Università degli Studi di Milano

Data Science and Economics

Unsupervised learning problem: obesity and lifestyle

Student:
Michele Bartesaghi
Registration number: T16334

ACADEMIC YEAR 2021-2022

Contents

Chapter 1

Abstract

The aim of this project is to identify different profiles of people, on the basis of the answers they gave to an online questionnaire about their physical condition and lifestyle. Moreover, this paper will try to proof that a sedentary lifestyle combined with low quality heating habits is much likely to belong to a person who is not winning the "battle of the bulge".

In order to do so, a dataset collecting over two thousands answers to the survey has been downloaded. The respondents were originally 458 people from Mexico, Peru and Colombia. Later, Fabio Mendoza Palechor and Alexis de la Hoz Manotas used an oversampling technique to create a larger dataset, were 77% of the data points are fictitious ([2]). The original data set comprises a sort of response variable, which has been removed to apply the major unsupervised learning techniques.

After a brief pre-processing, some data visualisation is displayed in order to grasp some relevant information about the people answering the questions. Afterwards, a factor analysis is carried out in an attempt to find out how variables are related to each other and at which extent they explain the variability of the data.

Finally, two different clustering techniques (i.e. hierarchical clustering and k-prototypes) will be applied to divide respondents into groups and study their characteristic, in order to check if they eventually meet the suppositions. On the whole the respondents' profiles are similar to the ones expected at the beginning, yet not much can be said about the impact of the different lifestyles, as most of the individuals gave similar answers to the same questions.

Chapter 2

The dataset

As anticipated before, the dataset contains 2111 records, each one representing a different person taking the survey, and 17 columns, one for each of the questions displayed below. Originally, the data points were only 23% of the actual ones, which have been obtained artificially, hence it is possible that results will be surprising in some fashion. Furthermore, both the survey and the dataset were created and elaborated with the intent of applying supervised learning techniques, in order to predict the respondent's obesity level, therefore the response variable has been removed before doing anything else. With this in mind, both the questions and the possible answers are reported below (between brackets one can read the corresponding renamed *variable* of the dataset, with its type, as well as how the answer has been originally encoded).

- What is your gender? (*gender, chr*)
  - Male ("Male")
  - Female ("Female")
- What is your age? (*age, num*)
- What is your height in metres? (*height, num*)
- What is your weight in kilograms? (*weight, num*)
- Has a family member suffered or suffers from overweight? (*ow_history, chr*)
  - Yes ("yes")
  - No ("no")

- Do you eat high caloric food frequently? (*caloric_food_freq, chr*)
  - Yes ("yes")
  - No ("no")
- Do you usually eat vegetables in your meals? (*vegetables_freq, num*)
  - Never (1)
  - Sometimes (2)
  - Always (3)
- How many meals do you have daily? (*meals_num, num*)
  - Between 1 and 2 (1)
  - Three (2)
  - More than three (3)
- Do you eat any food between meals? (*snacking_freq, chr*)
  - No ("no")
  - Sometimes ("Sometimes")
  - Frequently ("Frequently)
  - Always ("Always")
- Do you smoke? (*smoke, chr*)
  - Yes ("yes")
  - No ("no")
- How much water do you drink daily? (*water_intake, num*)
  - Less than a litre (1)
  - Between 1 and 2 litres (2)
  - More than two litres (3)
- Do you monitor the calories you eat daily? (*calories_monitoring, chr*)
  - Yes ("yes")
  - No ("no")

- How often do you have physical activity? (*PA_freq, num*)
  - I do not have (0)
  - 1 or 2 days (1)
  - 3 or 4 days (2)
  - 4 or 5 days (3)
- How much time do you do you use technological devices such as cell phone, video games, television, computer and others? (*technology_ daily_freq, num*)
  - 0-2 hours (0)
  - 3-5 hours (1)
  - More than 5 hours (2)
- How often do you drink alcohol? (*alcohol_freq, chr*)
  - I do not drink ("no")
  - Sometimes ("Sometimes")
  - Frequently ("Frequently")
  - Always ("Always")
- Which transportation do you usually use? (*mtrans, chr*)
  - Automobile ("Automobile")
  - Motorbike ("Motorbike")
  - Bike ("Bike")
  - Public Transportation ("Public_Transportation")
  - (Walking) ("Walking")

It is immediate to notice that this dataset presents data of mixed types, meaning that this project is likely to present the challenges that come naturally when dealing with both qualitative and quantitative variables. However, this situation is ordinary in real world cases. Even though many features has been encoded as numbers, probably it is more appropriate to transform them into factors, representing different levels, distinctly unordered categories of lifestyle and eating habits. Furthermore, the features height and weight are combined in order to work with fewer variables and create the variable called *bmi*, which is expected to be really influential in the definition of different

respondents' profiles. The $k-th$ individual's Body Mass Index (BMI) is defined as follows:

$$bmi_k = \frac{weight_k(kgs)}{(height_k(m))^2}.$$

After these transformations, there are 14 factor variables and only two numeric ones, namely *age*, stored as an integer, and *bmi*. Finally, observe that no NA's are present in the dataset.

2.1 Exploratory data analysis

Now a few relevant graphs will be shown, in order to see the relation between the various respondents' BMI and a few other answers that have been given.

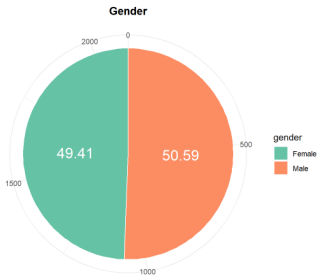


Figure 2.1: Gender distribution.

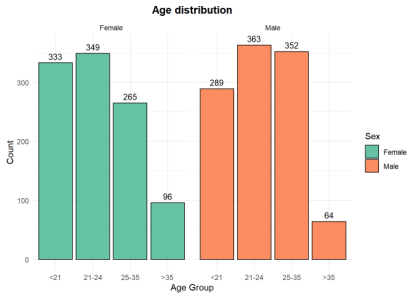


Figure 2.2: Age distribution according to gender.

As one can see, the gender distribution is very balanced, but this could be also due to the oversampling applied to the original data. However, the age distribution is strongly concentrated between 18 and 26 years old, meaning that the sample is not really representative of the entire population, and this should be kept in mind when drawing the conclusions.

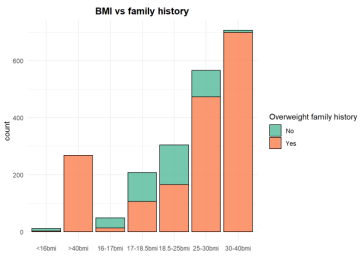


Figure 2.3: Relationship between BMI and family history of overweight problems.

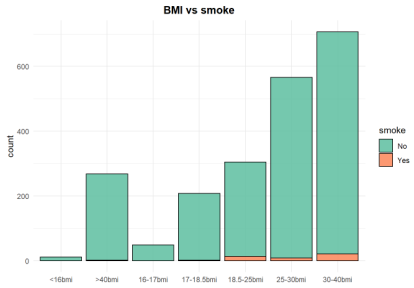


Figure 2.4: Relationship between BMI and smoking habits.

