Bachelor's Thesis Computing Science

Radboud University Nijmegen

Data analysis on motivation for lifestyle changes

Author:
Imke Huisink
s1037284

First supervisor/assessor:
dr. I. Wilmont

Second assessor:
dr. IMN Wortel

January 12, 2024

Abstract

Unhealthier lifestyles are getting more common which is causing a lot of different problems. The foundation 'Je Leefstijl Als Medicijn' offers information and other services on how to change your lifestyle and get healthier. People can sign up for the foundation via a registration form. This data is then stored in Hubspot. In this research, we looked into why people want to get started with lifestyle changes via a data analysis on the data from Hubspot. Our methods included Hubspot's built-in tools for analytics and basic descriptive statistics as well as a data analysis in Python. We looked at certain statistics, such as the mean, median, mode and standard deviation. We created graphs and applied the K-means and DBSCAN algorithms to find clusterings. We analysed the textual data by performing a word count. The analysis showed that getting healthier is the biggest motivation for getting started with lifestyle changes. This included reducing the effects of diabetes and the use of medication. Most of the members were overweight, but they did not explicitly mention losing weight as a motivation for signing up. COVID-19 was an unexpected motivator as well.

Contents

Chapter 1

Introduction

There is a medical crisis happening in the world and we all know it. News channels such as the BBC keep reporting about how most people are living an unhealthy lifestyle [6], showing the 'obesity problem' [2] and how the obesity rates have been increasing for decades. However, a solution has yet to be found.

Lifestyle has a big impact on our health. It is about diet and physical activity, but also environmental factors such as social influences and education on health [26]. Physical activity is associated with a higher quality of life and obesity with a lower quality of life [20]. Obesity is defined as a condition of abnormal or excessive body fat, to the extent that health may be impaired [29]. This excess weight can lead to different health problems, such as diabetes and heart diseases [21] [35]. Nonetheless, study also shows that these effects can sometimes be reversed [36]. A healthy lifestyle can improve or delay the deterioration of glucose tolerance [32], thus improving the effects of diabetes. In addition, exercise training improves glucose tolerance and insulin sensitivity independently of weight loss [29]. Thus, implementing a healthy lifestyle that involves exercise and a healthy diet can lead to weight loss and improve someone's overall health.

'Je Leefstijl Als Medicijn' (JLAM), translated 'Your Lifestyle As Medicine', is a foundation that helps people change their lifestyle and therefore get healthier [4]. They set up support groups for patients with certain diseases such as diabetes, rheumatism and obesity. People can ask questions, share information with each other, sign up to join the weekly weigh-ins and receive the newsletters. Some of the focus points of the foundation are nutrition and diet, exercise and sleep.

The foundation was founded about five years ago. They have gathered a lot of data over these years in three different digital locations. First of all, they use Facebook groups where people send messages and react to each other. Secondly, they have their own database where the measurements of

the weekly weigh-ins are stored. And lastly, they use Hubspot which stores the member data, so that is the data that people fill in when they sign up, what links the users click, any emails that were sent, etc.

The foundation is interested in a data analysis on all this data. They are interested in why people sign up and what their motivations are for changing their lifestyle. Why do people want to get healthier? What are they trying to achieve? Did the members get the results they were looking for after changing their lifestyle? Did they get the help they wanted and/or needed?

1.1 Related work

Study shows that behavioural treatments to manage obesity are effective in the short term, but result in gaining all this weight back and sometimes even more when the program is over [33] [18]. This is mostly about treatments that last for a specific amount of time, after which the participants are left to their own devices again. When checking in with the participants a few years after these treatments, we see that they gained the weight back. The reasons for this weight regain are unknown. It is possible that the behavioural treatments do not actually change the behaviours in the long term [10], which means that the participants slip back into old habits. Study also reports that long-term compliance with these behavioral treatments does result in long-term weight loss [27]. Therefore, these behavioural treatments in itself are not bad, but there should be extra tools that can help stick to the techniques in the long term.

If the goal is to lose weight and get healthier, we should be looking at goal achievement and motivation. Study shows that mastery goals, which deal with developing competence and building new skills, result in more motivational persistence and more perceived control [8]. Study also shows that autonomous motivation or internal motivation, which reflects personal interests and values, is related to goal process and external motivation is not [24]. Goals that involve overall health influence physical activity via internal motivation and weight loss goals have a negative effect on physical activity via external motivation [12]. Furthermore, an increased perceived behavioural control and internal motivation can reinforce behaviour change [34]. People are more likely to present behaviours that they feel like they have control over. Perceived control is seen as an important academic marker in achievement settings, as it affects motivation as well as achievement-striving [30]. Therefore, setting a realistic goal that focuses on learning and personal interests rather than performance should result in a higher internal motivation, which in turn can promote behaviour change.

Internal motivation is a very important factor in goal achievement. However, external influences can increase this internal motivation. External fac-

tors such as extra information on diet and weight loss or a financial incentive can lead to better results than without these extra resources [14]. Thus, providing the information and tools that people need to increase their internal motivation can lead to better results.

One of these external influences could be accountability to others instead of yourself. Study shows that higher accountability leads to healthier dietary choices and better psychological coping with weight management challenges [9]. Social support can positively influence one's motivation to change their lifestyle as they have someone to relate to or to commit to [31]. So identification with and commitment to others can lead to more motivation and better results.

Lastly, lifestyle is defined as a style or way of living [7]. Therefore, changing your lifestyle is meant as a long-term solution. Study shows that lifestyle interventions can help lose weight and reduce blood pressure [28]. Implementing public health programs has been shown to be very important [17]. Study also shows that identification with and acceptance of a new lifestyle leads to greater commitment [31]. Another study shows that group lifestyle intervention among people with Type 2 Diabetes leads to weight loss and medication reduction compared to standard medical nutrition therapy [11]. Furthermore, exercising and eating healthy for one day leads to enjoyment, which has a positive affect on the next day as people are expecting a more enjoyable day [15]. This is also known as the upward spiral theory of lifestyle changes. When a new healthy behaviour is a positive experience it creates non-conscious motives for that activity, which grow stronger over time and thus stimulate to keep behaving that way [16]. So small lifestyle changes can lead to better overall health. Study also shows that maintaining a healthy lifestyle can contribute to successful aging [25].

1.2 Research question and hypothesis

To help the foundation in finding an answer to these questions, our research question for this thesis is:

What motivates people to get started with lifestyle changes?

Some of the subquestions we will be answering are:

- What kind of people sign up for the foundation?
  Before we can say something about motivation, we need to know what personality traits contribute to this. We will be looking for the main characteristics of the members.
- What lifestyles and lifestyle patterns do people have at the time of signing up for the foundation?

We will be looking for patterns in the lives of the members to see if certain aspects of lifestyle play a part in motivation.

- What factors, such as existing medical problems or the risk of these problems, contribute to motivation?
- What do people expect from the foundation?
- What results do people hope to get through the foundation?
- How did people come in contact with the foundation?

There are many reasons why people would want to change their lifestyle. Based on the related work about goal achievement, medical problems and internal and external motivation as well as our knowledge about the foundation and our own experience with lifestyle changes, we expect these to be the main motivators among the members of the foundation:

- To lose weight
  We expect that the members of the foundation are overweight. As mentioned before, obesity is still a big problem and a healthy lifestyle can be a solution.
- To reduce overall health risks
  We expect that people want to reduce the risk of health issues such as cardiovascular diseases or diabetes. Preventing these problems is always better than experiencing the symptoms and having to undergo treatment.
- To reduce the effects of certain diseases
  We expect that the members are looking for ways to reduce or reverse the effects of certain diseases and health problems, such as heart diseases, high blood pressure or diabetes. We especially expect diabetes to be a recurring motivator, as it is one of the foundation's main focuses.
- To slow down the effects of ageing
  We expect that people want to age successfully without any of the big complications that can come with it. Therefore, we also expect the people who sign up for the foundation to be a bit older. Younger people tend to look for different resources such as a gym.
- To stick to their New Year's resolutions
  Most people have a New Year's resolution that is related to health, such as losing weight. We expect the foundation to notice this in the number of registrations in December and January.

However, these can all be related to one:

To improve their overall health

By implementing a healthy lifestyle, people's health will be less of an obstacle in daily life and thus the overall quality of life will improve. People will be able to participate in activities they enjoy while the risks and effects of health issues are decreased.

1.3 Outline

We start in Chapter 2 where we will discuss the preliminary knowledge needed to understand this paper. In Chapter 3 we will discuss the methods used in this paper. In Chapter 4 we will introduce the results and Chapter 5 will discuss these results and any limitations in this research. Lastly, Chapter 6 will present the conclusions.

Chapter 2

Preliminaries

In this Chapter, we elaborate on some topics that are required to understand our research. We start by explaining what Hubspot is and how it keeps track of some of the data. Secondly, we explain the BMI as it is an important measure in this research. Lastly, correlation is described.

2.1 Hubspot

Hubspot offers a Customer Relationship Management (CRM) platform where companies can gain insight into their customer relationships such as the behaviour of their customers, their sales activities and their marketing. It is a database containing all the relationships and processes of a company [5]. CRM is building a strategy to develop attachment to customers through studying customer needs and habits [22].

As Hubspot keeps track of a lot of different data from these customers, it also offers its clients the tools to create reports to analyse this data. Reports are graphs and tables that show a certain part of the data. These reports can then be collected in a dashboard to show related reports in one overview. Hubspot offers several templates for dashboards and reports that most companies are interested in, but clients can also create their own.

