Breast Health Monitoring & Awareness - Analysis

Prepared by

Romaric SALLUSTRE M.Sc.(C)
Data Analyst Internship
APRIL - AUGUST 2023

Mentor

Dr. Jean Christophe Thalabard M.D., Ph.D.

Supervisor

Mrs. Sweekrity Kanodia Ph.D.(C)

Submitted
in Partial Fulfillment of the Requirements for the degree of
Master of Science AIRE – Digital Sciences
at
Université Paris Cité – Learning Planet Institute

Table of Contents **1**
Introduction 2
Bibliographic Review 3
Overview of the Activities Undertaken 4
BHA Study 4
Objectives 4
Seintinelles study 5
Objectives 5
Additional work 5
Approaches and Execution 6
BHA_Study - Study Methodology 6
Seintinelles Study - Study Methodology 7
Breasties Board - Additional Work 8
Development of Symptoms_Algorithm - Additional Work 9
Results 9
BHA Study 9
Seintinelles Study 11
Breasties Board 15
Development of Symptoms Algorithm 16
Discussions & Conclusions 17
Interpretation 17
BHA Study: 17
Seintinelles Study: 17
Critical Lessons and Achievements from My Internship 18
Future Directions 19
Acknowledgements 19
APPENDIX 20
References 20
Additional Supporting Materials 21
Files and Notebooks 21

Introduction

Breast cancer is the most common cancer among women worldwide. In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. Nevertheless, when found early, and if adequate diagnosis and treatment are available, there is a good chance that breast cancer can be cured. In 2021, WHO [1] established the WHO Global Breast Cancer Initiative with the aim of reducing global cancer mortality by 2.5% per year, thus avoiding 2.5 million early deaths due to breast cancer between 2020 and 2040 in women under the age of 70 years. The three pillars of action for achieving this mortality reduction are:

- Health promotion for early detection: public health education to improve awareness of the signs and symptoms of breast cancer, and of the importance of its early detection and treatment.
- Timely diagnosis: public and health worker education on signs and symptoms of early breast cancer so that women are referred to diagnostic services when appropriate.
- Comprehensive breast cancer management: cancer management requires some level of specialized care, establishing centralized services, and treatment for breast cancer.

Project 'Breast Health and monitoring in Nepal' proposes to spread awareness and monitoring of breast health via an android application and some educational materials. The project BreMo aims to be the two pillars of breast cancer mortality reduction as defined by WHO - Health Promotion and Timely diagnosis. BreMo is a mhealth application which helps women monitor their breast health by:

1) Educating women about breast cancer - risk factors, prevention, symptoms and associated myths
2) Helping women perform self-check with BreMo's step by step guide
3) Easy symptom recording by providing a comprehensive list of symptoms to choose from and a feedback about their symptoms based on the selection
4) Tracking of changes in breasts over time
5) Setting reminders for monthly check as per convenience and according to women's menstruation cycle for best results.

To ensure behavioral change, the Theory of change methodology was followed to develop this app, which is a systematic approach used to map out the steps and causal relationships that explain how an intervention or project is expected to bring about specific outcomes and impacts. This app is only available in android for now. In a general context, the project comes under the 3rd sustainable development goals of Health & Well Being.

Figure 1: SDG Goal for BreMo

Bibliographic Review

This internship encompasses a diverse range of tasks centered around the analysis of data with ordinal, non-Gaussian, and categorical characteristics. The study's workflow is designed to facilitate a comprehensive understanding of various statistical techniques and methodologies, with a focus on Randomized Controlled Trials (RCTs)[2], ordinal regression, non-parametric tests (illustrated through examples), and dimension reduction techniques such as Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA). Additionally, the study involves employing thematic analysis to interpret and extract insights from the comments section.

To gain deeper insights into the interplay between dependent and independent variables, inherent randomness within the data needs to be explored. This is achieved through the application of Linear Mixed Models, which enable us to delve into both random and fixed effects.

Since the data gathered was transformed into categorical-ordinal data, specific methods were required to analyze them adequately. Thus, I familiarized myself with ordinal regression [3], which are basically statistical procedures designed to analyze data that do not meet the assumptions of normality or homogeneity of variance inherent in traditional parametric tests. Two non-parametric tests that I specifically used for the analysis were Kruskal-Wallis and Wilcoxon rank-sum tests [4]. I also spent time learning Linear Mixed Model Regression [5] which are extensively used for analysis of data collected via citizen science studies.

For exploration of data in sentinelles study, tools like Dimension reduction techniques [6] Principal Component Analysis (PCA, Multiple Correspondence analysis (MCA) were studied. PCA [7] [8] is employed to transform correlated variables into uncorrelated principal components and plot the components according to maximum Standard Deviation order.

Furthermore, I gained understanding of thematic analysis [9] is an exploratory method that identifies patterns, trends, and meanings within open-ended questionnaire responses like comments and feedback.

Overview of the Activities Undertaken

This internship focused on two main studies of this project

1. Breast health application study - to compare BreMo with existing apps
2. seintinelles study - to pre-test BreMo and questionnaire designed for planned study in Nepal

BHA Study

Objectives

There are two main principle objectives of this study included:

1. To analyze whether BreMo meets the current standards in comparison to the other apps (Stan Swasthya, Pink Pakistan, Daisy Wheel, Breast Check Now, Dear Mamma - apps were selected from a comprehensive database of breast health apps created by supervisor SK)
2. To compare BreMo with the other existing apps

Seintinelles study

Objectives

Seintinelles study is a pilot study to test the questionnaire designed for field study in Nepal and additionally get feedback about BreMo from a French perspective. Seintinelles is a group of people who have either lived with cancer or those passionate about the subject.