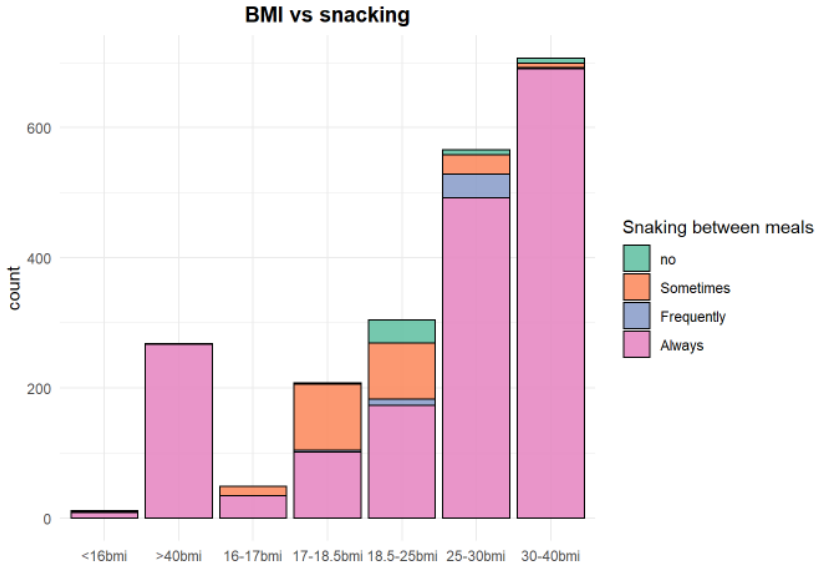


Figure 2.5: Relationship between BMI and snacking habits.

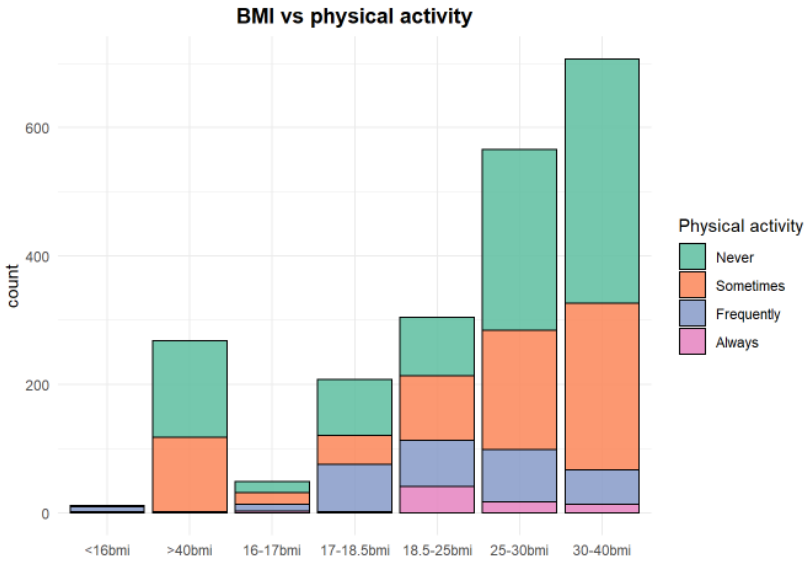


Figure 2.6: Relationship between BMI and physical activity.

In figure 2.3 it is interesting to notice how in the intermediate BMI ranges there are both people with and without at least a case of an overweight person in their family, whereas in the lowest BMI range almost only people

without overweight individuals in their family are found. At the other end of the scale only and solely people with a case of an obese relative fall in the ">40bmi" category. The graph showing the relationship between the BMI and smoking habits of the respondents is not really relevant, as there is a manifest disproportion between smokers and non-smokers in the dataset. In figure 2.5 it is clear that all the people with a BMI greater than 40 are always snacking between meals, increasing their calories intake, whilst one would not expect that people with a BMI of 16 or less (severely underweight) stated to eat either frequently or always between meals; perhaps, instead of eating proper meals, they just eat something quick, and that makes them constantly snacking. Ultimately, the last graph shows that most of the respondents never do physical activity, whilst an alarmingly little number of individuals are either frequently or always doing something in order to avoid a sedentary lifestyle.

In the figure below, one can see how the features that were initially numeric are distributed. No point will be treated as an outlier.

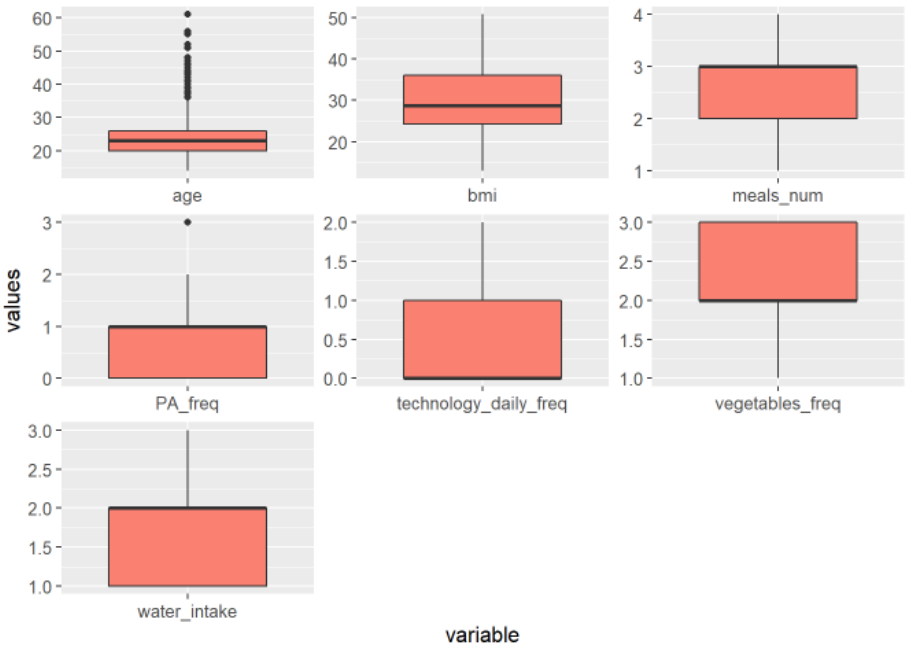


Figure 2.7: Variables distribution.

Finally, the correlation matrix shows neither positive nor negative correlation between any two of the variables.

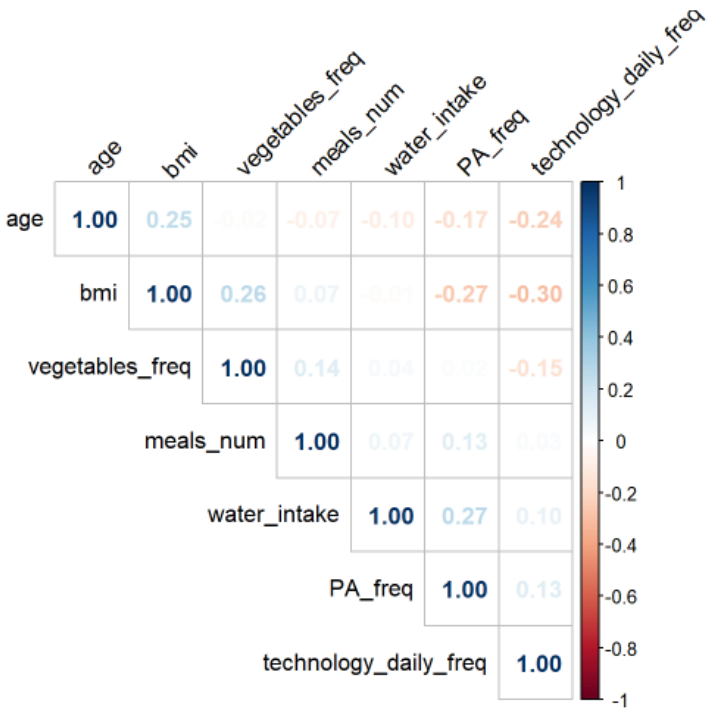


Figure 2.8: Pairwise correlation between numeric variables.