An example of this tracking of the data in Hubspot is their traffic analytics tool. This can be used to view your websites' traffic data such as sources, which is where on the internet the visitors came from, country, device type, etc. [1]. This is measured in sessions, where one session is defined as a series of analytics activities taken by a visitor on the website [3]. It expires after 30 minutes of inactivity. Users can select to show data for a certain time period, such as the last month, last year or all data.

2.2 BMI

The Body Mass Index (BMI) is a measure to classify underweight, overweight and obesity in adults. It is a simple index defined as the weight in kilograms divided by the square of the height in meters [29]. The World Health Organization (WHO) recommends a BMI between 18.5 and 25 kg/m^2 for optimal health.

2.3 Correlation

Correlation is the measure of association between two variables. For example, when the temperature increases, people will probably buy more ice cream. This correlation is measured with the correlation coefficient (r), which ranges between 1 and -1. The correlation gets weaker the closer r is to zero. When r is positive, it indicates a positive correlation, which means that when one variable increases the other one also increases. When r is negative, it means that there is a negative correlation, so when one variable increases the other one decreases.

2.4 K-means and DBSCAN clustering

Clustering, also known as cluster analysis, groups data objects without consulting class labels so it can be used to generate class labels for a group of data [19]. Groups with similar traits are divided and assigned into clusters. There are many different clustering algorithms, where K-means and DBSCAN are the most popular.

K-means is a centroid-based clustering algorithm, which means that each data point in the cluster is closer to the centre of that cluster than to the centres of the other clusters. Outliers are also included in the clusters. This algorithm can identify spherical clusters, which means that the data is distributed with equal space between each other in all directions. This algorithm is widely used and easy to understand.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm, which means that clusters are identified by areas where points are concentrated and separated by empty areas. Outliers are not assigned to a cluster in this algorithm. It can identify clusters of arbitrary shapes.

The clusterings made by these algorithms can be evaluated in many different ways. One way is by using the silhouette scores. The silhouette score tells us how close a data point is to its own cluster compared to the other clusters. This value ranges from -1 to 1, where 0.7 or higher means a strong clustering and 0.2 or lower a weak clustering.

Chapter 3

Methods

In this Chapter, we describe the methods we used in this research. Firstly, we discuss the preparatory work we did via Hubspot. Then we talk about how we extracted and preprocessed the data set we used for our Python analysis. Lastly, we talk about how we performed this data analysis.

3.1 Preparatory work

As Hubspot offers a CRM, we were interested in what information we could find through that platform to get a better understanding of Hubspot and the data. We were not looking for answers to our research question yet. We simply wanted to see what we and the foundation could gain from it. So we started by creating multiple reports via Hubspot.

In the templates, there was a report of the most viewed blog posts. The foundation might want to know this, so they can focus their services on these topics. It was also interesting to us, as the blog posts show what topics the members want to learn more about. We created this report for the past year and overall.

The second report we decided to take a further look at was a graph that shows which social platforms are most popular among the members. This might be interesting to the foundation, so they know what platform reaches most people.

Lastly, Hubspot offers a premade dashboard for web analytics with reports such as the number of sessions per country, information about session engagement and what devices people use to visit the website. We thought this could also be of interest to the foundation, so we also created these.

3.2 Data

The data in Hubspot that we were interested in for our research question are the contacts and all attributes that are known about them. The contacts

are all the contacts of the foundation. This includes all people who joined the support groups for diabetes, rheumatism, obesity, etc. as well as the people who signed up for the newsletter. There might be more data stored as well, such as the companies that the foundation works with, but as we were only interested in the members of the foundation, we excluded that from our data set.

When we view the data from the contacts in Hubspot, the contacts are represented in the rows and the attributes in the columns. The data can easily be exported to an Excel file when logged into Hubspot. By default, Hubspot will only export the columns that are in the current view. Choose 'All properties on record' for 'Properties included in export' to export all data.

3.2.1 Data set

As Hubspot stores all information about the contacts and their sales, there were a lot of columns that were not interesting to our analysis. Some examples are *Abandoned Cart Products*, *Average Order Value* and *Last Order Status*. Therefore, we first needed to select the columns that we were interested in. By looking at all the titles and entries of the data columns, we selected the columns that contained interesting information for this data analysis. A description of the data can be found in Table 3.1 of the appendix.

The data only contained a date of birth, but we were also interested in the age. So we wrote a function that converts the date of birth into the age and added this as a new column to our data set.

3.2.2 Text preprocessing

The registration form for the foundation includes questions that require a textual answer. We wanted to convert this textual data to numerical data so that we could easily analyse this data as well. This required us to do some preprocessing.

We converted all letters to lowercase and removed punctuation, enters and digits. Then we removed any stopwords using the stopword list that the Natural Language Toolkit (NLTK) offers. We also removed prefixes and suffixes by using stemming, which is also offered by the NLTK. Lastly, we removed any extra spaces.

3.2.3 Data description

Exporting the columns we were interested in from Hubspot and preprocessing this data gave us the data set we needed for our analysis. Table 3.1 shows a description of each column in this set.

Column	Description
Record ID	A unique ID for each contact in Hubspot
Gewicht	Body weight in kilogram
BMI	Body mass index
Geboortedatum	Date of birth
Age	Age in years
Uw vraag of opmerking	Any questions or things the contact wants to share
Vetpercentage	Body fat percentage
Gender	Gender which is defined as 'Male' or 'Female'
Waarover wilt u informatie ontvangen	Subjects the contact wants to receive information about
Create date	Creation date of the contact
Ervaring diensten JLAM	Experience with the services the foundation offers
Ik ontvang graag de nieuwsbrief van Stichting Je Leefstijl Als Medicijn	The contact would like to receive the newsletter of the foundation
Wat voor steun JLAM	The kind of support the contact expects from the foundation
Wat werkt voor jou	What kind of things work for the contact, e.g. diet, information, results, conversations, etc
Welke behoeften	Needs the contact might have
Lengte in cm	Length in centimeters
lengtekwadraat	Length squared
ZWEMMER	What other services the contact is using, e.g. joined a Facebook group, joined the program to measure their weight every Saturday, etc.
Last Viewed Webinar Name	The last Zoom webinar the contact viewed

Table 3.1: Summary of the data after preprocessing

3.3 Data analysis

Now that the data was ready to be used, we started our data analysis. We decided to use Python as it is a perfect fit for data analytics, due to its readability and extensive set of libraries among other things [13].

3.3.1 Basic information

We started our data analysis by getting to know some basic information. Table 3.2 of the appendix shows the functions we used to get this information.

Function	Description
`df.info()`	Prints a summary of the dataframe
`df.describe()`	Generates descriptive statistics
`df.isnull().sum()`	Count the amount of missing values per column
`df.value_counts()`	Count the frequency of distinct rows in a specific column

Table 3.2: Functions used to find basic information

3.3.2 Numerical data

Graphs

We were interested in what kind of people signed up for the foundation, so we wanted to take a look at the weight, height and age of the members. We were also interested in when the members of the foundation signed up. To analyse these numerical data, we created graphs.

We started with a scatter plot where we compared the weight to the height. This meant that we had to drop all entries that had a null value for any of these columns. Next, we grouped the data by gender so that the graph shows the gender of each plot point. Lastly, we added the upper and lower limits of the healthy BMI range and we added a line to indicate an obese BMI.

We also created a scatter plot to compare the weight to the age. We again dropped all entries with a null value and grouped the data by gender.

Lastly, we created two bar charts based on the creation date of the contact in Hubspot to see if we could find a pattern in the dates that people signed up for the foundation. One of the bar charts shows the count per year and the other one per month.

Clustering

For both scatter plots, we were interested to see if any groups could be identified. To do this, we used clustering. We decided to use the K-means and DBSCAN algorithms to see if we could identify groups.

Python provides a K-means library that we could use to perform the algorithm. First, we split the data into test and train data by using `train_test_split()`.

Then we normalized the data to make sure all features have the same scale so they are weighted the same in the algorithm. We also needed to choose the number of clusters so that we were not overfitting or underfitting the data with clusters. To choose the number of clusters we used the Silhouette method, which tests and calculates the results of each number of clusters that can be used and shows this in a graph. When the line starts to flatten it means that the algorithm is not improving anymore, so that number is the best number of clusters to choose. For both of the clusterings, the line did not start to flatten. We still chose six for the clustering on the weight and height and seven for the clustering on the weight and age to at least see what a clustering would look like for our data. Now we were able to perform the algorithm and show the results in a graph.

Python provides a DBSCAN library that provides all the functions we need to perform the DBSCAN algorithm. We needed to determine the minimum amount of points to form a cluster. There are multiple ways to do this. We chose to use a simple calculation for this, namely the minimum amount of points is twice the number of dimensions. As we were using two dimensions, we used four as the minimum amount of points. Then we needed to calculate the best value for the Epsilon variable in the algorithm, for which we used the Elbow method. We started by creating the nearest neighbours with our data set using the Nearest Neighbor library. Then we calculated the distances between these neighbors and plotted them in a graph. We chose the value of Epsilon where the graph shows the maximum curvature. For the clustering on the weight and height, the Epsilon was six and for the clustering on the weight and age, the Epsilon was thirteen. Then we performed the DBSCAN and showed the results in a graph.

To evaluate these clusterings, we used the silhouette scores. The silhouette score tells us how close a data point is to its own cluster compared to the other clusters. For each clustering, we created a graph that shows the silhouette score per number of clusters for the K-means clusters or Epsilon for the DBSCAN clusters.

