An analysis of the causes of retention rates in the Bump2Baby program

Quan Zhao and Michael Kerr
School of Mathematics and Statistics
Victoria University of Wellington, New Zealand

November 5, 2023

Abstract

Using data about individuals from the Bump2Baby app and their interactions with their health coach, we were able to identify clusters among the dataset which have significant and impactful traits which may explain the retention rates in the program per cluster. Through the usage of k-means clustering and survival analysis, it is possible to identify potential worrisome individuals who may be more at risk of not staying engaged with their coach. These individuals are described by cluster 4 and their top five most impactful features are medicine, obesity 2 (as defined by World Health Organization (2010)), advisor ID, number of diet entries into the app, and underweight. This cluster had a survival probability at 200 days, which is the approximate day of birth, of 16.4%. Comparing this to the best performing cluster, cluster 3, there is a noticeable difference in the top five most impactful features. For cluster 3, these are: number of alcohol entries into the app, number of messages received, maximum text message length, and age. This cluster had a much higher survival probability after 200 days with 99.2% of individuals staying engaged with their coach.

Contents

1 Introduction

Gestational diabetes and obesity are two common symptoms of having a child. According to the International Association of the Diabetes and Pregnancy Study groups, gestational diabetes affects one in five pregnancies (Sacks et al. (2012)) and weight gain after birth further worsens these odds (Lim et al. (2019)). Approximately half of these women will go on to develop type 2 diabetes within five to ten years along with their children who are more likely to have type 2 diabetes later in their life (Vounzoulaki et al. (2020)). The Bump2Baby project aims to prevent these risks by focusing on weight management during and after pregnancy. Individuals in this project are assisted by a personal health coach along with an accompanying app to track their weight and nutritional goals.

The aims of the following analyses in this report are to create clusters among the individuals and identify significant traits among each cluster. In doing so, we are able to determine what types of individuals have the highest retention rate in the project and the causes of these retention rates. The methods of analyses involved used medical information about the individual (illnesses, body mass index) and summary statistics of the messages between the individual and their coach.

Sections 3 to 4 describe the final dataset to be worked with and sections 5 to 6 describe the analyses and the key findings. The last sections discuss potential further work and an overall conclusion to this project.

2 Objectives and Scope

2.1 Objectives

The research is focused on achieving two primary objectives within the context of the Bump2Baby program, which is dedicated to fostering health and wellness among pregnant women to reduce the risks of gestational diabetes and obesity:

1. Group Differentiation and Analysis: To identify distinct groups or clusters of individuals within the program based on their engagement and behavioral data. This includes analyzing differences in how individuals interact with the Bump2Baby app and their coaches, as well as any other registered data that may influence the program's effectiveness.

2. Predictive Group Assignment: To develop a predictive framework that can accurately classify new participants into these identified groups, thus tailoring the program's interventions to fit the specific needs and behavioral profiles of the individuals.

2.2 Scope

The scope of this study is defined to include:

1. Cluster Identification: Utilizing k-means clustering and other statistical techniques to discover inherent groupings within the 367 individuals based on their app usage and interaction data.

2. Behavioral Data Analysis: Assessing and visualizing patterns of communication, app interaction, and self-registered information to understand the nuances of participant engagement.

3. Survival Analysis: Implementing Kaplan-Meier estimators to measure program retention rates, providing insights into how long individuals stay engaged with the program across different clusters.

4. Feature Impact Study: Performing Principal Component Analysis (PCA) to identify key features that characterize the different clusters, thereby aiding in the understanding of what drives engagement and retention.

5. Missing Data Management: Addressing missingness in the dataset through Multiple Imputation by Chained Equations (MICE), under the assumption that the data are Missing At Random (MAR).

6. Statistical Interpretation and Predictive Modeling: Interpreting the statistical results to understand group dynamics and to inform the development of a model for assigning new individuals to the appropriate groups for customized intervention strategies.