Chapter 3

Factor analysis

It is safe to say that a lifestyle is not something that can be defined unequivocally. While it is true that in this dataset there are not innumerable features, one could still ask himself whether it would be possible to investigate this latent concept of different lifestyles and eventually reduce to a slight extent the dimensionality of the data. Furthermore, with a factor analysis one could understand what the data points have in common and how variables are related. In this particular case it will be a sort of confirmatory factor analysis, meaning it is done in an attempt to test the common ideas about the relationship between the features.

Multiple Factor Analysis (MFA) is a multivariate data analysis method used to summarise and visualise a dataset in which individuals are described by a set of both qualitative and quantitative variables structured into groups, hence it is perfectly suitable for datasets collecting answers to a survey. In this particular case the variable order has been re-arranged to identify 5 different groups:

- physical status: comprising the respondents' BMI and age,
- eating habits: consisting of the answers regarding their relationship with food (*caloric_food_freq*, *vegetables_freq*, *meals_num*, *snacking_freq*, *water_intake*),
- everyday life habits: made of the features regarding their daily habits (*technology_daily_freq*, *PA_freq*, *mtrans*, *calories_monitoring*),
- vices: containing the answers about their relationship with alcohol and smoke,
- hereditary traits: comprising the respondents' gender and the familiar history of overweight cases.

MFA is an algorithm that could be seen as a meeting point between a Principal Component Analysis[1] (PCA) and a Multiple Correspondence Analysis[2] (MCA). In addition, it takes into consideration the contribution of all the groups of variables to define the distance between individuals. During the analysis, it is required to balance the contribution of each set of variables, therefore the features are weighted. The ones that are in the same group are normalised using the same weighing value, which can vary among different groups; more precisely, to each variable of a given group k, a weight equal to the inverse of the first eigenvalue of the group k is assigned. Notice that eigenvalues are associated with the proportion of variances retained by the different dimensions, so the more influential the group, the less the variables weigh. Ultimately, thanks to this analysis one could hope to identify some variables that are contributing to a great extent at explaining the variability of the data, thus to understand which are the crucial features to divide the observations into clusters.

First and foremost, in the figure below one can see the proportion of variance explained by each dimension, that is remarkably low, as ten dimensions are needed just to explain about 50% of the variability of the data points.

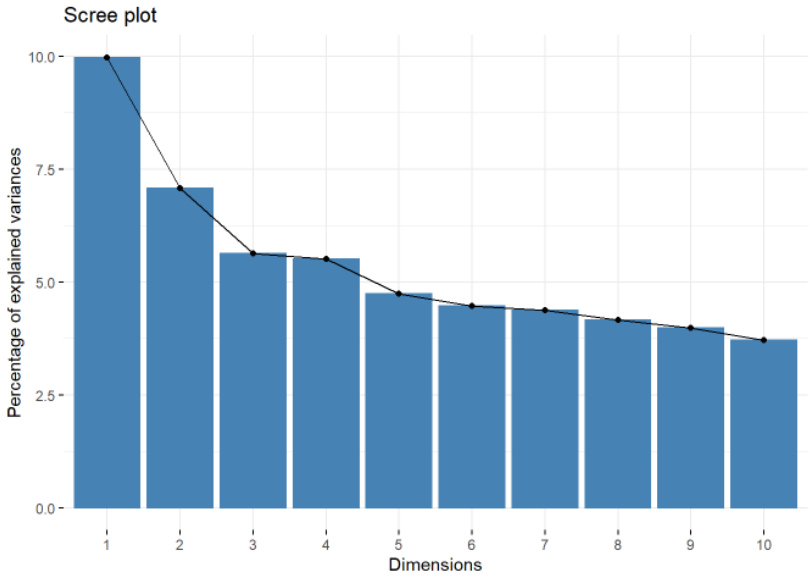


Figure 3.1: Visualisation of the percentage of variances explained.

[1]The most common dimensionality reduction method.

[2]Another technique to detect and represent underlying structures in a dataset on a lower dimensional space.

Then, it is important to understand how much each group contributes to each one of the two principal dimensions. As one can see in the figure below (3.2), the coordinates of both the eating habits and the everyday life habits groups on the first dimension are almost identical, meaning that they contribute similarly to the first dimension. However, physical status is the group that contributes more to this dimension. With regard to the second dimension, hereditary traits and and eating habits have the highest coordinates, indicating the highest contribution to the second dimension. In the other two graphs a more standard visualisation of these concepts can be seen.

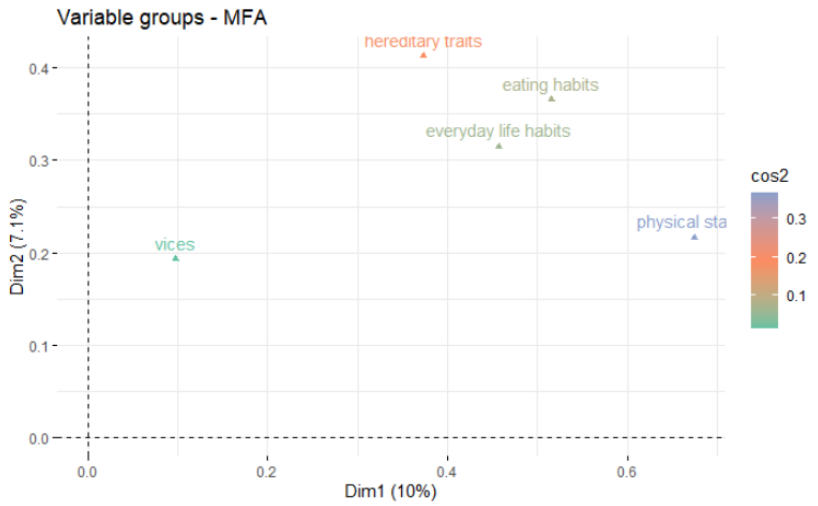


Figure 3.2: Contribution of groups to each of the two dimensions, coloured according to quality of their representation.

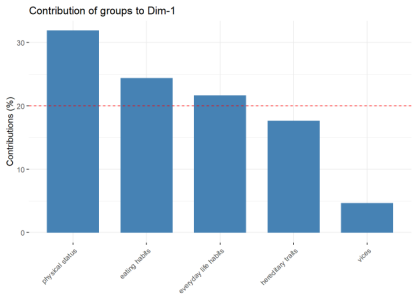


Figure 3.3: Contribution of groups to the first dimension.

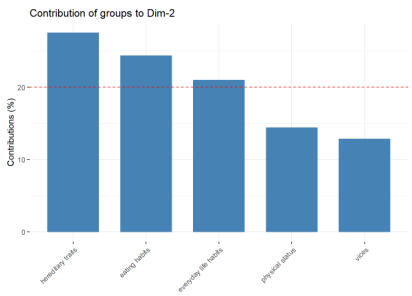


Figure 3.4: Contribution of groups to the second dimension.

According to this analysis one could expect that the variables in the most contributing groups might play an important role in the definition of the

different clusters. Moreover, it is possible to investigate the relationships between the variables, the correlation between the dimensions and the features, as well as the quality of the representation of the latter in the first two dimensions. To interpret the graphs below notice that:

- positively correlated variables are grouped together, whereas negatively correlated ones are positioned symmetrically with respect to the origin,
- the distance between each variable and the origin measures the quality of the variable on the factor map, namely how well it is represented,
- the closer a feature is to a dimension (axis), the more correlated they are.

