UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Data Analysis

Nutrition Study

Matija Jakovac, Dunja Petrović, Léo Serra

Barcelona

June 1, 2025

Index

1 Introduction

This project explores multivariate data analysis techniques to investigate complex relationships within a dataset titled Nutrition Study, which includes observations from 315 individuals. The data focuses on nutrition and health-related variables, providing a foundation for examining patterns and associations across multiple dimensions.

The goal of this analysis is to uncover underlying structures in the data, reduce dimensionality where appropriate, and generate insights that contribute to a deeper understanding of the nutritional and health profiles of the individuals studied.The analysis aims to not only describe the structure of the data but also identify key factors influencing outcomes, offering a comprehensive understanding of the studied phenomenon.

A variety of multivariate statistical techniques were applied to explore relationships in the dataset. **Principal Component Analysis (PCA)** was used to reduce dimensionality and identify key patterns in the filtered subset of variables. **Multidimensional Scaling (MDS)**, both classical and based on Gower distance, was performed to visualize the relative positioning of observations based on dissimilarities. **Multiple Correspondence Analysis (MCA)** was employed for exploring associations between categorical variables. **Cluster analysis**, including both hierarchical and non-hierarchical methods, was applied to uncover natural groupings within the data. **Discriminant analysis**, using both Linear (LDA) and Quadratic (QDA) methods, aimed to classify individuals based on dietary and lifestyle variables. **MANOVA** was conducted to assess whether multivariate means differed by smoking status and vitamin use, with an additional comparison using raw data. Finally, **Hotelling's T^2 test** was used to test for significant multivariate differences between two groups, providing a formal test of mean vector equality.

Variables in the Dataset:

- **ID** – Identifier number for each individual
- **Age** - Subject's age (in years)
- **Smoke** – Whether the individual currently smokes (Yes/No)
- **Quetelet** – Weight divided by height squared: weight\height^2
- **Vitamin** – Coded as: 1 = Regularly, 2 = Occasionally, 3 = No
- **Calories** – Daily calorie intake
- **Fat** – Grams of fat consumed per day
- **Fiber** – Grams of dietary fiber per day
- **Alcohol** – Number of alcoholic drinks consumed per week
- **Cholesterol** – Cholesterol intake (mg per day)
- **BetaDiet** – Beta-carotene from food (mcg per day)
- **RetinolDiet** – Retinol intake from food (mcg per day)
- **BetaPlasma** – Beta-carotene concentration in blood (ng/ml)
- **RetinolPlasma** – Retinol concentration in blood (ng/ml)
- **Sex** – Coded as Male or Female
- **VitaminUse** – Text-coded version of Vitamin (No, Occasional, Regular)
- **PriorSmoke** – Smoking history: 1 = Never, 2 = Former, 3 = Current

2 Preprocessing of the dataset

We first standardize our data types, turning text fields into factors and ensuring integer columns are treated as continuous numerics, then isolate only the truly numeric variables. This guarantees that subsequent correlation and KMO analyses run smoothly on a consistent, all-numeric matrix.

2.1 Examine the data

To make PCA and MDS informative, we must remove both redundant and uninformative variables. If we leave these redundancies in place, the first few dimensions will simply reflect the same shared variance over and over—resulting in cramped, uninterpretable plots. Redundant variables (those almost perfectly correlated) exaggerate the same signal, while uninformative ones (those sharing little variance with the rest) add noise. We'll therefore (a) trim by pairwise correlation, then (b) assess and drop low-shared-variance variables via KMO.

a) Correlation

We inspect the correlation matrix to flag pairs with |r| above our cutoff (e.g. 0.7) and then drop one member of each pair. By removing one variable from each such pair, we ensure that the remaining variables capture distinct sources of variation, letting PCA and MDS reveal genuine structure rather than echo the same signal.

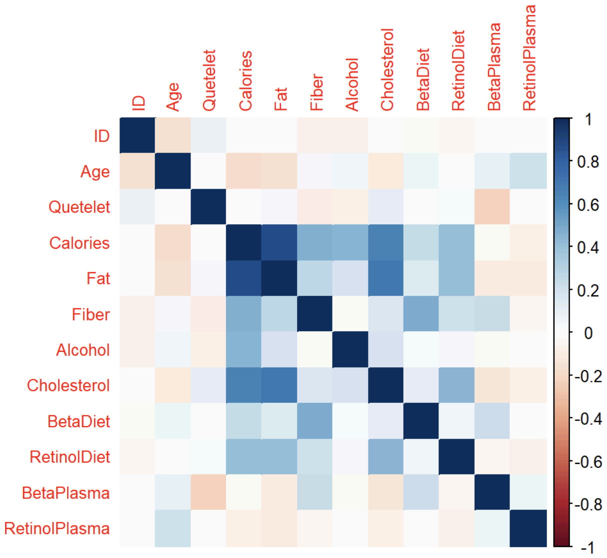


Figure 2.1.1.: Dataset correlation matrix

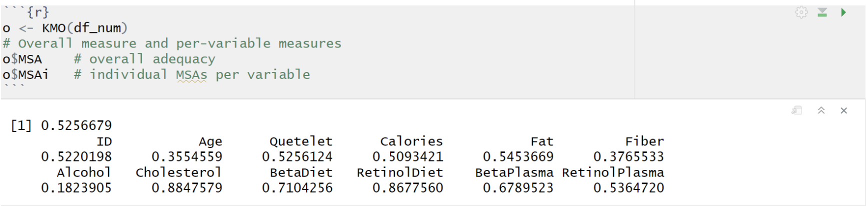
Looking at the correlation matrix you can see that the only really "hot" cell is the dark blue at Calories ↔ Fat ($r \gtrsim 0.8$), confirming those two move almost in lockstep. There are also mild positive correlations between Cholesterol and both Calories and Fat (mid-blue), but beyond that nearly every other pair sits close to zero, indicating most nutrients and biomarkers aren't strongly tied together. In short, Calories and Fat are essentially redundant, while the rest contribute largely independent signals.

b) KMO (Low-Shared-Variance Removal)

Next, we compute the Kaiser–Meyer–Olkin (KMO) statistic to quantify how much each variable’s variance is shared with the others.

