Analysis of the epidemiological structure of cancers in New Caledonia: a clustering approach

Naïma BECK, Axelle LE POUL

18 december 2025

Abstract

In this study, we analysed a dataset on cancers in New Caledonia between 2008 and 2020. Our goal was to identify typical patient or pathology profiles through clustering.

To do this, we performed several clustering tests after cleaning our data. We applied four algorithms: PCA, K-means, Hierarchical Clustering, and DBSCAN.

We were able to identify age as the main factor in the clusters. We found six different clusters.

Table of Contents

Table of Figures

I. Introduction

I.1 Epidemiological Context of Cancer in New Caledonia

Recently, New Caledonia, an archipelago in the South Pacific Ocean, has been in the news for its Kanak independence struggles and post-colonial tensions, which led to urban riots in May 2024. However, it remains a rich overseas territory that deserves attention beyond the headlines of sensationalist newspapers.

As proof, healthcare provision remains very unevenly distributed, with a high concentration of healthcare professionals in the South Province, mainly in the Nouméa metropolitan area, while the North and the Islands remain underserved (Cour des comptes, 2024).

At the same time, it appears that cancer is a major public health problem in New Caledonia, with the territory ranking sixth in the world for cancer incidence among women. More than half of cancers are diagnosed after the age of 60 : 68% in men and 53% in women. Furthermore, while the overall incidence is decreasing among men, particularly due to the decline in prostate cancer, it is increasing among women, especially for breast cancer in women aged 30 to 39. In addition, cancer remains the leading medical cause of death, with nearly 500 deaths per year. From 2018 to 2020, the region recorded an average of 1,055 new cases diagnosed and 452 deaths per year. But even more alarming is the fact that the number of cases is expected to increase by 23.5% between 2028 and 2030, reaching approximately 1,300 new diagnoses per year. (DASS-NC, 2020)

Since 1994, cancer has become a notifiable disease in this territory, resulting in the creation of massive databases.

I.2 Knowledge gap

In recent years, several descriptive studies have been conducted by various institutions. However, no in-depth analysis has been carried out using data partitioning, and none has identified the complex, non-linear interactions between variables. It therefore seems necessary to perform analyses using robust clustering algorithms. Here, clustering is crucial because it allows us to isolate whether these profiles are solely related to age or to other factors, and to go beyond simple observation and discover multidimensional epidemiological profiles.

I.3 Research Questions

We will therefore ask ourselves the following question : are there hidden patterns in the cancer data in NC? Can similar cases be grouped together to better understand the epidemiological distribution?

I.4. Methodological Overview

To address these research questions, we analyzed the DASS-NC cancer registry dataset from 2008 to 2020 through a three-step process: data exploration, rigorous cleaning including missing values management, and the sequential application of clustering algorithms.

II. Materials and Method

II.1. Dataset Description: The DASS-NC Registry

We chose a dataset that lists the cases of cancer diagnosed in New Caledonia between 2008 and 2020. This dataset contains 12 variables :

- **SEXE/sex** : Men or Women.
- **ANNÉE/year** : Year of diagnosis or death.
- **CLASSE_AGE/age class** : Age group in 25-year increments.
- **CODE_CIM10/cim10 code** : Diagnostic code or group of codes according to ICD-10.
- **libelle_CIM10/cim10 label** : Descriptive label associated with the ICD-10 code.
- **label** : Label associated with CODE_CIM10.
- **Indicateur/indicator** : Type of indicator analysed (Incidence or Mortality).
- **rang/rank** : Position of cancer in the ranking according to its frequency, by gender, year, age group and indicator.
- **frequence/frequency** : Ratio of the number of cancers to the total number of cancers, by gender, year, age group and indicator.
- **age_median/median age** : Median age at diagnosis or death.
- **nb_cas/case number** : Number of cases or deaths.
- **asr** : Age-standardised incidence or mortality rate.

Data have been treated for the last time on the 25th of April 2025. The reference of the dataset comes from https://gouvnc-dass.shinyapps.io/cancers_nc/.

II.2. Software Tools

We used the R language, and before performing any test on our dataset we had to make it clear and ready for interpretations. For this we first needed to install 2 packages :

- *Tidyverse* to manipulate the data and clear them. We used different commands such as select() to delete useless columns or lines, the pipe command to continue the operations.
- *Caret* to encode the data and prepare them. We principally used the command dummyVars() to encode categorical variables into indicator variables.

II.3. Data Cleaning and Categorical Encoding

After selecting the packages we needed we can see the different steps we went through to clean the data.

First, we displayed the data to have an overview of the dataset. Thanks to that we could start the cleaning and organisation of data. We started by deleting the unnecessary columns and encoding categorical variables. We had binary variables like the sex of the people and for categorical variables we used the dummy encoding.

II.3. Management of Missing Values and Data Imputation

After that, we had to handle the missing values. We counted the missing values for each column and we computed the proportion of missing values for each of them. Then we could delete the columns for which the missing values were too important like for 'age_median' column with 83% of missing values. Thus we deleted the lines where we still had missing values. After those operations we had 0 missing values. We could export the cleaned data and start our analysis on it.

II.4. Algorithmic Framework

To extract meaningful patterns from our multidimensional dataset, we implemented a robust sequence of algorithms, ranging from dimension reduction to density detection.

II.4.1. Principal Component Analysis (PCA) : Dimensionality Reduction

PCA is not a clustering algorithm per se, but rather an unsupervised approach used for data visualization and preprocessing (James and al., 2023). After applying one-hot encoding to our categorical variables, the number of dimensions increased significantly. PCA allows us to project this high-dimensional data into a reduced space (2D or 3D) while capturing as much variance as possible. This step is important for identifying whether natural clusters or scatter plots emerge visually before launching more computationally expensive partitioning algorithms.

II.4.2. K-means Clustering : Centroid-based Partitioning

K-means is one of the most established methods for dividing a dataset into K distinct, non-overlapping groups. The algorithm assigns each observation to a cluster in order to minimize intra-class variation, mostly the "squared Euclidean distance". Since the number of clusters is not known a priori, we used the "Elbow Method" and the "Silhouette score" to determine the optimal K, thus ensuring that the clusters are both compact and well separated.

II.4.3. Hierarchical Agglomerative Clustering (HAC) : Structural Dendrograms

Unlike K-means, hierarchical clustering does not require specifying a number of clusters in advance. It follows a bottom-up approach, where each observation starts in its own cluster.

Pairs of clusters are then merged based on their similarity until a single tree structure called a dendrogram is formed. (James and al., 2023)

For our epidemiological study, this is particularly useful because it allows us to visualize the hierarchical relationships between cancer types and age groups, revealing substructures that a "flat" partition might overlook.

II.4.4. DBSCAN : Density-based Spatial Clustering for Outlier Detection

To complete our analysis, we implemented DBSCAN (Density-Based Spatial Clustering of Applications with Noise). This algorithm identifies clusters as high-density regions separated by low-density areas. This choice is justified by two major factors: unlike K-means, DBSCAN can discover irregularly shaped clusters and, above all, it is extremely effective at identifying outliers.

In the context of cancer in New Caledonia, this allows us to isolate rare pathologies or atypical patient profiles that do not fit into major epidemiological trends.

III. Results

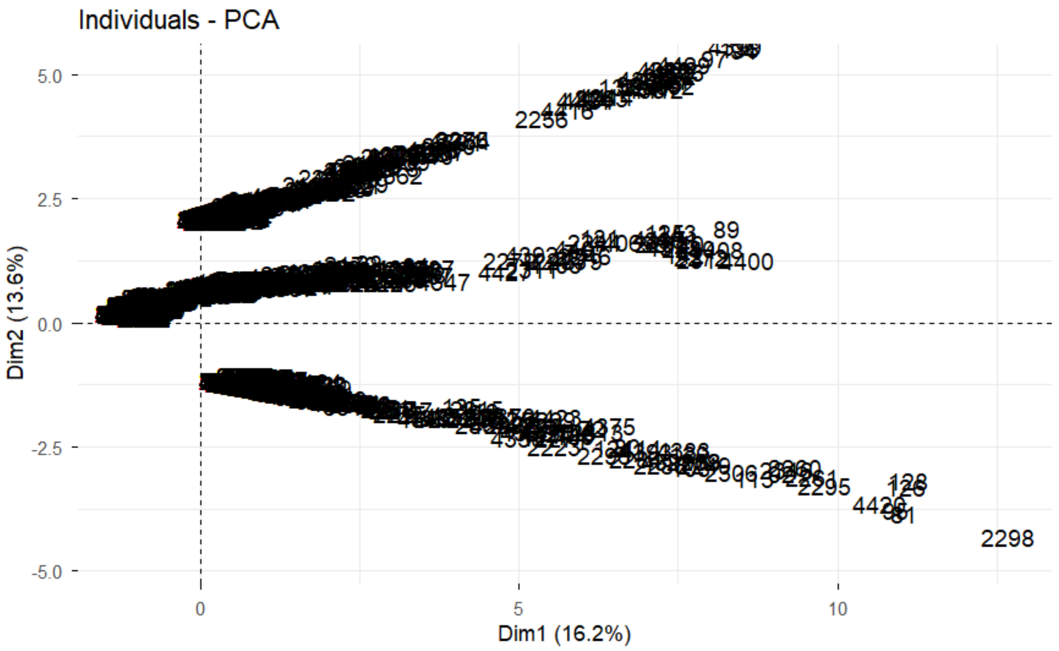


Figure 1 PCA score plot representing the projection of cancer data on the first two principal components

In figure 1, we have the results of our PCA test, after selecting the variables we wanted to keep : gender, number of cases, ASR rate, breast/prostate cancers (C50/C61) and all categories of age (CLASSE_AGE). Our PCA is distributed across two axes, Dim1 and Dim2.

We can see that the PCA is divided into three distinct profiles. These profiles are associated with the age of the patients. Axis 1 shows the incidence rate (number of cases and ASR) and the axis separates the pathologies between breast cancer and prostate cancer. This distribution confirms that age remains the dominant predictor of the structure of cancer data in New Caledonia between 2008 and 2020.

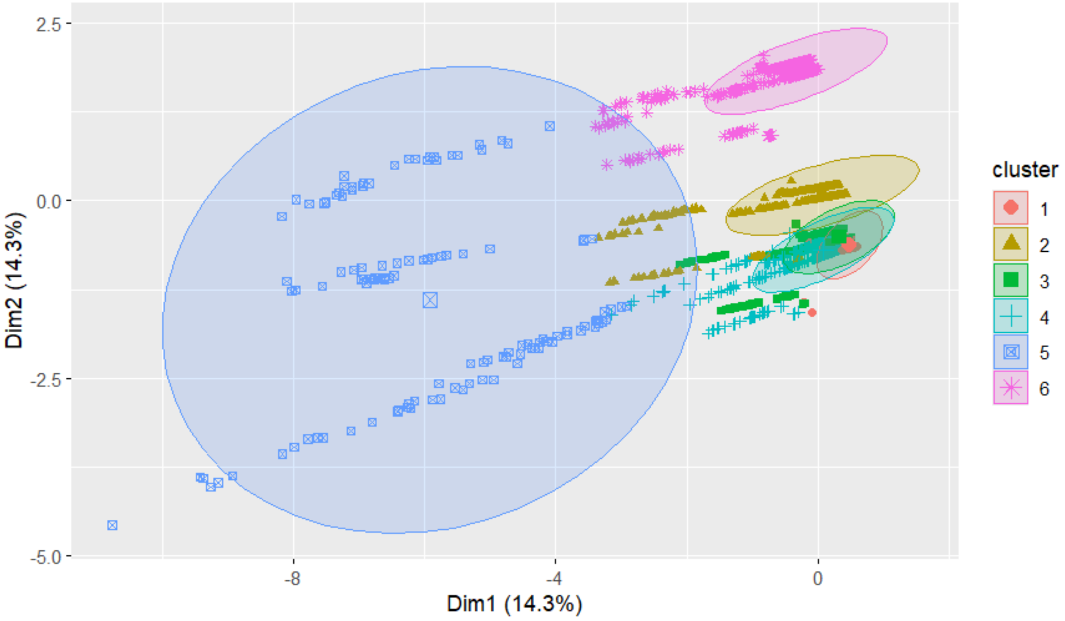


Figure 2 K-means clustering (k=6) visualization based on the first 7 principal components

In figure 2, the application of the K-means algorithm, with k = 6, based on PCA data, enabled accurate segmentation of the patient population. The analysis reveals that this classification is not limited to age groups alone. It generates subgroups within these groups, in particular by distinguishing between male cancers (prostate) and female cancers (breast) in the most affected age groups.