Therefore, one can conclude that the second dimension mostly represents the means of transportation, the gender and the number of meals, whereas the variables that are mainly correlated to the first dimension are perhaps the caloric food frequency of consumption and the technological devices daily usage. Similar considerations hold for the only two quantitative variables.

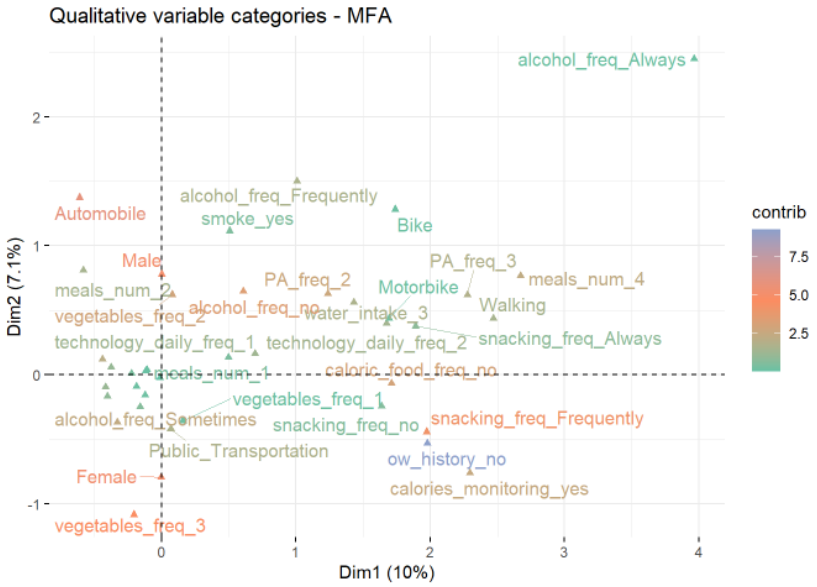


Figure 3.5: Relations among qualitative variables. Correlation between dimensions and qualitative variables, coloured according to their contribution.

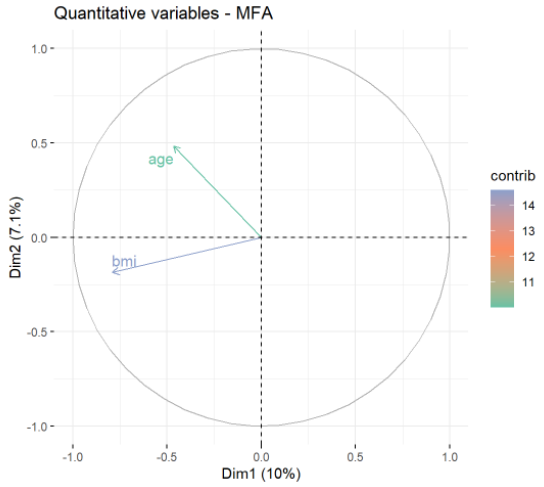


Figure 3.6: Relations among quantitative variables. Correlation between dimensions and quantitative variables, coloured according to their contribution.

Ultimately, it is possible to see how individuals with similar profiles are close to each other on the following factor map (3.7). As described before, the first dimension better represents individuals who neither eat caloric food frequently nor make excessive use of technology; in fact, for example, the respondent whose answer are stored in the 465 row reflects these habits.

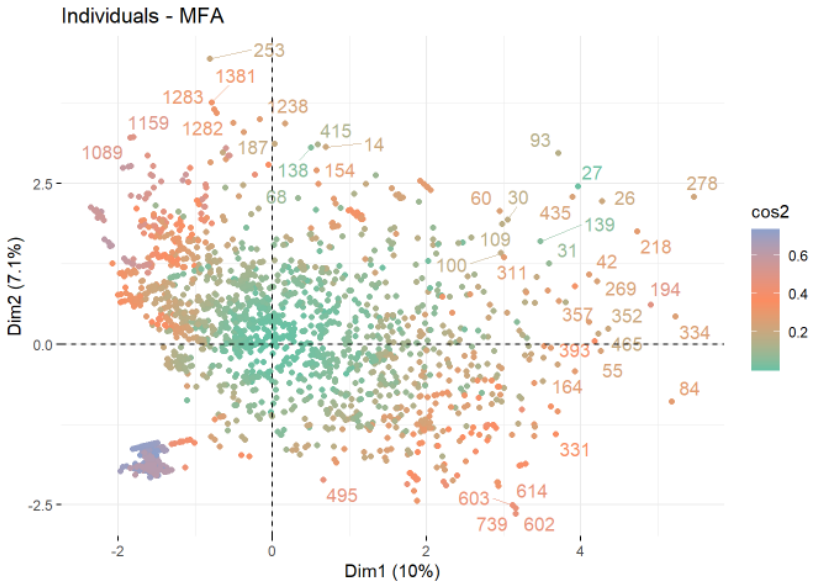


Figure 3.7: Relationship between individuals, coloured according to the quality of their representation.

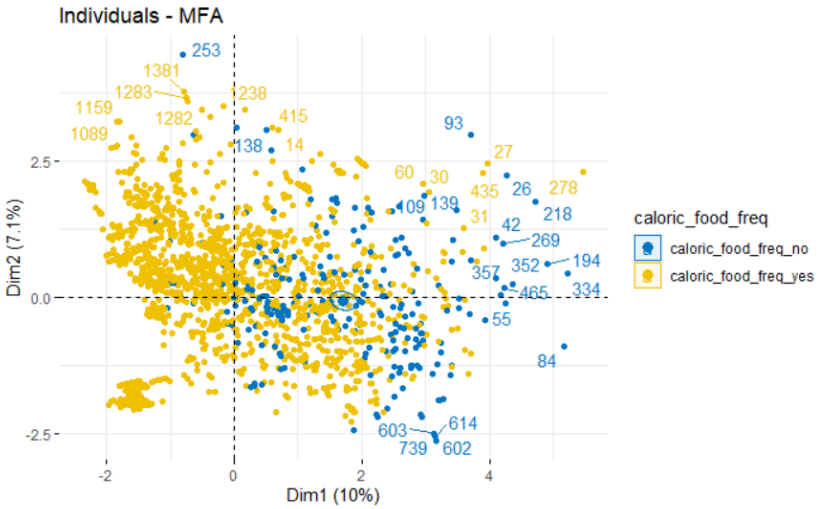


Figure 3.8: Relationship between individuals, coloured according to their answer about physical activity.

On the whole, it makes no sense to reduce the dimensionality of the dataset, as more then half of the variables are required to explain at least 50% of the variability of the data. Nevertheless, a few variables that might play an important role in clustering have been identified and the relationships between them have been highlighted.

Chapter 4

Clustering

As anticipated before, ideally the aim of this paper is to define four different respondents' profiles on the basis of the answers that have been given to the survey. Intuitively, it is natural to expect that BMI will be really incisive in the distinction of individuals, as well as their gender, height, and their sedentary lifestyle. More precisely, it makes sense to expect these different clusters:

- "Healthiest": people with a healthy BMI, prevalently young, with a rather active lifestyle as well as laudable eating habits;
- "Healthy": people with a borderline high BMI, conducting an occasionally active lifestyle with healthy eating habits;
- "Unhealthy": individuals with a high BMI, bad eating habits and a sedentary lifestyle;
- "At risk": respondents with a really excessive BMI and "self-damaging" daily routines.