Data Collection and Management

BHA Study

BHA study is a comparison study between various apps and BreMo where each participant reviewed one of the current apps (Dear Mamma, Breast Check Now, Daisy wheel, Stan Swasthya and Pink Pakistan) with "BreMo". A questionnaire-based blind study was conducted with a small group of women sampled using convenience sampling. A questionnaire-based randomized trial was conducted (15 women) to test BreMo with5 existing apps. Questions on four main features - self-check, symptom recording, symptom tracking, reminder were included in the questionnaire along with questions about user perception of data management, login, and language, and security.

Seintinelles study

Questionnaire was adapted from the Comprehensive Breast Cancer Knowledge Test (CBCKT) to include questions about breast cancer knowledge. Additional questions about breast self-check were included. Questionnaire also included questions about app UI and usability, was adapted

from mHealth App Usability Questionnaire (MAUQ).

The datasets in both cases were organized on Google Sheets. Each row in the dataset corresponds to a single survey response, capturing the information provided by an individual participant in both cases. On the other hand, each column represents a specific attribute or variable related to the survey questions, capturing various aspects of the data being collected. By leveraging Google Sheets as a data storage and organization platform, it was easy to input, edit, and analyze survey data collaboratively. This allowed us to have efficient data manipulation, statistical analysis, and the identification of trends and patterns that can contribute to meaningful insights and research findings.

Overview of Tools Used

The programming languages used for this project were Python on Google Colab and R on R studio.. Processed and analyzed data are stored in the Github repository via Visual Studio code. A combination of Python and R programming languages was utilized to leverage their respective strengths in data analysis and statistical modeling. Python provides robust data manipulation and machine learning capabilities, while R excels in statistical analysis and visualization. Google Colab and Datacamp enabled collaborative and scalable data analysis, and hence was chosen to work together with SK .

Approaches and Execution

BHA_Study - Study Methodology

Following an in-depth review of the data, the methodologies applied in the Breast Health App (BHA) study were distinctly defined. The BHA study is structured around a unique approach, where each participant evaluated the BreMo application in comparison to other existing apps, generating 15 rows of data.

The participants within the age group of 15-40 years old were recruited. Each participant was assigned to review one of the discussed apps and BreMo. To ensure no bias against BreMo, participants were made unaware that they were all assigned to review BreMo (app 2) along with another app. For them, BreMo was one of the existing apps that was randomly assigned to them. In the end, each app (excluding BreMo) was reviewed by three independent participants and BreMo was reviewed by all of them.

The evaluation of the apps was conducted through a paired comparison approach where each participant was assigned a pair consisting of BreMo and one of the five pre-selected apps. This systematic evaluation process was repeated 15 times in total, providing a comprehensive and reliable analysis of the apps' performance.

Initial preprocessing of this data was conducted using the Natural Language Toolkit (NLTK)

library. Descriptive statistics were initially performed. The answers on 4 important features (breast self-examination, reminder functions, symptom tracking, and symptom recording) were transformed into a matrix. Within this matrix, all features related to were converted into binary values (Table 1).

App_ID in Table 1 signifies - id allocated to each app (1 to 5), and User_ID signifies id allocated to each user (1 to 15).

The "Total_score'' variable signifies the total number of features correctly identified by each participant.. Further analysis was performed on this matrix. Ordinal regression models were implemented in R with the MASS and Lme4 packages, as well as in Python with Statsmodel. Each of these approaches offered unique advantages in the analytical process, contributing to a comprehensive and robust analysis of the dataset.

	User_ID	App_ID	f1	f2	f3	f4	f5	TotalScore
1	1	2	1	0	1	1	5	3
2	1	1	0	1	1	0	5	2
3	2	2	1	1	1	1	4	4
4	2	1	1	1	0	1	3	3
5	3	2	1	1	1	1	4	4
6	3	1	1	1	0	1	3	3
7	4	3	1	0	0	1	5	2
8	4	1	1	1	1	1	5	4
9	5	3	1	0	0	1	5	2
10	5	1	1	1	1	1	4	4
11	6	3	1	0	0	1	5	2
12	6	1	1	1	1	1	3	4
13	7	4	1	1	1	1	5	4
14	7	1	1	1	1	1	3	4
15	8	4	0	1	1	1	3	3
16	8	1	1	1	0	1	4	3
17	9	4	1	1	1	1	5	4
18	9	1	1	0	0	1	3	2
19	10	5	0	1	0	0	5	1
20	10	1	1	1	1	1	3	4
21	11	5	1	1	0	0	3	2
22	11	1	1	1	1	1	4	4
23	12	5	1	1	1	1	5	4
24	12	1	1	1	1	0	5	3
25	13	6	1	0	0	0	3	1
26	13	1	1	0	1	1	5	3
27	14	6	1	0	0	1	3	2
28	14	1	0	1	1	1	4	3
29	15	6	1	1	0	1	5	3
30	15	1	0	0	0	1	3	1

Table 1: Table of scores appointed to each feature (0 or 1) based on whether a particular feature (f1 - BSE, f2 - symptom recording, f3 - symptom tracking, f4 - reminder) was correctly identified by the user (User_ID) for each app of the two tested by each of them (App_ID)

The table displays scores assigned to specific features (like breast self-examination, symptom recording, symptom tracking, and reminders) for each user (identified by User_ID) and each of the two tested apps (identified by App_ID). These scores are either 0 or 1, indicating whether the symptoms are selected No(0) or yes(1) of a particular feature in each app they were using. This table helps assess how well users recognized these features in the two different apps during the testing process.