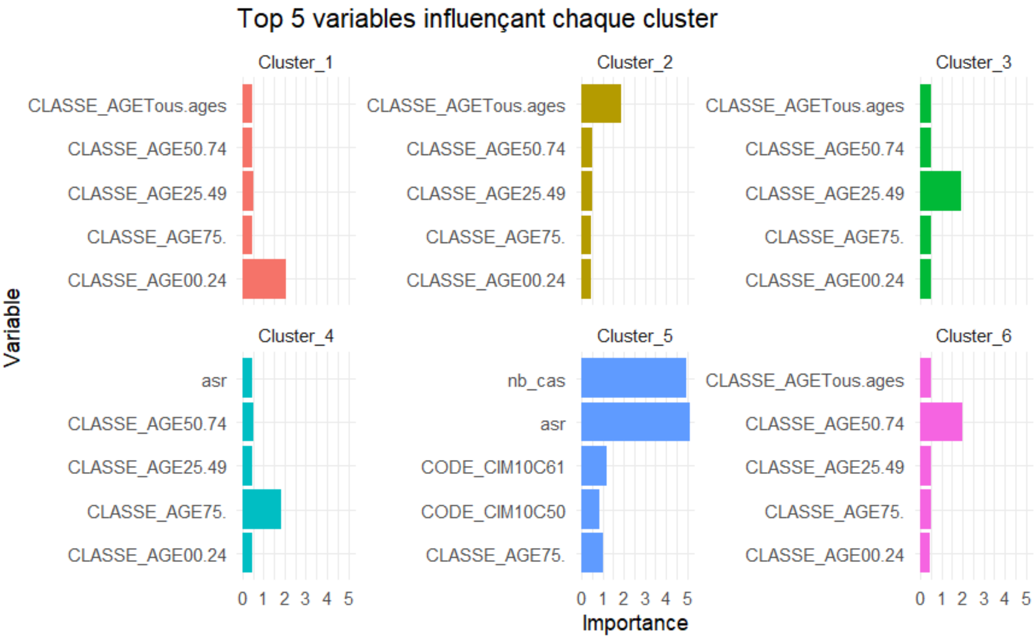


Figure 3 Variable importance plot: Top 5 contributing factors for each K-means cluster

In figure 3, we created a graph showing which factor was decisive in selecting which cluster a patient would belong to. We can see that most clusters are due to age and one is due to the severity of the cancer.

While age is the discriminating factor for the majority of profiles (biological segmentation), grouping based on case volume (nb_cas) and incidence rate (asr) makes it possible to isolate the major pathologies that represent the highest epidemiological weight for the territory. This distinction is crucial because it separates the care needs associated with aging from the prevention needs associated with the most common cancers.

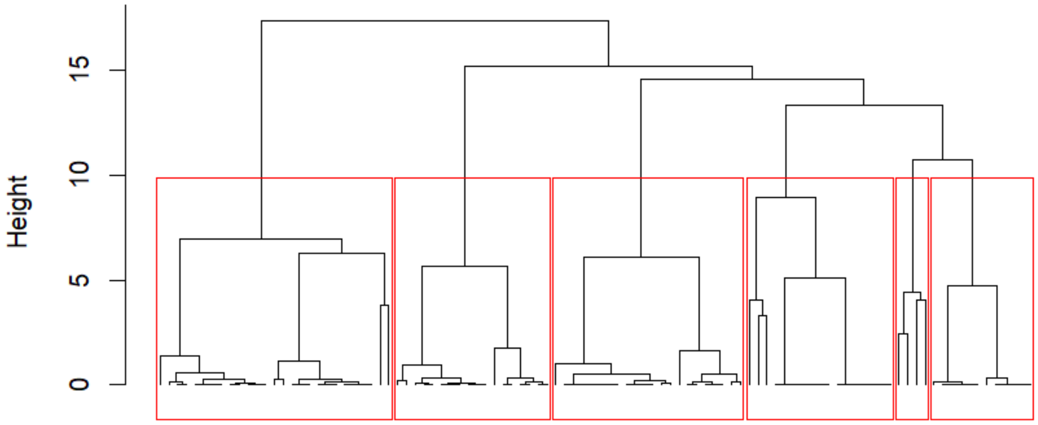


Figure 4 Hierarchical Agglomerative Clustering (HAC) dendrogram using Ward's linkage

In figure 4, we then created a dendrogram using the data obtained with PCA. Here, we are looking for clusters due to a certain characteristic of the individuals. To do this, we used Ward's method, which seeks to minimise variance within each group. This creates very compact and balanced clusters, which is ideal for defining precise patient profiles.

This dendrogram gives us a more refined result than that obtained using PCA. Here, we obtain six statistically differentiated groups (height of the upper branches), which tells us that more precise divisions can be made for cancer prediction (type of cancer, gender, etc.).

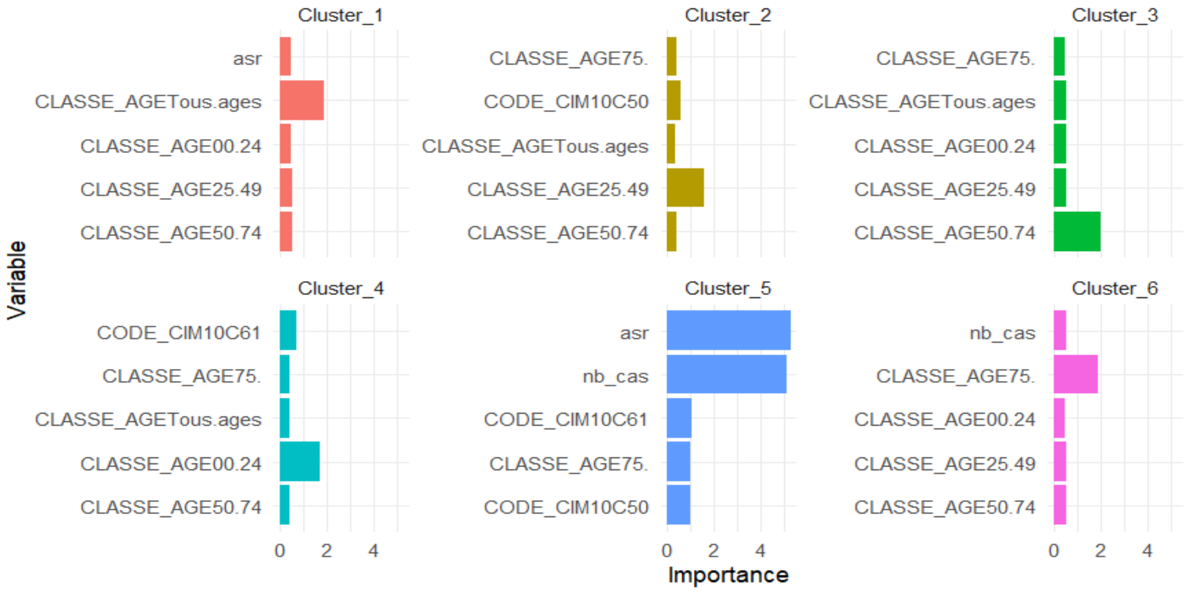


Figure 5 Profile characterization: Top 5 influential variables per hierarchical cluster

In figure 5, we also created a graph selecting the major characteristic of each cluster and obtained results similar to those of k-mean clustering. Age explains who is in the group, but the importance of nb_cas/asr explains the magnitude of the problem for that group.

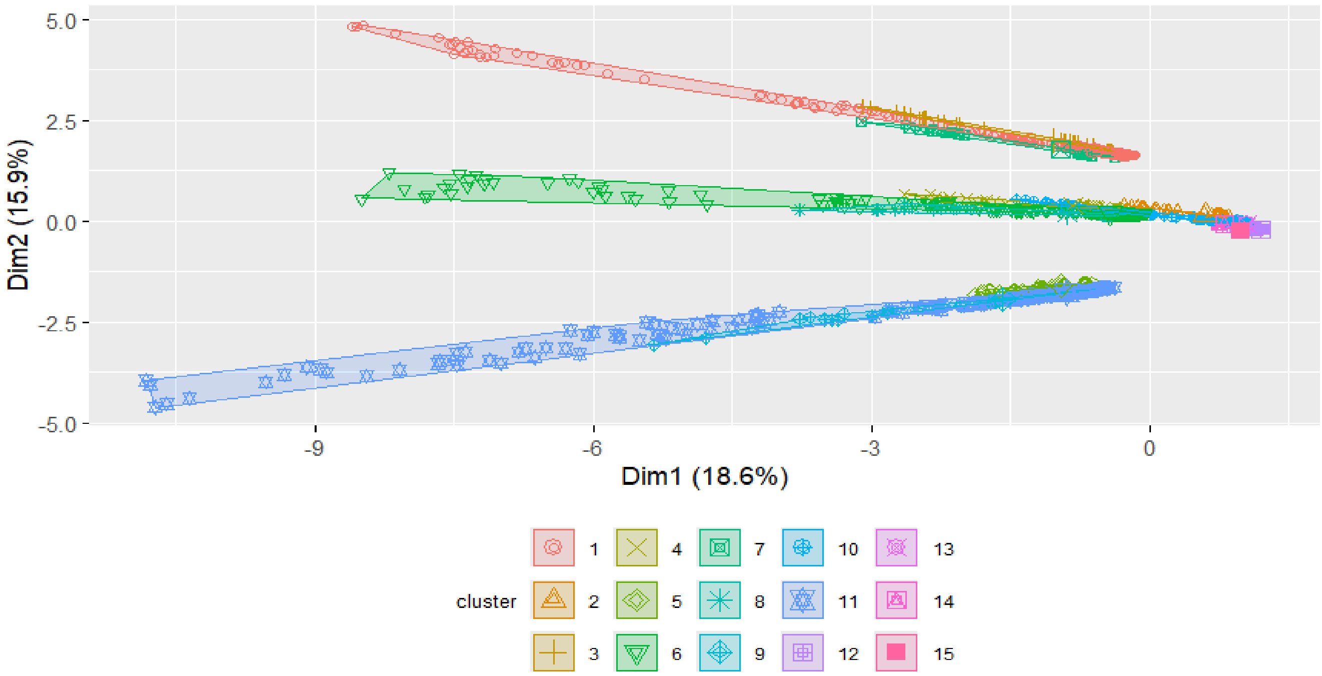


Figure 6 Density-based clustering (DBSCAN) on 7 PCs with an epsilon value of 3

Finally, in figure 6 we performed a DBSCAN, again using the values obtained by PCA. The clustered central points show the most common and similar types of cancer in New Caledonia. This confirms that the data form a stable structure and can be used for prediction.

The DBSCAN outliers are not errors, but atypical cancer cases in New Caledonia. This can be illustrated by a type of cancer that normally affects elderly people, but which appears here in a 20-year-old patient, or an extremely rare cancer for which there were only one or two cases between 2008 and 2020.

This method complements K-means by showing that, while typical profiles do exist, a significant proportion of cancer cases remain unique and require an individualised monitoring approach.

IV. Discussion/summary

Here, all our statistical tests reveal important information. First, the combination of PCA and k-means allowed us to interpret the raw data into more precise groups, isolating in particular cancers with high epidemiological weight (asr and nb_cas). It is therefore thanks to this visualisation that we were able to determine comprehensible groups.

The results obtained using hierarchical clustering are similar to those obtained using k-means clustering and reinforce our understanding of the data.

DBSCAN allows us to identify outliers that represent specific cancer cases. This makes it possible to separate major trends (common cancers) from rare cases that require different investigation.

