitative responses to numerical sentiment values, enabling the analysis of emotional patterns. Following, the responses are processed to generate sentiment scores for each user, which are then used in further statistical and machine-learning analyses. Table 10 below visualizes how responses are processed (grouped and ranked) in question 4 (what are you doing?). For question 1 (how are you feeling?) the ranking was: Good - 3, Okay - 2, Not Okay - 1 while question 5 (who are you with?) was separated into two: alone or not alone. The remaining questions (2 - further description of how you are feeling and 3 - why) are presented in the Appendix.

Ranking	**Activity**	**Category**
1	Work	Outdoor
1	Sports	
1	Being outside	
0.75	Something else	Leisure
0.75	Homework	
0.75	Gaming	
0.5	*nan*	Essential
0.5	Eating/drinking	
0.5	Visiting	
0.25	Relaxing	Passive
0.25	Nothing	
0.25	Watching TV	

Table 3: Question 4 Response Ranking

Furthermore, the mean and variance for all five question responses across the specified number of days are calculated. This statistical quantification helps understand the distribution and variability of user responses. Moreover, the data is grouped by user ID and date components and aggregated to calculate mean values for the responses. The filtered and pivoted data is then used to compute the mean and standard deviation for each question response, providing insights into the central tendency and dispersion of the data.

4.2 Clustering

The first step is identifying the optimal number of clusters using the elbow method. This method involves running the k-means clustering algorithm on the dataset with a range of different cluster counts and recording the inertia (sum of squared distances from each point to its assigned cluster center) for each count. Plotting the inertia values against the number of clusters reveals an 'elbow point' where the rate of decrease sharply slows down, indicating the optimal number of clusters.

Once the optimal number of clusters is determined, the k-means clustering algorithm is applied to the scaled dataset to partition it into distinct clusters. Each data point is assigned to a cluster, and the resulting cluster labels are added to the dataset for further analysis.

To evaluate the performance of the clustering, a classification report is generated using a machine learning model (XGBoost). The dataset is split into training and testing sets, and a classifier is trained on the training data. The classifier's performance is then evaluated on the test data using metrics such as accuracy, precision, recall, and F1 score. Feature importance is also assessed to understand which variables contribute most to the classification.

Dimensional reduction techniques, such as t-SNE (t-Distributed Stochastic Neighbor Embedding), are applied to the dataset to reduce the number of dimensions while preserving the structure of the data.t-SNE is a powerful tool for visualizing high-dimensional data. As described in [5],t-SNE is an iterative non-linear dimensional reduction tool that helps in separating non-linear data, by creating a probability distribution for similar objects within the data and measuring the distance between data points to represent them in a low-dimensional space. This helps in visualizing the high-dimensional data in a three-dimensional space. The reduced data is then plotted, allowing for a clear visualization of the clusters.

To further analyze the emotional and activity patterns, word clouds are generated for each cluster. This involves transforming qualitative responses into textual data and creating visual representations of the most common words or phrases within each cluster. This provides insights into the dominant themes and characteristics of each cluster.

4.3 Markov Chain

The employment of a Markov chain model involves constructing a transition matrix, visualizing the transitions, and analyzing the properties of the chain to understand the long-term behavior of the system.

To begin, the Markov chain model is constructed by first initializing a transition table that tracks the number of transitions from one state to another. The data is then prepared by creating unique user identifiers and sorting the dataset by date for each user. This step ensures that the transitions are chronologically accurate. The transition table is populated by iterating over each user's data and calculating the time shift between entries. For each entry, if the time shift matches a predefined interval (e.g., seven days), the transition from the current state to the next state is recorded. This process creates a detailed account of how users move between states over time.

Next, the transition matrix is derived from the transition table. The matrix represents the probabilities of transitioning from one state to another, calculated by normalizing the counts of transitions. To improve readability and ensure non-zero probabilities, the values in the matrix are adjusted and rounded. Visualization of the transition matrix is achieved using a directed graph. Nodes represent the states, and directed edges represent the transitions with weights corresponding to the transition probabilities. Edge labels indicate the transition probabilities for the understanding of the dynamics between states.

To further analyze the Markov chain, the transition matrix is converted into a directed graph, allowing for the identification of strongly connected components. These components represent

potential recurrent states, which are states that the system tends to revisit over time. Additionally, the stationary distribution of the Markov chain is computed. This involves finding the eigenvector corresponding to the eigenvalue of one, which represents the long-term stable distribution of states. The stationary distribution shows the likelihood of the system being in each state over an extended period.