Later in this report, one will see if the data will confirm or disproof this hypothesis.

4.1 Hierarchical clustering

While the dimensionality reduction techniques aim at representing the observations in a low-dimensional space in a way that explains the variance to a good extent, clustering looks for homogeneous subgroups among data points, in other words it seeks to find clusters so that the observations within each of them are similar. It is clear that on the basis of the type of data one is dealing with, *similar* means something different.

In reality, in hierarchical clustering the number of clusters desired is not known in advance, as it is a method that returns a dendrogram[1] that allows to see at once all the possible clusters that could be obtained as the number of groups varies from 1 to n, where n is the number of observations. The most common type of hierarchical clustering is called *agglomerative*, because the resulting dendrogram is built starting from the leaves (data points) and combining them into subgroups, on the basis of a certain distance matrix. More rigorously, the algorithm adheres to the following steps:

- it starts with each point in its own cluster,
- it identifies the two closest[2] clusters to merge them in a single one,
- it repeats until all the points are in a single cluster.

There are different types of linkage:

- Complete: computes all the pairwise dissimilarities between observations of two different clusters and record the largest of them (maximal inter-cluster dissimilarity);
- Single: computes all the pairwise dissimilarities between observations of two different clusters and record the smallest of them (minimal inter-cluster dissimilarity);
- Average: computes all the pairwise dissimilarities between observations of two different clusters and record the average of them (mean inter-cluster dissimilarity);
- Centroid: computes the dissimilarity between the centroids[3] of two different clusters.

[1] A tree-like visual representation of the data points.
[2] Whatever closest means, based on the type of distance considered.
[3] Vectors containing the means of the variables for the observations in that cluster.

- Ward: computes the total within-cluster variance and minimises it; at each step it finds the pair of clusters that leads to the minimum increase in total within-variance[4] after merging.

In this case, the last method leads to the best result in terms of clusters division, and it is applied together with Gower's distance as a measure of dissimilarity.

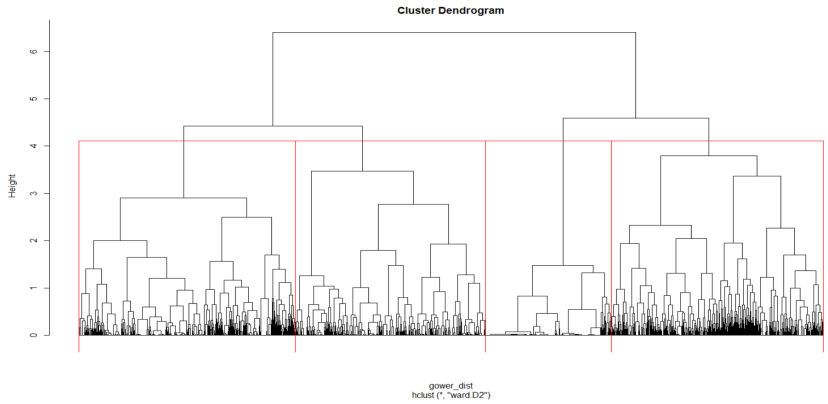


Figure 4.1: Hierarchical clustering on mixed-type data using Gower's distance and Ward's linkage method.

The Gower's distance has been chosen as it allows to measure how different two records are, which makes it particularly useful in this case of mixed-type data. The records may be a combination of numeric, categorical or text data, and the resulting distance is always a number between 0 and 1. Formally, the distance is defined as follows: for each feature $k = 1, \dots, p$, a score $s_{ijk} \in [0, 1]$ is defined[5]. If x_i and x_j are close to each other along feature k, then the score s_{ijk} is close to 1. Conversely, if they are distant along feature k, the score s_{ijk} is close to 0. The type of feature k determines how the score is computed. Let also δ_{ijk} be defined as follows: if x_i and x_j can be compared along feature k, then $\delta_{ijk} = 1$, else $\delta_{ijk} = 0$. Then, Gower's distance is just the average of the scores:

$$S_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}\delta_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}.$$

This is how the scores are computed for each type of feature:

[4]Measure of the variability of the observations within each cluster, that is a measure of compactness.

[5]Dissimilarity between the i-th and the j-th observation along the k-th feature.

- Quantitative variables: $s_{ijk} = 1 - \dfrac{|x_{ik} - x_{jk}|}{R_k}$, where R_k is the range of the feature k.
- Qualitative variables: $s_{ijk} = \mathbb{1}\{x_{ik} = x_{jk}\}$.
- Dichotomous variables: given two individuals, the absence of both is not considered as a match. More precisely

i	1	1	0	0
j	1	0	1	0
s_{ijk}	1	0	0	0

i	1	1	0	0
j	1	0	1	0
δ_{ijk}	1	1	1	0

Observe that often $\sqrt{1 - S_{ij}}$ is returned, and this is called a distance because it is non-negative, symmetric and it satisfies the triangle inequality.

cluster	gender	age	bmi	oh	mtrans	meals	PA	cal_food
1	Female	21	23.50	yes	Public_Trans	3	0	yes
2	Male	22	28.30	yes	Public_Trans	3	0	yes
3	Male	30	31.10	yes	Automobile	3	0	yes
4	Female	24	40.50	yes	Public_Trans	3	0	yes

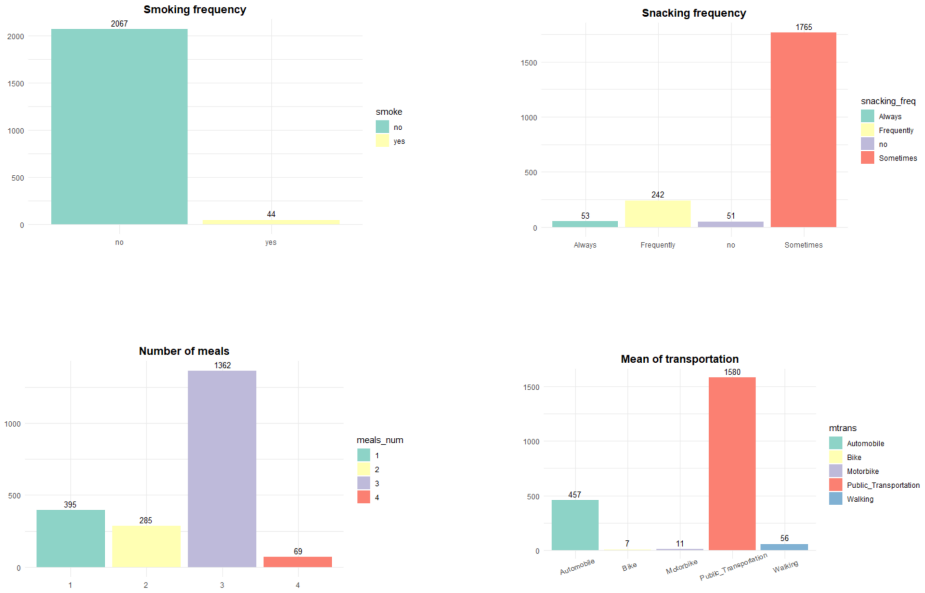
smoke	snacking	vegetables	water	cal_monitoring	tech	alcohol
no	Sometimes	2	2	no	0	no
no	Sometimes	2	2	no	0	Sometimes
no	Sometimes	2	2	no	0	no
no	Sometimes	3	2	no	0	Sometimes

Apparently, there are two main subgroups, based on the gender of the members. Each one of them is then divided into two clusters, one with a lower mean age and the other one with people who are older on average[6]. Furthermore, in both the clusters with younger people, the BMI is lower, especially in the first one, prevalently made of female individuals. Observe how alarming is the fact that in every cluster the majority of the individuals has had, or currently has, an overweight relative.