- Overall KMO assesses whether the dataset as a whole is suitable for factor-based methods (PCA/MDS)—values above \~0.7 are desirable.
- Individual MSAi scores tell us which variables contribute little common variance: those with MSAi below our threshold (e.g. 0.6) are effectively noise in a factor model.

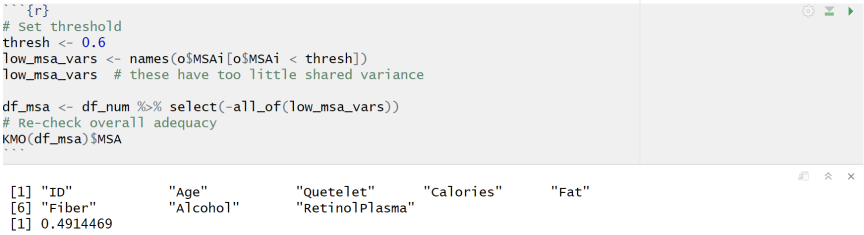
By dropping any variable with a low MSAi, we concentrate on features that meaningfully co-vary with the rest, boosting the clarity and stability of subsequent PCA and MDS results.



~~~
```{r}
o <- KMO(df_num)
# Overall measure and per-variable measures
o$MSA    # overall adequacy
o$MSAi   # individual MSAs per variable
```
~~~

```
[1] 0.5256679
           ID           Age      Quetelet      Calories           Fat         Fiber
    0.5220198     0.3554559     0.5256124     0.5093421     0.5453669     0.3765533
      Alcohol   Cholesterol      BetaDiet   RetinolDiet    BetaPlasma RetinolPlasma
    0.1823905     0.8847579     0.7104256     0.8677560     0.6789523     0.5364720
```

Figure 2.1.2: KMO



~~~
```{r}
# Set threshold
thresh <- 0.6
low_msa_vars <- names(o$MSAi[o$MSAi < thresh])
low_msa_vars  # these have too little shared variance

df_msa <- df_num %>% select(-all_of(low_msa_vars))
# Re-check overall adequacy
KMO(df_msa)$MSA
```
~~~

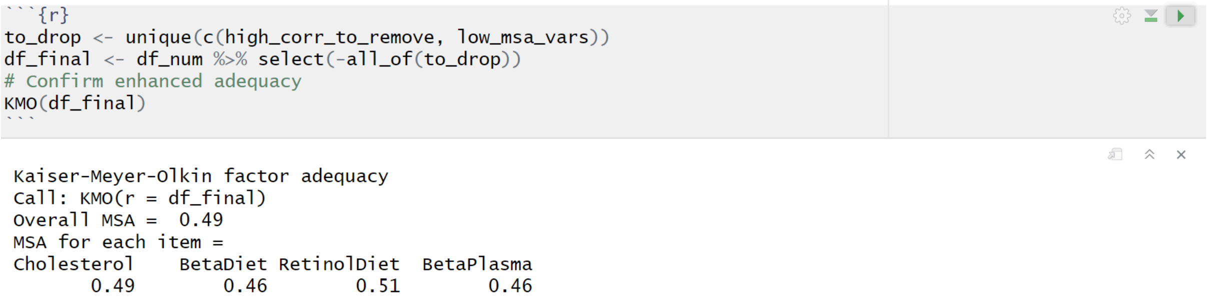
```
[1] "ID"            "Age"           "Quetelet"      "Calories"      "Fat"
[6] "Fiber"         "Alcohol"       "RetinolPlasma"
[1] 0.4914469
```

Figure 2.1.3: Reducing the dataset by shared variance

Even after dropping those low-MSAi variables, the overall KMO remains below 0.6—so we still don’t have ideal sampling adequacy. For now, we’ll proceed with this reduced set and see whether the cleaned data yields any interpretable structure in the PCA and MDS.

2.2 Final Subset After Correlation + MSA Filtering

At this stage, we merge our two exclusion lists, low-MSAi variables and highly collinear pairs, to arrive at a concise, robust set of features. This compact, well-behaved subset is now ideal for running clear, interpretable PCA and MDS.



~~~
```{r}
to_drop <- unique(c(high_corr_to_remove, low_msa_vars))
df_final <- df_num %>% select(-all_of(to_drop))
# Confirm enhanced adequacy
KMO(df_final)
```
~~~