5 Results

5.1 Clustering Data

This section discusses the findings from the wrangled data that was used for clustering. The dataset consists of information from 598 individual users.

The birth year distribution of the users shows a range from 1955 to 2013, with the mean birth year being approximately 1996. The standard deviation is 11.5 years, indicating a wide age range among users. The quartiles show that 50% of the users were born between 1991 and 2005, with the median birth year being 2001. This distribution suggests that the majority of users are young adults.

The gender distribution among the responses is predominantly female, with 69.6% identifying as "she". Males constitute 27.9% of the users, while those identifying as "them" and "not say" make up 1.7% and 0.8%, respectively. This skew towards female users is significant and might influence the overall clustering results, as gender can affect emotional and activity patterns.

The distribution of daily steps among the responses shows variability across the clusters. The box plot below (figure 5) illustrates the median, quartiles, and outliers for daily steps for each cluster label.

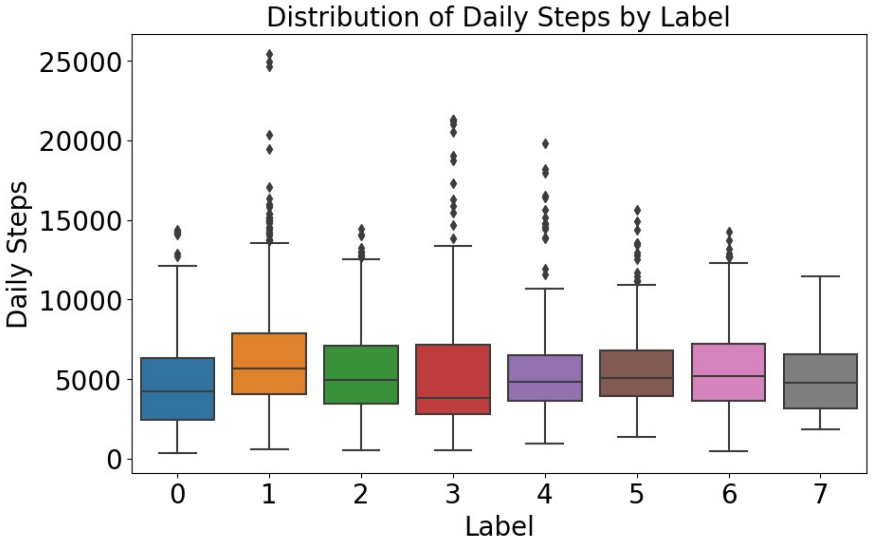

Figure 5: Distribution of Daily Steps

The plot reveals that: clusters have different median values for daily steps, there are noticeable outliers with high step counts in several clusters and some clusters (e.g., Cluster 3 and Cluster 4) have higher variability in daily steps compared to others.

5.2 Elbow Method for Optimal Clusters

Using the elbow method to determine the optimal number of clusters, an observation was made that the inertia significantly decreases to around 8 clusters, after which the rate of decrease slows down. This suggests that an optimal number of clusters for this dataset is around 8, balancing between granularity and generalizability.

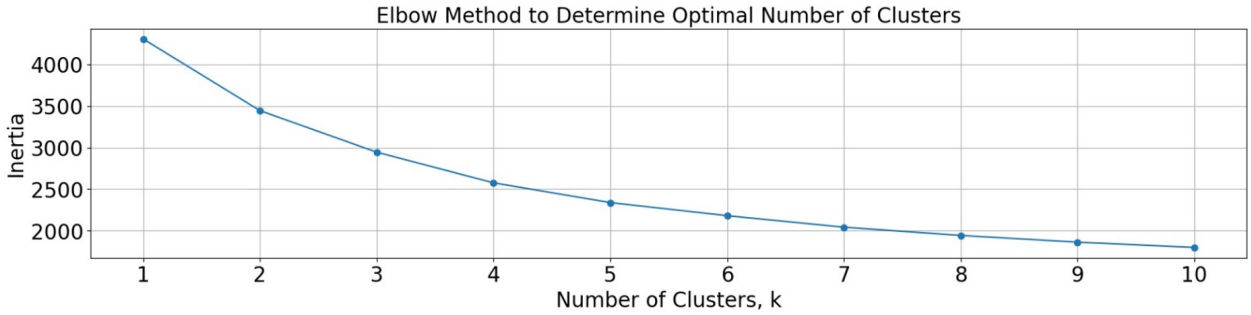

Figure 6: Elbow Method

Figure 6 shows the elbow plot, where the x-axis represents the number of clusters, and the y-axis represents the inertia. The point where the curve bends ("elbow point") indicates the optimal number of clusters.