However, the analysis has some limitations. Firstly, when preparing the data, we had to remove the age_median variable (83% of data missing), which inevitably had an impact on the accuracy of the data and therefore the analysis. In addition, we were unable to use the raw data because we had too much information and were therefore limited to a PCA with the first seven components, which also reduced the accuracy of the analysis.

Our analysis of the data has enabled us to move towards a more stratified approach to data management. We can now break down cancer prevention in New Caledonia according to individuals' ages, the severity of cancer and specific types of cancer. This makes it possible to create preventive pathways that vary according to age group, but also to highlight the scale of investment needed to treat these cancers. Outliers can also be used to highlight certain forms of cancer that could become significant or may have an environmental link.

V. References

Agence Sanitaire et Sociale de la Nouvelle-Calédonie. (2022). *Baromètre Santé Adulte 2021-2022: résultats descriptifs* [Rapport]. ASS-NC. Consulté à l'adresse : https://www.santepourtous.nc/images/st-barometre-sante-adulte/st-pdf/st-rapport-baromtre-v10-lq.pdf

Cour des comptes. (2024). *L'organisation territoriale des soins de premiers recours : Cahier territorial de la Nouvelle-Calédonie*. Cour des comptes. Consulté à l'adresse https://www.vie-publique.fr/files/rapports/fichiers_joints/294141-Nouvelle-Caledonie.pdf

Direction des Affaires Sanitaires et Sociales de Nouvelle-Calédonie. (2024). *Rapport sur l'épidémiologie des cancers en Nouvelle-Calédonie, incidence et mortalité 2018-2020*. Gouvernement de la Nouvelle-Calédonie. Consulté à l'adresse https://dass.gouv.nc/actualites/le-rapport-2018-2020-du-registre-du-cancer

Ferlay, J., Ervik, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2024). *New Caledonia: GLOBOCAN 2020 cancer statistics* [Fiche pays]. International Agency for Research on Cancer. Consulté à l'adresse https://gco.iarc.who.int/media/globocan/factsheets/populations/540-new-caledonia-fact-sheet.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning: with Applications in R* (2nd ed., Corrected printing June 2023). Springer.

VI. Appendices

Data Cleaning and Preparation for Cancer Epidemiology in New Caledonia

Naïma Beck and Axelle Le Poul

```r
#install.packages(c("tidyverse", "factoextra", "cluster", "caret"))
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   4.0.0     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(cluster)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

1. Dataset and data overview

```r
epidemiology_cancers_nc_raw <- read.csv("../data/raw/epidemiologie_cancers_nc.csv", stringsAsFactors =
head(epidemiology_cancers_nc_raw)
```

```
##   id_row  SEXE ANNEE CLASSE_AGE CODE_CIM10
## 1    151 Homme  2012  Tous ages        C16
## 2    169 Homme  2018  Tous ages        C16
## 3    203 Homme  2019  Tous ages        C16
## 4   3443 Homme  2013  Tous ages        C22
## 5   3426 Homme  2015  Tous ages        C22
## 6    189 Homme  2017  Tous ages        C22
##                                                   libelle_CIM10
## 1                                   Tumeur maligne de l'estomac
## 2                                   Tumeur maligne de l'estomac
## 3                                   Tumeur maligne de l'estomac
## 4 Tumeur maligne du foie et des voies biliaires intrahépatiques
## 5 Tumeur maligne du foie et des voies biliaires intrahépatiques
## 6 Tumeur maligne du foie et des voies biliaires intrahépatiques
##                                     label indicateur rang frequence age_median
## 1                                 Estomac  Incidence    5       5.2         63
## 2                                 Estomac  Incidence    5       4.6         62
## 3                                 Estomac  Incidence    6       3.8         60
## 4 Foie et voies biliaires intra-hépatiques  Mortalité    2       9.2         67
## 5 Foie et voies biliaires intra-hépatiques  Mortalité    2      11.3         63
## 6 Foie et voies biliaires intra-hépatiques  Incidence    6       4.1         59
##   nb_cas   asr
## 1     22 15.54
## 2     21 12.11
## 3     19 10.50
## 4     19 12.76
## 5     25 15.91
## 6     19 10.97
```

```
unique(epidemiology_cancers_nc_raw$CLASSE_AGE)
```

```
## [1] "Tous ages" "25-49"     "50-74"     "75+"       "00-24"
```

```
unique(epidemiology_cancers_nc_raw$indicateur)
```

```
## [1] "Incidence" "Mortalité"
```

```
colnames(epidemiology_cancers_nc_raw)
```

```
##  [1] "id_row"        "SEXE"          "ANNEE"         "CLASSE_AGE"
##  [5] "CODE_CIM10"    "libelle_CIM10" "label"         "indicateur"
##  [9] "rang"          "frequence"     "age_median"    "nb_cas"
## [13] "asr"
```

```
# On part du dataset original "epidemiology_cancers_nc_raw"
cim10_mapping <- epidemiology_cancers_nc_raw %>%
  select(CODE_CIM10, libelle_CIM10) %>%
  distinct()  # On ne garde qu’une ligne par code unique

cim10_mapping
```

```
##    CODE_CIM10                                                 libelle_CIM10
## 1         C16                                   Tumeur maligne de l'estomac
## 2         C22 Tumeur maligne du foie et des voies biliaires intrahépatiques
## 3         C43                                     Mélanome malin de la peau
## 4         C50                                        Tumeur maligne du sein
## 5         C53                             Tumeur maligne du col de l'utérus
## 6         C54                           Tumeur maligne du corps de l'utérus
## 7         C61                                 Tumeur maligne de la prostate
## 8         C73                                 Tumeur maligne de la thyroïde
## 9      C00-13
## 10     C00-97
## 11     C18-20
## 12     C33-34
## 13     C91-95
## 14 C82-86,C96
## 15        C15                                 Tumeur maligne de l'oesophage
## 16        C25                                    Tumeur maligne du pancréas
## 17        C56                                    Tumeur maligne de l'ovaire
## 18        C62                                   Tumeur maligne du testicule
## 19        C67                                   Tumeur maligne de la vessie
## 20     C23-24
## 21     C30-32
## 22     C64-65
## 23     C70-72
## 24    C88,C90
```

2. Data cleaning and preparation

2.1 Deleting unnecessary columns

```
epidemiology_cancers_nc <- epidemiology_cancers_nc_raw %>%
  select(-id_row, -label, -libelle_CIM10)
```

2.2 Encoding categorical variables

Encodaging of binary variables : 'sexe' and 'indicateur'

```
epidemiology_cancers_nc$SEXE <- ifelse(epidemiology_cancers_nc$SEXE == "Homme", 1, 0) # 1: Homme, 0:Fem

epidemiology_cancers_nc$indicateur <- ifelse(epidemiology_cancers_nc$indicateur == "Mortalité", 1, 0) #
```

Dummy encoding

It is encoding categorial variables into numerical variables.

```
categorical_vars <- c( "CLASSE_AGE", "CODE_CIM10")
col_encoded <- dummyVars(" ~ .", data = epidemiology_cancers_nc[, categorical_vars]) %>%
  predict(newdata = epidemiology_cancers_nc[, categorical_vars]) %>%
  as.data.frame()
```

combine the encoded columns and the numerical columns

```
numeric_vars <- epidemiology_cancers_nc %>%
  select(-all_of(categorical_vars))
epidemiology_cancers_nc <- cbind(numeric_vars, col_encoded)

head(epidemiology_cancers_nc)
```

```
##   SEXE ANNEE indicateur rang frequence age_median nb_cas   asr CLASSE_AGE00-24
## 1    1  2012          0    5       5.2         63     22 15.54               0
## 2    1  2018          0    5       4.6         62     21 12.11               0
## 3    1  2019          0    6       3.8         60     19 10.50               0
## 4    1  2013          1    2       9.2         67     19 12.76               0
## 5    1  2015          1    2      11.3         63     25 15.91               0
## 6    1  2017          0    6       4.1         59     19 10.97               0
##   CLASSE_AGE25-49 CLASSE_AGE50-74 CLASSE_AGE75+ CLASSE_AGETous ages
## 1               0               0             0                   1
## 2               0               0             0                   1
## 3               0               0             0                   1
## 4               0               0             0                   1
## 5               0               0             0                   1
## 6               0               0             0                   1
##   CODE_CIM10C00-13 CODE_CIM10C00-97 CODE_CIM10C15 CODE_CIM10C16
## 1                0                0             0             1
## 2                0                0             0             1
## 3                0                0             0             1
## 4                0                0             0             0
## 5                0                0             0             0
## 6                0                0             0             0
##   CODE_CIM10C18-20 CODE_CIM10C22 CODE_CIM10C23-24 CODE_CIM10C25
## 1                0             0                0             0
## 2                0             0                0             0
## 3                0             0                0             0
## 4                0             1                0             0
## 5                0             1                0             0
## 6                0             1                0             0
##   CODE_CIM10C30-32 CODE_CIM10C33-34 CODE_CIM10C43 CODE_CIM10C50 CODE_CIM10C53
## 1                0                0             0             0             0
## 2                0                0             0             0             0
## 3                0                0             0             0             0
## 4                0                0             0             0             0
## 5                0                0             0             0             0
## 6                0                0             0             0             0
##   CODE_CIM10C54 CODE_CIM10C56 CODE_CIM10C61 CODE_CIM10C62 CODE_CIM10C64-65
## 1             0             0             0             0                0
## 2             0             0             0             0                0
## 3             0             0             0             0                0
## 4             0             0             0             0                0
## 5             0             0             0             0                0
## 6             0             0             0             0                0
##   CODE_CIM10C67 CODE_CIM10C70-72 CODE_CIM10C73 CODE_CIM10C82-86,C96
## 1             0                0             0                    0
## 2             0                0             0                    0
## 3             0                0             0                    0
## 4             0                0             0                    0
## 5             0                0             0                    0
## 6             0                0             0                    0
##   CODE_CIM10C88,C90 CODE_CIM10C91-95
## 1                 0                0
## 2                 0                0
## 3                 0                0
## 4                 0                0
## 5                 0                0
## 6                 0                0
```

2.3 Handling missing values

To see how many missing values we have in the dataset :

Counting missing values

```
sum(is.na(epidemiology_cancers_nc))
```

```
## [1] 5496
```

They are a lot of missing values in this data set, let's look at the proportion of missing values in the columns so we can take a decision.

number of rows(observations) in the dataset

```
nrow(epidemiology_cancers_nc)
```

```
## [1] 6480
```

missing values per column

```
na_per_column <- colSums(is.na(epidemiology_cancers_nc))
na_pct_per_column <- colSums(is.na(epidemiology_cancers_nc)*100/nrow(epidemiology_cancers_nc))

cat("--- Number of NA per colonne ---\n")
```

```
## --- Number of NA per colonne ---
```

```
head(sort(na_per_column, decreasing = TRUE))
```

```
## age_median  frequence       SEXE      ANNEE indicateur       rang
##       5381        115          0          0          0          0
```

```r
cat("\n--- Pourcentage of NA per colonne (%) ---\n")
```

```
##
## --- Pourcentage of NA per colonne (%) ---
```

```r
head(sort(na_pct_per_column, decreasing = TRUE))
```

```
## age_median  frequence       SEXE      ANNEE indicateur       rang
##  83.040123   1.774691   0.000000   0.000000   0.000000   0.000000
```

We can see that in the column 'age_median' we have 83% of missing values. So we may delete this column. And for the column 'frequence' we only have 2% of missing values so we may delete the observations with missing frequence.

We delete the columns 'age_median'

```r
epidemiology_cancers_nc <-  epidemiology_cancers_nc %>%
select(-age_median)
```

We delete the rows where the frequence is missing

```r
epidemiology_cancers_nc <- epidemiology_cancers_nc[complete.cases(epidemiology_cancers_nc), ]
```

verification of the number of missing values in the data set