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = df_final)
Overall MSA =  0.49
MSA for each item = 
Cholesterol    BetaDiet RetinolDiet  BetaPlasma 
       0.49        0.46        0.51        0.46 
```

2.2.1 Final subset of dataset

3 PCA on the Filtered Subset

Principal Component Analysis (PCA) is a dimensionality-reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated "principal components," each capturing a descending proportion of the dataset's total variance. By projecting high-dimensional data onto these new axes, PCA helps us uncover the most salient patterns or clusters and visualize complex relationships in just two or three dimensions.

We've seen that applying PCA to the raw dataset didn't yield much insight—many variables were either redundant or contributed very little shared variance, so the resulting components were hard to interpret. Thus, we first filtered out variables with low sampling adequacy and those that were nearly collinear, producing a leaner subset. Running PCA on this refined set now produces components that more clearly reflect underlying dietary, biomarker, and lifestyle differences among our subjects.

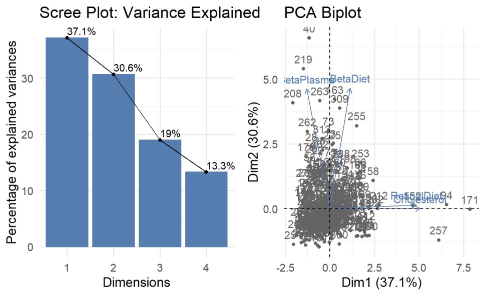


Figure 3.1 : PCA Scree Plot and Biplot

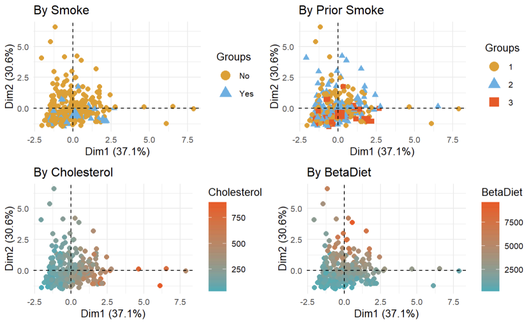


Figure 3.2 : Individual PCA results

From these plots we can see that PC1 now explains 37.1% of the variance, while PC2 explains 30.6%. Taken together, the first two components capture nearly 68% of all variability in our filtered dataset.

In the biplot of PC1 vs. PC2, most individual points remain tightly clustered around the origin, but a few outliers extend far along PC1. The variable arrows show that **Cholesterol**, **RetinolDiet**, and **RetinolPlasma** point almost purely along the positive PC1 axis, indicating that those variables dominate the "unhealthy-diet" direction. By contrast, **BetaDiet** and **BetaPlasma** point straight up along PC2, confirming that PC2 represents a "beta-carotene" axis orthogonal to PC1. In other words, people with high retinol metrics score high on PC1, whereas those rich in beta-carotene score high on PC2.

When we color individuals by **Smoke**, the separation along PC1 becomes striking: non-smokers (gold circles) cluster near PC1 ≈ 0 or negative, while smokers (blue triangles) extend rightward into large positive PC1 values. This shows that the "unhealthy-diet" dimension aligns almost exactly with current smoking status.

Coloring by PriorSmoke shows no clear overall gradient along PC1, although individuals with PriorSmoke = 3 (red squares) do overlap heavily with current smokers on the right side. For the other PriorSmoke levels, there is no consistent pattern along PC1.

Turning to **Cholesterol**, we see a continuous shift of point color from teal (low cholesterol) on the left side of PC1 to burnt orange (high cholesterol) on the right. This confirms that PC1 essentially ranks individuals by their cholesterol levels (and related fat/calorie measures), with higher cholesterol clustered in the same positive-PC1 region as smokers.

Finally, when colored by **BetaDiet**, points with the highest beta-carotene intakes appear at top-center of the plot (high PC2, near PC1 ≈ 0), whereas lower BetaDiet values cluster below. The vertical spread along PC2 confirms that individuals rich in beta-carotene remain distinct from those on the "unhealthy" PC1 axis, illustrating again that PC2 captures vitamin-A precursor intake independently from fat/cholesterol.

In summary, these updated PCA plots show:

- **PC1 (37.1%)** is dominated by Cholesterol and Retinol variables, and it cleanly separates current smokers (and those with high PastSmoke) from non-smokers.

- **PC2 (30.6%)** continues to represent a beta-carotene gradient (BetaDiet, BetaPlasma), largely orthogonal to PC1.

The groupings by **Smoke**, **PriorSmoke**, **Cholesterol**, and **BetaDiet** each align strongly with the respective component that most directly reflects those measures, confirming that the filtered PCA has successfully isolated two key, interpretable axes of variation.

4 MDS (Classical & Grower)

4.1 Multidimensional Scaling (Classical)

As you can see in the following plot, applying classical MDS (using Euclidean distance) to our filtered numeric dataset still fails to reveal any clear grouping or pattern:

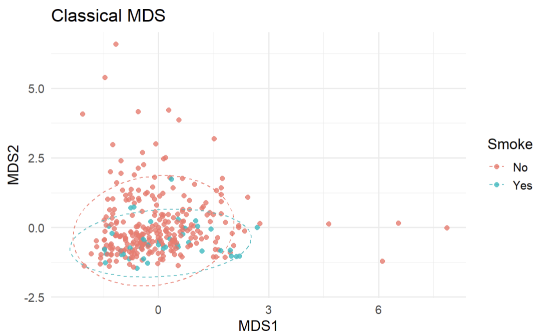


Figure 4.1.1 Classical MDS

That lack of structure, despite having removed redundant and low-variance variables, suggests that Euclidean distance over purely numeric columns does not capture the mixed nature of our data.

4.2 Multidimensional Scaling (Grower)

That is why we have chosen to compute a **Gower distance** (which handles both numeric and categorical variables) and re-run classical MDS on the full set of mixed-type features. By including categorical factors alongside scaled numeric measures, Gower-based MDS produces a more interpretable map in which groupings by smoking status, vitamin use, sex, and other factors become evident.

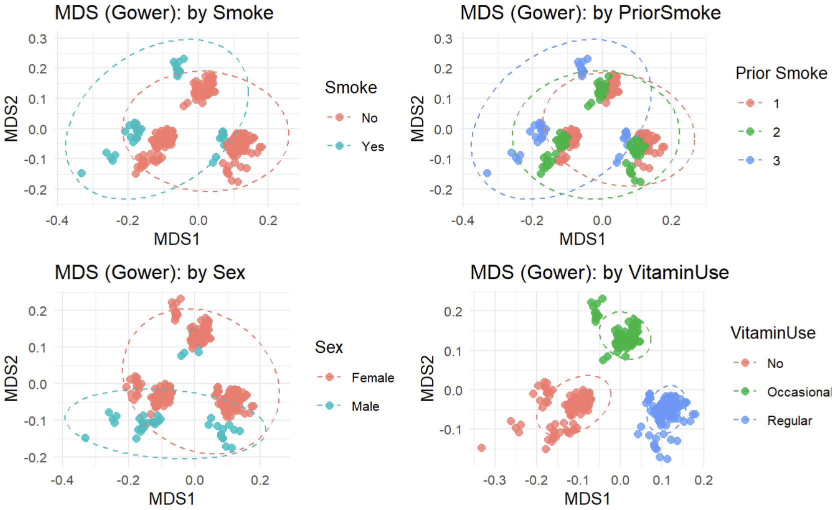


Figure 4.2.1 : Gower MDS

When we map everyone into two-dimensional Gower-MDS space, so that both their numeric and categorical traits feed into the same plot, several clear patterns emerge.

Current smokers and non-smokers form two mostly non-overlapping clusters along MDS1, smokers occupy the right-hand side of the plot, while non-smokers remain on the left, with only a few points in between. This indicates that tobacco use aligns very strongly with the combined dietary, biomarker, and categorical variables used to compute Gower distances.

Plotting the same points by PriorSmoke level, we observe that those with the highest PriorSmoke score (blue circles) again cluster mainly on the far right, overlapping heavily with today's smokers, while those with the lowest PriorSmoke (peach circles) lie on the left. In between, the middle PriorSmoke group (green circles) sits closer to the center. Although the bands overlap somewhat more than the pure Smoke map, it remains apparent that past smoking history aligns closely with current smoking status and its associated diet/biomarker signature.

In the lower-left panel, points colored by Sex reveal a subtler shift. Females form a slightly tighter, more centrally located ellipse, mostly toward the top of the plot, whereas males are more dispersed, particularly along the bottom half and toward the right side. This suggests that male participants show greater diversity in their combined dietary/biomarker/smoking profiles, while females tend to be more homogenous. Sex does not drive a complete split, but it does add a secondary layer of separation within the two main clusters.

Finally, VitaminUse levels now define two nearly orthogonal axes of variation. Along MDS2, Occasional users (green) stand apart from both Regular (blue) and Non-users (peach), forming a middle band that is clearly separated vertically. Meanwhile, MDS1 drives a “goodness” gradient: Non-users cluster on the left (lower MDS1), Occasional users occupy the center, and Regular users lie on the right (higher MDS1). In other words, taking vitamins regularly corresponds with the most “positive” combined dietary/biomarker/smoking profile (highest MDS1), while those who never use supplements occupy the opposite end. The Occasional group sits squarely between them, confirming that vitamin-use frequency is its own coherent dimension, just as strong as smoking status, in our Gower-based MDS representation.

Together, these four MDS maps show that using Gower distance on our mixed numeric + categorical data uncovers clear, meaningful groupings. Smoking status drives the primary axis (MDS1), and PriorSmoke clusters tie in directly to that same dimension. Sex adds a subtler layer of separation, males display broader dispersion, females remain more clustered, while VitaminUse creates a nearly orthogonal gradient. In short, by combining lifestyle, dietary, and biomarker variables into one distance measure, we reveal well-defined patterns that would be invisible if we examined only numeric data.

5 Multiple Correspondence Analysis (MCA)

Before applying the MCA, it is essential to create a dataset that includes only categorical variables, as the initial step of MCA involves one-hot encoding these variables into a binary indicator matrix.
Next step is defining the Burt and Indicator matrices. A Burt matrix is a special type of matrix that summarizes the relationships between all pairs of categorical variables in a dataset. An indicator matrix transforms a dataset of categorical variables into a binary format where each column represents one category (level) of a variable and each row represents one individual (observation). The cell value is 1 if the individual belongs to that category, 0 otherwise. Then, MCA is applied to the Indicator matrix.