3 Data Description

3.1 Feature construction and consideration

The World Health Organization (2010) defines body mass index (BMI) as BMI $= \frac{\text{weight (kg)}}{[\text{height}(m)]^2}$. These indexes are then further categorized into: Underweight, Normal, Pre-obese, Obesity 1, Obesity 2 and Obesity 3. Each individual had their BMI calculated by using their weight and height that was entered at the start of the study. Each individual's change in BMI across the study has not been considered.

Across the study, each individual has been asked three times whether or not they are going to breastfeed. The final time they were asked is the value used in the analysis with the rest excluded.

Features *not* considered in the analysis are:

- Time data such as dates.
- Weight milestones, while very interesting to analyse on their own, will only complicate the analysis. It is not easy to determine when exactly the weights are taken over the study.
- All but one of the variables that are highly correlated with eachother, and variables that likely carry the same information such as daysOnProgram and daysOnPlatform.
- Assumed information such as gender.
- Forum activity due to lack of response.

3.2 Citizen Registration Analysis

3.2.1 Start end registrations

The *start end registrations* table aggregates data for each registration type of citizens. Despite the table containing 1048575 rows, a significant majority (1047272 rows) have their 'Type' set as "nan", indicating that these citizens have not specified a registration Type.

Only 357 unique citizens have defined a registration type. Fig 1

The distribution of the count of citizens against the number of types selected by them is illustrated in Fig 2.

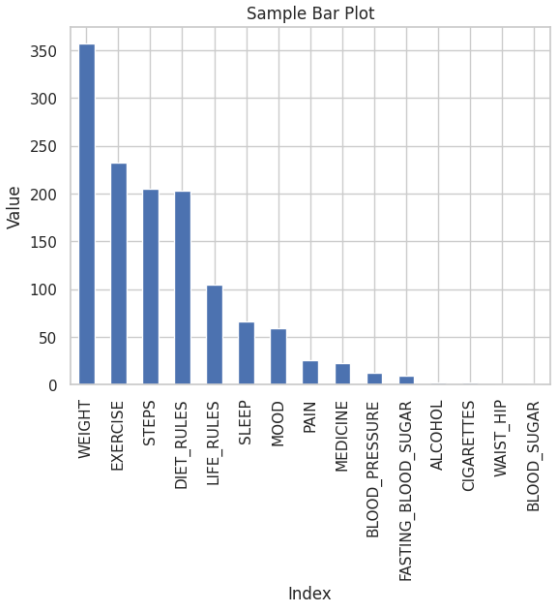

Figure 1: Count of types

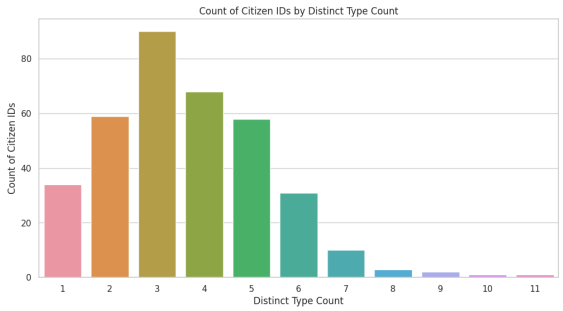

Figure 2: How many citizens select how many types in start_end_registrations

3.2.2 Registrations per Goal

The *registrations per goal* table aggregates registration data by the goal of each registration type for citizens.

Notably, a single registration type can be associated with multiple goals. There are 362 unique citizens with registration types in both the *registrations per goal* and *goal and registration* tables, differing from the distribution in *start_end_registrations*.

3.2.3 Goal and Registration

The *goal and registration* table provides detailed registration data for each type and its associated goals for citizens.

3.2.4 Missing Citizens in Start end registrations

Upon analysis, it was discovered that four citizens, namely 831841, 909811, 955121, and 1022212, were absent in the *start end registration* table but present in both the *registrations*

per goal and *goal and registration* tables.

3.2.5 Table Selection for Aggregation at the Citizen Level

The *Start end registrations* table appears to be the most suitable for representing citizen registration activities since it is already aggregated at the registration type level, although, it lacks information on four citizens and excludes the "EXERCISE" category. because of, the *registrations per goal* table detailed goal-wise data for each registration type might result in a substantial aggregation workload. The *goal and registration* table's data is deemed too intricate and is thus not considered for this phase.