Measures of central tendency and correlation

We calculated the mean, median and mode of the numerical data by using the predefined functions that are in the pandas library. We set the parameters to skip null values and include only numerical values.

The minimum and maximum of each column were also calculated using the predefined functions from the pandas library. The range was calculated using this minimum and maximum.

We calculated the correlation between weight, height, fat percentage, BMI and age by using the predefined function from the pandas library. We

set the method to ‘pearson’.

Multiple choice questions

The columns ‘Waarover wilt u informatie ontvangen’, ‘Ik ontvang graag de nieuwsbrief van Stichting Je Leefstijl Als Medicijn’, ‘ZWEMMER’ and ‘Last Viewed Webinar Name’ were answers to multiple choice questions. The answers were predefined and thus we could simply count the amount of distinct values to see how people answered these questions.

3.3.3 Textual data

The columns ‘Uw vraag of opmerking’, ‘Ervaring diensten JLAM’, ‘Wat voor steun JLAM’, ‘Wat werkt voor jou’ and ‘Welke behoeften’ were answers to open questions and thus required the previously mentioned preprocessing of textual data.

For each of these columns, we did a word count to see which words were used the most. To visualize this, we created a word cloud using the WordCloud library.

Chapter 4

Results

In this Chapter, we will show the results of our research. Firstly, we show the results we got from the preparatory work for which we used the Hubspot tools. Then we show the results of our own data analysis in Python.

4.1 Preparatory work

4.1.1 Blog posts

The most viewed blog posts of all time and of last year can be found in Table 4.1 and 4.2 of the Appendix. It shows that overall people were very interested in COVID-19 and the positive effects of lifestyle on the immune system. However, as COVID-19 was no longer a big health scare for us, we can see that this blog post was less interesting last year. People were mostly interested in the review of 'VET belangrijk', which is a book about nutrition, fat burning and secretly fattening foods. Furthermore, we see that people are also interested in the blog post about sugar as a bad guy instead of salt and a research about breast cancer and fasting.

The full tables are shown in Tables A.1 and A.2 of the Appendix. Overall, we see that people are mostly interested in articles about diet rather than specific illnesses.

Blog post	**Views**
Covid 19 en de positieve effecten van leefstijl op je weerstand	24860
Niet zout, maar suiker is de boosdoener - Internist Yvo Sijpkens	8590
Recensie 'VET Belangrijk' Mariëtte Boon & Liesbeth van Rossum	4607
Totaal	56236

Table 4.1: Top three blog posts of all time

Blog post	Views
Recensie 'VET Belangrijk' Mariëtte Boon & Liesbeth van Rossum	843
Borstkanker en vasten. Hoopgevende resultaten uit studie door LUMC	500
Niet zout, maar suiker is de boosdoener - Internist Yvo Sijpkens	314
Totaal	2374

Table 4.2: Top three blog posts between 11-01-2022 until 10-31-2023

4.1.2 Social interactions by network

The amounts of interactions per network can be seen in Figure A.1 of the Appendix. We see that the foundation's Facebook page is used the most, with their LinkedIn company page as a close second. Twitter is used the least.

4.1.3 Sessions per country

Table A.3 of the Appendix shows the amount of sessions and the percentage of sessions from new visitors per country for the top ten countries. Table 4.3 shows the top three of these countries. We found that most visitors of the website are from the Netherlands. People from Belgium are also quite interested.

Country	Sessions	New session in %
Netherlands	968,205	80.12
Belgium	100,357	87.16
United States	5,757	95.80

Table 4.3: Top three sessions per country

However, the most interesting to see is that there are also visitors from non-Dutch-speaking countries while all the foundation's services are in Dutch. This might be because certain words are the same in different languages and thus people can still find their website when they search for these words.

It is also interesting to note that the percentage of new sessions is roughly the same for each country. The United States has the highest percentage in these ten countries, which means that most of these people do not come back after already having visited the website. This makes sense, as the site is probably not usable for them. However, Switzerland and France have a rather low percentage. For some reason, one in four people does come back to the foundation's website.

Generally, we see that people from all around the world are interested in improving their lifestyle. It might be interesting to the foundation to also

start providing their information in English so their services can help more people.

4.1.4 Session engagement rates

Table A.4 of the Appendix shows the different traffic sources and their engagement. The bounce rate is the percentage of visitors that leave immediately without clicking anything else or visiting a second page on the site.

We see that overall the bounce rates are quite high and the average session lengths are no longer than two minutes. Visitors also do not visit many pages in one session. This can be due to a lot of different reasons, such as visitors not finding the information they were looking for. The page views per session for traffic from search engines are quite low and the bounce rates quite high.

The highest bounce rate and lowest average session length and page views per session is paid social. Companies can pay for social campaigns to get more engagement from their visitors and reach more people. For the foundation, we do not know if these campaigns reached more people, but we do know that they did not result in more engagement.

Referrals have one of the lowest bounce rates and one of the highest average session length and page views per session. This makes sense as people deliberately click the link to visit this page and thus will engage with it more.

4.1.5 Device breakdown

In Figure A.2 of the Appendix we see a breakdown of the kinds of devices that the visitors use. Most of the visitors use their mobile phones to view the website. Therefore, it might be a good idea for the foundation to pay extra attention to their website's layout for mobile phones.

4.2 Data analysis

4.2.1 Basic information

The total amount of entries in the data set that we worked with was 15,858. This data set was exported from Hubspot at 10-10-2023. Any members that registered after this date are not taken into account in this analysis.

Missing values

We noticed that there were a lot of null values in the data set. Table 4.4 shows the number of missing values, the number of filled-in values and the corresponding percentage of missing values per column. As we can see, only

Column	Number of missing values	Number of values filled-in	Number of missing values in %
Record ID	0	15,858	0.00
Gewicht	15,110	748	95.28
BMI	15,638	220	98.61
Geboortedatum	15,820	38	99.76
Age	15,820	38	99.76
Uw vraag of opmerking	14,807	1,051	93.37
Vetpercentage	15,394	464	97.07
Gender	14,247	1,611	89.94
Waarover wilt u informatie ontvangen	15,019	839	94.71
Create date	0	15,858	0.00
Ervaring diensten JLAM	15,825	33	99.79
Ik ontvang graag de nieuws-brief van Stichting Je Leefstijl Als Medicijn	15,847	11	99.93
Wat voor steun JLAM	15,825	33	99.79
Wat werkt voor jou	15,825	33	99.79
Welke behoeften	15,829	29	99.82
Lengte in cm	15,470	388	97.55
lengtekwadraat	15,470	388	97.55
ZWEMMER	15,562	296	98.13
Last Viewed Webinar Name	14,096	1762	88.89
Total amount of entries	15,858		

Table 4.4: Missing values in the data set

the record ID and the creation date have no missing values, as these are automatically registered for each new entry in Hubspot.

The other columns all have a very high percentage of missing values. No column has a percentage lower than 88%. This means that we have very little data to analyse compared the the data we could have had and any results are still only for a small subset of the total entries.

4.2.2 Numerical data

Weight compared to height

Figure 4.1 shows a graph that plots the weight against the height. The data points are grouped per gender. The green lines indicate the upper and lower limits of the healthy BMI range. The red line indicates an obese BMI.

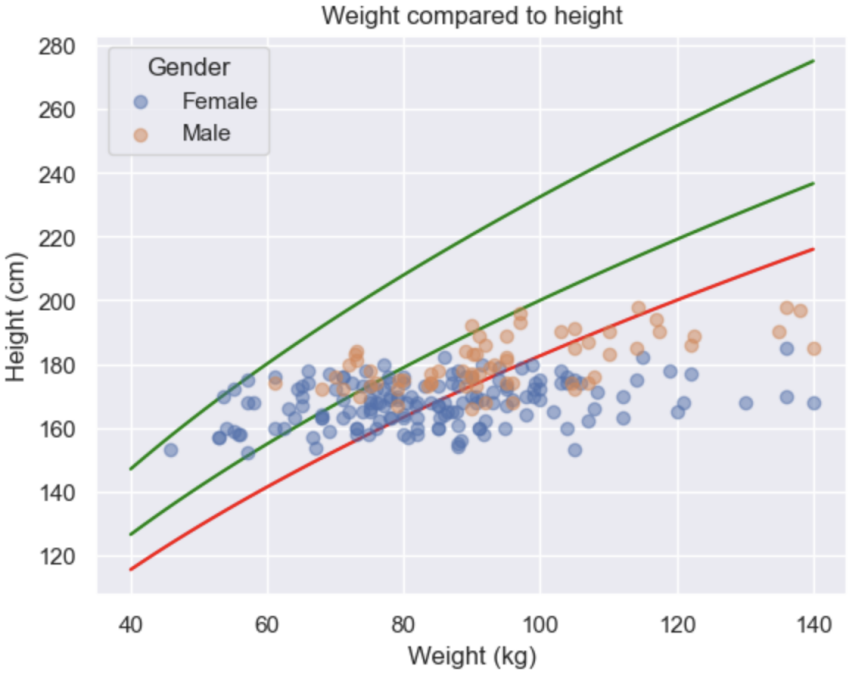
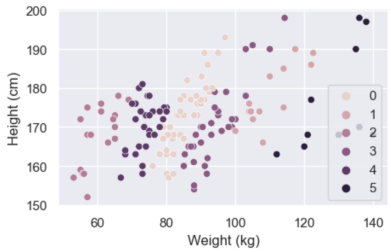
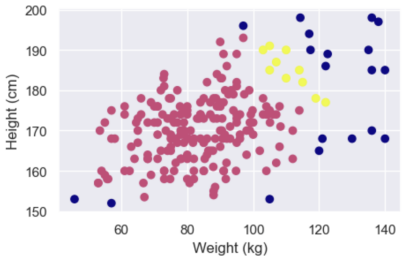


Figure 4.1: Weight compared to height

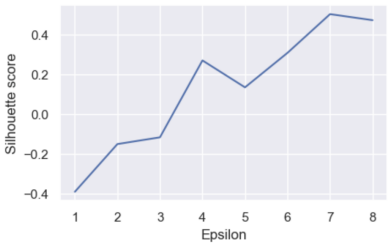


(a) Clustering after applying K-means on the weight and height

(b) Silhouette score per number of clusters for K-Means clustering on the weight and height



(c) Clustering after applying DBSCAN on the weight and height



(d) Silhouette score per epsilon for DBSCAN clustering on the weight and height

Figure 4.2: Clustering on the weight and height

Data points that are in between the green lines have a healthy BMI. Points that are to the right of these lines have a higher BMI and thus are overweight. Data points that are to the right of the red line have a higher BMI than 30, which means that they are obese.