5.3 Cluster Analysis

The cluster analysis revealed distinct groups characterized by varying emotional states and levels of engagement. The clusters are as follows:

Cluster 0: Low Feelings - Disengaged - Negative - Passive The 1156 Users in this cluster predominantly report negative emotions and a sense of disengagement. The word cloud for this cluster highlights terms such as "Not_Okay," "PassiveAc," and "DisengagedEm," suggesting that these users frequently experience low energy and negative emotions. The average age and daily steps of users in this cluster are respectively 23.5 and 4713, while the medians are respectively 19 and, 3471. The Top four reasons for feeling a certain way (question 3) in cluster 0 were "I don't know", "something else", "okay night", and "stress". This cluster's characterization can guide the development of targeted interventions aimed at increasing engagement and improving mood.

Cluster 1: Low Feelings This cluster consists of 2274 users, making it the biggest cluster. It is similar to Cluster 0 but lacks additional distinguishing characteristics. The primary feature of this group is the low emotional state, with terms like "Not_Okay" and "PassiveAc" appearing prominently in the word cloud. The average age and daily steps of users in this cluster are respectively 24.4 and 6226, while the medians are respectively 20 and 5167. The feeling that repeated the most in this cluster was "Tired", with a significant number of 8055 repetitions, as the runner-up was "Calm" with 5647 times. Another answer that repeated itself was "Satisfied", which symbolizes a neutral feeling and disengagement. It is seen as though this cluster is answering that they are not okay in question 1 but then move on to give more neutral answers for the other questions. In addition, users who were alone and not alone in this cluster were evenly distributed. Furthermore, stress is the fifth reason for users in this cluster for the way they are feeling. after "something else", "conversation", "okay night" and "I don't know". Interventions for this group may focus on general mood enhancement strategies.

Cluster 2: Moderate Feelings - Disengaged - Neutral - Alone Users in this cluster report moderate feelings but are often disengaged and feel alone. There are 1333 users on this cluster, 847 of them

are female, 457 male, 25 them, and 4 prefer not to say their gender. The average age and daily steps of users in this cluster are respectively 33 and 5728, while the medians are respectively 36 and 5169. The word cloud includes terms like "NeutralSt" and "Alone," indicating a need for social engagement and emotional support. The emotions that repeated the most in this cluster were "tired", "calm", "satisfied, and "indifferent", while the reasons that repeated the most were "okay night", "something else", "agenda" and "rest". Programs that encourage social interaction and community building could be beneficial for this group.

Cluster 3: Moderate Feelings - Calm - Neutral - Passive - Alone This cluster is the smallest cluster with only 490 users, the average age and daily steps of users are respectively 36 and 6163, while the medians are respectively 35 and 3415. This cluster has an almost even distribution of males (236) and females (247) and it includes users with moderate feelings who are calm, tired, satisfied, and neutral but also passive, worried, and alone. The most significant answer to question 3 in this cluster about why a user is feeling X is "something else" with 2324 times. The word cloud highlights terms such as "CalmEm" and "NeutralSt." This group's characteristics suggest a need for activities that can help them become more active and socially connected.

Cluster 4: Moderate Feelings - Neutral - Essential - With Someone The 567 Users in this cluster have moderate feelings and are often with someone. The average age and daily steps of users in this cluster are respectively 25.3 and 6329, while the medians are respectively 19 and 4896. Additionally, this cluster has 441 females, 126 males and no users who identify as them or prefer not to state their gender. The word cloud shows terms like "EssentialAc" and "WithSomeone," suggesting that these users value essential activities and companionship. The common answers for question 2 (further description of a feeling) are "tired", "happy", "calm", and "satisfied", while for question 3 the answer was not given for 2406 of the times, making it the most common answer, as "something else", "holiday" and "good night" come in the second, third and fourth places. Interventions could focus on enhancing these aspects and promoting activities that they find meaningful.

Cluster 5: Moderate Feelings - With Someone This cluster is characterized by moderate feelings and being with someone. There are 1477 users The word cloud highlights "WithSomeone" and "PositiveSt," indicating that social support plays a significant role in their emotional state. Their average and median daily steps are respectively 5653 and 4730, while their average age is 25.6 and the median age is 20. This cluster consists of 1033 females, 435 males, 6 them, and 3 who preferred not to mention a specific gender Moreover, the answers show that the most common answers to questions 2 and 3 are "tired", "calm", "satisfied", "happy", "indifferent", "okay night", "conversation", "something else", "I don't know", and "fun task".Strengthening social bonds and support networks could be a key focus for interventions.