Seintinelles Study - Study Methodology

The Sentinelles study, which began in June 2023, had a dual purpose: to test a questionnaire for a field study in Nepal and to evaluate the BreMo application among French women. So far, 64 responses have been collected. The study's main focus was on examining the BreMo app's effectiveness and understanding how well women knew about breast cancer and self-examination.

To make sense of the data, we used various techniques to simplify it. We had to deal with different kinds of answers, like 'yes,' 'no,' and 'do not know.' We transformed these into numerical values, where 'yes' became 1, 'no' became 2, and 'do not know' became missing data (NA). This helped us analyze the information better.

For the missing data, we didn't ignore them; instead, we used a method called Multiple Imputation to fill in the gaps. It's like restoring a broken puzzle. This technique used the information we already had to guess what the missing answers could be, completing our dataset.

Next, we explored dimensionality reduction techniques. Principal Component Analysis (PCA) helped us understand Likert scale data, which uses a range from 1 to 7 to rate the application. It helped uncover hidden patterns while keeping the original data intact.

Multiple Correspondence Analysis (MCA) was like a magic trick for our 'yes,' 'no,' and 'do not know' data. It grouped similar responses together, giving us insights into general trends, risk factors, symptoms, self-examination habits, screening patterns, and common misconceptions. It's like sorting puzzle pieces into categories. Additionally, we creatively explored comments from the app_comment section, identifying keywords and themes, which we carefully documented in the appendix as part of our thematic analysis.

Results

BHA Study

15 participants completed the study, with a 42% dropout rate. After pre-processing of the data obtained, descriptive statistics were performed. The results of descriptive statistics are provided with respect to BSE - Breast Self Examination (Figure 2.a), recording (Figure 2.b), reminder (Figure 2.c), graphics (Figure 2.d), and content (Figure 2.d). In these graphs, ‘Other apps’ were cumulative responses by participants for other apps tested except for BreMo. The results suggest that BreMo is comparable to the existing apps and might even be preferred (67%) over other apps, especially for self-check (80%), and graphics (67%).

A mixed model was used to find the effect of fixed variable App_ID (that is different apps used for the study including BreMo) and the random effect from different participants on the total score for each app (Table 1). The total score was calculated by summing the score given to each

of the four features (self-check, recording, tracking, and reminder) based on if it is available in the app or not. Here, the reference was set to the BreMo app, and all the coefficients of all other apps were calculated with respect to it.

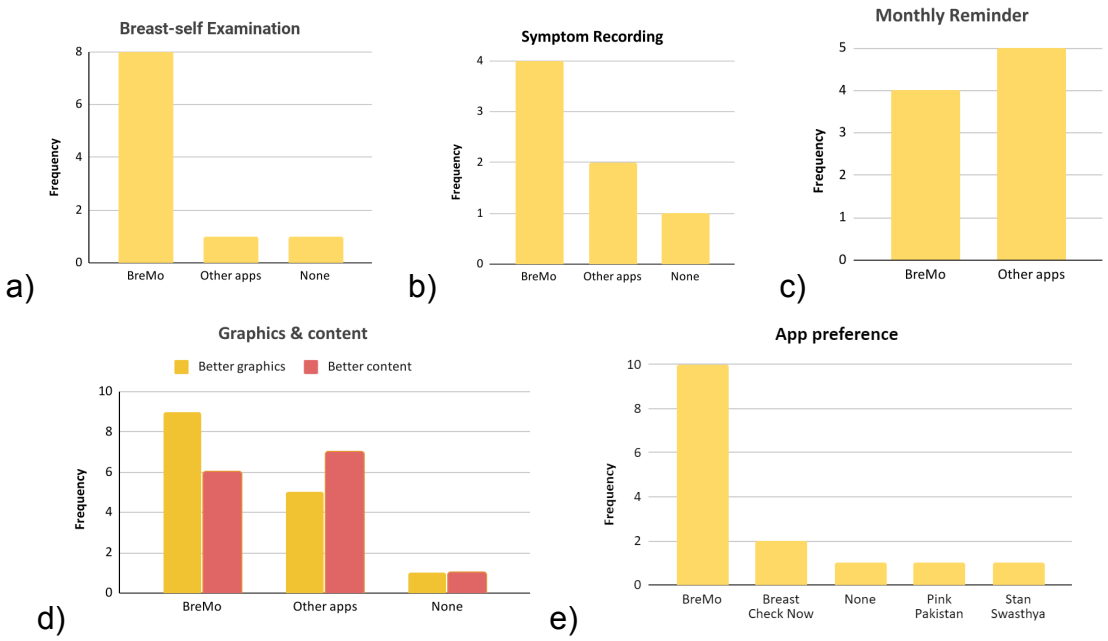

Figure 2:Features of BreMo compared with other apps: a) self-examination, b) symptoms recording, c) reminder, d) graphics and content, and e) app preference

Results of Linear Mixed models indicated variance due to random effect to be almost 0. The cumulative link mixed model resulted in similar observations. Similar results were observed of the linear mixed regression model in python with statsmodels. TotalScore is the total number of features selected by each individual. App_ID showcases the applications ranked from 1 to 6 where 6 is BreMo. At last, User_ID are the individuals. The dependent variable "TotalScore" is significant, but neither the "App_ID" nor the "User_ID" (as a random effect) seem to have a strong or statistically significant influence on the "TotalScore." The "Group Var" indicates that there is unexplained variability between groups.

```
Model:            MixedLM Dependent Variable: TotalScore
No. Observations: 30      Method:             REML
No. Groups:       6       Scale:              0.8547
Min. group size:  3       Log-Likelihood:     -44.5062
Max. group size:  15      Converged:          Yes
Mean group size:  5.0
-----------------------------------------------------
              Coef.  Std.Err.   z    P>|z| [0.025 0.975]
-----------------------------------------------------
Intercept      3.620    0.606  5.973 0.000  2.432  4.807
App_ID        -0.174    0.170 -1.019 0.308 -0.508  0.160
1 | User_ID   -0.026    0.048 -0.547 0.585 -0.121  0.068
Group Var      0.243    0.439
```

Figure 3: Mixed Linear regression results using package 'statsmodel' in Python

Owing to the negligible random effect from participants reviewing two different apps, the analysis was directed to ordinal regression models using "polr" package in R, as shown in Figure 3. It is evident from these observations that BreMo (reference) was similar to all the apps except App 2 (Dear Mamma) and App 4 (Breast Check Now) which are significantly different.