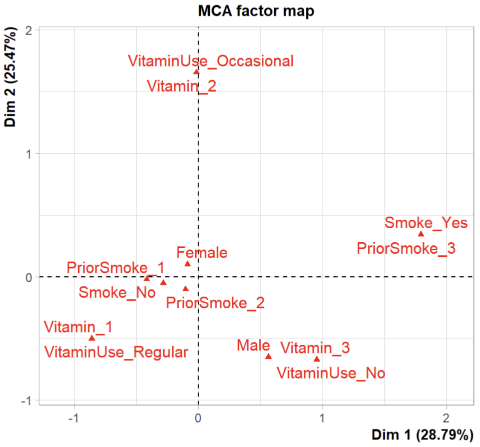


Figure 5.1 : MCA factor map

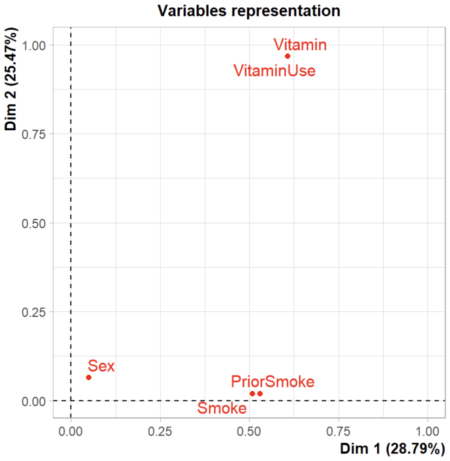


Figure 5.2 : Variables representation

The two MCA factor maps provide a visualization of relationships between categorical variables and individual observations. The left graph displays the variable categories, with Dimension 1 (explaining 28.79% of variance) distinguishing a lifestyle gradient: on the right side are categories such as Smoke_Yes, PriorSmoke_3, Vitamin_3, VitaminUse_No, and Male, while the left groups Smoke_No, PriorSmoke_1, Vitamin_1, VitaminUse_Regular, and

Female. Dimension 2 (25.47% variance) separates categories related to vitamin use frequency, with Vitamin_2 and VitaminUse_Occasional at the top and Vitamin_3 and VitaminUse_No at the bottom. Clusters suggest that smoking males who are prior heavy smokers tend not to use vitamins regularly, while non-smoking females are more likely regular vitamin users. The right graph plots individuals in this same space, where dense clusters correspond to groups sharing similar profiles: those on the far right align with smoking, male, and low supplement use categories, whereas those on the left align with non-smoking, female, and regular vitamin use categories.

To decide which dimensions to retain for further interpretation in MCA, we compare each dimension's percentage of explained inertia to a threshold based on the structure of the data. In this case, we have 5 categorical variables (J = 5), which together produce 13 indicator variables through one-hot encoding (K = 13). The average expected inertia per dimension under the null hypothesis is given by 100/(K - J). Therefore, we keep dimensions whose explained inertia exceeds 12.5%, as they capture more variance than expected by chance.

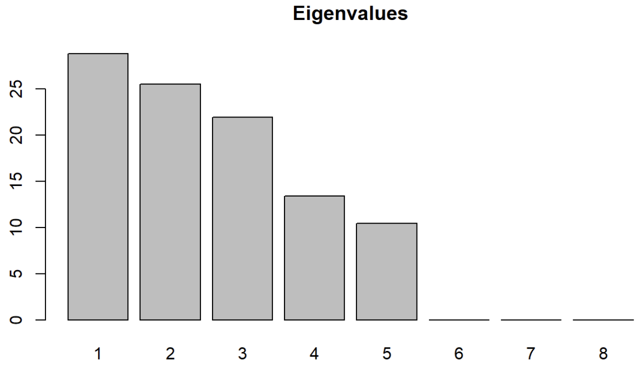


Figure 5.3 : MCA Eigenvalues

Based on the calculation made above, it is indicated we should keep dimensions 1 to 4.

Closer examination of dimensions reveal that the first 4 dimensions capture meaningful associations between categorical variables related to smoking habits, vitamin use, and sex. Dimension 1 is primarily driven by VitaminUse, Vitamin, PriorSmoke, and Smoke, suggesting that these factors jointly explain a large portion of the variation, with strong separation between users and non-users of vitamins, and smokers vs. non-smokers. Dimension 2 is almost entirely dominated by VitaminUse, particularly the "Occasional" and "Regular" categories, indicating that this variable differentiates individuals along a distinct axis. Dimension 3 again highlights PriorSmoke, Smoke, and VitaminUse, but with a different pattern, emphasizing previous smoking behavior. Dimension 4 is influenced most strongly by PriorSmoke and Sex, with moderate contribution from VitaminUse. Overall, the analysis shows that smoking history and vitamin consumption are the most discriminative variables across the dimensions, while Sex plays a more secondary but still relevant role.

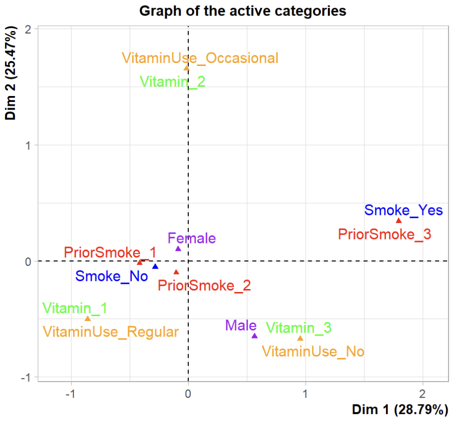


Figure 5.4 : MCA Graph of the active categories

6 Cluster Analysis

6.1 Hierarchical clustering

Hierarchical clustering is performed on scaled numeric data using a distance matrix which is computed with Euclidean distance. The dendrogram was plotted using Ward's method.

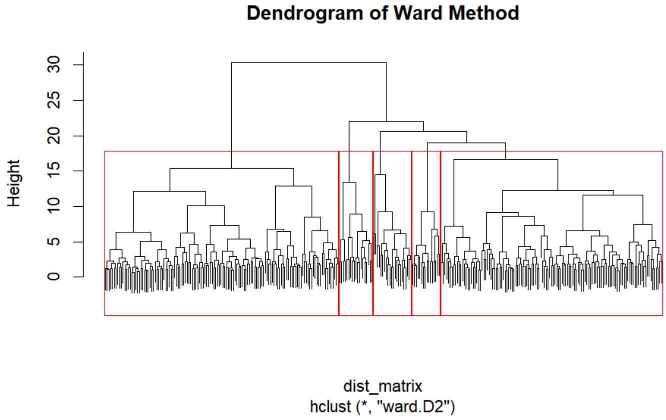


Figure 6.1: Dendrogram of Ward's Method

Examining the given dendrogram we look for the largest "vertical jump" to determine the number of clusters in the dataset. Following that we would choose *k*=2 but choosing two clusters can be too simplistic and misleading. Most of the time that choice is too broad and defeats the purpose of clustering where the goal is to identify groups with minimal internal variation. Second largest jump occurs at ***k*=5** dividing the dendrogram into five different clusters (Figure 6.1).

6.2 Non-hierarchical clustering

Based on the dendrogram, we can estimate a suitable number of clusters and proceed with non-hierarchical clustering (K-Means).. This approach provides insight into which numeric variables contribute most to the differences between clusters.