For a more comprehensive representation of citizen registration activities, We aggregated in registration type level for each citizen based on "valueCount" in the *Start end registrations* table.

3.3 Text Messages Analysis

We have consolidated the text message data for individual citizens and performed an aggregation based on four primary characteristics: intervention by the citizen in the message, the auto-generation of the message, the presence of a video in the message, and the length of the message text.

3.3.1 Features Description

1. **count_intervention**: This feature represents the number of times a citizen intervened in the message.
2. **count_autogenerated**: This denotes the number of messages that were auto-generated.
3. **count_withvideo**: This feature indicates the number of messages that contain a video.
4. **min_messagetext_length**: This represents the minimum length of the message text among all the messages for a given citizen.
5. **avg_messagetext_length**: This provides the average length of the message text for a given citizen.

6. **max_messagetext_length**: This denotes the maximum length of the message text among all the messages for a given citizen.

These features provide a comprehensive overview of the text messaging behavior and content of individual citizens. By analyzing these features, we can gain insights into the communication patterns, preferences, and tendencies of the citizens.

4 Missing values and MICE

Missing values are often a problem when working with datasets. Some techniques to handle missing values include ignoring them, or imputing them. To impute means to replace these cells with missing values with another value. The problem now is to determine the most appropriate value to give the cell or, in a more appropriate and principal phrasing, determine the most appropriate value to give the individual. This distinction is important as it gives emphasis on the need for context when imputing missing values.

Single imputation encompasses several methods of imputation including mean imputation or regression imputation (fit $Y = \widehat{\beta}X + \epsilon$, where Y is your missing variable and X is the rest of your variables). Jadhav et al. (2019) says that it is assumed that the single imputation is the correct one precision is overstated, but there can never be absolute certainty about validity of imputed values. Rubin (1987) developed a method for averaging the outcome across multiple imputed datasets which has been further developed into R's package mice.

Multiple imputation in chained equations (MICE) is a popular method of imputation largely because of its ability to efficiently compute many different variations of these missing values and find the values which best fit. Each iteration of imputation generates complete datasets based on regression models of each variable where these regression models can be specified, for example, as logistic, poisson or even random forest (Azur et al. (2011)). In our MICE model, each variable was fitted to a random forest regression model and imputed using these values. A major assumption that MICE makes, however, is that the data is missing at random.

4.1 Missingness

According to Arnold (2023), there are three main categories of missingness: Missing completely at random (MCAR), missing at random (MAR) and non-ignorable non-response (or missing not at random; MNAR). MCAR means that there is some computable probability that a value is missing, and MNAR means that the reason why a value is missing is due to the outcome of interest. MAR means that the reason for why the value is missing is *entirely* dependent on the auxiliary variables. Multiple imputation makes the assumption that the missing values are MAR.

It can be argued that the data is not actually missing at random as missingness in variables a lot of the time comes down to the location of the individuals (Kang (2013)): some sites where the study was taken may have different health coach techniques, and therefore different coach performance. Our data does slightly include this information as the variable advisorId encodes site information in its first digit. Further investigation into missingness is encouraged.

5 Statistical Models and Techniques Used

5.1 Clustering

We endeavor to cluster citizens based on their messaging behavior and associated features so we adopted the k-means clustering algorithm. To facilitate a more intuitive visualization of these clusters, we employed the t-SNE (t-distributed Stochastic Neighbor Embedding) technique.

A noteworthy observation from the clustering results is the emergence of a distinct cluster, differentiated from the others. A deeper analysis revealed that members of this cluster uniformly exhibited the characteristic of having the highest maximum message length. Specifically, the longest messages dispatched by these individuals are notably extensive in character count compared to their counterparts in other clusters.

After a meticulous preprocessing and normalization phase of the features, and subsequent iterative experimentation with diverse cluster counts, a consensus was reached that a quintet of clusters—five in total—offered the most balanced distribution. This determination was substantiated by the insights gleaned from the T-SNE plot in Figure 3.