Seintinelles Study

The Seintinelles study dataset predominantly comprises categorical variables, and its exploration followed a meticulous three-step approach. Initially, K-Means Clustering was applied to the Likert scale data to unveil inherent patterns and groupings within the categorical responses. Subsequently, Multiple Correspondence Analysis (MCA) was conducted, facilitating a comprehensive examination of relationships between variables. Additionally, Principal Component Analysis (PCA) was specifically formulated for the Likert scale data to extract key components and latent structures, simplifying data interpretation. To ensure completeness, missing 'do not know' values within the knowledge section were addressed using the MissMDA package in conjunction with the MCA analysis. This comprehensive approach deepened our understanding of complex relationships within categorical variables and revealed latent structures in the Likert scale data, enriching the insights drawn from the Seintinelles study dataset.

1. **K-Means Clustering**

   K-means clustering was chosen for this specific context because it offers a straightforward and effective way to group the data into meaningful clusters based on the user feedback features provided. In the seintinelle study, we wanted to understand how users perceived and rated different aspects of an application and questionnaire. K-means is well-suited for this task because it doesn't require any assumptions about the underlying data distribution, and it can handle both numerical and categorical

features. We formulated k-means for the likert data which is basically the app review. The choice of K-means also aligns with the goal of creating six distinct clusters, which can help categorize users with similar feedback patterns. These clusters are essential for gaining insights into user preferences and identifying areas where the application may need improvement.

```
km.out <- kmeans(df, 6, nstart = 20)
km.out
```

```
## K-means clustering with 6 clusters of sizes 1, 19, 14, 3, 10, 14
##
## Cluster means:
##        use organisation acceptable      bse recording timeBSE    track
## 1 4.000000     4.000000   6.000000 6.000000  1.000000     5.0 6.000000
## 2 6.894737     6.736842   7.000000 6.789474  6.789474     7.0 7.000000
## 3 6.285714     6.142857   6.714286 6.714286  6.714286     7.0 6.428571
## 4 6.333333     6.666667   7.000000 5.666667  6.666667     7.0 7.000000
## 5 6.400000     5.400000   5.100000 6.200000  5.900000     6.4 5.900000
## 6 6.000000     5.785714   6.357143 6.357143  5.928571     6.5 6.714286
##   comfortRecrd rightTimeBSE  addInfo deleteData acceptableInfo intrusive
## 1     4.000000     5.000000 7.000000   6.000000            5.0       1.0
## 2     6.947368     6.842105 6.842105   7.000000            7.0       1.0
## 3     7.000000     2.714286 6.214286   6.857143            7.0       1.0
## 4     7.000000     4.000000 5.000000   6.333333            5.0       1.0
## 5     6.000000     6.300000 6.200000   6.300000            6.5       2.4
## 6     6.642857     5.714286 6.428571   5.500000            6.5       1.0
##   useAgain expectation   useful    enjoy satisfaction recommend
## 1 1.000000    5.000000 5.000000 4.000000     4.000000  6.000000
## 2 6.736842    6.842105 6.947368 6.947368     6.842105  6.947368
## 3 6.285714    6.214286 6.785714 6.785714     6.785714  6.357143
## 4 5.666667    4.333333 4.666667 4.666667     4.333333  4.666667
## 5 4.300000    5.300000 6.000000 5.700000     5.500000  4.500000
## 6 6.500000    5.214286 6.500000 6.357143     6.214286  6.428571
```

Figure 4: K-Means Clustering with optimal clusters results using package 'k-means' in R

The K-means clustering analysis was performed on a dataset with six clusters of varying sizes: 19, 3, 1, 10, 14, and 14. The sizes determine the total number of data points present in each cluster from 1 to 6. Each cluster is characterized by its mean values across 19 different features, such as "use," "organisation," "acceptable," and others. The clustering vector assigns each data point to one of these six clusters, indicating the cluster to which it belongs.

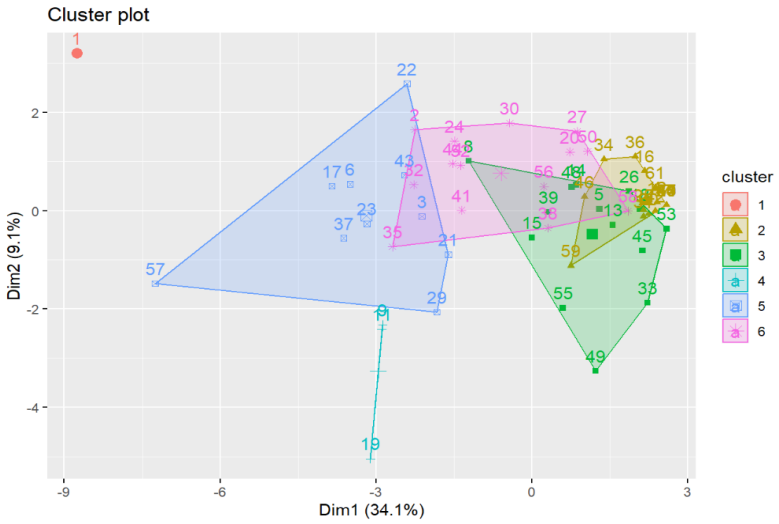

Figure 5: Cluster Plot using package 'NbClust' in R

The within-cluster sum of squares (WSS) for each cluster reflects the variability within that cluster, while the between-cluster sum of squares (BSS) measures the separation between clusters. From the Likert data, the majority of the data points were effectively grouped into clusters 1, 4, and 6, which exhibit relatively low within-cluster variation. Cluster 1 stands out as having the highest WSS, suggesting some variation within it. Cluster 3 contains only one data point, making its WSS zero. The other clusters, 2 and 5, represent smaller subsets of the data with fewer data points.