We can see that many of the members are overweight. Some are even obese. The highest male weight in this graph is 140 kg, with a BMI of 35.7. The highest female weight is also 140 kg, with a BMI of 40.9.

Figure 4.2a shows the clustering after applying the K-Means algorithm. We see that the data points are separated from left to right, so it does not look like the clusters are any interesting groups we did not know about.

Figure 4.2b shows the silhouette scores per number of clusters. It is the graph that was created to choose the number of clusters. We see that the silhouette score is above 0.5, which means that there is a reasonable clustering. However, as said before, the score does not seem to flatten so we cannot say with certainty that the clustering is reasonable.

Figure 4.2c shows the clustering after applying the DBSCAN algorithm. We see that there is one very clear cluster in the middle of the graph. The outliers seem to form a cluster as well. This clustering looks more reasonable than the K-means clustering. However, the clusters are still very close to each other. This means that we probably did not find specific groups among the members that we did not yet know about.

Figure 4.2d shows the silhouette score per epsilon. As we see, the silhouette score does not exceed 0.5. Therefore, this clustering is weak.

Weight compared to age

Figure 4.3 shows the weight against the age. We noticed that there were very few members who filled in their weight, date of birth and gender. In total, there were only 28 people who filled this in and could be plotted in the graph.

The data points seem to be scattered around the entire graph. We cannot identify a trend or pattern.

Figure 4.4a shows the clustering after applying the K-Means algorithm. The clustering does not seem to be very strong, as the number of clusters is very high for the number of data points.

Figure 4.4b shows the silhouette scores per number of clusters. It is the graph that was created to choose the number of clusters. We see that the silhouette score decreases rapidly when Epsilon is eleven. When Epsilon is lower than eleven, the silhouette score is between 0.45 and 0.55. This means that the clustering is not strong and probably also not reasonable.

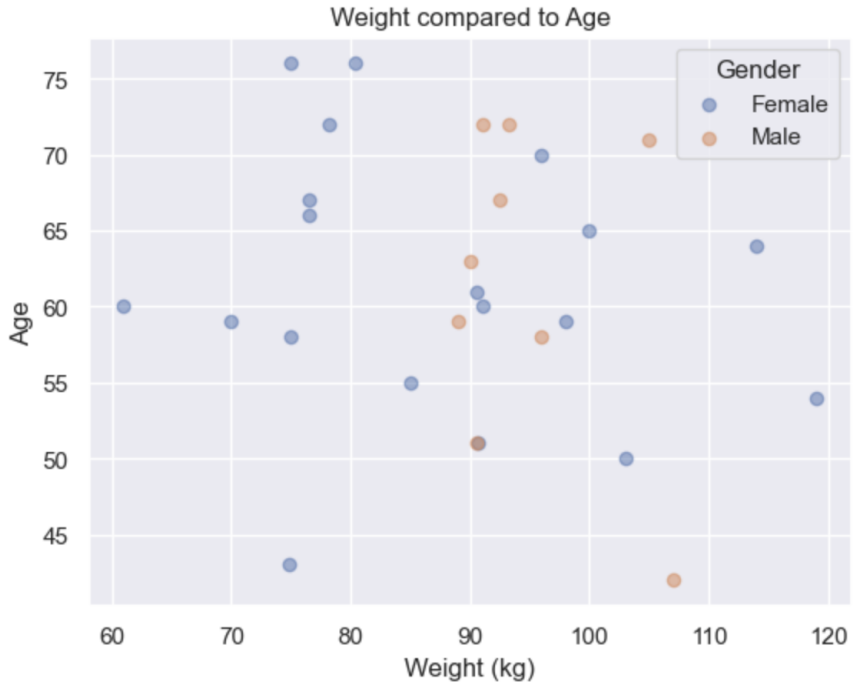
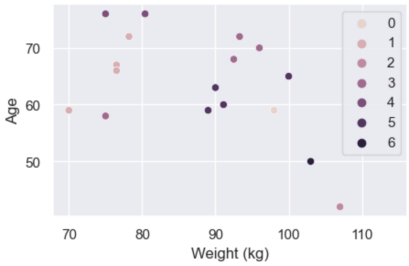
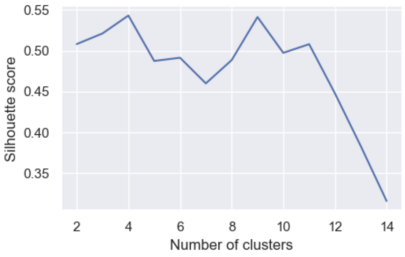


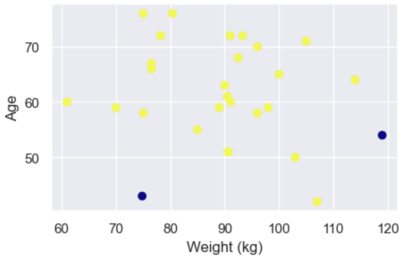
Figure 4.3: Weight compared to age



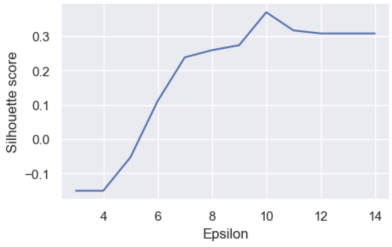
(a) Clustering after applying K-means on the weight and age



(b) Silhouette score per number of clusters for K-Means clustering on the weight and age



(c) Clustering after applying DBSCAN on the weight and age



(d) Silhouette score per epsilon for DBSCAN clustering on the weight and age

Figure 4.4: Clustering on the weight and age

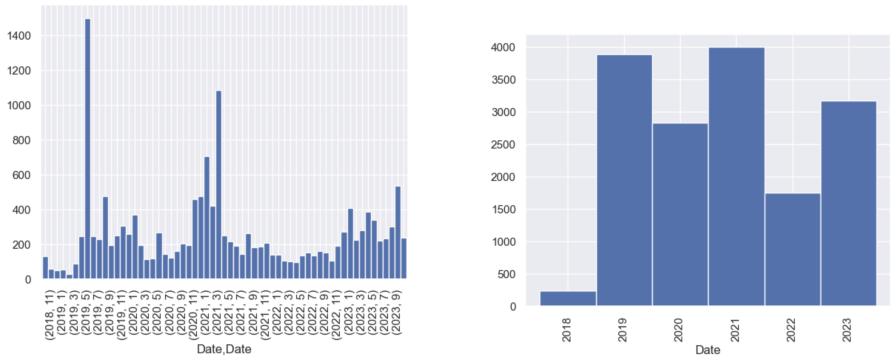
Figure 4.4c shows the clustering after applying the DBSCAN algorithm. We see that there is one very clear cluster in the middle of the graph with two outliers. However, there are too few data points to call this a good clustering.

Figure 4.4d shows the silhouette score per epsilon. As we see, these silhouette scores do not exceed 0.5. Therefore, the clustering is weak.

Creation date

Figure 4.5a shows the amount of newly created contacts in Hubspot per month. We see that May 2019 had the most registrations for the foundation and March 2021 had the second most registrations. There does not seem to be a pattern in which months are popular. Each year, different months resulted in high registrations.

Figure 4.5b shows the amount of newly created contacts in Hubspot per year. The foundation was founded in July 2018, which explains why 2018 is so much lower than the other years. 2021 has the highest amount of new members, with 2019 as a close second. 2022 seems to have been a low year overall, which can also be seen in Figure 4.5a



(a) Amount of new contacts in Hubspot per month

(b) Amount of new contacts in Hubspot per year

Figure 4.5: Creation date graphs

Statistics

The mean, median and mode are measures of central tendency that can help understand the distribution in a data set [23]. The mean is the average of the set, the median is the value that has the middle position when the set is sorted in ascending order and the mode is the most frequently occurring value in the set.

	Weight (kg)	**BMI**	**Fat percentage (%)**	**Height in cm**	**Age**
Mean	86.36	29.64	35.43	172.15	61.18
Median	83.25	29.05	35.00	172.00	60.00
Mode	80.00	21.50 22.92 24.41 25.06 25.59 28.52 29.05 33.20	40.00	168.00	51.00
Minimum	45.80	18.51	7.00	152.00	42.00
Maximum	742.00	49.60	60.00	203.00	80.00
Range	696.20	31.09	53.00	51.00	38.00
Standard deviation	31.36	5.62	8.82	9.49	9.22

Table 4.5: Statistics

Table 4.5 shows the mean, median and mode of the numerical columns in our data set. The BMI has multiple values for the mode, as they all occurred equally often.