These results are quite satisfying, yet there is something unusual in the similarity of the most frequent answers given within each group, as they show little or no difference at all between distinct clusters. The reason why the four clusters have so many common modes, even though they appear to be

[6]In the table the median of the age is reported.

well distinct in terms of physical appearance, lies in the nature of the answers. In the figures below one can see a few examples of how unbalanced the frequencies are, so even if people are well divided into these four clusters due to their physical status, they seem to share the same habits. Notice that the questionnaire is definitely quick to complete, so any kind of respondent fatigue, which is a widespread phenomenon that occurs when survey participants become tired of the task, should be excluded. Perhaps, the original participants tended to select a sort of neutral midpoint as an answer and this led to the unbalanced distribution of answers, or maybe this could also be a result of the oversampling done by the owners of the data.



4.2 K-prototypes

Now another clustering method will be used in an attempt to retrieve four different profiles, in order to compare the results with the ones of the hierarchical clustering.

K-prototypes is an algorithm that, similarly to k-means, iteratively recomputes cluster prototypes and reassigns clusters to each observation in the dataset. This algorithm is chosen because it can handle mixed-type data, as the clusters are assigned using the simple distance

$$d(x, y) = d_{euclid}(x, y) + \lambda d_{simplematch}(x, y),$$

where the *simple matching* coefficient is the simplest way of measuring similarity, being 1 in case of exact match, 0 otherwise and λ is a positive parameter to trade off between Euclidean distance and simple matching coefficient. Cluster prototypes are computed as cluster means for numeric variables and modes for factors.

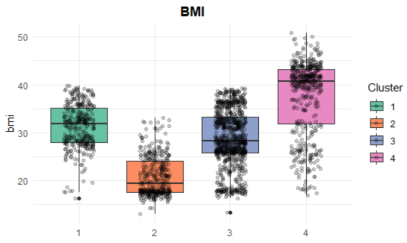
clusters	Freq	gender	age	bmi	oh	cal_food	vegetables	meals
1	347	Male	29	31.33	yes	yes	2	2
2	385	Female	20	20.78	no	yes	2	3
3	905	Male	24	28.76	yes	yes	2	3
4	474	Female	24	37.55	yes	yes	3	3

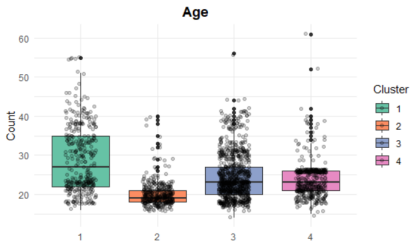
snacking	smoke	water	cal_monit	PA	tech	alcohol	mtrans
S/times	no	1	no	0	0	no	Public_Trans
S/times	no	1	no	0	1	S/times	Public_Trans
S/times	no	2	no	0	0	S/times	Public_Trans
S/times	no	2	no	0	0	S/times	Public_Trans

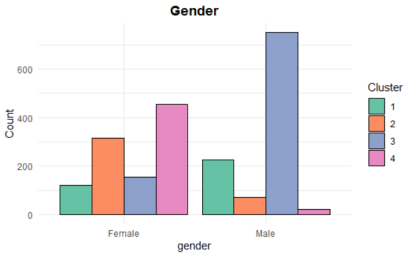
The results[7] in term of clusters division are really similar to the ones observed with the hierarchical clustering approach. In addition, there is a cluster with no family history of overweight individuals, in which mostly females with an healthy BMI are included. It is interesting to explore further these subgroups. They are all prevalently made up of non-smokers, eating highly caloric food frequently, using public means of transportation and occasionally drinking alcohol as well as snacking. Notice also that nearly none of the respondents monitors his calories intake. However, there are a few interesting variables that can give an insight into the composition of the different clusters.

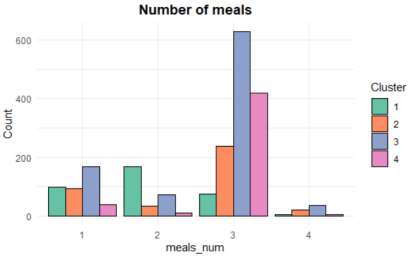
[7] In this case the arithmetic mean of the age is computed, instead of using the median.

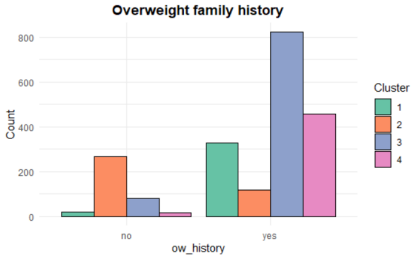
4.2 K-prototypes

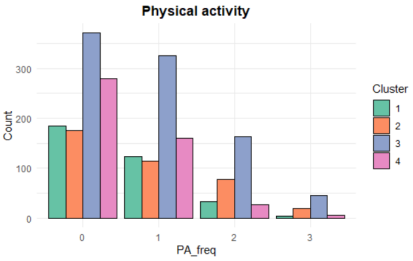


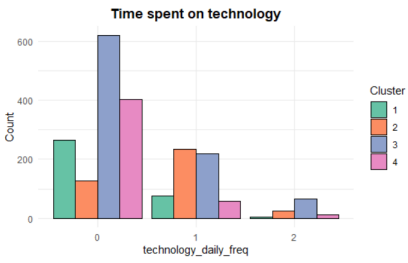


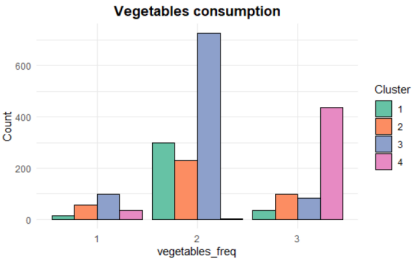


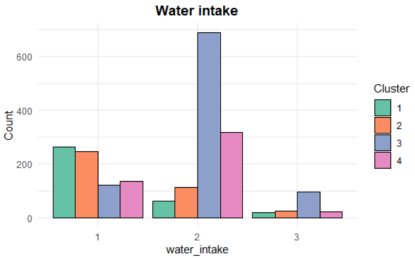


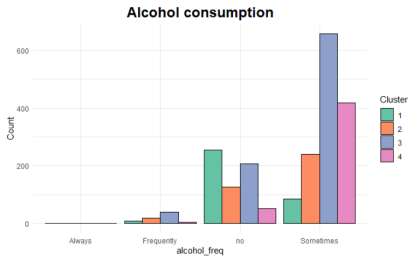












All of the figures above help define better the respondents' profiles in each cluster, and this will be done in detail in the conclusive chapter. Ultimately, in the figure below one can see the points divided into the four clusters that have just been obtained. Observe that the Rtsne() function has been used, which essentially takes a high-dimensional dataset and reduces it to a low-dimensional graph that retains a lot of the original information. Since it has been showed that each principal component only explains a tiny portion of the variances, this graph is not expected to present a perfect division between points, however it still manages to achieve a reasonable result.