Overall, the K-means clustering analysis resulted in six distinct clusters, with clusters 1, 4, and 6 being more prominent in terms of data points and relatively lower within-cluster variation. This clustering helped us identify patterns and relationships within the Likert data, enabling further exploration. The optimal cluster for the data is chosen as 6. A Rmarkdown file was created with the K-Means for more reference is attached to the appendix.

2. Multiple Correspondence Analysis

MCA is performed with the questions on knowledge about breast cancer . To avoid overlapping, data is broken down into small groups and then MCA is applied to each group.The plot below shows the dimension plot of the variables. The results of 1-way ANOVA tests between various variables and categorical variables. The output includes R-squared values (R2), p-values, and estimates for the link between the variables and the categories of the categorical variables.

```
Link between the variable and the categorical variable (1-way anova)
                ======================================
                                    R2      p.value
                 braIrrit     0.4772607 6.765028e-09
                 phy_Activity 0.4766739 6.988746e-09
                 painlessLump 0.4641048 1.390963e-08
```

wounds	0.4538064	2.415743e-08
ovrwght	0.4120135	2.049528e-07
preg	0.3989411	3.877883e-07
bfeeding	0.3143450	1.766298e-05
lumps	0.2538127	2.053939e-04
mammo	0.1985808	1.629045e-03
norisk	0.1420246	1.176982e-02

Link between variable and the categories of the categorical variables

	Estimate	p.value
phy_Activity=phy_Activity_	0.64766430	8.845478e-10
painlessLump=painlessLump_	1.26809613	1.590987e-09
wounds=wounds_	0.60884057	1.067910e-08
ovrwght=ovrwght_	0.42303750	5.290234e-08
preg=preg_	0.40622916	8.040192e-08
bfeeding=bfeeding_	0.39639998	5.467487e-06
braIrrit=braIrrit_Yes	0.49961452	1.655988e-04
braIrrit=braIrrit_	0.09393926	2.720845e-04
mammo=mammo_	0.86149751	5.236110e-04
lumps=lumps_	0.35713864	1.469710e-03
norisk=norisk_	0.31629033	2.912796e-03
lumps=lumps_No	-0.39600438	7.099575e-04
bfeeding=bfeeding_Yes	-0.28564653	2.561597e-04
wounds=wounds_No	-0.11290093	2.757441e-05
preg=preg_Yes	-0.29308196	1.997048e-06
painlessLump=painlessLump_painlessLump	-0.69237877	1.804867e-06
phy_Activity=phy_Activity_Yes	-0.40302655	5.265871e-08
ovrwght=ovrwght_Yes	-0.45878860	3.821990e-08
braIrrit=braIrrit_No	-0.59355378	3.878502e-09

Table 2: MCA Dimensionality Summary

In Dimension 1 (Dim 1), we delve into the intricate relationships between our variables and the categorical variable through a one-way analysis of variance (ANOVA). This dimension uncovers the degree to which each variable's variance is influenced by the categorical variable. Notably, 'Pressure' and 'Auxiliary' emerge as standout contributors, elucidating approximately 56.25% and 55.52% of their respective variances, accompanied by remarkably low p-values, signifying exceptionally significant associations. Additionally, several other variables, including 'Practice,' 'Positions,' 'Mammo,' 'Phy_Activity,' 'Bfeeding,' 'KnowBSE,' 'Norisk,' and 'Step2move,' reveal noteworthy links to the categorical variable. These connections exhibit varying degrees of explained variance and statistical significance.

This description provides a more detailed overview of the relationships observed in Dimension 1, Dimension 2,3 and 4 were also studied in detail to offer a clearer perspective on the

significance and implications of these associations. From MCA, we can see how different individuals or variables relate to each other in terms of categorical data. It helps you see patterns, clusters, or associations among them, making it easier to understand the relationships within the data