Cluster	Size	Calories	Fat	Fiber	Alcohol	Cholesterol	BetaDiet	BetaPlasma
1	39	**2770**	**126**	16	**4.34**	**394**	1919	140
2	**121**	**1475**	**64.6**	**9.9**	2.29	202	1018	165
3	47	1945	79.4	16.5	3.3	250	3889	244
4	92	1634	69.5	12.5	**1.98**	221	2237	191
5	15	2072	83	**18.3**	3.09	261	**6593**	**345**

Table 6.1: K-Means means of clusters

Cluster 1 – High-calorie / high-fat eaters: very energy-dense diets with the most fat, alcohol, and cholesterol.
Cluster 2 – Low-intake group: consistently lowest calories, fat, fibre, and β-carotene, indicating generally sparse diets, largest number of people
Cluster 3 – High-fibre, carotene-rich: oldest group, highest fibre and β-carotene from plentiful fruit, veg, and whole grains.
Cluster 4 – Moderate, low-alcohol: nutrient intakes near average but markedly minimal alcohol consumption.
Cluster 5 – Carotene boosters: exceptionally high β-carotene and fibre, suggesting heavy use of carotene-rich foods or supplements.

7 Discriminant Analysis

7.1 LDA on clean data

After verifying the two key prerequisites for Linear Discriminant Analysis—**within-group multivariate normality** and **homogeneity** of covariance matrices—none of the raw nutrient variables satisfied the Shapiro–Wilk normality criterion. Because all violated normality and were strongly right-skewed, every variable was log-transformed; a subset (*Calories*, *Fat*, *Fiber*, *Cholesterol* and *BetaDiet*) was then retained because, in their transformed form, they met both the Shapiro–Wilk and Box's M requirements.

Applying LDA to predict three categories of vitamin use (**No**, **Occasional**, **Regular**) gives the following summary:

VitaminUse	No	Occasional	Regular
Prior prob.	0.35	0.26	0.39
CCR (class)	0.48	0.02	0.64

Table 7.1: LDA results

	No	Occasional	Regular
No	53	3	54
Occasional	34	2	46
Regular	41	3	78

Table 7.2: LDA contingency table

The overall **correct-classification rate (CCR)** is **0.424**, only modestly higher than the random-chance benchmark (**PA = 0.342**), indicating limited discriminative power. Performance is acceptable for **Regular** users (64 % correct) and passable for **Non-users** (48 %), but essentially fails for the **Occasional** group (2 %), suggesting that the selected nutrients do not form a distinctive profile for intermittent vitamin use.

7.2 QDA on raw data

Since the raw data violated the assumption of homogeneity of covariance matrices, Quadratic Discriminant Analysis (QDA) was applied. Although the variables also failed to meet the multivariate normality assumption, QDA was conducted for comparative purposes.

VitaminUse	**No**	**Occasional**	**Regular**
Prior prob.	0.35	0.26	0.39
CCR (class)	0.55	0.49	0.68

Table 7.3: QDAresults

	No	**Occasional**	**Regular**
No	69	23	34
Occasional	24	46	23
Regular	17	13	65

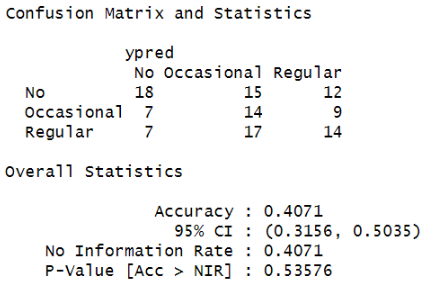
Table 7.4: QDA contingency table

Prior probabilities remained consistent with class distribution as the same entries were used as in LDA. The **CCR** improved relative to LDA indicating better identification across all groups—particularly for ***Occasional*** users, where performance improved substantially. Misclassifications were notably lower for the *Occasional* group compared to LDA, reducing from 98% misclassified to about 43%. This suggests that QDA better captured the distinct covariance structures among classes.

While both LDA and QDA were constrained by violations of normality, QDA performed notably better in practice. The overall classification accuracy was higher under QDA, particularly for the *Occasional* group, which LDA struggled to distinguish (CCR of 2%). QDA's flexibility in allowing class-specific covariance matrices likely contributed to its improved performance. These results highlight that, even under imperfect assumptions, QDA can provide a **more accurate** classification model when group variances differ significantly.

7.3 Naive-Bayes

In accordance with the Naive Bayes assumption of conditional independence among predictors given the class, the variable *Calories* was removed from the raw dataset due to its high correlation with *Fat*, which could violate this assumption. The Naive Bayes model was then trained on a subset comprising 70% of the data and subsequently validated on the remaining 30% test set to evaluate its predictive performance.