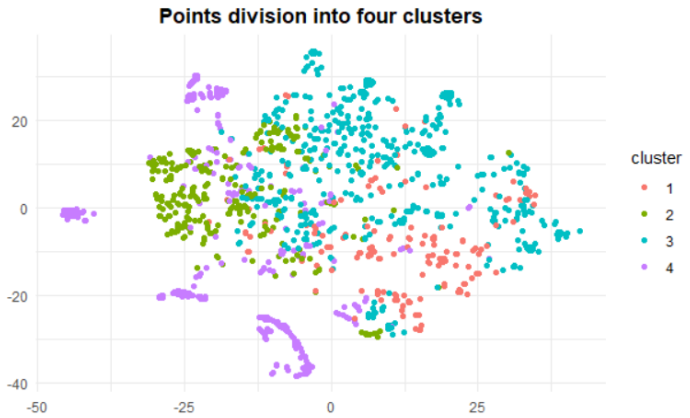


Figure 4.2: Clusters division obtained with the K-prototypes algorithm, visualised in a lower dimensional space thanks to the Rtsne() function.

Chapter 5

Conclusions

As repeatedly stated, the aim of this project is to define four different respondents' profiles, according to their physical status, their eating habits and their lifestyle. Both with mixed-type and categorical-only data, interesting results have been achieved and they appear to confirm the hypothesis made at the beginning of Chapter 4. According to the k-prototypes approach on mixed-type data, it is possible to say that these profiles have been identified:

- Cluster 1 (347): The least numerous one, largely made up of males around 30 year old, with an average BMI greater than 30, meaning they are at level I obesity. They usually consume two meals per day, and sometimes snack. Most of them has a case of overweight in their family history and they are prevalently inactive, which could explain why they have such an high BMI, given they only have 2 meals per day. Finally, they spend less than 2 hours on technological devices, which is rare nowadays, and they drink an insufficient amount of water with respect to their size, therefore they are likely to be badly hydrated. ("Unhealthy")

- Cluster 2 (385): Comprising mostly very young females with an healthy average BMI of about 21, notwithstanding their sedentary lifestyle and the higher number of meals with respect to Cluster 1. They have no overweight cases in their family, which seems to confirm the importance of the familiar environment and partly explains the fact that they might have a faster metabolism and good genetics, of course thanks to their young age, too. In this cluster there is the majority of people spending more time on technological devices compared to the others. ("Healthiest")

- Cluster 3 (905): The most numerous one, made up chiefly of male

individuals who are younger than the ones in the first group, with a BMI that classifies them as overweight. They usually have three meals as well, with occasional snacks an extremely low physical activity rates. Similarly to the other respondents in the first cluster, they spend less time on technological devices. ("*almost* Healthy")

- Cluster 4 (474): Comprising mainly female individuals with an extremely high BMI, therefore risking their own life, as being obese at level III means depriving your lifespan of entire decades and increasing the chance of suffering from vital organs diseases. Basically none of these individuals comes from a family without an overweight case, but unexpectedly, a larger number of people from this subgroup declared to be sometimes more active with respect to people from the first and the second cluster ("At risk").

Notice how there is not a cluster in which the majority of people is underweight, the absence of which could be interpreted in two different ways: an optimistic one and a more realistic one. The first suggests that fewer people are underweight nowadays, yet, on the other hand, it is widespread knowledge that eating disorders are constantly increasing; conversely, the second one confirms the excessive and alarming percentage of the population living its life as a supersized individual. Ultimately, it is safe to say that almost each one of the profiles that were expected to emerge, actually appeared from this analysis. Furthermore, also the hierarchical clustering delineated almost the same subgroups, only highlighting slightly different habits in terms of alcohol consumption and daily dose of water intake within each cluster.

The reason why both eating habits and lifestyles are quite similar between groups with significantly different BMI's might lie in people's unconsciousness about how easy it is to exceed with food and how hard it is not to be sedentary (working out a few times per week has nothing to do with sedentary lifestyles, which are more about the NEAT and how each individual moves throughout the day ([3])). This assumption seems to be confirmed by the fact that the majority of the individuals are overweight or obese, a worrying trend that has slowly begun to interest every single country, in particular the less developed ones in which people cannot afford to lead a healthy lifestyle, especially in terms of diet.

Appendix A

Categorical data

Before drawing the conclusions of this work, every variable is transformed into a factor. All the answers to the multiple choice questions are labelled with a representative string, and also original open-ended questions about age, height and weight are replaced by two multiple choice questions, as if the respondent had the possibility to choose both an age and a BMI range, with five and six categories respectively:

- What is your age?
  - Between 11 and 21
  - Between 21 and 30
  - Between 31 and 40
  - Between 41 and 50
  - Between 51 and 61
- What is your Body Mass Index[1]?
  - Less than 18.5 (*Underweight*)
  - Between 18.5 and 25 (*Normal*)
  - Between 25 and 30 (*Overweight*)
  - Between 30 and 35 (*Obesity level I*)
  - Between 35 and 40 (*Obesity level II*)
  - More than 40 (*Obesity level III*)

[1] Within brackets the idea behind the division, not shown in the hypothetical survey

This is done in order to apply the k-modes algorithm, that is intended to work with categorical variables, as well as to understand from a different perspective how influential the BMI category is in the definition of the different clusters. First and foremost, in the figure below one can see the relationship between all the categories of the different variables, as a result of a homogeneity analysis, whose objective is to make a joint plot in a p-dimensional space of the categories of all variables[2]. Considerations similar to the ones done in Chapter 2 hold; according to the graph, for example, sporty people tend to drink more than 3 litres of water per day, which makes sense.

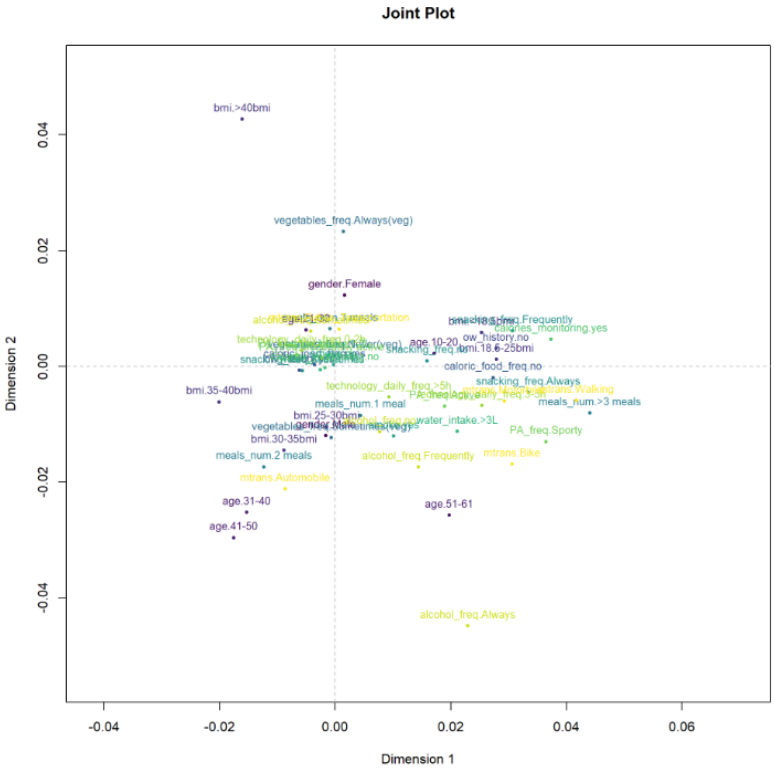


Figure A.1: Similarity between all the categories of all the variables.

Then, in order to apply the k-modes algorithm with the best k[3], the elbow method is used. This technique consists in plotting the within cluster difference, that is a measure of the variability of the observations within each

[2]This is done using the homals() function, that performs a multiple correspondence analysis, similar to the one that has been reported before. In this case $p = 2$.