```r
sum(is.na(epidemiology_cancers_nc))
```

```
## [1] 0
```

```r
head(epidemiology_cancers_nc)
```

```
##   SEXE ANNEE indicateur rang frequence nb_cas   asr CLASSE_AGE00-24
## 1    1  2012          0    5       5.2     22 15.54               0
## 2    1  2018          0    5       4.6     21 12.11               0
## 3    1  2019          0    6       3.8     19 10.50               0
## 4    1  2013          1    2       9.2     19 12.76               0
## 5    1  2015          1    2      11.3     25 15.91               0
## 6    1  2017          0    6       4.1     19 10.97               0
##   CLASSE_AGE25-49 CLASSE_AGE50-74 CLASSE_AGE75+ CLASSE_AGETous ages
## 1               0               0             0                   1
## 2               0               0             0                   1
## 3               0               0             0                   1
## 4               0               0             0                   1
## 5               0               0             0                   1
## 6               0               0             0                   1
##   CODE_CIM10C00-13 CODE_CIM10C00-97 CODE_CIM10C15 CODE_CIM10C16
## 1                0                0             0             1
## 2                0                0             0             1
## 3                0                0             0             1
## 4                0                0             0             0
## 5                0                0             0             0
## 6                0                0             0             0
##   CODE_CIM10C18-20 CODE_CIM10C22 CODE_CIM10C23-24 CODE_CIM10C25
## 1                0             0                0             0
## 2                0             0                0             0
## 3                0             0                0             0
## 4                0             1                0             0
## 5                0             1                0             0
## 6                0             1                0             0
##   CODE_CIM10C30-32 CODE_CIM10C33-34 CODE_CIM10C43 CODE_CIM10C50 CODE_CIM10C53
## 1                0                0             0             0             0
## 2                0                0             0             0             0
## 3                0                0             0             0             0
## 4                0                0             0             0             0
## 5                0                0             0             0             0
## 6                0                0             0             0             0
##   CODE_CIM10C54 CODE_CIM10C56 CODE_CIM10C61 CODE_CIM10C62 CODE_CIM10C64-65
## 1             0             0             0             0                0
## 2             0             0             0             0                0
## 3             0             0             0             0                0
## 4             0             0             0             0                0
## 5             0             0             0             0                0
## 6             0             0             0             0                0
##   CODE_CIM10C67 CODE_CIM10C70-72 CODE_CIM10C73 CODE_CIM10C82-86,C96
## 1             0                0             0                    0
## 2             0                0             0                    0
## 3             0                0             0                    0
## 4             0                0             0                    0
## 5             0                0             0                    0
## 6             0                0             0                    0
##   CODE_CIM10C88,C90 CODE_CIM10C91-95
## 1                 0                0
## 2                 0                0
## 3                 0                0
## 4                 0                0
## 5                 0                0
## 6                 0                0
```

3. Exporting cleaned data

```
write.csv(epidemiology_cancers_nc,"../data/processed/epidemiologie_cancers_nc_clean.csv",
          row.names = FALSE)
```

PCA and Clustering Applied to Cancer Epidemiological Data in New Caledonia

Naïma Beck and Axelle Le Poul

```r
library(factoextra)
```

```
## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(dbscan)
```

```
##
## Attaching package: 'dbscan'

## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```r
library(cluster)
```

1. Importation of the cleaned dataset

```r
epidemiology_cancers_nc <- read.csv("../data/processed/epidemiologie_cancers_nc_clean.csv", stringsAsFa
head(epidemiology_cancers_nc)
```

```
##   SEXE ANNEE indicateur rang frequence nb_cas   asr CLASSE_AGE00.24
## 1    1  2012          0    5       5.2     22 15.54               0
## 2    1  2018          0    5       4.6     21 12.11               0
## 3    1  2019          0    6       3.8     19 10.50               0
## 4    1  2013          1    2       9.2     19 12.76               0
## 5    1  2015          1    2      11.3     25 15.91               0
## 6    1  2017          0    6       4.1     19 10.97               0
##   CLASSE_AGE25.49 CLASSE_AGE50.74 CLASSE_AGE75. CLASSE_AGETous.ages
## 1               0               0             0                   1
## 2               0               0             0                   1
## 3               0               0             0                   1
## 4               0               0             0                   1
## 5               0               0             0                   1
## 6               0               0             0                   1
##   CODE_CIM10C00.13 CODE_CIM10C00.97 CODE_CIM10C15 CODE_CIM10C16
## 1                0                0             0             1
## 2                0                0             0             1
## 3                0                0             0             1
## 4                0                0             0             0
## 5                0                0             0             0
## 6                0                0             0             0
##   CODE_CIM10C18.20 CODE_CIM10C22 CODE_CIM10C23.24 CODE_CIM10C25
## 1                0             0                0             0
## 2                0             0                0             0
## 3                0             0                0             0
## 4                0             1                0             0
## 5                0             1                0             0
## 6                0             1                0             0
##   CODE_CIM10C30.32 CODE_CIM10C33.34 CODE_CIM10C43 CODE_CIM10C50 CODE_CIM10C53
## 1                0                0             0             0             0
## 2                0                0             0             0             0
## 3                0                0             0             0             0
## 4                0                0             0             0             0
## 5                0                0             0             0             0
## 6                0                0             0             0             0
##   CODE_CIM10C54 CODE_CIM10C56 CODE_CIM10C61 CODE_CIM10C62 CODE_CIM10C64.65
## 1             0             0             0             0                0
## 2             0             0             0             0                0
## 3             0             0             0             0                0
## 4             0             0             0             0                0
## 5             0             0             0             0                0
## 6             0             0             0             0                0
##   CODE_CIM10C67 CODE_CIM10C70.72 CODE_CIM10C73 CODE_CIM10C82.86.C96
## 1             0                0             0                    0
## 2             0                0             0                    0
## 3             0                0             0                    0
## 4             0                0             0                    0
## 5             0                0             0                    0
## 6             0                0             0                    0
##   CODE_CIM10C88.C90 CODE_CIM10C91.95
## 1                 0                0
## 2                 0                0
## 3                 0                0
## 4                 0                0
## 5                 0                0
## 6                 0                0
```

```
colnames(epidemiology_cancers_nc)
```

```
##  [1] "SEXE"                 "ANNEE"                "indicateur"
##  [4] "rang"                 "frequence"            "nb_cas"
##  [7] "asr"                  "CLASSE_AGE00.24"      "CLASSE_AGE25.49"
## [10] "CLASSE_AGE50.74"      "CLASSE_AGE75."        "CLASSE_AGETous.ages"
## [13] "CODE_CIM10C00.13"     "CODE_CIM10C00.97"     "CODE_CIM10C15"
## [16] "CODE_CIM10C16"        "CODE_CIM10C18.20"     "CODE_CIM10C22"
## [19] "CODE_CIM10C23.24"     "CODE_CIM10C25"        "CODE_CIM10C30.32"
## [22] "CODE_CIM10C33.34"     "CODE_CIM10C43"        "CODE_CIM10C50"
## [25] "CODE_CIM10C53"        "CODE_CIM10C54"        "CODE_CIM10C56"
## [28] "CODE_CIM10C61"        "CODE_CIM10C62"        "CODE_CIM10C64.65"
## [31] "CODE_CIM10C67"        "CODE_CIM10C70.72"     "CODE_CIM10C73"
## [34] "CODE_CIM10C82.86.C96" "CODE_CIM10C88.C90"    "CODE_CIM10C91.95"
```

2. Full dataset

```
epidemiology_cancers_nc2 <- epidemiology_cancers_nc %>%
  select(-"ANNEE", -"rang", -"frequence", -"CLASSE_AGETous.ages")
```

We are doing this to remove non-informative and redundant variables so the PCA focuses only on meaningful epidemiological features instead of noise or duplicated information.

2.1 PCA

Normalisation

Now all variables have the same scale.

```
epicancer_scaled <- scale(epidemiology_cancers_nc2)
```

```
ncol(epidemiology_cancers_nc2)
```

```
## [1] 32
```

ncol < 50 so it's ok

test for pca

```
# Inf
sum(is.infinite(epicancer_scaled))
```

```
## [1] 0
```

So we can continue.

PCA applied

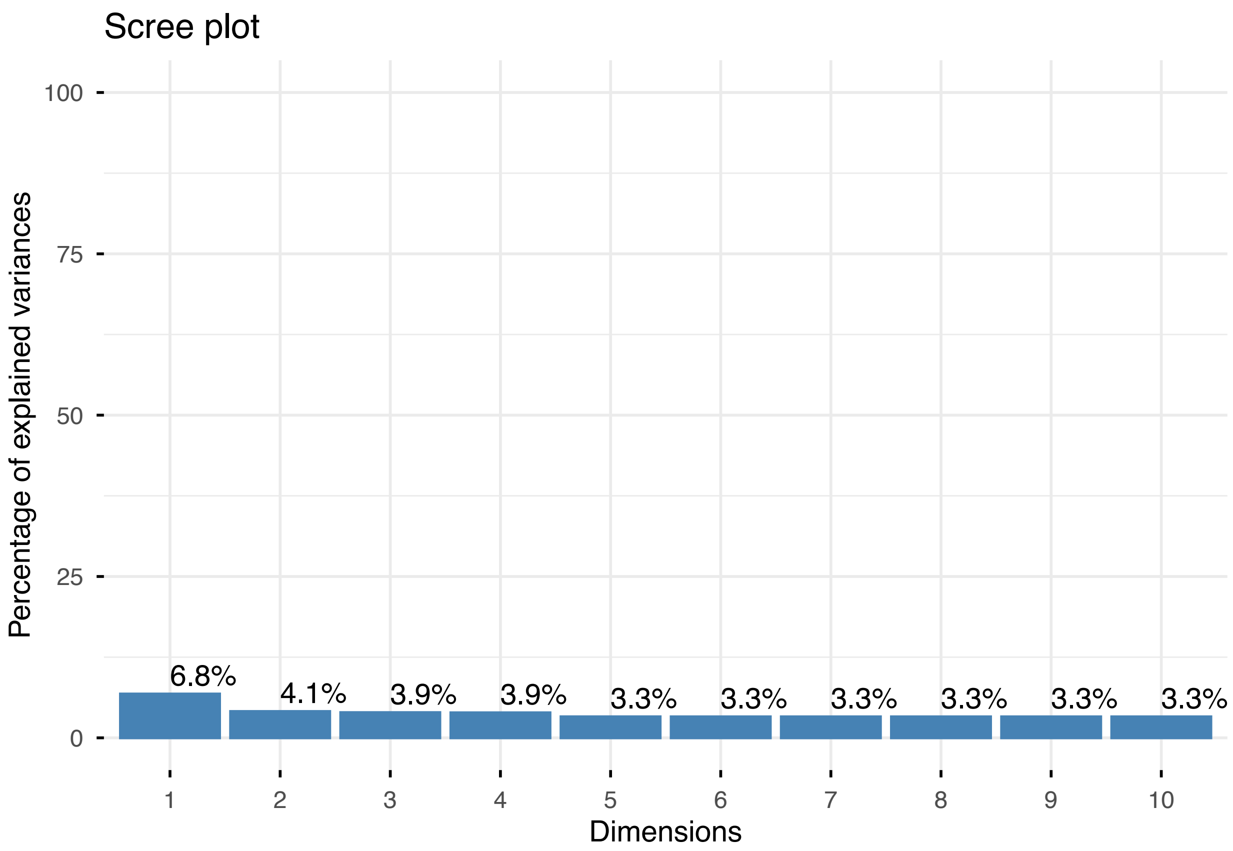
```r
pca_result_cancer <- prcomp(epicancer_scaled, center = TRUE, scale = TRUE)

summary(pca_result_cancer)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.47427 1.14277 1.11877 1.11474 1.02211 1.02152 1.02149
## Proportion of Variance 0.06792 0.04081 0.03911 0.03883 0.03265 0.03261 0.03261
## Cumulative Proportion  0.06792 0.10873 0.14784 0.18668 0.21932 0.25193 0.28454
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.02149 1.02149 1.02149 1.02149 1.02149 1.02149 1.02149
## Proportion of Variance 0.03261 0.03261 0.03261 0.03261 0.03261 0.03261 0.03261
## Cumulative Proportion  0.31715 0.34976 0.38236 0.41497 0.44758 0.48019 0.51279
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.02149 1.02149 1.02149 1.02149 1.02149 1.02149 1.02149
## Proportion of Variance 0.03261 0.03261 0.03261 0.03261 0.03261 0.03261 0.03261
## Cumulative Proportion  0.54540 0.57801 0.61062 0.64322 0.67583 0.70844 0.74105
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     1.02149 1.02149 1.02149 1.02149 1.02149 1.00631 0.99161
## Proportion of Variance 0.03261 0.03261 0.03261 0.03261 0.03261 0.03165 0.03073
## Cumulative Proportion  0.77366 0.80626 0.83887 0.87148 0.90409 0.93573 0.96646
##                           PC29    PC30    PC31      PC32
## Standard deviation     0.74202 0.55838 0.45929 7.445e-15
## Proportion of Variance 0.01721 0.00974 0.00659 0.000e+00
## Cumulative Proportion  0.98366 0.99341 1.00000 1.000e+00
```

Scree plot