```
Confusion Matrix and Statistics

            ypred
             No Occasional Regular
  No         18         15      12
  Occasional  7         14       9
  Regular     7         17      14

Overall Statistics

               Accuracy : 0.4071
                 95% CI : (0.3156, 0.5035)
    No Information Rate : 0.4071
    P-Value [Acc > NIR] : 0.53576
```

Figure 7.1: Confusion matrix and Statistics

The Naive Bayes classifier achieved an overall accuracy of 40.7%, which matches the No Information Rate, indicating that the model performs **no better** than random guessing. The high p-value further confirms that the difference is not statistically significant. Misclassifications were common across all classes, especially between *Occasional* and *Regular* users, suggesting poor separation between groups. Overall, the model shows **limited** predictive value on the test data and may require better feature selection, preprocessing, or an alternative classification approach.

8 MANOVA

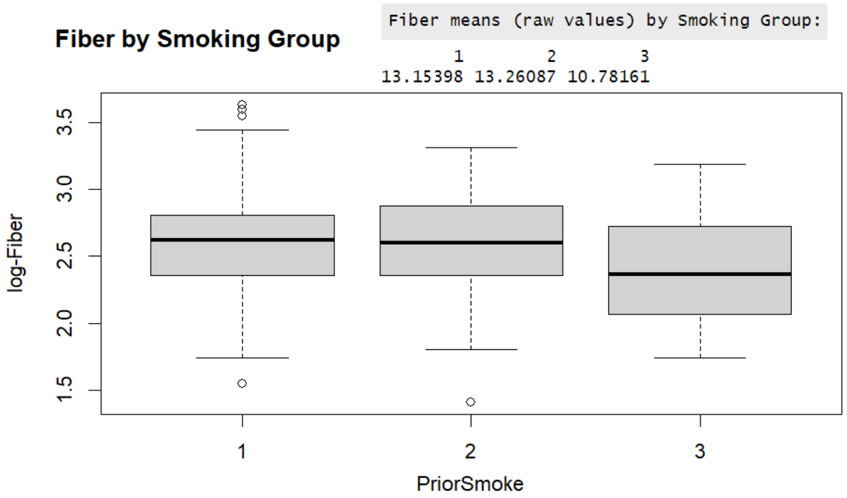
As MANOVA relies on two key assumptions—multivariate normality and homogeneity of covariance matrices—these were tested and addressed as needed. As also shown before, testing revealed that none of the raw variables satisfied the assumption of multivariate normality. Consequently, a logarithmic transformation was applied to correct the observed skewness. Following transformation, a subset of variables (*Calories*, *Fat*, *Fiber*, *Cholesterol*, and *BetaDiet*) was identified as meeting both assumptions and was selected for inclusion in the MANOVA model.

8.1 *PriorSmoke*

A MANOVA was conducted to assess whether dietary variables could differentiate individuals based on their prior smoking status (*Never - 1*, *Former - 2*, or *Current - 3* smokers). Among the variables analyzed, only *Fiber* demonstrated a statistically significant difference across groups ($p \approx 0.005$), while *Fat* approached significance with a *p*-value of approximately 0.057. All other variables showed no meaningful association with smoking status.

Post-hoc Tukey comparisons for Fiber revealed significant differences between *Current* **smokers** and both *Never* and *Former* smokers (adjusted $p < 0.05$), indicating that individuals who currently smoke consume significantly less fiber. This likely reflects broader dietary differences associated with smoking behavior.

Although Fat did not reach statistical significance, adjusted *p*-values between *Current* smokers and *Never* **smokers** were near 0.1, suggesting a potential trend. These findings align with the broader notion that non-smokers tend to follow **healthier lifestyles** compared to current smokers.



```
Fiber means (raw values) by Smoking Group:
       1        2        3 
13.15398 13.26087 10.78161 
```

Figure 8.1: Log-Fiber by Smoking Status boxplot

Never smokers appear to follow a healthier dietary pattern compared to both former and current smokers. They consume significantly more fiber—about 3 grams more, 13g compared to 10g, per day than current smokers—which may reflect a more balanced and health-conscious diet. Additionally, they tend to consume up to 7 grams less fat, 67g compared to 74g, than those with a history of smoking, further supporting the link between non-smoking and healthier lifestyle habits.

8.2 *VitaminUse*

A MANOVA was conducted using the same set of dietary variables to evaluate whether nutrient intake differed significantly across levels of vitamin use, categorized as *No*, *Occasional* and *Regular use*. The purpose was to explore whether patterns of vitamin supplementation are associated with broader differences in dietary habits.

The results of the MANOVA were **not statistically significant**, with all *p*-values exceeding 0.15. The lowest *p*-value was observed for **BetaDiet**, which is plausible given the strong link between vitamin use and β-carotene intake. However, even this relationship did not reach statistical significance. This outcome illustrates an important point: **MANOVA does not always give significant findings**, and the absence of group-level differences suggests that, in this case, vitamin use may not be strongly associated with distinct dietary patterns.

8.3 Raw data

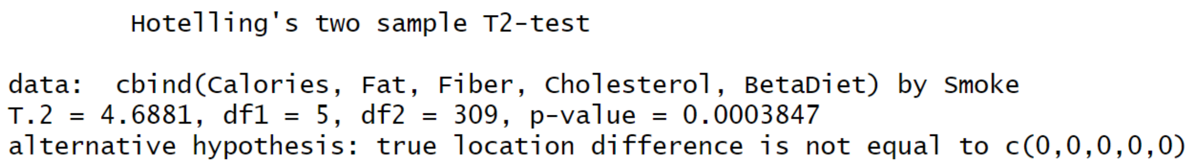
A MANOVA was also performed on the raw (untransformed) data, despite the fact that it did not meet the assumption of multivariate normality and only marginally satisfied the assumption of homogeneity of covariance matrices. This analysis was conducted for comparative purposes, as it retains the original scale of the variables. The results were largely consistent with prior findings, reinforcing the observation that **non-smokers tend to follow healthier lifestyles**, while **vitamin use** did not show significant differences in dietary patterns.

One notable, though not statistically significant, result was observed for **Quetelet index (BMI)**, which yielded a *p*-value around 0.1. This variable measures body mass and is often used as a proxy for body fat, where higher values typically indicate poorer health outcomes. The largest difference in BMI was between **non-smokers** and **current smokers**, further supporting the trend that non-smokers generally have **healthier dietary habits**.

9 Hotelling’s T2 Test

Similar to MANOVA, Hotelling’s T2 test relies on two assumptions: multivariate normality and homogeneity of covariance matrices. So, for conducting the Hotelling’s T2 test, the same subset of variables (*Calories*, *Fat*, *Fiber*, *Cholesterol*, and *BetaDiet*) was selected. This test is appropriate only when the grouping variable is categorical with exactly two categories. That’s why only two categorical variables (Sex and Smoke) were selected as grouping factors, respectively.

9.1 *Smoke*



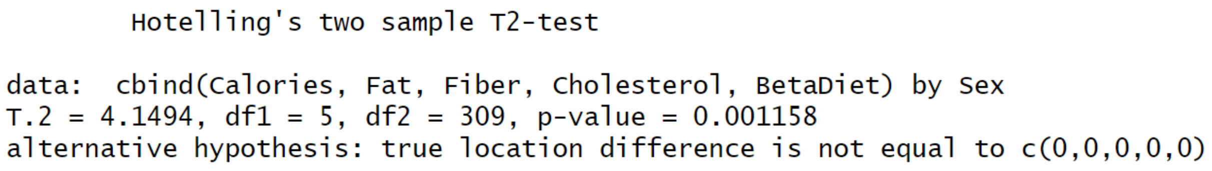
```
        Hotelling's two sample T2-test