Cluster 6: Good Feelings - Neutral This cluster is relatively large as there are 2342 users who report good feelings and tend to be neutral. Their average and median daily steps are respectively 5799 and 5219, while their average age is 25 and median age is 22. The word cloud includes terms like "Good" and "NeutralSt" suggesting a stable emotional state with most people feeling calm and happy for reasons of being on holiday, having a fun task, having a conversation or having a good night. Maintaining this stability and providing resources to handle potential stressors can help this group.

Cluster 7: Good Feelings - Energetic - Positive - With Someone This group consists of 783 users (610 females, 173 males) who are characterized by good feelings, high energy, and positive emotions, often with someone. Their average and median daily steps are respectively 5473 and 4313, while their average age is 21 and median age is 17. The word cloud features terms like "EnergeticEm" and "PositiveSt," reflecting their active and positive outlook. A clear statement of this cluster is that they do not at all feel nervous, anxious, worried, angry, dissatisfied, and so on, this is seen by 0 times those answers are given within this cluster. the feeling "Energetic" is in fourth place regarding

question 2 with 902 times it was answered, the first three are "happy", "calm", and "satisfied". The reasoning for users in this cluster for their feeling are having a good night, doing a fun task, or being on holiday. Encouraging continued engagement in positive activities and social interactions can support their well-being.

5.4 t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction, particularly well-suited for the visualization of high-dimensional datasets. Introduced by Laurens van der Maaten and Geoffrey Hinton in 2008, t-SNE converts high-dimensional data into a two or three-dimensional map, making it easier to visualize patterns and structures within the data [6]. This method builds upon the earlier Stochastic Neighbor Embedding (SNE) technique, offering improvements that mitigate common issues such as the crowding problem and optimization difficulties.

The clusters discovered for Feelee data are visualized using the t-SNE plot, providing a clear representation of the separation between different emotional states and engagement levels.

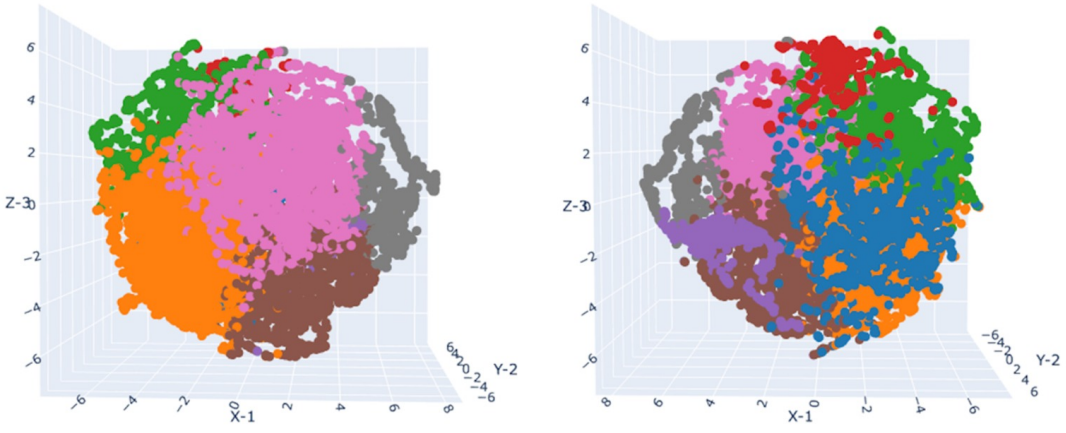

Figure 7: t-SNE

The t-SNE plot (Figure 7) shows well-defined clusters, with different colors representing the different clusters. Each point in the plot corresponds to a user, and their position relative to others indicates their similarity in terms of emotions and engagement metrics.

5.5 Classification Results

The classification model exhibited high performance with the following metrics:

Metric	Value
Accuracy	0.9511
Precision	0.9515
Recall	0.9511
F1 Score	0.9509

Table 4: Classification Metrics

The high accuracy, precision, recall, and F1 - score indicate that the model is effective in

correctly identifying and classifying the emotional states and engagement levels of users. This high performance is crucial for the effectiveness of the Feelee app in providing accurate and reliable feedback to its users.

The following figure 8 shows the relative importance of each feature used in the model, indicating that the most significant features contributing to the model's performance were 'Q5_mean', 'Q1_mean', and 'Q5_std'. These features highlight the importance of specific emotional metrics in predicting user states.