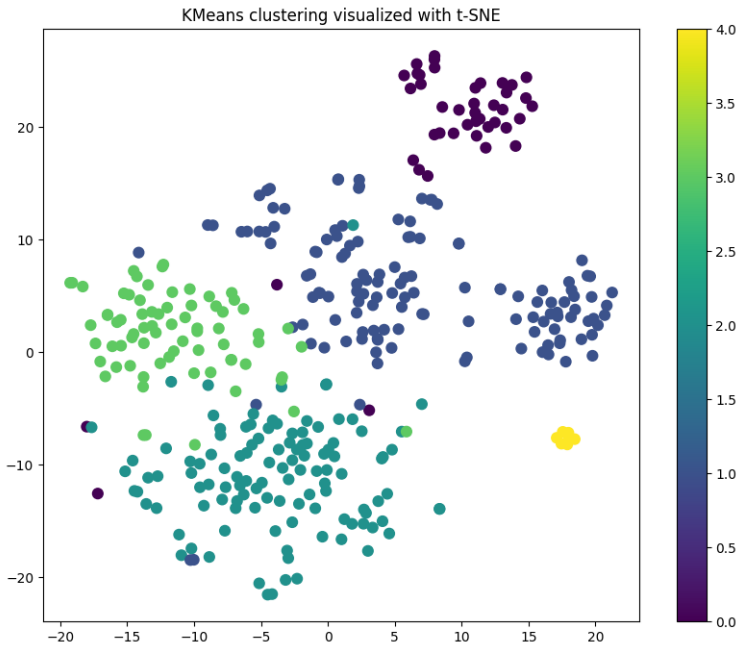

Figure 3: Count of types

5.2 PCA Analysis

Principal Component Analysis (PCA) is a statistical procedure that orthogonally transforms the original variables of a dataset into a set of linearly uncorrelated variables known as Principal Components (PCs). These PCs are ordered such that the first few retain most of the variation present in all of the original variables. PCA is frequently employed for dimensionality reduction, especially in contexts where data variables are highly correlated.

In the context of our study, PCA was not just a tool for dimensionality reduction but was instrumental in discerning the key features characterizing each cluster. Our methodology involved executing PCA individually for each cluster, aiming to identify the components that cumulatively explained 95% of the variance.

Subsequent to the PCA, we computed the cumulative loadings for all features across the identified principal components. Here, 'cumulative loading' refers to the aggregated contribution of each feature across the components, providing an indication of the significance of each feature in explaining the variance within the cluster.

To elucidate the relative importance of the original features, they were ranked based on their cumulative loadings. This ranking facilitated the identification of features that were

most explanatory for each cluster.

Visually, our findings are presented in two distinct plots. Figure 4, a radar plot, provides a comprehensive view showcasing the impact of all features for each cluster. In contrast, Figure 5 focuses on the top 5 most impactful features for each cluster, offering a more concentrated perspective. This latter plot is particularly illuminating as it accentuates the distinguishing features between clusters, providing clear differences in their characteristics.