The range is calculated by subtracting the minimum value from the maximum value of a data set. The standard deviation shows how distributed the data is compared to the mean. A low standard deviation means that the data is all very close to the mean. A high standard deviation means that the data is more spread out.

We see that on average the members are overweight. Their BMI and fat percentage are higher than healthy ranges. The maximum weight found in the data set is probably a mistake as 742 kg is a very high weight for a person. This also affects the standard deviation. The fat percentage is spread out quite a lot and thus has a high standard deviation.

The people who sign up for the foundation are mostly older. The youngest person that filled in their date of birth is 42, so as far as we can see there are no teenagers or young adults interested in the foundation.

	Weight	BMI	Fat percentage	Height in cm	Age
Weight	1	0.84	0.44	0.46	-0.17
BMI	0.84	1	0.48	-0.08	-0.30
Fat percentage	0.44	0.48	1	-0.22	0.38
Height in cm	0.46	-0.08	-0.22	1	0.1
Age	-0.17	-0.30	0.38	0.10	1

Table 4.6: Correlation between numerical values

Table 4.6 shows the correlation between the numerical data. There is a significant positive correlation between weight and BMI, which makes sense as the BMI is calculated using weight. However, there are no other significant correlations in this data set. Oddly, there is no correlation between height and BMI as height is also used to calculate BMI.

Statistics grouped by gender

Table 4.7 shows the statistics on the numerical data grouped by gender. Not everyone filled in their gender, so part of the data is dropped after it is grouped.

		Weight (kg)	BMI	Fat percentage (%)	Height in cm	Age
Male	**Mean**	93.30	28.78	31.70	180.72	61.33
	Median	91.00	28.61	32.50	181.00	63.50
	Mode	90.00	22.60	33.00	178.00	51.00
	Minimum	61.00	20.15	25.00	157.00	42.00
	Maximum	140.00	40.91	42.00	203.00	72.00
	Range	79.00	20.76	17.00	46.00	30.00
	Standard deviation	17.88	3.34	4.85	8.22	9.48
Female	**Mean**	83.95	29.94	38.46	167.77	60.23
	Median	82.00	29.36	40.00	168.00	59.50
	Mode	75.00	25.59	45.00	168.00	59.00
	Minimum	45.80	18.51	22.00	152.00	43.00
	Maximum	140.00	49.60	51.00	185.00	76.00
	Range	94.20	31.09	29.00	33.00	33.00
	Standard deviation	17.97	6.03	7.59	6.72	8.66

Table 4.7: Statistics grouped by gender

We can see that a lot of different people sign up for the foundation. Most of them are overweight, but there are also a few people that are underweight. The BMI and fat percentage range quite a lot. Men should have a fat percentage between 11% to 21% and women a fat percentage between 24% to 35%. Especially the women are distributed quite a lot for the fat percentage. Their standard deviation is higher than the men's standard deviation.

Table 4.8 shows the correlation between the numerical values grouped by gender. Again, there is a significant positive correlation between weight and BMI.

For the men there also seems to be a negative correlation between fat percentage and age. Thus if someone gets older, the fat percentage decreases. However, there is too little data to confirm that this is a serious correlation and not just a coincident.

		Weight	**BMI**	**Fat percentage**	**Height in cm**	**Age**
Male	**Weight**	1	0.86	-0.09	0.60	-0.30
	BMI	0.86	1	-0.18	0.11	-0.73
	Fat percentage	-0.09	-0.18	1	0.03	-0.91
	Height in cm	0.60	0.11	0.03	1	0.52
	Age	-0.30	-0.73	-0.91	0.52	1
Female	**Weight**	1	0.92	0.65	0.31	-0.17
	BMI	0.92	1	0.47	-0.08	-0.17
	Fat percentage	0.65	0.47	1	0.07	0.25
	Height in cm	0.31	-0.08	0.07	1	-0.17
	Age	-0.17	-0.17	0.25	-0.17	1

Table 4.8: Correlation between numerical values grouped by gender

Multiple choice questions

Table 4.9 shows how the multiple choice questions have been answered. We can see that people are very interested in information about COVID-19, as well ass recipes for keto meals.

Secondly, there are very little people that answered the question about the newsletter of the foundation, so we cannot gain any useful information from that.

There are some people that are a member of the foundation's FaceBook group for diabetes. The founder of the foundation has experience with diabetes himself, so it makes sense that people also joined the FaceBook group.

Column	Answer	Count	Total count
Waarover wilt u informatie ontvangen	Ik wil graag het COVID-19 e-book ontvangen	282	839
	Ik wil het keto recepten e-book ontvangen; Ik wil graag het COVID-19 e-book ontvangen	255	
	Ik wil het keto recepten e-book ontvangen	254	
	Ik wil graag het COVID-19 e-book ontvangen; Ik wil het keto recepten e-book ontvangen	48	
Ik ontvang graag de nieuwsbrief van Stichting Je Leefstijl Als Medicijn	Ja	10	11
	Nee	1	
ZWEMMER	Lid FB groep diabetes	155	296
	Ex Zwemmer	71	
	Actief Zwemmer	70	
Last Viewed Webinar Name	Metabole disfunctie door internist Yvo Sijpkens	381	1762
	Periodiek vasten	306	
	Via leefstijl op weg naar gezond en geluk	256	
	Leefstijlgeneeskunde Waarom en hoe?	238	
	Pijn, leefstijl en emoties. Emotionele leefstijl, de sleutel naar herstel	234	
	Hoe leefstijl actief bijdraagt aan gezondheid en herstel	106	
	Wijkaanpak in Gezond Wijdemeren	102	
	Leefstijlapotheker Anne-Margreeth Krijger	94	
	Een koolhydraatarme leefstijl in behandeling van diabetes type 1. Dokter Ian Lake	45	

Table 4.9: Multiple choice questions answers

Lastly, people are most interested in metabolic dysfunction. Metabolic dysfunction is when the body's metabolism is not working properly. The metabolism is a complicated process that involves many different chemicals,

organs and other things. When the body is unable to produce these chemicals or when the organ is failing, the metabolism is not functioning properly. Diabetes is a form of metabolic dysfunction, as the body is not able to produce enough insulin and thus the metabolism is failing. The second most popular webinar is about intermittent fasting, which we also already saw in the top blog posts. The least interesting webinar is about low-carb lifestyles to treat diabetes type 1.

4.2.3 Textual data

The textual data is shown in Word Clouds. The bigger the word is in a Word Cloud, the more often the word is mentioned in the corresponding column.

Uw vraag of opmerking

Figure 4.6 shows the answers to the column 'Uw vraag of opmerking', which is the question where people can leave any remarks or questions that they have for the foundation, in a Word Cloud. This was the most answered question of the questions with a textual answer.

We can see that 'diabetes' and 'lifestyle' are used a lot. 'Medicine' and 'healthy' are also quite big. Other smaller words are insulin, nutrition, weight, food and sugar.

Figure 4.6: A Word Cloud showing the most used words in 'Uw vraag of opmerking'

Table 4.10 shows the top ten most used words in 'Uw vraag of opmerking'. The most interesting words in this list are diabetes, lifestyle and healthy. Both medicine and medication are mentioned as well. People write that they want to reduce their medication use. Furthermore, it is interesting

to see that 'year' is the second most used word in this column. People write about how old they are or how long ago they were diagnosed with a medical disease. This means that there is some extra data hidden in these answers that could be interesting. However, we were unable to extract this.

Word	**Original word**	**Translation**	**Count**
diabetes		diabetes	398
jar	jaar	year	392
grag	graag	gladly	212
leefstijl		lifestyle	204
sind	sinds	since	163
jullie		you	154
gezond		healthy	151
medicatie		medication	150
medicijn		medicine	148
wer	weer	again	137

Table 4.10: Top ten words used in 'Uw vraag of opmerking'

Ervaring diensten JLAM

Figure 4.7 shows a Word Cloud of 'Ervaring diensten JLAM', which is the experiences of the members with the services the foundation provides. We see that 'good' and 'positive' are big. 'Zer' comes from 'zeer', which means 'very'. That looks like the biggest word in the cloud. People are emphasizing their words. Other words that stand out are nice, accessible, pleasant, informative, group, enormous and lifestyle.

Table 4.11 shows the top ten words used in 'Ervaring diensten JLAM'. Table 4.4 tells us that this column was filled in by only 33 people, which explains why the top word only has a count of six. However, the people who did fill in the question had a positive experience with the services of the foundation. As the counts are very low, we can say that there were very few negative answers as at least one negative word would have been in this top ten.

People might not feel the need to write a positive review about the foundation, but their experience also was not bad enough to write a bad review. This lack of feedback can still be interpreted as feedback.

Figure 4.7: A Word Cloud showing the most used words in 'Ervaring diensten JLAM'

Word	Original word	Translation	Count
zer	zeer	very	6
goed		good	4
positief		positive	3
mooi		nice	3
toegank	toegankelijk	accessible	3
ervar	ervaring	experience	3
webinar		webinar	3
informatief		informative	2
stichting		foundation	2
waardevol		valuable	2

Table 4.11: Top ten words used in 'Ervaring diensten JLAM'

Wat voor steun JLAM

Figure 4.8 shows a Word Cloud of 'Wat voor steun JLAM', which describes what kind of support the people expect from the foundation. We see that 'human' and 'information' are very big. Other words that stand out are 'knowledge', 'inspiring', 'better', 'help' and 'website'.