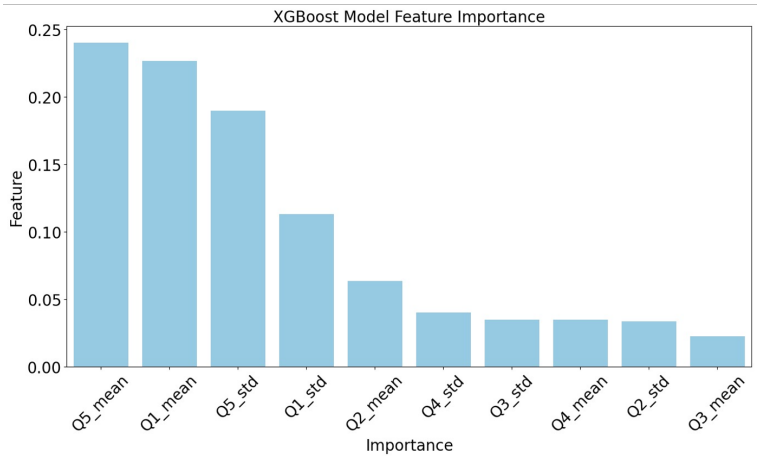

Figure 8: Feature Importance

5.6 Word Clouds for Cluster Characteristics

Word clouds were generated for each cluster to visualize the predominant emotional states and activities within each group. These visualizations confirm the descriptive labels assigned to each cluster, with terms like "Good", "Okay", "Alone", "Passive", and "Energetic" prominently appearing, reflecting the emotional and activity states of users in each cluster. The figure 9 below shows an example of a clear word cloud (the remaining word clouds can be found in the Appendix 12).

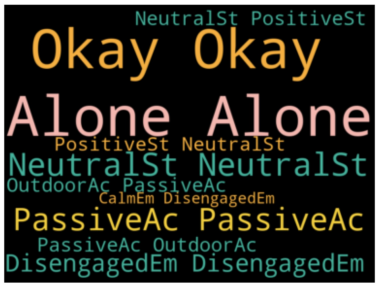

Figure 9: Word Cloud

The word clouds provide a quick visual summary of the most common words and phrases associated with each cluster, highlighting the key characteristics and differences between them.

5.7 Markov Chain Transition Analysis

The Markov chain transition diagram illustrates the probabilities of transitioning between different emotional states over time. The transition matrix shows high probabilities for certain transitions,

such as from moderate to good feelings or from neutral to disengaged states, indicating common emotional trajectories among users. This analysis helps in understanding the dynamics of user emotions and can be used to predict future states based on current observations.