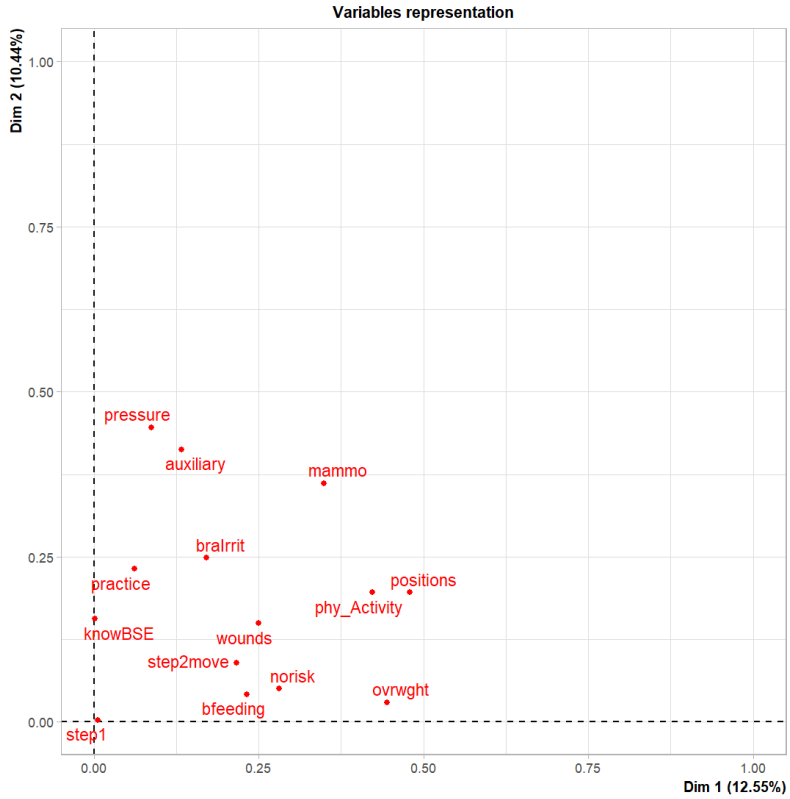

Fig 6: Variables Representation Plot after Imputation

The presented plot provides a comprehensive overview of individuals' data, demonstrating the impact of the missing data treatment using the MissMDA package.

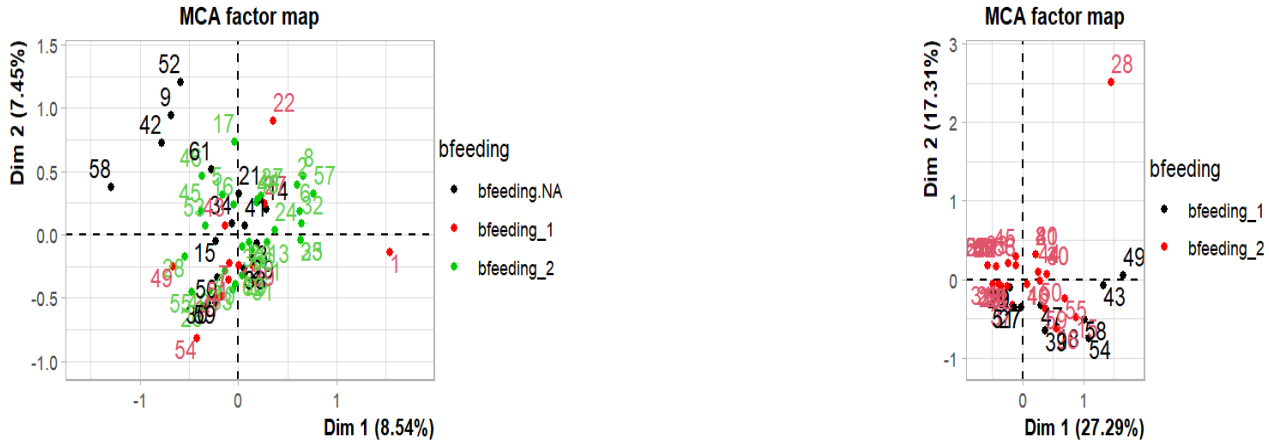

Fig 7: Comparison plot of Before and After Imputation of MCA

Additionally, it extends its insights by incorporating visualizations of various dimensions related to the "bfeeding" variable. We can compare the plot before and after imputation of missing values i.e Do not know. From the graphs, we can say that the cumulative proportion of variance increases from 16.58% to 44.6% From which we can conclude the contribution of the variable to the dataset. Likewise, we can do this for all the variables using the option Habillage from the MissMDA package.

Additional Work

I helped in the development of the decision tree based algorithm for the prediction of a prognosis based on symptoms selected. I also analyzed a study done by Bachelors' students as part of their SCORE project under the supervision of Sweekrity Kanodia - review of educational material - the breasties board. A pre-, post survey was done to analyze the effect of breasties boards in increasing awareness about breast cancer and sentiment towards the board.

Breasties Board

The study was undertaken by the FDV Bachelor's students for enhancing awareness on Breast Cancer. It presents a comprehensive examination of the efficacy of an educational board called the Breasties Board designed specifically to educate individuals about breast cancer and promote breast self-examination (BSE). Pre and post survey was conducted by these students to evaluate its impact on knowledge acquisition, and attitudes towards breast self-examination, among participants from LPI via a questionnaire in google forms. The surveys include questions on knowledge of Breast cancer, self-check and about the use of the board. Pre - survey was conducted with interested participants and then they were exposed to the board. A post survey was conducted right after the participants played with the board. Follow-up survey was also conducted 2 weeks after the initial post-survey to evaluate the retention in knowledge and attitude towards the board.

The three questionnaires are amalgamated for conducting a comparative analysis using the designated questions. The reference of the excel is given in the appendix. The three questionnaires are combined and calculated using COUNTIF function and each question is rated for the categories of initial, intermediate and final survey.