```r
library(factoextra)

fviz_eig(pca_result_cancer,
         addlabels = TRUE,
         geom = "bar",
         ylim = c(0, 100))
```

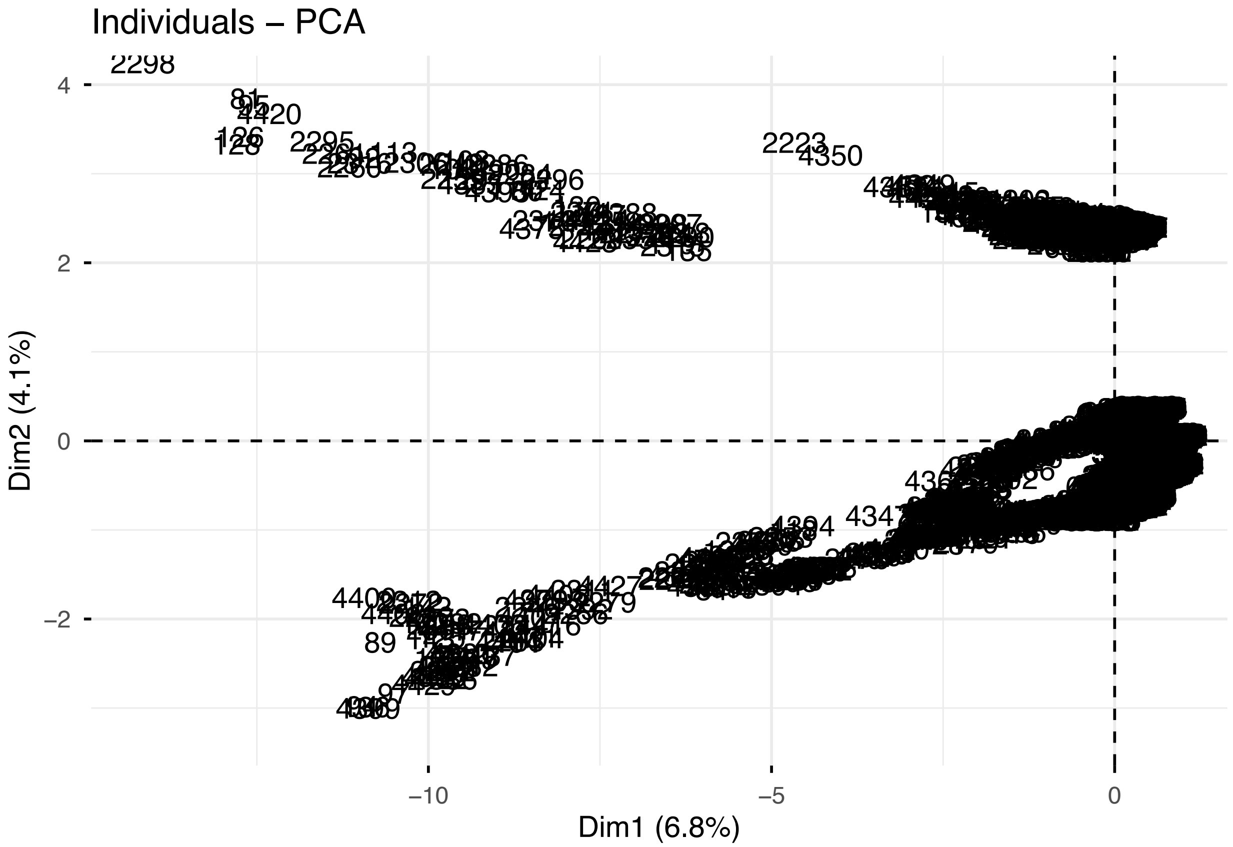
```
## Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
## Ignoring empty aesthetic: `width`.
```



The PCA results show that the structure of the dataset is weakly organized, with no dominant pattern explaining a substantial portion of the variance. The first principal component accounts for less than seven percent of the total variance, and adding several more components still captures only a modest amount of information, indicating that the variables are largely uncorrelated. This behavior is expected given that most variables are binary indicators representing mutually exclusive cancer sites or age groups, combined with a few continuous variables such as ASR. Such data provide little shared variance for PCA to extract, causing the method to distribute information across many dimensions rather than concentrating it in the first few components. Consequently, the PCA is not particularly effective at reducing dimensionality or revealing clear latent structures in this context. Nonetheless, it confirms that the dataset is highly sparse and that patterns such as sex-specific cancer types or age-related distributions arise more from categorical distinctions than from continuous correlations.

Biplot

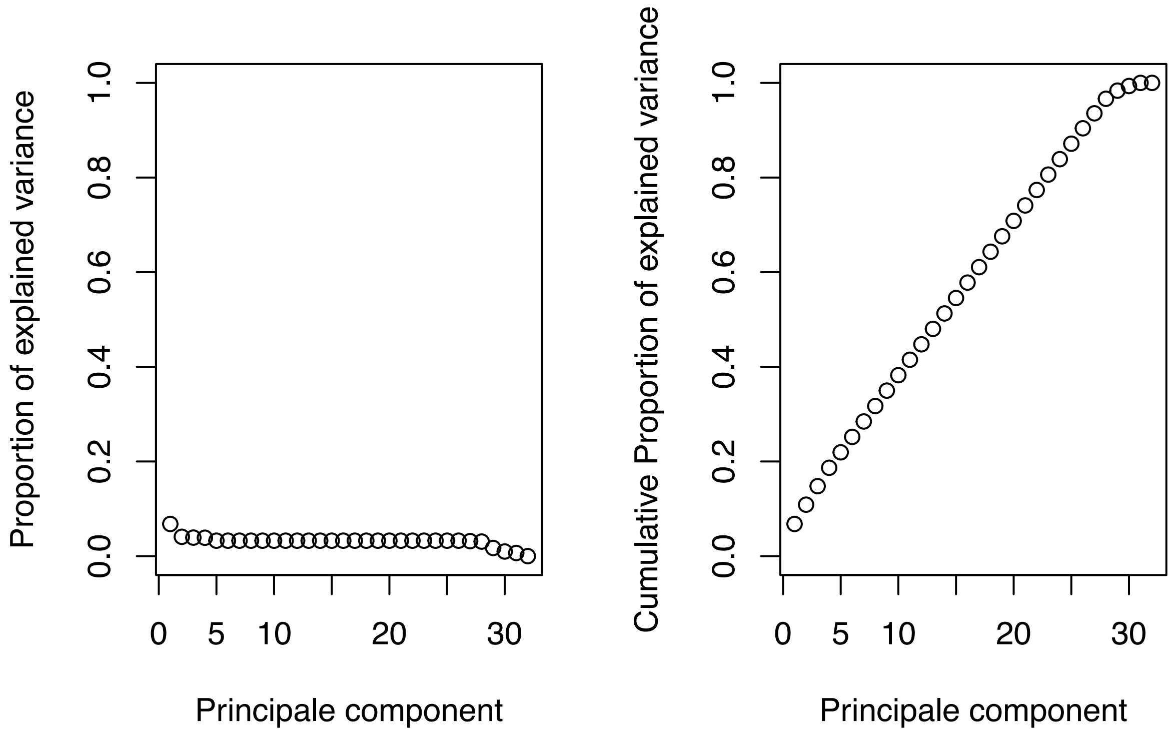
```r
fviz_pca_ind(pca_result_cancer, geom="text")
```



The PCA biplot appears cluttered and compact because the first components explain only a very small proportion of the total variance, and the dataset contains many weakly correlated variables. As a result, no dominant structure emerges, the projection collapses into a dense cloud, and the variable vectors become too small and dispersed to interpret meaningfully.

Variance and cumulative variance

```r
variance <- pca_result_cancer$sdev^2
pve <- variance / sum(variance)

par(mfrow = c(1, 2))


plot(pve, xlab = "Principale component",
     ylab = "Proportion of explained variance",
     ylim = c(0, 1), type = "b")

plot(cumsum(pve), xlab = "Principale component",
     ylab = "Cumulative Proportion of explained variance",
     ylim = c(0, 1), type = "b")
```



```
par(mfrow = c(1,1))
```

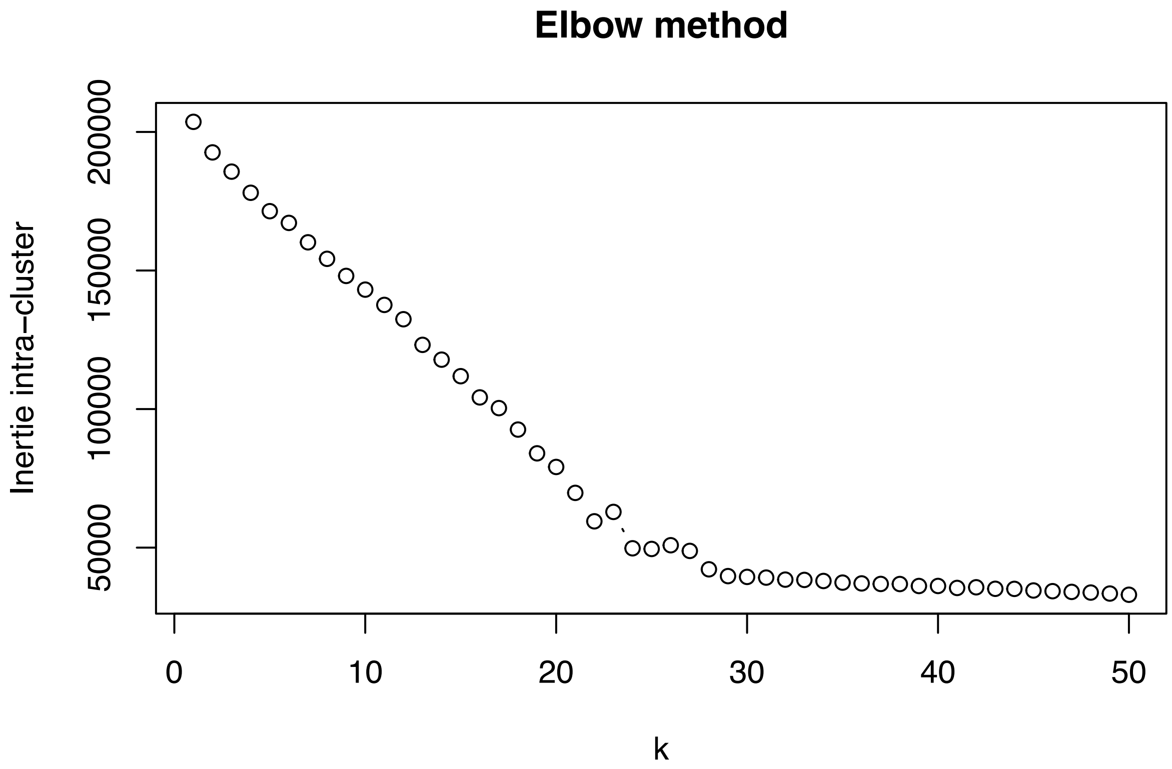
The plots show that proportion of explained variance is weak for each PC, there is no dominant axis. The cumulative explained variance is rather linear, that means there is no strong structure in the data.

2.2 K-means

Elbow method to choose k

First, we are doing it without the PCA because using PCA make us lose informations.