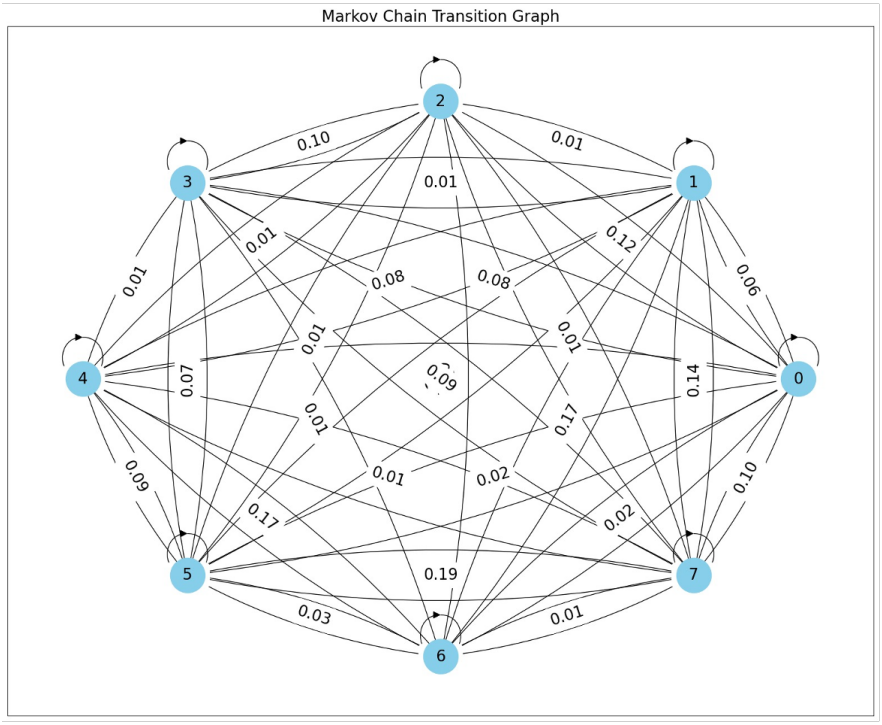

Figure 10: Markov Chain

Figure 10 presents the Markov chain diagram, where nodes represent different emotional states and edges represent the transition probabilities between these states. The Markov Chain transition analysis reveals several interesting observations about the emotional and engagement states of Feelee app users. Clusters exhibit high self-transition probabilities, meaning users tend to remain in the same state. However, excluding self-transitions and conditional probabilities, we observe that Cluster 0 (Low Feelings - Disengaged - Negative - Passive) transitions to Cluster 1 about one-third of the time. Cluster 1 (Low Feelings) transitions to Cluster 6 one-third of the time. Cluster 3 (Moderate Feelings - Calm - Neutral - Passive - Alone) transitions to Cluster 2 60% of the time. Cluster 4 (Moderate Feelings - Neutral - Essential - With Someone) transitions to Cluster 5 or Cluster 7 40% of the time. Cluster 6 (Good Feelings - Neutral) transitions to Cluster 1 one-third of the time. Cluster 7 (Good Feelings - Energetic - Positive - With Someone) transitions to Cluster 6 40% of the time. The most interesting pattern is the oscillation between Cluster 1 and Cluster 6, where users move from low feelings to good feelings and back, suggesting a recurring cycle between these states. The original transition matrix and the conditional transition matrix are presented in Table 11 and Table 12, respectively

6 Conclusions and Recommendations

The project has been structured to effectively answer the research question "How many user profiles exist and what are their characteristics?". After carefully analyzing and interpreting the user data from the Feelee app, a data pipeline was created to transform individual entries into a 7-day data

set. Key steps included pivoting data to align responses over time, quantifying sentiment scores from categorical data, and calculating measures as mean and variance of quantifiable feelings.

Moving forward, the 7-day data set based on emotions was used to deploy k-means clustering. This provided the final clusters of users with similar characteristics. One of the most interesting findings was the high performance in the classification model used to evaluate the quality of clusters and the feature importance. The accuracy, precision, and recall of the classification model were more than 95%, showing that it is possible to use the features of the questionnaire to deduce the clusters in which a user is situated. This suggests that the model can robustly predict the cluster or the label of each user.

Furthermore, the lack of significance of physical activity in the analysis was a surprising finding. Despite the known correlation between regular exercise and happiness, there was no consistent indication of this across the clusters. It was found that users in "okay"-clusters exhibit a mixed level of physical activity, which can be lower than those in "not-okay" clusters and higher than those in "good" clusters. This highlights the complexity of factors influencing well-being and suggests that other variables beyond physical activity like the answers to the questionnaire may play a more important role in determining happiness within specific clusters.

Additionally, the application of Markov chain modeling enabled the understanding of transitional patterns within clusters. By analyzing the state dynamic of the resulting Markov Chain, a high probability cycle between cluster 1 and cluster 6 was discovered. This shows that a user can move from a "good" cluster to a "not-okay" cluster recurrently.

Finally, it is crucial to highlight that the results of the clustering process are highly dependent on the interpretation of feelings. The decision of how to rank and group feelings fundamentally alters the clusters. Therefore, interpreting and categorizing feelings with the right tools and advice is vital to ensure meaningful and reliable clusters.

6.1 Limitations

Next, the limitations encountered will be discussed. One of the biggest limitations was the distribution of data following the data-wrangling process. Out of approximately 3000 users, the top 50 users with the most entries contributed to about 50% of the entire clustering data. It is important to note that many users were filtered out due to the requirements of having at least seven days of entries.

The distribution of our data after the data-wrangling process presents a problem in two ways. Firstly, this means that the clustering methods are heavily influenced by the behaviors and patterns of the top 50 users. This would imply that the clustering analysis may not be representative of the broader user base, leading to biased conclusions. Secondly, the model may be overfitted to these specific users, rather than providing generalized insights about the population. In summary, the heavy reliance on a small subset of users poses a substantial limitation for the project, because it affects the reliability and applicability of the results.

The next limitation encountered was the interpretation of feelings. To transform categorical data into numerical data, the feelings have been ranked and grouped according to the feedback of professionals. However, a completely different meaning can be given to the analysis by changing the scale of the ranking, and consequently altering the interpretability of results. The weights and significance assigned to each feeling could be impacted by these changes, altering highly the outcomes of clustering and Markov Chain analysis. Therefore, validation of the ranking and grouping process by a professional is needed to ensure that the analysis remains accurate and that the results are reliable and meaningful.