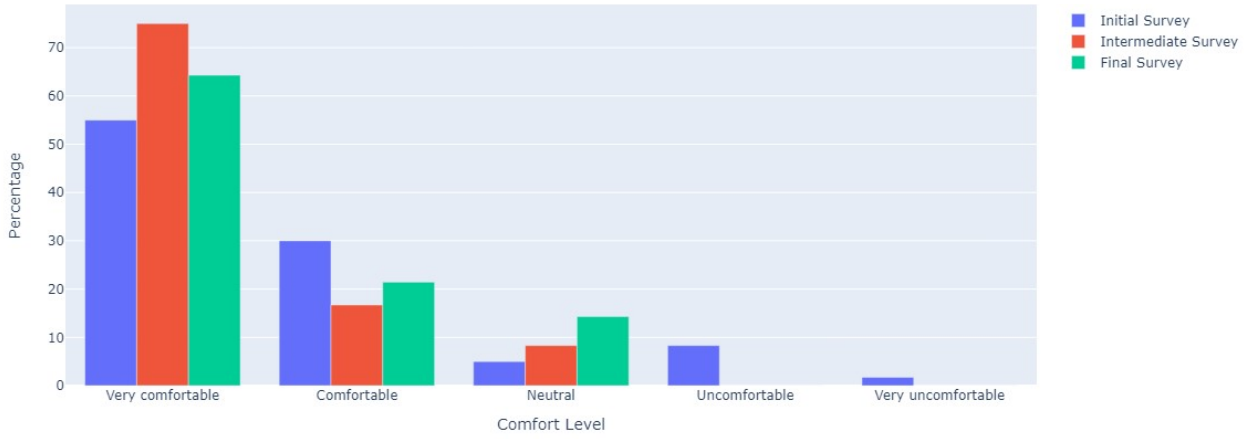

Figure 8: Comfort level from the surveys in discussing about self-check

Development of Symptoms Algorithm

A decision tree algorithm [10] was developed to classify a list of breast symptoms into categories: "Healthy" indicating good health, "Need to follow-up" suggesting rechecks in following months, and "Visit a doctor" indicating symptoms needing immediate medical help. This algorithm aims to assist individuals in gauging the urgency of seeking medical attention based on their symptoms.

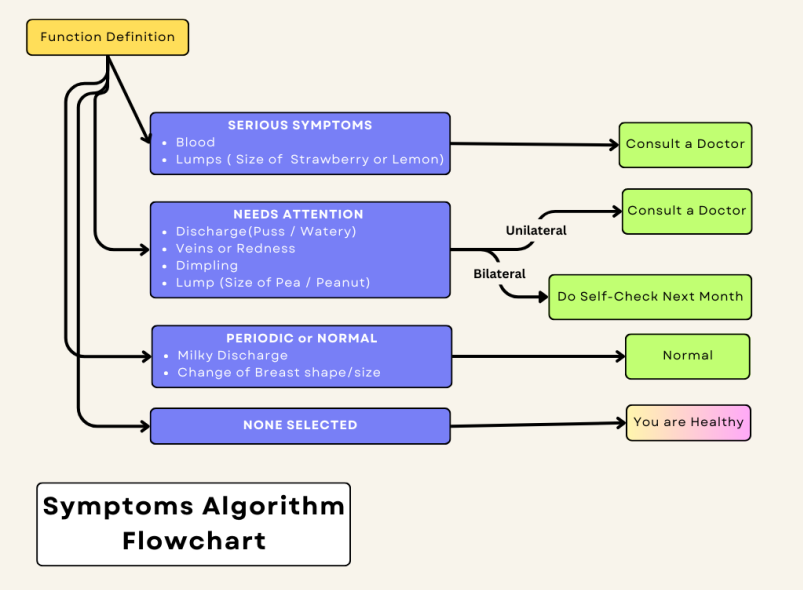

Figure 9: Algorithm Flowchart for breast_Symptoms

The algorithm outlined in this demonstration effectively processes symptoms and provides a classification based on the given input. It defines a function named "classify_breast_disease_symptoms," which takes two essential parameters: "symptoms," a list containing symptom codes, and "breast_side," a string specifying whether the symptoms relate to the "Unilateral" or "Bilateral" side of the breast. The symptom classification process involves

categorizing them into four conditions: "Serious," "Needs Attention," "Could be Periodic or Normal," and "None Selected." If a symptom falls under the "Needs Attention" category, the algorithm further examines whether it is "Unilateral," prompting users to "Consult a Doctor," or "Bilateral," suggesting a "Self-Check Next Month." For symptoms classified as "Serious," it is crucial to visit a doctor urgently. Once the algorithm evaluates the symptoms and applies these classification rules, it generates a message based on the classification. For symptoms classified under Periodic or Normal, These symptoms can occur due to changes of your body. If the user selects "None" or "Do not wish to answer," the algorithm provides a reassuring message, stating "Normal healthy breasts."

Discussions & Conclusions

Summary

BHA Study:

The significance of the application BreMo lies in its unique features and the tracking of symptoms, setting it apart from existing apps in the market. Notably, BreMo's features and symptoms tracking screen have been found to offer distinct advantages when compared with other applications.

However, it's essential to acknowledge certain limitations within this study. One limitation pertains to the availability of data. The study was conducted with a minimum amount of data, which may have influenced the depth and scope of the analysis. Additionally, the study relied on a questionnaire with a substantial number of open-ended questions, which may have posed challenges in terms of data collection and analysis.

Seintinelles Study:

The Seintinelles study embarked on a comprehensive data analysis journey, with the primary objective of gaining insights for betterment of the questionnaire and application. The initial approach involved dimensionality reduction using Principal Component Analysis (PCA). However, it became evident that PCA was suitable only for numeric data.

To overcome this limitation, analysis was switched to Multiple Correspondence Analysis (MCA), a technique better suited for categorical data. Subsequently, the focus shifted to interpreting the clusters within the data. K-Means clustering was employed to identify optimal clusters and facilitate anomaly detection, adding depth to the analysis.

In a nutshell, the Seintinelles study was like a deep dive into data using advanced techniques such as MCA, PCA, and K-Means clustering. These methods helped them uncover valuable insights from the data. However, it's important to recognize a limitation they faced—time

constraints. Due to limited time, they couldn't explore and interpret the data as thoroughly as they would have liked. The sentinelles study has completed the initial stage of evolving the right techniques of analysis and will continue for the further extensive analysis.
For more reference all the relevant documents related to this study are attached to my GitHub repository LINK.