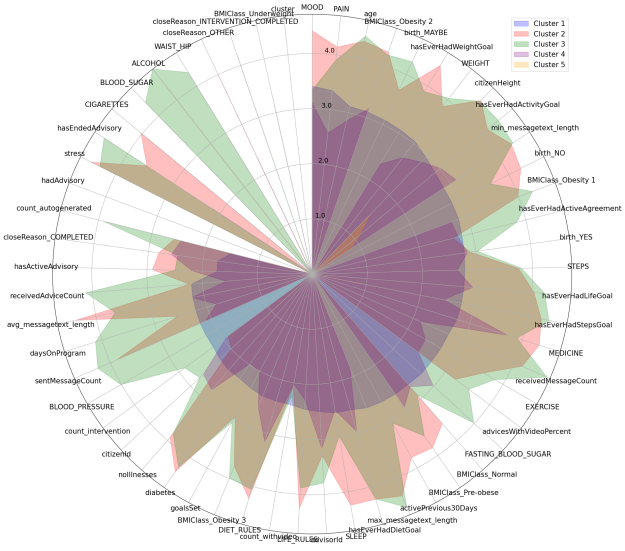

Figure 4: Impact of all features for each cluster

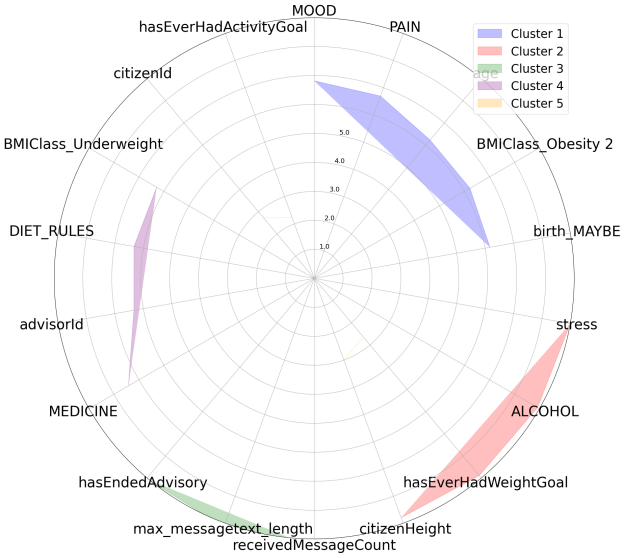

Figure 5: Impact of top 5 features for each cluster

5.3 Survival analysis

After clustering, some measure of determining the performance of each cluster with respect to how long the clusters remained in the program is desired. Survival analysis is perfect for this, where survival is measured by how long each citizen stays in the program. Given the data, survival probabilities up to a time t (days) can be estimated using a Kaplan-Meier estimator. This survival estimator takes the form:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where t_i is a day when an individual was censored or left the program, d_i is the number of people who left the program at t_i, and n_i is the number of people still in the program.

5.3.1 Censorship

The Kaplan-Meier estimator has the ability to account for events that are invalid. For example: an individual leaves the program due to an outside reason and not of their own accord (which is the outcome of interest: voluntarily leaving the program). An individual could be asked why they left the program, or the reason could be given by the coach, or there could simply be no reason why an individual left. These responses (or lack thereof) were recorded and individuals were deemed uncensored/censored according to the reasons given. The following examples give a general criteria for censorship:

- An individual left the program and there is no reason given: censored.
- An individual left the program but did not communicate when exactly they left: censored.
- An individual left the program and communicated when exactly they left: not censored.
- An individual completed the program: not censored.

5.3.2 Results

Cluster 3 has a much higher retention rate and far less wide confidence intervals than the other clusters and cluster 2 has a much more average retention rate than the rest of the

clusters. Cluster 1 does have higher survival probabilities than 4 and 2, but its confidence intervals are very, very wide. The cluster of primary concern is cluster 4 where individuals in this cluster leave the program at a much faster rate than all other clusters.

Day	**Cluster**	**Survival probability**	**95% CI**
50	1	97.6%	(93.0%, 100%)
	2	88.1%	(82.9%, 93.6%)
	3	100.0%	(100.0%, 100.0%)
	4	73.2%	(60.8%, 88.2%)
200	1	92.3%	(84.3%, 100%)
	2	74.2%	(66.9%, 82.3%)
	3	99.2%	(97.6%, 100.0%)
	4	16.4%	(7.4%, 36.1%)
350	1	76.4%	(61.4%, 94.9%)
	2	63.6%	(55.0%, 73.6%)
	3	91.7%	(87.0%, 96.8%)
	4	0%	(NA, NA)
500	1	49.5%	(28.7%, 85.4%)
	2	27.1%	(10.8%, 68.1%)
	3	54.6%	(45.9%, 65.0%)
	4	0%	(NA, NA)