6.2 Recommendations for Further Studies

Finally, the recommendations for the continuation of the project will be laid out. The first suggestion regards question 4 of the Feelee app. It is suggested that the questions "What are you doing?", which has resulted in mainly three different responses: "relaxing", "nothing", and "something else", be either removed or rephrased to obtain more specific and insightful data. The reason for deleting or rephrasing this question is that the answers currently provided by users do not enrich the analysis, but rather they introduce ambiguity.

The removal of question 4 is proposed to simplify the journaling process of Feelee. This change is intended to increase user engagement due to the reduced completion time of the survey, and overall time consumption in the app. Alternatively, rephrasing this question could lead to more valuable insights. The question could be rephrased to "What activity were you engaged in before completing this questionnaire?" or "What were you doing 15 minutes before?". This change is intended to gather more detailed information about why the user is feeling a certain way.

The next recommendation concerns the grouping and ranking of feelings/emotions within the Feelee application. Initially, this was accomplished by combining insights from relevant literature and input from professionals to categorize and rank these feelings. However, as previously mentioned, the results of the cluster analysis are highly dependent on these groupings and rankings. Thus, it is recommended that Feelee revisit and edit these groupings to better align with the specific needs and objectives.

Finally, to improve the quality and quantity of data collected, it is advised for Feelee to implement a reward system. By offering incentives, Feelee can encourage a longer engagement and more frequent usage of the application. This could include implementing features like loyalty programs, achievement badges, or other forms of recognition to encourage users to keep engaging with the app.

References

[1] Marie Dasborough et al. "Measuring Emotion: Methodological Issues and Alternatives". In: *Research Companion to Emotion in Organizations*. 2008. DOI: 10.4337/9781848443778.00021.

[2] Alaa Alslaity and Rita Orji. "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions". In: *Behaviour & Information Technology* 43 (2022), pp. 1–26. DOI: 10.1080/0144929X.2022.2156387.

[3] A. Hazra and N. Gogtay. "Biostatistics Series Module 4: Comparing Groups - Categorical Variables". In: *Indian Journal of Dermatology* 61.4 (July 2016), pp. 385–392. DOI: 10.4103/0019-5154.185700.

[4] Saikat Basu et al. "Statistical Analysis of Emotional Response through Physiological Signals". In: Apr. 2021, pp. 189–201. ISBN: 978-981-33-4083-1. DOI: 10.1007/978-981-33-4084-8_18.

[5] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[6] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

7 Appendix

Exploratory Data Analysis

Age group	Number of entries
9 - 20	25337
21 - 30	15878
31 - 40	9125
41 - 50	5933
51 - 60	2341
61 - 70	613
71 - 81	44

Table 5: Data Exploration - Age groups and Users

	Good	Okay	Not okay
count	8405	8621	4113
mean	6099	5658	5248
std	4478	4381	4261
min	1	0	8
25%	2775	2392	2031
50%	5252	4806	4271
75%	8367	7898	7390
max	36672	36672	33702