Critical Lessons and Achievements from My Internship

My internship experience was a rich source of learning and personal growth, yielding several key takeaways:

1. Proficiency in the R programming language.
2. The importance of thorough data examination prior to analysis.
3. The ability to construct algorithms using a structured approach, incorporating sequencing, selection, and iteration—a skill honed during the development of the symptoms tracking algorithm.
4. Competence in generating well-structured documentation using R markdown files and LaTeX documents.
5. A comprehensive understanding of ethical considerations in the context of Ph.D. research.
6. Versatility in employing diverse data preparation methods.

Additionally, my contribution to the publication of the BHA study on IAFOR, combined with my active role in preparing the forthcoming publication for the Seintinelles study, has significantly elevated my competence and comprehension within the realm of academic article composition and publication. For your convenience, I have attached the archival publication link for the BHA study LINK.

Future Directions

Here are some enhanced suggestions for the future of this important awareness project BreMo are:

- Processing with the metadata in the Sentinelles study for having more valuable insights
- Improvement of the algorithm for showing the significance of predicting the symptoms.
- Well structured questionnaire for simplifying the data preprocessing

Acknowledgements

In conclusion, I wish to express my profound gratitude to those who have played pivotal roles in my transformative internship journey. Foremost, I extend my deepest thanks to my mentor, Dr. Jean-Christophe Thalabard, whose unwavering support, invaluable guidance, and wealth of

knowledge as a Biostatistician and expertise in Medical Sciences have left an indelible mark on my professional development. Dr. Thalabard's grounded nature served as a remarkable example of professionalism, and his mentorship has been instrumental in shaping my comprehension of various facets of academia. I am truly privileged to have had the opportunity to learn under his tutelage.

Equally deserving of my heartfelt appreciation is my supervisor, Mrs. Sweekrity Kanodia, whose unwavering dedication and commitment have provided me with invaluable direction for my career. Mrs. Kanodia's extensive knowledge and her patience in reiterating concepts and generously sharing her expertise has been a cornerstone of my learning experience. Her willingness to assist me in resolving complex analytical issues by sitting down and working through errors exemplifies her outstanding mentorship. I am profoundly grateful for the wisdom and perspective she has graciously imparted.

The knowledge, skills, and experiences I have acquired during this internship have truly transformed my outlook, and I attribute a significant portion of this growth to the exceptional mentorship and unwavering support of Dr. Jean-Christophe Thalabard and Mrs. Sweekrity Kanodia. Armed with the invaluable support and experiences garnered during this internship, I eagerly anticipate continuing my academic journey with newfound confidence and boundless enthusiasm.

APPENDIX

References

01. NCDs. (n.d.-b). Breast Cancer Awareness Month 2022. World Health Organization - Regional Office for the Eastern Mediterranean. https://www.emro.who.int/fr/noncommunicable-diseases/campaigns/breast-cancer-awareness-month-2022.html
02. De Jong, M. J., Van Der Meulen-De Jong, A. E., Romberg-Camps, M., Becx, M., Maljaars, J., Cilissen, M., Van Bodegraven, A. A., Mahmmod, N., Markus, T., Hameeteman, W., Dijkstra, G., Masclee, A., Boonen, A., Winkens, B., Van Tubergen, A., Jonkers, D., & Pierik, M. J. (2017b). Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. The Lancet, 390(10098), 959–968. https://doi.org/10.1016/s0140-6736(17)31327-2
03. Frank, E., & Hall, M. (2001b). A simple approach to ordinal classification. In Lecture Notes in Computer Science (pp. 145–156). https://doi.org/10.1007/3-540-44795-4_13

04. Nahm, F. S. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean Journal of Anesthesiology, 69(1), 8. https://doi.org/10.4097/kjae.2016.69.1.8
05. Introduction to linear mixed models. (n.d.-b). https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/
06. Pramoditha, R. (2023, August 24). 11 Dimensionality reduction techniques you should know in 2021. Medium. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b
07. Plotting PCA (Principal Component Analysis). (n.d.). https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html
08. Keita, Z. (2023). Principal component analysis in R tutorial. https://www.datacamp.com/tutorial/pca-analysis-r
09. Dovetail Editorial Team. (2023). Thematic Analysis: A Step-by-Step Guide. dovetail.com. https://dovetail.com/research/thematic-analysis/
10. Saini, A. (2023). Decision Tree Algorithm – A Complete Guide. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#:~:text=A%20decision%20tree%20algorithm%20is,each%20node%20of%20the%20tree.

Additional Supporting Materials

BHA Study:

The markdown file is formulated in R for documenting the ordinal regression is attached here as a LINK.

Seintinelles:

A. The thematic analysis is attempted for the comment section of the survey. The link of the analysis is as follows Sentinelles data analysis .
B. The K-Means Clustering is documented as a R markdown file LINK.
C. The MissMDA infused MCA file is attached here LINK.

Breasties Board:

The below picture will depict the count of the results from the questions via “COUNTIF” function. Likewise, this was applied to all the needed questions.

A. Google sheets - responses
B. Google Colab - Analysis

B4 =COUNTIF('Copy #1'!E2:E61;"Never")/COUNTA('Copy #1'!E2:E61)*100

	A	B	C	D
1		Before	After #2	After #3
2				
3	Do you self-check your breasts?			
4	Never	45,0	33,3	57,1
5	Occasionally	48,3	66,7	28,6
6	Monthly	6,7	0	14,3

Files and Notebooks

You can find all the notebooks and visualizations in my GitHub Repo. Please feel free to explore them to gain a deeper understanding of the data and the project's insights. Your visit to the repository will provide you with a richer perspective on the data and its visual representation.