Figure 4.8: A Word Cloud showing the most used words in 'Wat voor steun JLAM'

Table 4.12 shows the top ten words used in 'Wat voor steun JLAM'. Table 4.4 tells us that only 33 people filled in this question as well. Therefore, the total counts per word are very low. However, the people who did fill in this question were mostly expecting information, help and inspiration.

Word	**Original word**	**Translation**	**Count**
mens		human	7
informatie		information	6
all	alle	all	4
geeft		give	4
goed		good	4
beter		better	3
wim		wim	3
inspireert		inspire	3
help		help	3
ander		other	3

Table 4.12: Top ten words used in 'Wat voor steun JLAM'

Wat werkt voor jou

Figure 4.9 shows the answers given to 'Welke behoeften', which is what kind of things work for the person, such as diet, conversations, etc. We see that 'health' is the biggest word in the Cloud. Other interesting words are 'moving', 'lifestyle' and 'knowledge'.

Figure 4.9: A Word Cloud showing the most used words in 'Wat werkt voor jou'

Table 4.13 shows the top ten words used in 'Wat werkt voor jou'. Again, Table 4.4 tells us that this question has only been filled in by 33 people. However, we see that most people mention 'healthy' in their answers. Other interesting words are 'stay', 'with', 'busy' and 'keep'. People might have used the phrase 'bezig blijven met', which translates to 'keeping busy with'. Overall we see that a healthy lifestyle works for people. They want to keep moving and also involve their diet.

Word	**Original word**	**Translation**	**Count**
gezond		healthy	16
werkt		works	8
blijv	blijven	stay	7
kennis		knowledge	6
beweg	bewegen	moving	6
mee		with	6
leefstijl		lifestyle	5
bezig		busy	5
eten		food or eating	5
houd		keep	4

Table 4.13: Top ten words used in 'Wat werkt voor jou'

Welke behoeften

Figure 4.10 shows a Word Cloud of 'Welke behoeften', which describes what needs the members have. Again, 'human' is a very big word, together with

'webinar', 'good' and 'person'.

Figure 4.10: A Word Cloud showing the most used words in 'Welke behoeften'

Table 4.14 shows the top ten used words in 'Welke behoeften'. Table 4.4 tells us that this question was answered by only 29 people. However, the people that did fill it in are interested in health. We see that people also explicitly mention the webinars.

Word	**Original word**	**Translation**	**Count**
mens		human	6
goed		good	4
person	persoon	person	4
per		per	4
webinar		webinar	4
gezond		healthy	3
werkt		works	3
behoeft	behoeftes	needs	3
bijvoorbeeld		for example	3
nam	nemen	take	3

Table 4.14: Top ten words used in 'Welke behoeften'

Chapter 5

Discussion

In the Discussion, we first discuss the results, including getting back to the existing literature. Lastly, we discuss any limitations we encountered in our research and we give a recommendation for possible future work.

5.1 Interpretation of the results

The results show that there are many different reasons why people want to get started with lifestyle changes. Some were in line with our hypothesis, while others were new. There were also some expected reasons that were not actually a motivation.

5.1.1 Reflecting on our hypothesis

We expected that the members of the foundation were overweight and wanted to lose weight. This turned out to be partially true. Most of the people who filled in their weight are overweight and the average BMI and fat percentage are too high. However, the people did not mention that they wanted to lose weight. Therefore, weight in itself was not a reason for the members of the foundation to start with lifestyle changes and this part of our hypothesis was incorrect. We were also unable to identify a clustering when we plotted the height against the weight.

Secondly, we expected that people want to reduce the risk at certain health issues, such as diabetes or cardiovascular problems. This turned out to be correct, as diabetes is a recurring factor in a lot of the questions. First of all, 'diabetes' is the most used word in 'Uw vraag of opmerking'. As this is also the most answered textual question, it is a big factor among the people who filled it in. Secondly, some people joined the FaceBook group that is specifically for people with diabetes. Lastly, people were interested in the webinar about metabolic dysfunction and diabetes is a metabolic disorder,

so it is also recurring in this question. However, other health diseases were not mentioned at all or as much as diabetes. This is probably because the foundation focused a lot on diabetes when they started. The founder of the foundation had diabetes due to his unhealthy lifestyle and by implementing a new, healthier lifestyle he was able to revert this. Therefore, people who have diabetes and are looking for ways to reduce the effects will probably end up at the foundation.

Furthermore, the results show that people were interested in reducing the medication that they were using. The founder of the foundation was able to reverse the effects of diabetes through his lifestyle and study shows that this can indeed be done [36] [32]. This part of our hypothesis was correct.

We also expected that people were interested in reducing their risks for certain health diseases, such as cardiovascular issues or diabetes. However, our results show that this is not correct. Words as 'prevention' or 'risk' were not used in the textual questions. Most of the members were already experiencing these problems and were looking for ways to reduce the effects.

Additionally, we expected new year's resolution to be a reason to start with lifestyle changes. However, the results show that there is no increase in member registrations around December or January every year. The graph of the creation date shows that there is no pattern in the number of sign-ups per month. Therefore date is not a motivator.

Lastly, we expected that people wanted to age without any complications that can come with it and thus we expected the members to be a bit older. This was partly correct. The average age was 61.18 years, so the members who filled in their date of birth are a bit older. There were also no teenagers that signed up and filled in their date of birth. This is in line with our hypothesis. However, the members did not mention age as a motivation. The complications that can come with getting older are not limiting the members as much as we were expecting. We were also unable to identify a clustering when we plotted the age against the weight.

All of the reasons that we mentioned in our hypothesis could be related back to one reason:

To improve their overall health

Our results show that this is indeed a big motivation to start with lifestyle changes. 'Healthy' was used the most in the textual questions. The top ten blog posts were also covering a lot of different aspects of health. All of the other reasons that we mentioned in this section are specific reasons that contribute to improving someone's overall health. Our hypothesis was correct.

5.1.2 Unexpected results

Our results show that COVID-19 was another reason why people wanted to get started with lifestyle changes. The most viewed blog post of all time is an article about COVID-19 and the positive effects of lifestyle on the immune system. A lot of people were also interested in the COVID-19 e-book. However, in the past year this blog post has been viewed way less than the years before. COVID-19 used to be a big motivation when the pandemic was happening, but it is no longer a reason to start with lifestyle changes. This was not in our hypothesis, but it is an interesting result.

5.1.3 Existing literature

Our results show that people are looking for a support group. Words like 'help', 'accessible', 'informative', 'human' and 'inspire' were used in the textual questions. They show that people are reaching out to the foundation for help. They are looking for information on healthy lifestyles and help on how to start implementing this. Study showed us that extra resources can lead to better results [14].

These words also show that people are looking for external motivation, such as accountability from others. Study showed that accountability leads to healthier dietary choices [9]. The FaceBook group for diabetes that some of the members joined is a support group where people help each other. Study showed that a support group can positively influence motivation for lifestyle changes [31]. The foundation provides these support groups and gives the members the accountability they are looking for.

Furthermore, our results show that people want to reduce their medication usage. When people need less medication, it means that the effects of the corresponding medical disease have decreased as well. Study showed us that a healthy lifestyle can sometimes reverse the effects of health problems [36] [32]. Diabetes was the most recurring medical disease among the members. Study showed that lifestyle intervention can lead to medication reduction among people with Type 2 Diabetes [11].

The services that the foundation offers, such as support groups and information through blog posts and webinars, can be seen as extra tools that can help stick to a healthier lifestyle in the long term. When people start to fall back into their old habits, they can reach out to the foundation and the other members to give them tips and get them back on track. Study showed that long-term compliance to behavioural treatments results in long-term weight loss [27].

Lastly, most people want to improve their overall health rather than lose weight. Study showed that these goals have a positive effect on physical activity due to higher internal motivation [12]. This can then reinforce behaviour change [34]. Therefore, the members of the foundation have realistic

goals that are more likely to be achieved. They set themselves up for success instead of failure.

5.2 Limitations

5.2.1 Data

A big limitation in this data analysis is the data itself. When people sign up for the foundation, they have to fill in their information. However, none of this information is mandatory. Additionally, all people who only sign up for the newsletter do not have to fill in these questions, but their data does get stored in Hubspot. This resulted in a lot of null values in the data set. When we were looking at specific characteristics, such as gender or age, there was a lot less data to work with. Therefore, the findings of this analysis do not represent the entire data set, but only the entries that had the specific information. We tried to restore some of the data, for example by using weight and height to calculate the BMI if it was missing. However, there were no entries for which this could be applied as the weight and height would both be empty as well. So it unfortunately did not create more data to work with.

In addition to this, we noticed that people wrote long texts in the comments where they hid important information. For example, they mentioned their age in the comments, but they left their date of birth empty. In this data analysis, we calculate the age by using the date of birth so these entries are missing. The same could be said about their gender, as for some entries it can be derived from what they wrote. However, we were unable to find a way to extract this information from these comments.

Another limitation of the data is that it is in Dutch, while most Python functions are created for English data sets. The NLTK did offer a list with Dutch stopwords as well as a stemmer, but we also noticed that there were a lot of spelling mistakes in the the textual data. Python offers various spell checkers which can identify misspelled words and correct them, but Dutch is not a supported language for these functions. Therefore, we could not apply them and thus might have missed some instances of words.