data:  cbind(Calories, Fat, Fiber, Cholesterol, BetaDiet) by Smoke
T.2 = 4.6881, df1 = 5, df2 = 309, p-value = 0.0003847
alternative hypothesis: true location difference is not equal to c(0,0,0,0,0)
```

Figure 9.1.1 : Hotelling’s T2-Test results for *Smoke*

Hotelling's T² test was conducted to evaluate whether there is a difference in the multivariate means of the nutritional variables (Calories, Fat, Fiber, Cholesterol, BetaDiet) between smokers and non-smokers. The null hypothesis states that the mean vectors of these variables are equal across the two groups, while the alternative hypothesis states that at least one mean differs. The test yielded T² = 4.6881 with degrees of freedom 5 and 309, resulting in a p-value of 0.0003847. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in the combined nutritional profile between smokers and non-smokers.

9.2 *Sex*



```
        Hotelling's two sample T2-test

data:  cbind(Calories, Fat, Fiber, Cholesterol, BetaDiet) by Sex
T.2 = 4.1494, df1 = 5, df2 = 309, p-value = 0.001158
alternative hypothesis: true location difference is not equal to c(0,0,0,0,0)
```

Figure 9.1.1 : Hotelling’s T2-Test results for *Sex*

Similarly, Hotelling’s T² test was applied to compare the multivariate means of *Calories*, *Fat*, *Fiber*, *Cholesterol*, and *BetaDiet* between males and females. The test produced a T² statistic of 4.1494 with degrees of freedom 5 and 309, and a p-value of 0.001158. Given that the p-value is below 0.05, we reject the null hypothesis, indicating significant differences in the multivariate nutritional measures between males and females.

10 Conclusion

Applying a variety of multivariate techniques to the *Nutrition Study* dataset several consistent insights were reached. Dimensionality-reduction methods (PCA and Gower-based MDS) revealed two orthogonal dietary axes: an “unhealthy, high-fat/cholesterol” gradient tightly aligned with current smoking, and a “β-carotene/fibre” gradient associated with vitamin use and fruit-and-vegetable intake. Multiple Correspondence Analysis reinforced these findings on the categorical side, showing that smoking history and frequency of vitamin supplementation are the dominant drivers of categorical separation, with sex adding a subtler secondary layer.

Both hierarchical and k-means clustering converged on five interpretable dietary profiles, ranging from high-calorie/high-fat to carotene-rich eaters. Discriminant methods confirmed that classification is possible but imperfect: LDA struggled with the *Occasional* vitamin-use group, whereas QDA—allowing class-specific covariance—improved accuracy, especially for that middle category. Naive Bayes, even after removing highly correlated predictors, failed to outperform chance, highlighting the limitations of the conditional-independence assumption for this dataset.

MANOVA showed that prior smoking status affects fibre intake (and marginally fat), whereas vitamin use does not materially alter dietary means. Hotelling’s T^2 tests showed large multivariate differences between smokers and non-smokers and between males and females. Taken together, these analyses give an inference: current smoking is the single strongest lifestyle marker of an unhealthful dietary pattern, while routine vitamin supplementation appears to coexist with, rather than drive, healthier eating.

11 Bibliography

- Universitat Politècnica de Catalunya. *Multivariate Analysis Course Notes*. Atenea Virtual Campus. Accessed May 2025. https://atenea.upc.edu/course/view.php?id=96829

- Lock, R., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Lock5Data: Datasets for Statistics: Unlocking the Power of Data*. R package version [retrieved via CRAN documentation]. Dataset: `NutritionStudy`. Available at: https://search.r-project.org/CRAN/refmans/Lock5Data/html/NutritionStudy.html