```
set.seed(123)
inertie <- numeric(50)

for (k in 1:50) {
  km <- kmeans(epicancer_scaled, centers = k, nstart = 25)
  inertie[k] <- km$tot.withinss
}

plot(1:50, inertie, type = "b",
     xlab = "k",
     ylab = "Inertie intra-cluster",
     main = "Elbow method")
```

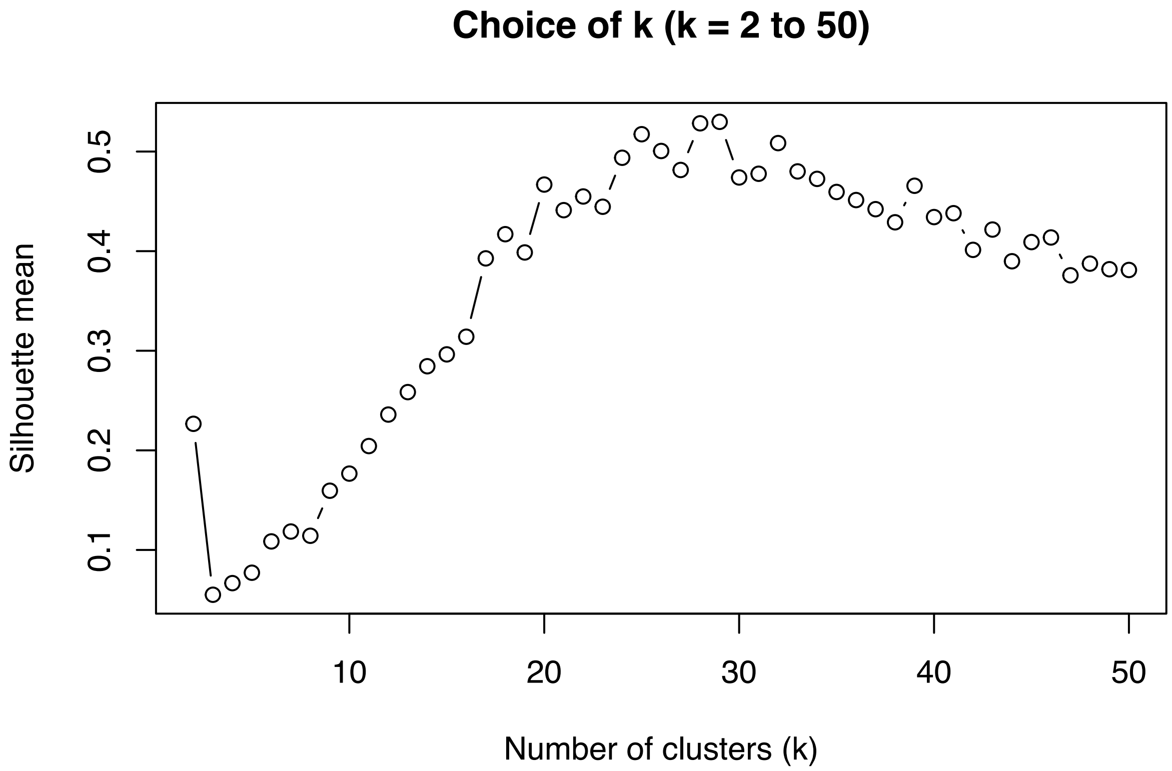


Here we don’t have a significant result for the elbow method so we’re gonna try with the silhouette method

```
sil_avg <- sapply(2:50, function(k) {
  km <- kmeans(epicancer_scaled, centers = k, nstart = 25)
  mean(silhouette(km$cluster, dist(epicancer_scaled))[,3])
})
```

Visualisation

```
plot(2:50, sil_avg, type="b",
     xlab="Number of clusters (k)",
     ylab="Silhouette mean",
     main="Choice of k (k = 2 to 50)")
```

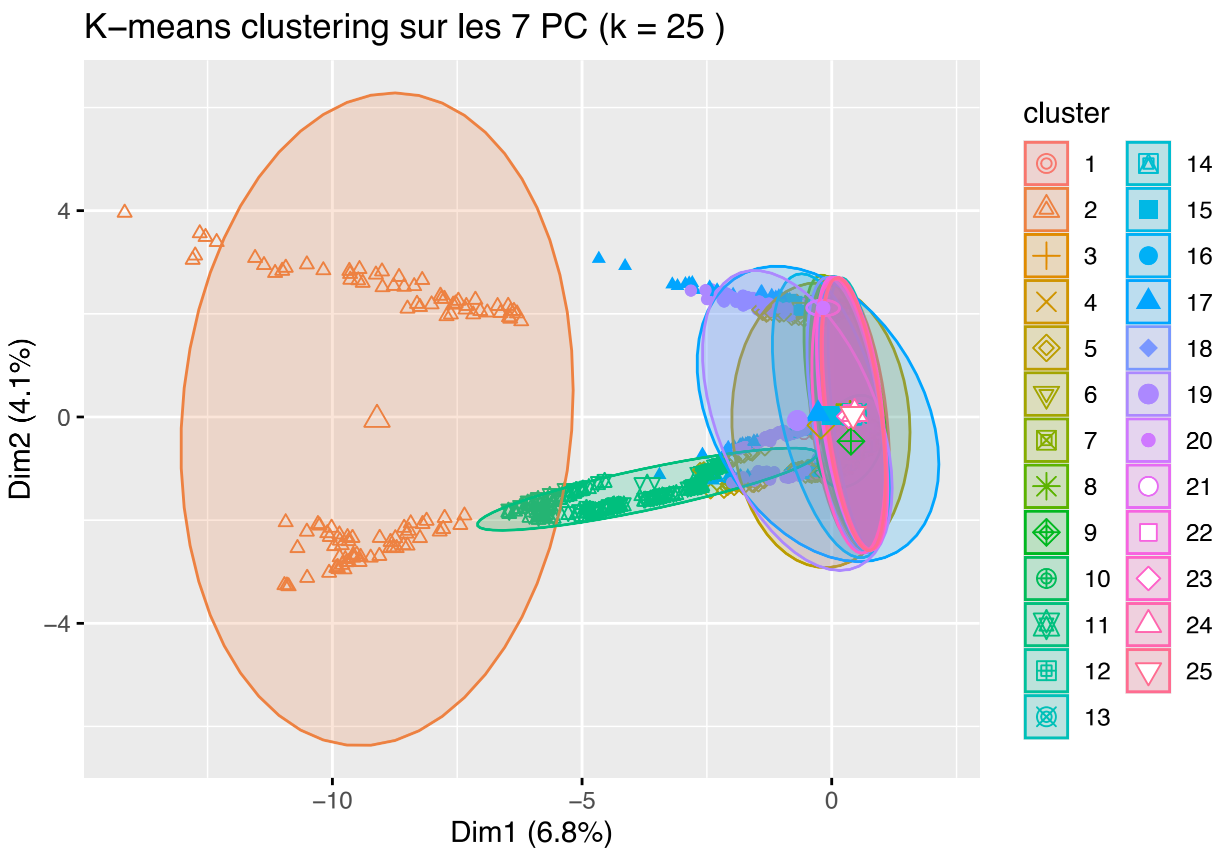


Best k

```r
k_values <- 2:50
best_k <- k_values[which.max(sil_avg)]

cat("Best k :", best_k, "\n")
```

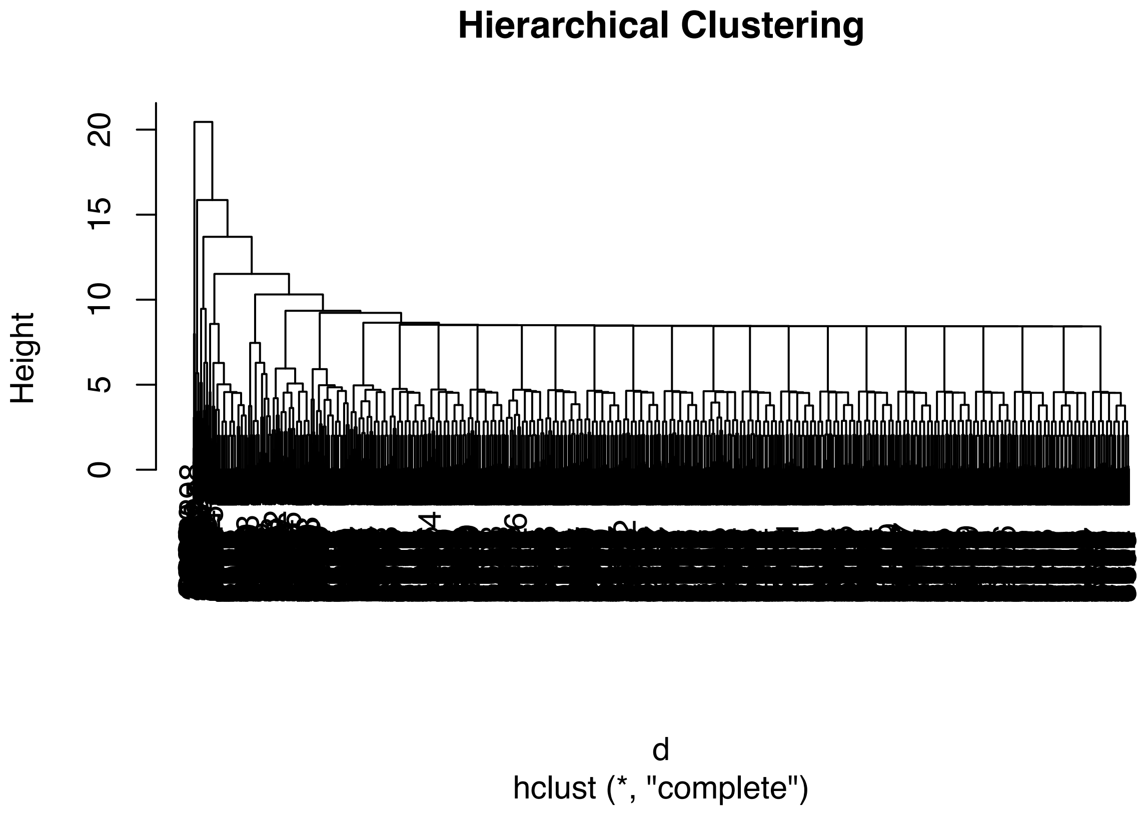
```
## Best k : 29
```

```r
k_optimal <- 25

set.seed(123)
km_result <- kmeans(epicancer_scaled, centers = k_optimal, nstart = 25)

#km_result$cluster

fviz_cluster(km_result, data = epicancer_scaled,
             geom = "point",
             ellipse.type = "norm",
             main = paste("K-means clustering sur les 7 PC (k =", k_optimal, ")"))
```



2.3 Hierarchical clustering

```
d <- dist(epicancer_scaled, method = "euclidean")

hc <- hclust(d, method = "complete")

plot(hc, main = "Hierarchical Clustering")
```



2.4 DBSCAN

We choosed to stop doing the Clustering with this dataset because the plot are not relevant.

3. Dataset reduced to breast and prostate cancers

```
cols_pca <- c("SEXE", "nb_cas", "asr",
              "CODE_CIM10C50",  # breast cancer
              "CODE_CIM10C61",  # prostate cancer
              grep("CLASSE_AGE", colnames(epidemiology_cancers_nc), value = TRUE)
              )
```

```
epicancer_pca <- epidemiology_cancers_nc[, cols_pca]
```

3.1 PCA

Normalisation