Table 1: The probability of surviving up to day t

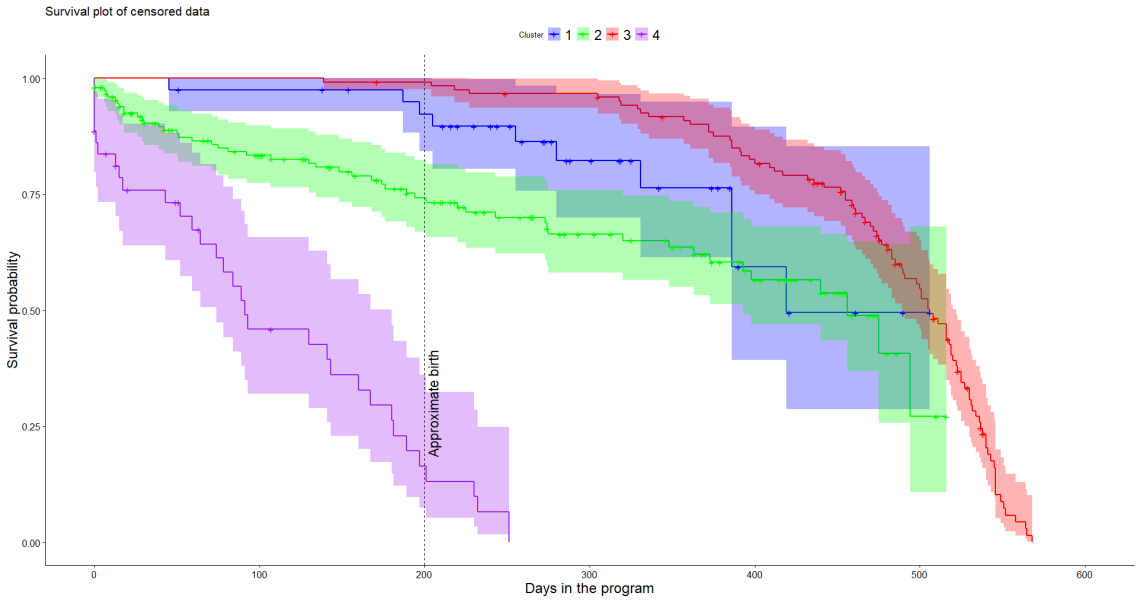

Figure 6: Kaplan-Meier curve of survival probabilities per cluster

6 Key Findings

6.1 Most impactful features

Table 2 shows the top 5 most impactful features for each cluster.

An examination of the top features for each cluster reveals intriguing patterns and insights. Cluster 3 is particularly noteworthy. Citizens encompassed within this cluster tend to remain in the program for extended durations, receiving messages with higher frequency and greater length. An interesting feature influencing this cluster is the registration type labeled 'Alcohol'. This prominence of 'Alcohol' as a defining feature warrants a deeper exploration to understand its implications and potential underlying causes. The 'Alcohol' feature is a count of the number of times a user has entered alcohol related entries into their app. This may mean that individuals in Cluster 3 tend to focus on their own alcohol habits which is important during pregnancy.

Conversely, Cluster 4 exhibits a contrasting behavior. Citizens here often exit the program prematurely. Influential factors for this cluster encompass attributes like BMI, DIET, and MEDICINE. A point of particular concern is the prominence of individuals labeled as 'BMI underweight'. Such individuals might necessitate specialized attention and interven-

tions to ensure their well-being and sustained participation.

It's imperative to understand the nuances of feature loadings. All the top features identified are in a positive direction. A feature's high loading across all principal components, especially in the context of Cluster 2, signifies its pivotal role in determining the direction of the principal components for the data points within that cluster. However, it's crucial to differentiate between the importance and magnitude of a feature. While the loading elucidates the significance of a feature in capturing the variance or structure of the data, it doesn't provide insights into its magnitude. Thus, even if a feature has the highest loading on the principal components, it doesn't inherently imply that its average or cumulative values will be the most pronounced across all clusters.