We would recommend the foundation to make some of the questions mandatory. Then these questions will have to be filled in before people can send their answers. The data set will for sure contain a value in those columns. However, we do understand that this is personal data that people might not want to share. Another solution could be to add a small disclaimer on what the data will be used for. Explaining why the foundation would need that data might lead to more people filling it in and thus more data to analyse. Secondly, the foundation can in this disclaimer also ask not to write long texts in 'Uw vraag of opmerking' if this information can also be given via a different question. Lastly, the foundation might think about adding a

question asking for the age as many people do not fill in their date of birth, but do add their age in their remarks.

5.2.2 Methods

This research is limited to our knowledge and skills in data analyis. We used Hubspot as an analysis tool, but also applied our own Python analysis. These were the methods we thought about and were able to apply, but there might be different analysis tools that can also be applied on our data. Someone with more knowledge on data analysis in Python might know different techniques or ways to make the tools we used better and gain more information from it. This might show new insights or results that we did not get with our analysis.

Our methods and knowledge were a big limitation when we tried to restore the missing values in the data. There are ways to restore missing data, such as substituting the missing values with reasonable guesses. However, most of our data set was lacking so we felt like we did not have enough knowledge to properly restore this.

Lastly, we tried to use the tools Hubspot offers for analysing data but this is a very extensive system. There might be useful graphs that we missed because we did not understand Hubspot well enough.

5.3 Future research

First of all, we think that performing this data analysis again in a few years could be very interesting, as the foundation will have gathered more data by then. We might be able to see more patterns, such as in the graph of the creation date.

Secondly, future work could look into the handling of missing values or applying techniques we were unable to use.

Future work could also explore Hubspot more. We think that there is still some interesting information that can be presented through Hubspot's tools.

Lastly, there was one subquestion that we were unable to answer completely. Our results do not show how people came in contact with the foundation. The members might have mentioned it in the textual questions, but it did not come forward after our word count. The foundation might be interested in conducting an additional questionnaire to their members to get this information for future research.

Chapter 6

Conclusions

The results of our data analysis show that there are multiple reasons why people get started with lifestyle changes.

First of all, we found that most people who filled in their weight were overweight. The overall BMI and fat percentage were too high for both men and women. However, people did not explicitly mention that they wanted to lose weight. We were unable to find a clustering when we plotted the height against the weight.

Furthermore, the members mention diabetes a lot. It is the most used word in 'Uw vraag of opmerking', which was also the most answered textual question in our data set. Some of the people also joined the FaceBook group for diabetes and people were interested in the webinar about metabolic dysfunction, which was related to diabetes.

We also found that people are interested in reducing their use of medication. This is of course related to medical diseases, as related work showed us that lifestyle changes can reverse the effects of diabetes.

'Healthy' was the most recurring word in the textual questions. So overall, people wanted to improve their health. Our hypothesis was correct about this.

COVID-19 also turned out to be a motivator for people to sign up for the foundation. The most viewed blog post of all time is about COVID-19 and most people were interested in the COVID-19 e-book. However, the blog post is viewed way less this past year so it looks like it no longer is a big motivation. This was not in our hypothesis, but it is an interesting result.

We expected age to be a bigger motivation than it turned out to be. Teenagers do not seem to be interested in the foundation, as the average age was 61. However, age itself is not a motivator for the members to get started with lifestyle. We were also unable to find a clustering when we plotted the age against the weight.

We were also unable to identify patterns in the registration dates of the members. December and January did not result in more people signing up, so New Year's resolutions are not a motivation. Our hypothesis was wrong about this.

Lastly, our results did not show how people got in contact with the foundation. We know their main reasons and motivations for contacting the foundation, but not how they learned about the foundation and why they chose this instead of other resources.

Bibliography

[1] Analyze your site traffic with the traffic analytics tool. https://knowledge.hubspot.com/reports/analyze-your-site-traffic-with-the-traffic-analytics-tool. Hubspot's documentation about their analytics tool.

[2] How bad is our obesity problem. https://www.bbc.com/news/health-53514170. BBC news article about tackling the obesity problem in England.

[3] Hubspot analytics and google analytics don't match. https://knowledge.hubspot.com/reports/analyze-your-site-traffic-with-the-traffic-analytics-tool. Hubspot's documentation about their analytics tool.

[4] Je leefstijl als medicijn website. https://jeleefstijlalsmedicijn.nl. The foundation's website.

[5] Manage your CRM database. https://knowledge.hubspot.com/get-started/manage-your-crm-database. Hubspot's documentation about managing a CRM database.

[6] Most adults living unhealthy lifestyles. https://www.bbc.com/news/health-46439892. BBC news article about how most adults are living unhealthy lifestyles.

[7] Oxford dictionary. https://www.oed.com/dictionary. Oxford dictionary.

[8] Cristina Maria Bostan, Alexandru-Cosmin Apostol, Răzvan-Lucian Andronic, Tudor Stanciu, and Ticu Constantin. Type of goals and perceived control for goal achievement over time. the mediating role of motivational persistence. *Acta Psychologica*, 231:103776, 2022.

[9] Karishma Chhabria, Kathryn M Ross, Shane J Sacco, and Tricia M Leahey. The assessment of supportive accountability in adults seeking obesity treatment: Psychometric validation study. *Journal of Medical Internet Research*, 22(7):e17967, 2020.

[10] Zafra Cooper, Helen A Doll, Deborah M Hawker, Susan Byrne, Gillie Bonner, Elizabeth Eeley, Marianne E O'Connor, and Christopher G Fairburn. Testing a new cognitive behavioural treatment for obesity: A randomized controlled trial with three-year follow-up. *Behaviour research and therapy*, 48(8):706–713, 2010.

[11] Linda M Delahanty, Kristen M Dalton, Bianca Porneala, Yuchiao Chang, Valerie M Goldman, Douglas Levy, David M Nathan, and Deborah J Wexler. Improving diabetes outcomes through lifestyle change–a randomized controlled trial. *Obesity*, 23(9):1792–1799, 2015.

[12] Tonya Dodge, Deepti Joshi, Malak Alharbi, and Brad Moore. Effect of physical activity goals on aerobic physical activity: testing the mediating role of external and internal motivation. *Psychology, Health & Medicine*, 27(6):1296–1310, 2022.

[13] Maria Cristina Enache et al. Data analysis with pandas. *Economics and Applied Informatics*, (2):69–74, 2019.

[14] Jean L Forster, Robert W Jeffery, Thomas L Schmid, and F Matthew Kramer. Preventing weight gain in adults: a pound of prevention. *Health Psychology*, 7(6):515, 1988.

[15] Barbara L Fredrickson, Cara Arizmendi, and Patty Van Cappellen. Same-day, cross-day, and upward spiral relations between positive affect and positive health behaviours. *Psychology & Health*, 36(4):444–460, 2020.

[16] Barbara L Fredrickson and Thomas Joiner. Reflections on positive emotions and upward spirals. *Perspectives on psychological science*, 13(2):194–199, 2018.

[17] Guillermo García-Pérez-de Sevilla, Enrique Alonso Pérez-Chao, Helios Pareja-Galeano, Eva María Martínez-Jiménez, Beatriz Sánchez-Pinto-Pinto, Carlos Romero-Morales, et al. Impact of lifestyle on health-related quality of life among young university students: a cross-sectional study. *Sao Paulo Medical Journal*, 139:443–451, 2021.

[18] K Gunnar Götestam. A three year follow-up of a behavioral treatment for obesity. *Addictive Behaviors*, 4(2):179–183, 1979.

[19] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[20] Erin Hoare, Nicholas Crooks, Joshua Hayward, Steven Allender, and Claudia Strugnell. Associations between combined overweight and obesity, lifestyle behavioural risk and quality of life among australian regional school children: baseline findings of the goulburn valley health

behaviours monitoring study. *Health and quality of life outcomes*, 17:1–10, 2019.

[21] HB Hubert. The importance of obesity in the development of coronary risk factors and disease: the epidemiologic evidence. *Annual review of public health*, 7(1):493–502, 1986.

[22] M Janakova. Crm & social networks. *Academy of strategic management journal*, 17(5):1–15, 2018.

[23] Arjun Khorana, Ayoosh Pareek, Matthieu Ollivier, Sophia J Madjarova, Kyle N Kunze, Benedict U Nwachukwu, Jón Karlsson, Erick M Marigi, and Riley J Williams III. Choosing the appropriate measure of central tendency: mean, median, or mode? *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(1):12–15, 2023.

[24] Richard Koestner, Nancy Otis, Theodore A Powers, Luc Pelletier, and Hugo Gagnon. Autonomous motivation, controlled motivation, and goal progress. *Journal of personality*, 76(5):1201–1230, 2008.

[25] Michelle A Lee-Bravatti, H June O'Neill, Renee C Wurth, Mercedes Sotos-Prieto, Xiang Gao, Luis M Falcon, Katherine L Tucker, and Josiemer Mattei. Lifestyle behavioral factors and integrative successful aging among puerto ricans living in the mainland united states. *The Journals of Gerontology: Series A*, 76(6):1108–1116, 2021.

[26] Rebeka Lekše, Dijana Godec, and Mirko Prosen. Determining the impact of lifestyle on the health of primary school children in slovenia through mixed membership focus groups. *Journal of Community Health*, pages 1–13, 2023.

[27] Leonard S Levitz et al. Weight loss five years after behavioral treatment. 1980.

[28] Yunfei Li, Akira Babazono, Aziz Jamal, Ning Liu, Takako Fujita, Rui Zhao, Yukari Maeno, Ya Su, Lifan Liang, and Lan Yao. The impact of lifestyle guidance intervention on health outcomes among japanese middle-aged population with metabolic syndrome: A regression discontinuity study. *Social Science & Medicine*, 314:115468, 2022.