```
epicancer_scaled <- scale(epicancer_pca)
```

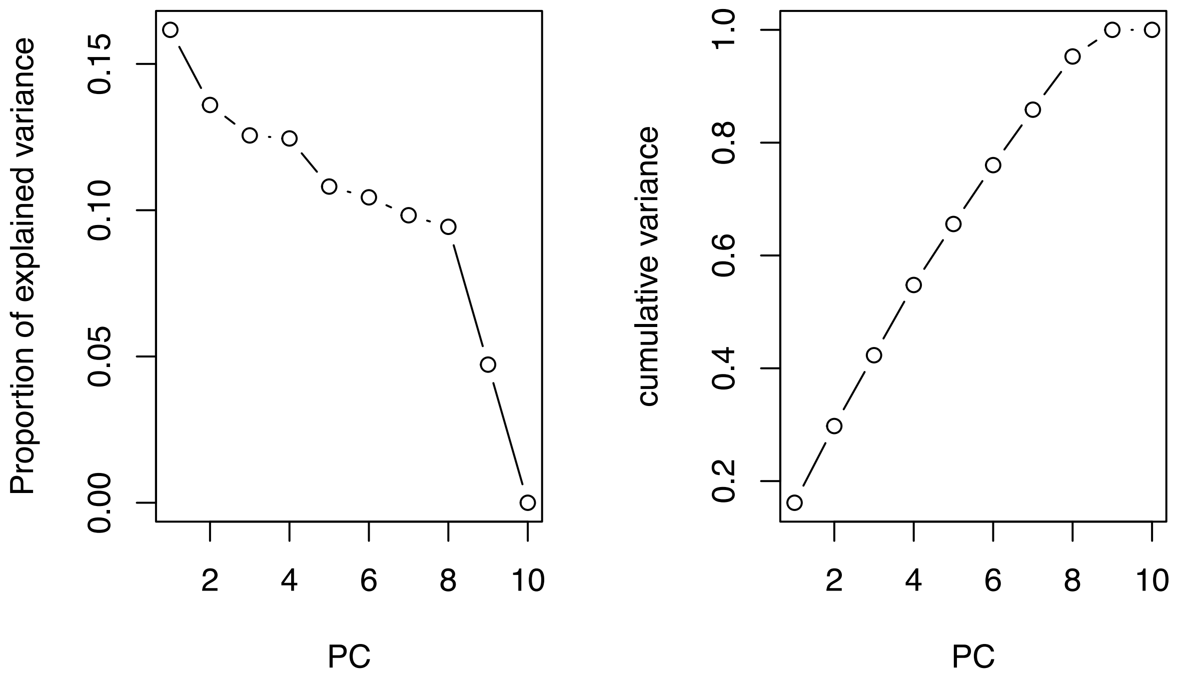
PCA applied

```
pca_result <- prcomp(epicancer_scaled, center = TRUE, scale. = FALSE)

summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.2714 1.1660 1.1206 1.1159 1.0396 1.0218 0.99131
## Proportion of Variance 0.1616 0.1360 0.1256 0.1245 0.1081 0.1044 0.09827
## Cumulative Proportion  0.1616 0.2976 0.4232 0.5477 0.6558 0.7602 0.85845
##                            PC8     PC9      PC10
## Standard deviation     0.97124 0.68714 1.557e-14
## Proportion of Variance 0.09433 0.04722 0.000e+00
## Cumulative Proportion  0.95278 1.00000 1.000e+00
```

Scree plot and cumulative variance

```
par(mfrow = c(1, 2))

variance <- pca_result$sdev^2
pve <- variance / sum(variance)

plot(pve[1:10], type="b", xlab="PC", ylab="Proportion of explained variance")

plot(cumsum(pve[1:10]), type="b", xlab="PC", ylab="cumulative variance")
```



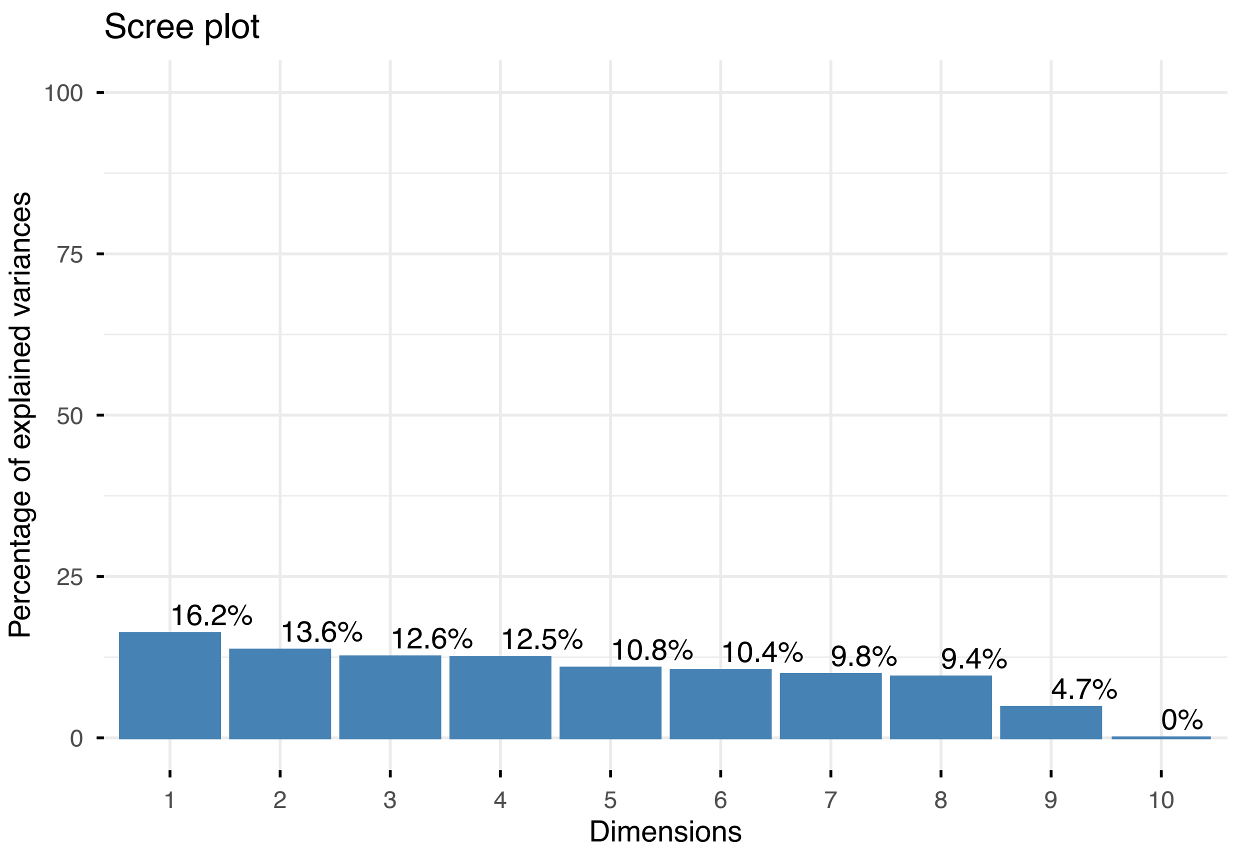
```
par(mfrow = c(1,1))
```

In this second set of PCA results, the first principal component explains a larger share of variance (16.16%), with PC2 at 13.60%, PC3 at 12.56%, and so on. The cumulative variance reaches 100% by PC9, suggesting that fewer components are needed to capture most of the dataset's variation in this analysis.

Scree plot

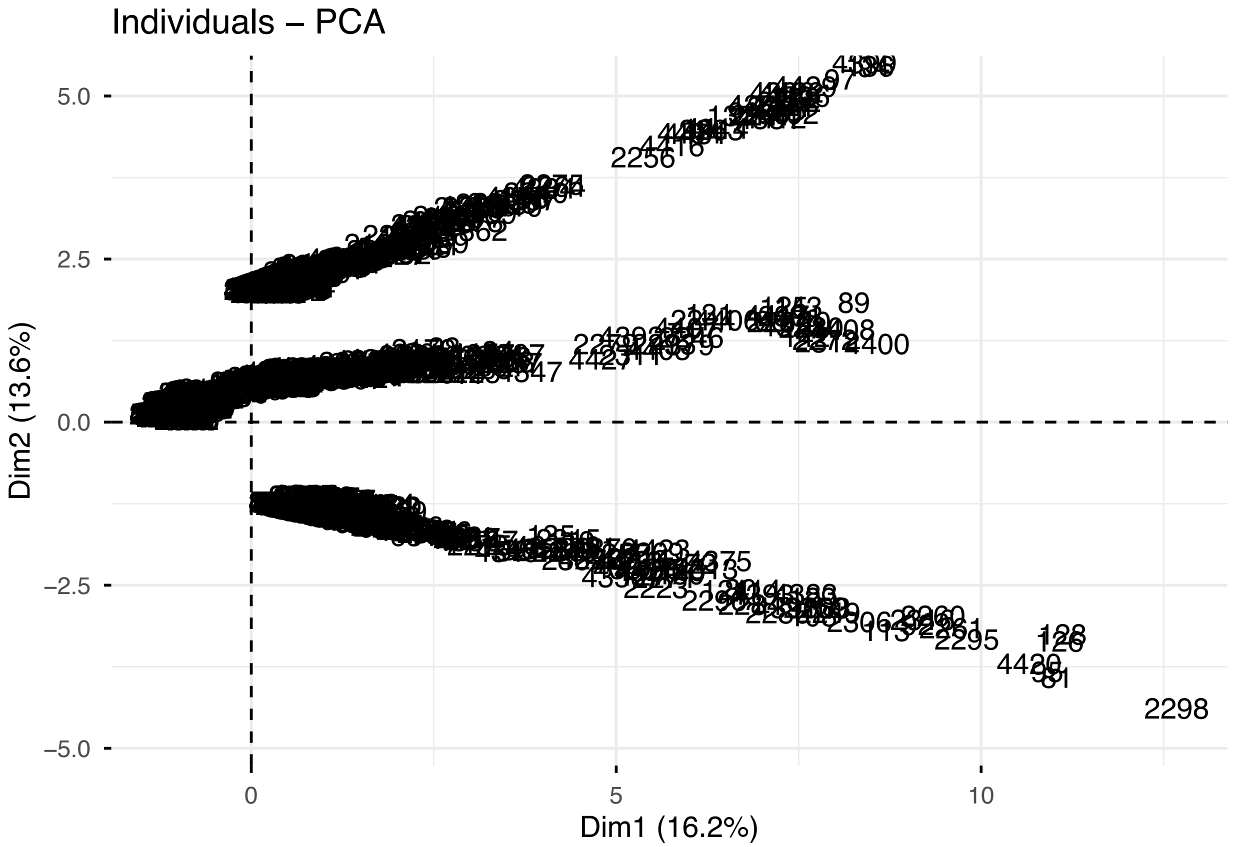
```
library(factoextra)

fviz_eig(pca_result,
         addlabels = TRUE,
         geom = "bar",
         ylim = c(0, 100))
```

```
## Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
## Ignoring empty aesthetic: `width`.
```



Biplot