[3]Even if there is no best k in a clustering problem, as it depends also on personal objectives.

cluster. The "optimal" value of k is the one associated with a low within difference and at the same time it is such that $k+1$ does not significantly reduce the within difference, thus forming an "elbow" in the graph.

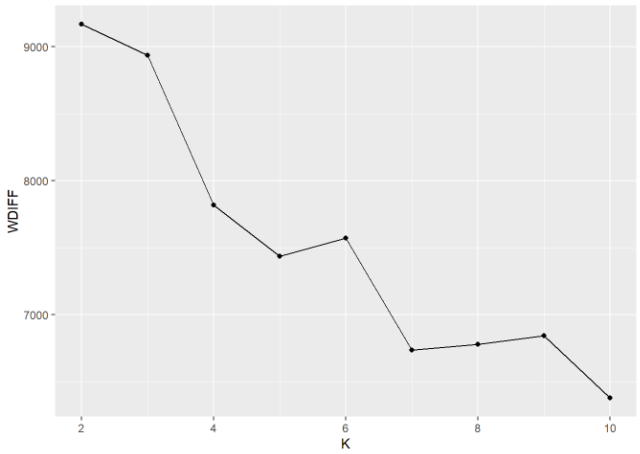


Figure A.2: Possible elbow where k=5.

Hence, the k-modes algorithm is applied with $k = 5$. In order to visualise the resulting clusters and the role of the individuals' physical status in this division, the two initial variables *height* and *weight* are plotted and coloured on the basis of the cluster to which each person with a certain pair of height and weight belongs. It is safe to state that k-modes does not achieve a remarkable result (figure A.3) and, in the table below, one could also see how clusters are not significantly distinct from each other.

Figure A.3: Data points divided into 5 clusters obtained via k-modes algorithm.

cluster	gender	age	bmi	oh	cal_food	vegetables	meals	smoke
1	Male	21-30	25-30bmi	yes	yes	S/times	3 meals	no
2	Female	21-30	>40bmi	yes	yes	Always	3 meals	no
3	Male	31-40	35-40bmi	yes	yes	S/times	3 meals	no
4	Male	10-20	25-30bmi	yes	yes	S/times	3 meals	no
5	Male	21-30	35-40bmi	yes	yes	S/times	3 meals	no

snacking	water	cal_monit	PA	tech	alcohol	mtrans
S/times	1-2L	no	Inactive	>5h	no	Public_Trans
S/times	<1L	no	Inactive	0-2h	S/times	Public_Trans
S/times	1-2L	no	Inactive	0-2h	S/times	Automobile
S/times	1-2L	no	Active	0-2h	no	Public_Trans
S/times	1-2L	no	Moderately active	0-2h	S/times	Public_Trans

Therefore a hierarchical clustering using Gower's distance and Ward linkage method is performed on the same categorical data and the visualisation gets slightly better (figure A.4). In fact, we can see a well distinct cluster and even though the other ones are overlapped, there is still a bit of coherence that seems to confirm the major role played by the physical status in the identification of the different respondents' profiles.

cluster	gender	age	bmi	oh	mtrans	meals	PA
1	Male	21-30	25-30bmi	yes	Public_Trans	3 meals	Inactive
2	Female	10-20	18.5-25bmi	no	Public_Trans	3 meals	Inactive
3	Male	31-40	35-40bmi	yes	Automobile	3 meals	Inactive
4	Female	21-30	>40bmi	yes	Public_Trans	3 meals	Inactive

cal_food	smoke	snacking	vegetables	water	cal_monit	tech	alcohol
yes	no	S/times	S/times	1-2L	no	0-2h	S/times
yes	no	S/times	S/times	<1L	no	0-2h	S/times
yes	no	S/times	S/times	<1L	no	0-2h	S/times
yes	no	S/times	Always	1-2L	no	0-2h	S/times

Figure A.4: Data points divided into 4 clusters obtained via hierarchical clustering.

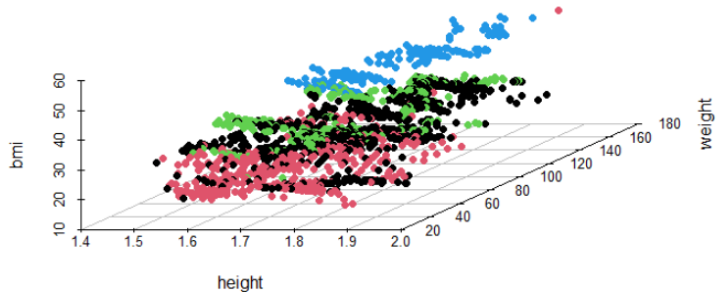
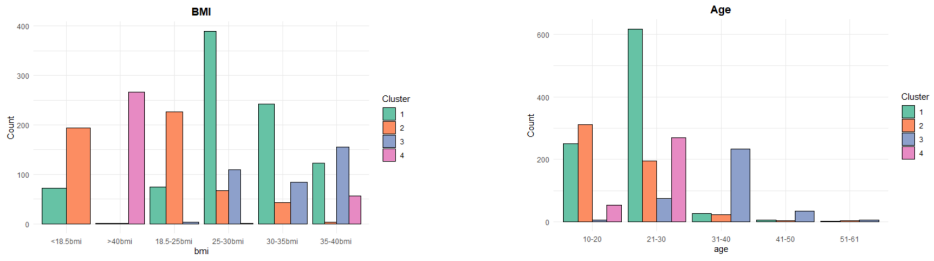


Figure A.5: Data points divided into 4 clusters obtained via hierarchical clustering, visualised in space.

The profiles obtained according to this division in subgroups are coherent with the ones delineated by the hierarchical clustering performed on the mixed-type data. In the following page, the same variables are reported, so that a direct comparison can be performed.



BMI is still a crucial feature that is taken into account when dividing the respondents into clusters. In the following graphs one can observe at a glance a few relevant differences between distinct groups: for example, the absence of male individuals in the fourth cluster, which comprises obese females only, as opposite to the first cluster prevalently made of males; on top of that, morbidly obese women using only public means of transport are found in the fourth subgroup, whereas people in the third cluster mostly use cars.

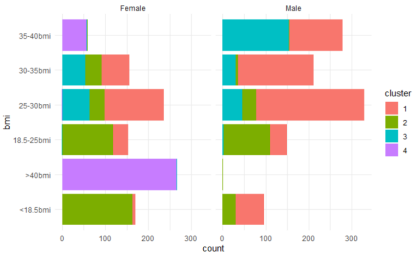


Figure A.6: Bmi division in clusters, according to gender.

Figure A.7: Bmi division in clusters, according to means of transportation.

On the whole, categorical data proved how the features that commonly considered as related by people, are actually "similar". K-modes proved to be not the best method to reveal the four different profiles this paper has been looking for, but the hierarchical clustering has confirmed the ones that have been found in the mixed-type data approach, with subtle differences.

Bibliography

[1] E. De-La-Hoz-Correa *et al.*, *Obesity Level Estimation Software based on Decision Trees*, Journal of Computer Science 2019, 15(1): 67.77

[2] F. Mendoza Palechor, A. de la Hoz Manotas, *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*, Data in Brief, Volume 25, 2019

[3] Chung N, Park MY, Kim J, Park HY, Hwang H, Lee CH, Han JS, So J, Park J, Lim K., *Non-exercise activity thermogenesis (NEAT): a component of total daily energy expenditure.* J Exerc Nutrition Biochem, 2018 Jun 30; 22(2): 23-30.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, Springer.