Cluster	**1st**	**2nd**	**3rd**	**4th**	**5th**
1	MOOD	PAIN	age	Obesity 2	Birth (Maybe)
2	stress	ALCOHOL	hasEverHadWeightGoal	MOOD	Height
3	ALCOHOL	Message count (received)	Max message length	hasEndedAdvisory	Age
4	MEDICINE	Obesity 2	advisor ID	Count of diet entries	Underweight
5	Height	citizen ID	Previous weight goal?	Previous activity goal?	Age

Table 2: The top 5 most impactful features per cluster

7 Discussion

7.1 Boundaries of the study

The study is limited to the data provided by the Bump2Baby app and does not include external variables that could impact the health outcomes of the participants. The predictive modeling is based on the assumption that app engagement patterns are indicative of program adherence and outcomes.

7.2 Conclusion

There are two clusters of significance: cluster 3: the cluster with retention rates within the program and cluster 4: the opposite. What is noticed in cluster 4 is the two features that make the most impact are related to an individual's BMI; Cluster 4's top five most impactful features include Obesity 2 and Underweight BMI classes. The DIET_RULES feature, which is about the amount of times an entry was added into the app about their diet, is also related

to this. It seems that individuals in cluster 4 may either have weight problems or their BMI is actually related to their ethnicity (Kirby et al. (2012)).

Cluster 3, the cluster with the highest retention rate, is most related to interactivity with the app. They seem aware of the impacts of alcohol on the child during pregnancy (Dejong et al. (2019)) and are the most responsive to their health coach with the highest text message length. This is indicative of the mothers within this cluster tending to be more concerned with the health of their child.

Among the other clusters, clusters 1 and 2, there are also noticeable patterns. These clusters tend to be associated with mental health where there is high amounts of entries into the app about mood, pain and stress. The retention rates of these clusters are not of massive concern and are likely representative of the wider population.

7.3 Future Extensions

The study paves the way for future research, which may include more comprehensive data collection, such as additional health markers and external factors that could influence the program's effectiveness. Further investigation could also refine predictive modeling techniques to enhance the personalization of the program for future participants.

There are many other clustering techniques than k-means clustering which may perform better and/or differently to k-means. Tuning the hyperparameters in these models and, in particular, tuning the number of clusters. There are also other ways of measuring survival such as the Nelson-Aalen estimator, and performing an in-depth log-rank test of each cluster to determine if certain clusters are really just the same.

References

Arnold, R. (2023). STAT392: Sample Surveys. `https://homepages.ecs.vuw.ac.nz/~rarnold/STAT392/SampleSurveysBook/_book/nonresponse.html#types-of-nonresponse`.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, **20**(1), 40–49.

Dejong, K., Olyaei, A., & Lo, J. O. (2019). Alcohol use in pregnancy. *Clinical obstetrics and gynecology*, **62**(1), 142–155.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, **33**.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, **64**(5), 402–406.

Kirby, J. B., Liang, L., J., C. H., & Wang, Y. (2012). Race, place, and obesity: the complex relationships among community racial/ethnic composition, individual race/ethnicity, and obesity in the United States. *American journal of public health*, **102**(8), 1572–1578.

Lim, S., Versace, V. L., O'Reilly, S., Janus, E., & Dunbar, J. (2019). Weight change and cardiometabolic outcomes in postpartum women with history of gestational diabetes. *Nutrients*, **11**(4).

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley.

Sacks, D. A., Hadden, D. R., Maresh, M., Deerochanawong, C., Dyer, A. R., Metzger, B. E., Lowe, L. P., Coustan, D. R., Hod, M., Oats, J. J., Persson, B., Trimble, E. R., & for the HAPO Study Cooperative Research Group (2012). Frequency of Gestational Diabetes Mellitus at Collaborating Centers Based on IADPSG Consensus Panel–Recommended Criteria: The Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study. *Diabetes Care*, **35**(3), 526–528.

Vounzoulaki, E., Khunti, K., Abner, S. C., Tan, B. K., Davies, M. J., & Gillies, C. L. (2020). Progression to type 2 diabetes in women with a known history of gestational diabetes: systematic review and meta-analysis. *BMJ*, **369**.

World Health Organization (2010). A healthy lifestyle - WHO recommendations. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.