[29] World Health Organization. Obesity: preventing and managing the global epidemic: report of a who consultation. 2000.

[30] Raymond P Perry. Perceived (academic) control and causal thinking in achievement settings. *Canadian Psychology/Psychologie Canadienne*, 44(4):312, 2003.

[31] Sabrina K Schmidt, Liv Hemmestad, Christopher S MacDonald, Henning Langberg, and Laura S Valentiner. Motivation and barriers to maintaining lifestyle changes in patients with type 2 diabetes after an intensive lifestyle intervention (the u-turn trial): a longitudinal qualitative study. *International journal of environmental research and public health*, 17(20):7454, 2020.

[32] RW Simpson, JE Shaw, and PZ Zimmet. The prevention of type 2 diabetes—lifestyle change or pharmacotherapy? a challenge for the 21st century. *Diabetes research and clinical practice*, 59(3):165–180, 2003.

[33] Peter M Stalonas, Michael G Perri, and Alan B Kerzner. Do behavioral treatments of obesity last? a five-year follow-up investigation. *Addictive behaviors*, 9(2):175–183, 1984.

[34] Liam Staunton, Paul Gellert, Keegan Knittle, and Falko F Sniehotta. Perceived control and intrinsic vs. extrinsic motivation for oral self-care: a full factorial experimental test of theory-based persuasive messages. *Annals of Behavioral Medicine*, 49(2):258–268, 2015.

[35] Kelly M West and John M Kalbfleisch. Influence of nutritional factors on prevalence of diabetes. *Diabetes*, 20(2):99–108, 1971.

[36] Birgit-Christiane Zyriax and Eberhard Windler. Lifestyle changes to prevent cardio-and cerebrovascular disease at midlife: a systematic review. *Maturitas*, 167:60–65, 2023.

Appendix A

Appendix

Blog post	Views
Covid 19 en de positieve effecten van leefstijl op je weerstand	24860
Niet zout, maar suiker is de boosdoener - Internist Yvo Sijpkens	8590
Recensie 'VET Belangrijk' Mariëtte Boon & Liesbeth van Rossum	4607
Borstkanker en vasten. Hoopgevende resultaten uit studie door LUMC	2632
Een gezonde leefstijl is niet alleen een kwestie van willen	2462
Behandeling tegen kanker is niet compleet zonder leefstijladviezen	2348
GLI: Gecombineerde Leefstijl Interventie of Grote Leefstijl Illusie	1972
De focus bij hart- en vaatziekten meer op de metabole kant	1561
Recensie 'Slapen is niets doen' - Aline Kruit	1517
Waarom is ander gedrag vasthouden zo lastig? Dr Pepijn van Empelen TNO	1186
Van insuline kun je doodziek worden - William cortvriendt	1132
Nationaal leefstijlplatform wil de zorg hervormen	1095
Heeft wondermiddel vasten echt zo een heilzame werking?	1058
Ramadan actie voor moslims met diabetes type 2	828
Artikel in het NRC met Hanno Pijl Jaap Seidell en Wim Tilburgs	208
	166
Test blog	10
	4
Your Blog Post Title Here...	0
Borstkanker en vasten. Hoopgevende resultaten uit studie door het LUMC	0
Totaal	56236

Table A.1: Top blog posts of all time

Blog post	Views
Recensie 'VET Belangrijk' Mariëtte Boon & Liesbeth van Rossum	843
Borstkanker en vasten. Hoopgevende resultaten uit studie door LUMC	500
Niet zout, maar suiker is de boosdoener - Internist Yvo Sijpkens	314
Recensie 'Slapen is niets doen' - Aline Kruit	156
Behandeling tegen kanker is niet compleet zonder leefstijladviezen	119
Van insuline kun je doodziek worden - William cortvriendt	101
Covid 19 en de positieve effecten van leefstijl op je weerstand	100
Waarom is ander gedrag vasthouden zo lastig? Dr Pepijn van Empelen TNO	59
GLI: Gecombineerde Leefstijl Interventie of Grote Leefstijl Illusie	59
Een gezonde leefstijl is niet alleen een kwestie van willen	37
Artikel in het NRC met Hanno Pijl Jaap Seidell en Wim Tilburgs	28
Heeft wondermiddel vasten echt zo een heilzame werking?	23
De focus bij hart- en vaatziekten meer op de metabole kant	23
Nationaal leefstijlplatform wil de zorg hervormen	12
Ramadan actie voor moslims met diabetes type 2	828
	0
Test blog	0
Your Blog Post Title Here...	0
Borstkanker en vasten. Hoopgevende resultaten uit studie door het LUMC	0
Totaal	2374

Table A.2: Top blog posts between 11-01-2022 untill 10-31-2023

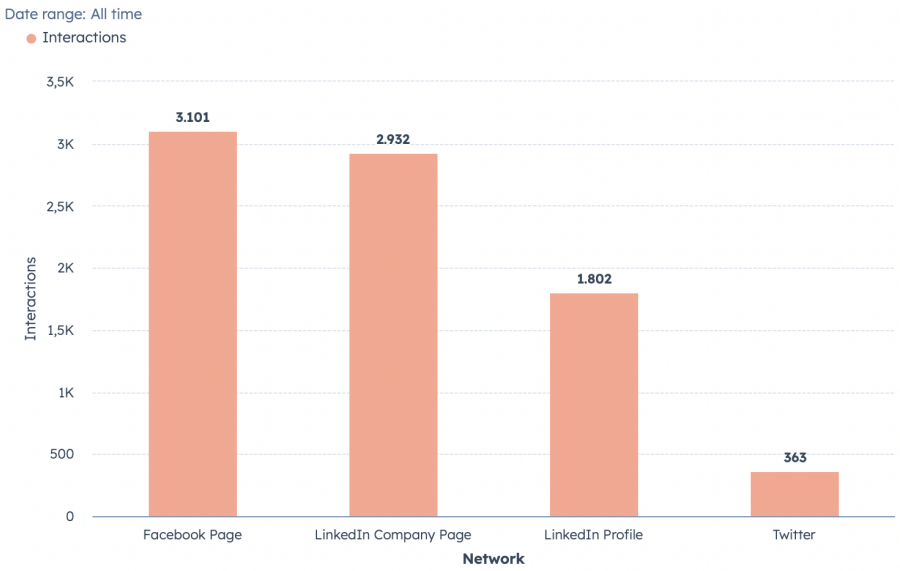


Figure A.1: Social interactions by network

Country	Sessions	New session in %
Netherlands	968,205	80.12
Belgium	100,357	87.16
United States	5,757	95.80
Germany	4,509	85.01
Spain	4,236	80.19
France	4,198	79.70
Switzerland	1,416	77.26
Italy	1,219	81.13
United Kingdom	1,209	87.26
Portugal	939	82.22

Table A.3: Sessions per country

Source	Description source	Bounce rate in %	Average session length in seconds	Page views per session
Organic search	Traffic from non-paid search results in known search engines	84.12	69	1.42
Direct traffic	Traffic that don't have an indication of their source	79.93	113	2.60
Organic social	Traffic from social media websites or apps	81.63	72	1.49
Paid search	Traffic from paid search campaigns	89.33	41	1.21
Email marketing	Traffic from emails	73.89	120	1.57
Referrals	Traffic from external sites that link to your website	74.90	124	2.22
Other campaigns	Traffic from traffic URLs created in Hubspot	66.13	83	1.62
Paid social	Traffic from a paid social campaign	90.00	47	1.20

Table A.4: Session engagement rates

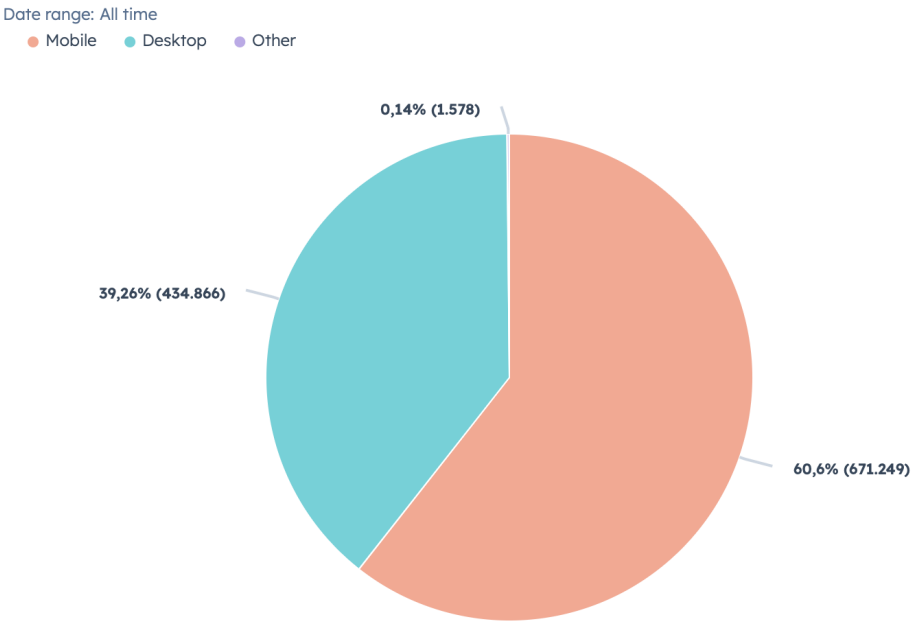


Figure A.2: Device breakdown