Table 6: Descriptive Statistics for Overall Feeling Groups

Statistical Data Analysis

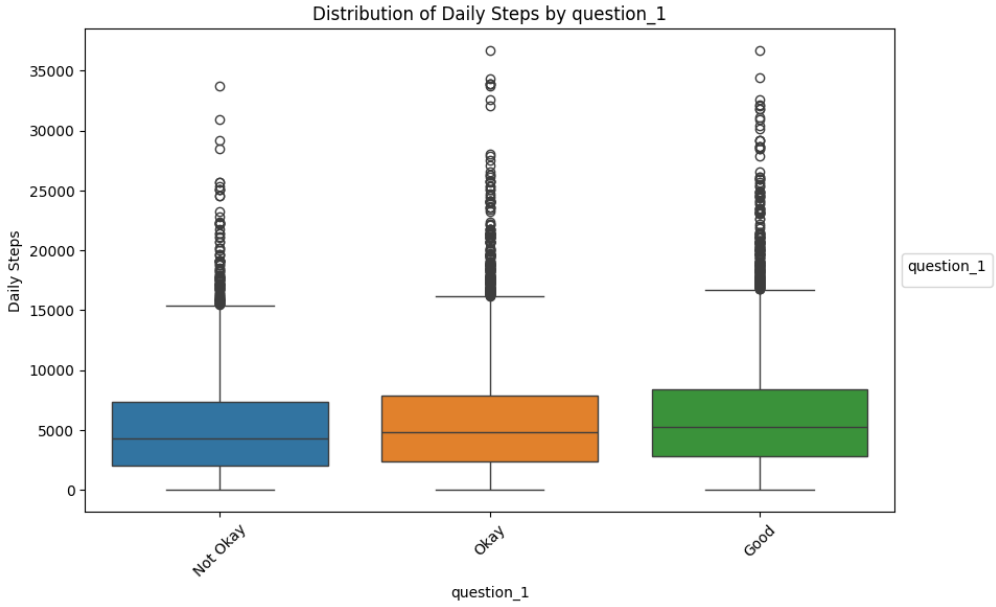

Figure 11: Boxplot of Daily Steps per Question 1 Answers

	She	He	Them	Not Say
count	13682	6722	691	44
mean	5495	6448	4227	3739
std	4224	4746	3600	3190
min	0	0	80	202
25%	2321	2946	1107	946
50%	4659	5578	3705	3080
75%	7616	9092	6015	5017
max	36672	34405	23819	16017

Table 7: Descriptive Statistics for Gender Groups

	Young adults	Adults
count	12641	8498
mean	5788	5701
std	4371	4463
min	0	0
25%	2473	2470
50%	5000	4716
75%	8092	7843
max	36672	34405

Table 8: Descriptive Statistics for Age Groups

Methodology

Ranking	Feeling	Category
1	Joyful	Energetic
1	happy	
1	Energetic	
1	Festive	
0.8	Surprised	Calm
0.8	Calm	
0.8	Proud	
0.8	Satisfied	
0.8	Interested	
0.6	Loved	Affection
0.6	In love	
0.6	Admiration	
0.4	Tired	Disengaged
0.4	Bored	
0.4	Tiresome	
0.4	Indifferent	
0.4	Confused	
0.4	Gloomy	
0.4	Sad	
0.2	Nervous	Anxious
0.2	Anxious	
0.2	Irritated	
0.2	Woried	
0	Dissatisfied	Angry
0	Angry	
0	Frustrated	
0	Embarrassed	
0	Disappointed	

Table 9: Question 2 Response Ranking

Ranking	Reasoning	Category
1	Compliment	Positive
1	Fun task	
1	Good news	
1	Good night	
1	Party	
1	Sports	
1	Self-care	
1	Holiday	
1	Rest	
1	Meet someone	
1	Conversation	
1	New relationship	
0.5	NA	Neutral
0.5	Something else	
0.5	Surprised	
0.5	Agenda	
0.5	Agenda	
0.5	Don;t know	
0.5	Okay night	
0.5	Mediocre task	
0.5	Mediocre news	
0.5	Environment	
0.5	Full agenda	
0	Stupid task	Negative
0	Sick	
0	Stress	
0	Bad news	
0	Lonely	
0	Bad night	
0	Argument	

Table 10: Question 3 Response Ranking

Result Analysis

Figure 12: Word Clouds for Clusters

	0	1	2	3	4	5	6	7
0	0.52	0.17	0.12	0.05	0.02	0.08	0.04	0.01
1	0.10	0.46	0.09	0.01	0.01	0.14	0.19	0.01
2	0.08	0.11	0.58	0.10	0.00	0.01	0.12	0.01
3	0.12	0.03	0.29	0.50	0.00	0.00	0.04	0.01
4	0.02	0.03	0.01	0.01	0.57	0.17	0.03	0.17
5	0.06	0.17	0.01	0.01	0.12	0.41	0.16	0.06
6	0.02	0.17	0.07	0.01	0.03	0.10	0.51	0.09
7	0.00	0.03	0.00	0.01	0.07	0.08	0.14	0.67

Table 11: Original Transition Matrix

	0	1	2	3	4	5	6	7
0	0.00	0.35	0.24	0.10	0.05	0.17	0.08	0.01
1	0.18	0.00	0.17	0.01	0.02	0.25	0.35	0.01
2	0.20	0.27	0.00	0.24	0.00	0.01	0.28	0.01
3	0.23	0.07	0.59	0.00	0.00	0.00	0.08	0.03
4	0.05	0.06	0.02	0.01	0.00	0.40	0.07	0.39
5	0.10	0.30	0.01	0.01	0.21	0.00	0.27	0.11
6	0.05	0.35	0.14	0.02	0.05	0.20	0.00	0.18
7	0.00	0.10	0.00	0.02	0.22	0.24	0.41	0.00

Table 12: Conditional Transition Matrix