```
fviz_pca_ind(pca_result, geom="text")
```

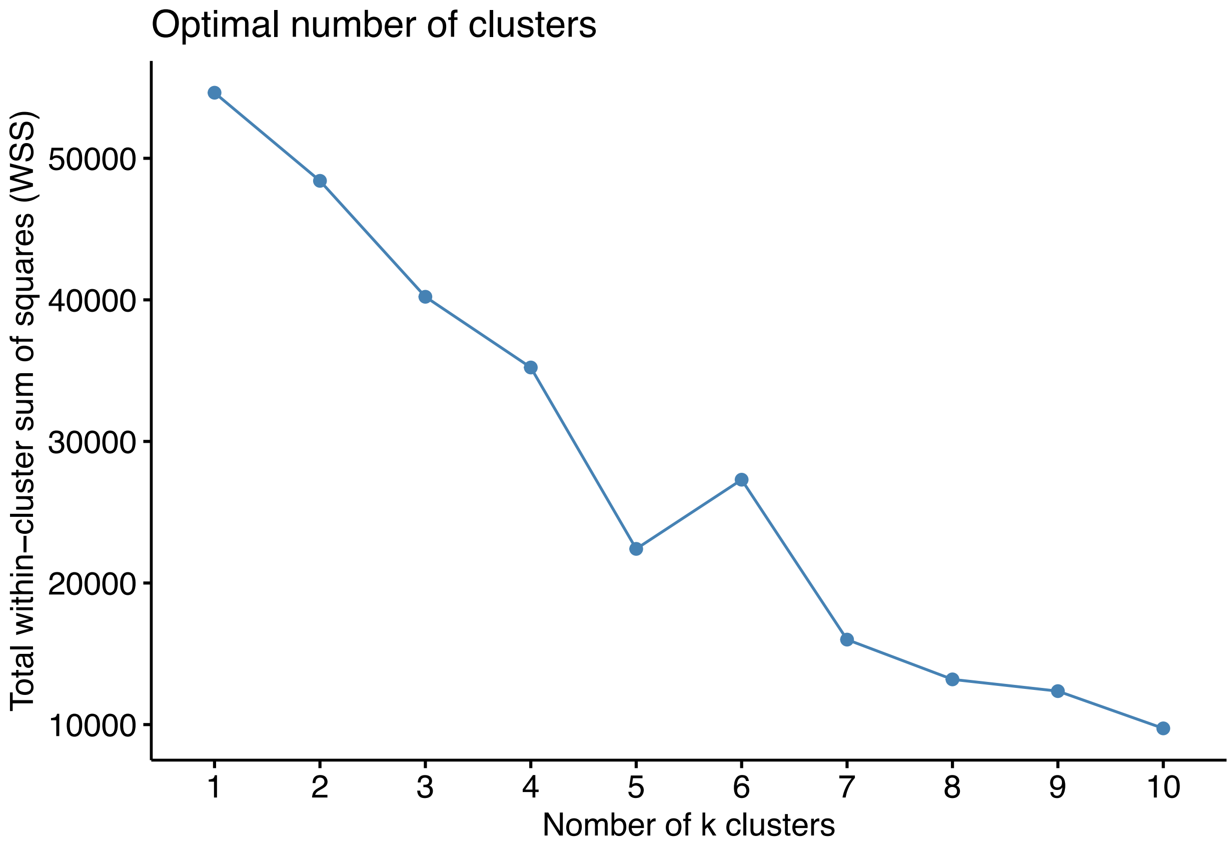


3.2 K-means

Elbow method to choose k

We do the albow again with the new data to see if we find a smaller one

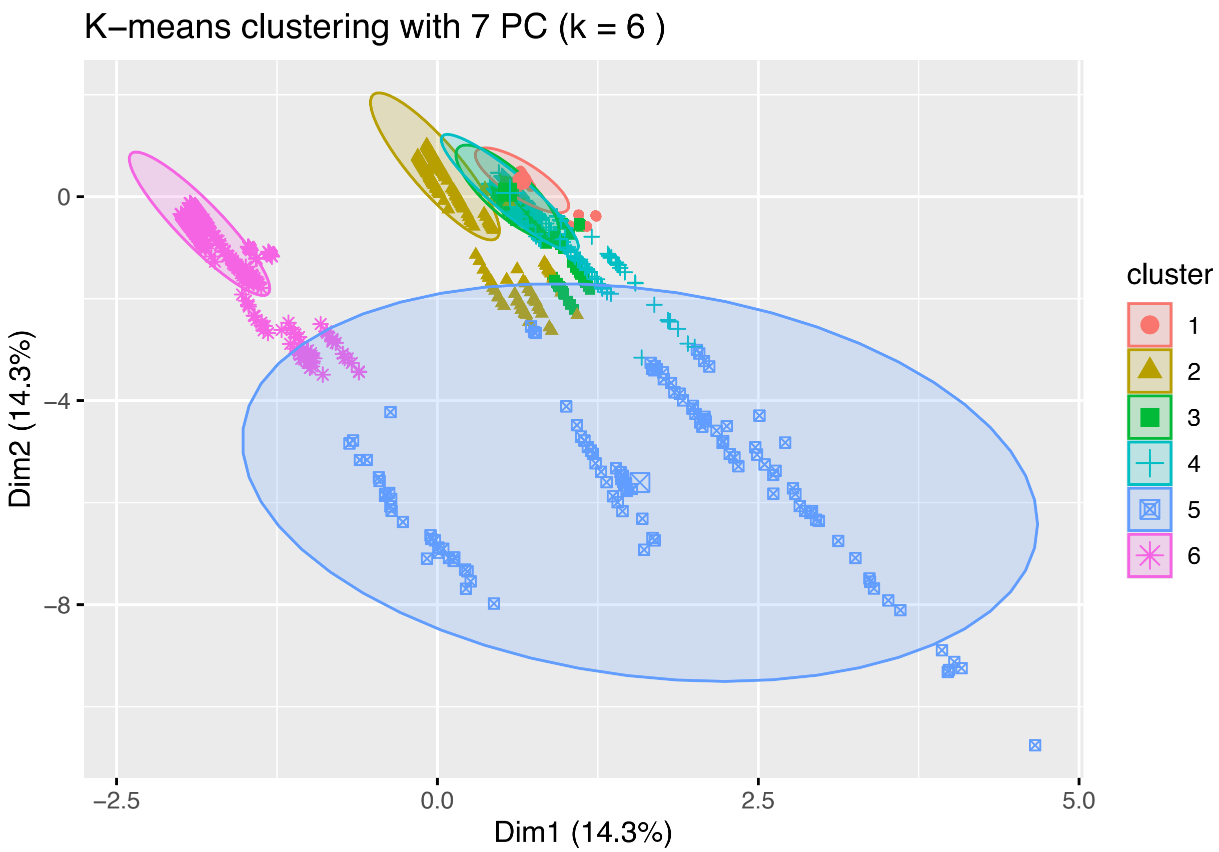
```r
pca_scores7 <- pca_result$x[, 1:7]  # 7 first PC (=85%)

set.seed(123)

fviz_nbclust(pca_scores7, kmeans, method = "wss") +
  labs(x = "Nomber of k clusters",
       y = "Total within-cluster sum of squares (WSS)")
```



So, we can choose k=6

Visualization

```r
k_optimal <- 6

set.seed(123)
km_result <- kmeans(pca_scores7, centers = k_optimal, nstart = 25)

#km_result$cluster

fviz_cluster(km_result, data = pca_scores7,
             geom = "point",
             ellipse.type = "norm",
             main = paste("K-means clustering with 7 PC (k =", k_optimal, ")"))
```



### Informations

```
km_result$centers
```

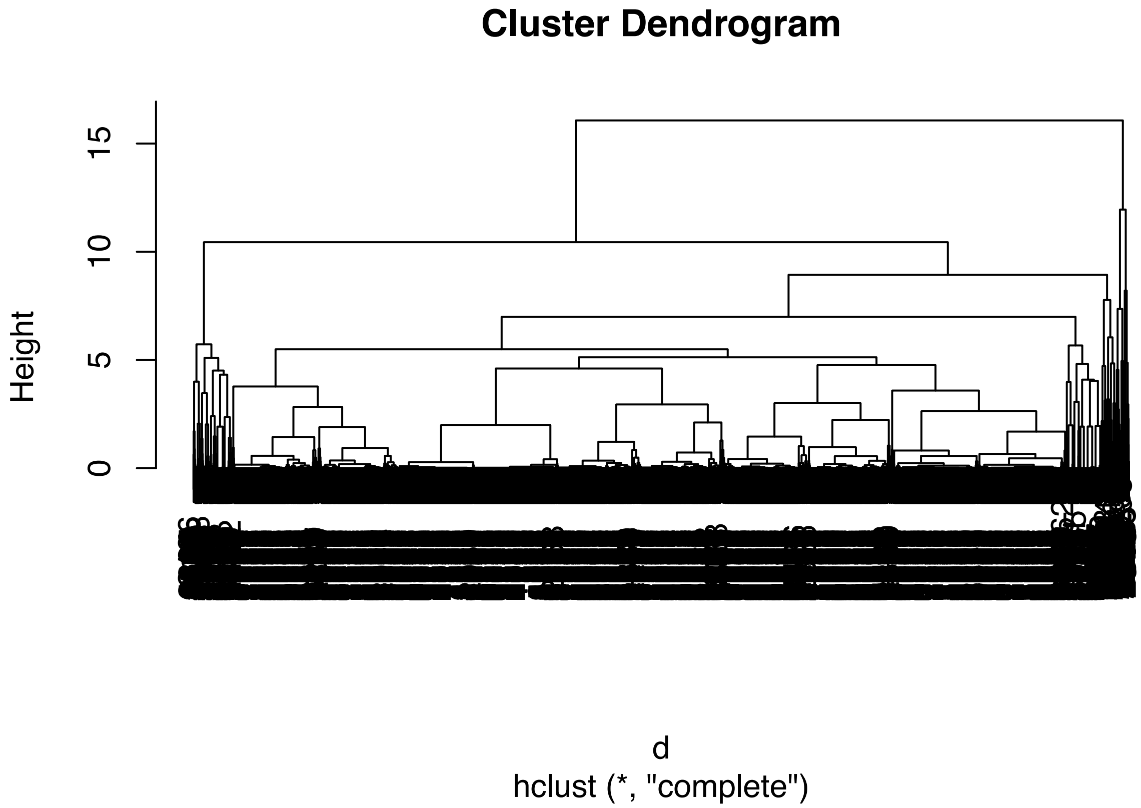
```
##          PC1         PC2           PC3        PC4        PC5         PC6
## 1 -1.1488806 -0.27183612  1.339998e-13 -1.6679582  1.0289511 -0.05774108
## 2  0.3862857  1.73288456 -1.020495e+00 -0.3228440 -0.6456435  0.09494856
## 3 -0.9094124 -0.08012292 -3.702574e-01  1.8479623  0.7290936 -0.03883402
## 4  0.7093069 -1.67103626 -7.224341e-01 -0.1489650 -0.8893431 -0.03147384
## 5  6.6928296 -0.08522424 -1.090466e-01  0.1539991  2.9907530 -0.01576371
## 6  0.2943296  0.24466641  2.113186e+00  0.1177908 -0.4728805  0.03095517
##          PC7
## 1  0.4401092
## 2 -0.2618771
## 3  0.1371260
## 4 -0.2306425
## 5  1.2107508
## 6 -0.1775724
```

```
pca_loadings <- pca_result$rotation[, 1:7]  # PC1 to PC7
pca_loadings
```

```
##                                  PC1          PC2           PC3          PC4
## SEXE                      0.04659327 -0.021421941  1.749729e-14 -0.066277189
## nb_cas                    0.55182661  0.278752205  9.223137e-14  0.031724377
## asr                       0.59366043 -0.207006542  1.009356e-14  0.022785326
## CODE_CIM10C50        0.03525420  0.027220948  6.603995e-15  0.005353084
## CODE_CIM10C61        0.04918498 -0.009565325  6.955544e-15  0.005890238
## CLASSE_AGE00.24     -0.33925261 -0.095435185  5.690549e-14 -0.639387776
## CLASSE_AGE25.49     -0.28448416 -0.029799367 -1.491087e-01  0.750448990
## CLASSE_AGE50.74      0.13650736  0.097328507  8.510149e-01  0.050624037
## CLASSE_AGE75.        0.30674334 -0.642826449 -2.909361e-01 -0.056074119
## CLASSE_AGETous.ages  0.16873776  0.667427574 -4.109701e-01 -0.127753131
##                             PC5          PC6         PC7
## SEXE                 0.43111049 -0.038215279 -0.85412011
## nb_cas               0.35636707 -0.045529669  0.17413093
## asr                  0.32927046 -0.002857545  0.11810672
## CODE_CIM10C50        0.08723079 -0.719805566  0.27619249
## CODE_CIM10C61        0.18046027  0.689478951  0.26993885
## CLASSE_AGE00.24      0.45443719 -0.026400765  0.21377979
## CLASSE_AGE25.49      0.34112322 -0.018810177  0.07056264
## CLASSE_AGE50.74     -0.18111340  0.013095103 -0.07326491
## CLASSE_AGE75.       -0.34204621 -0.009380459 -0.09024957
## CLASSE_AGETous.ages -0.25666364  0.040582039 -0.11342475
```

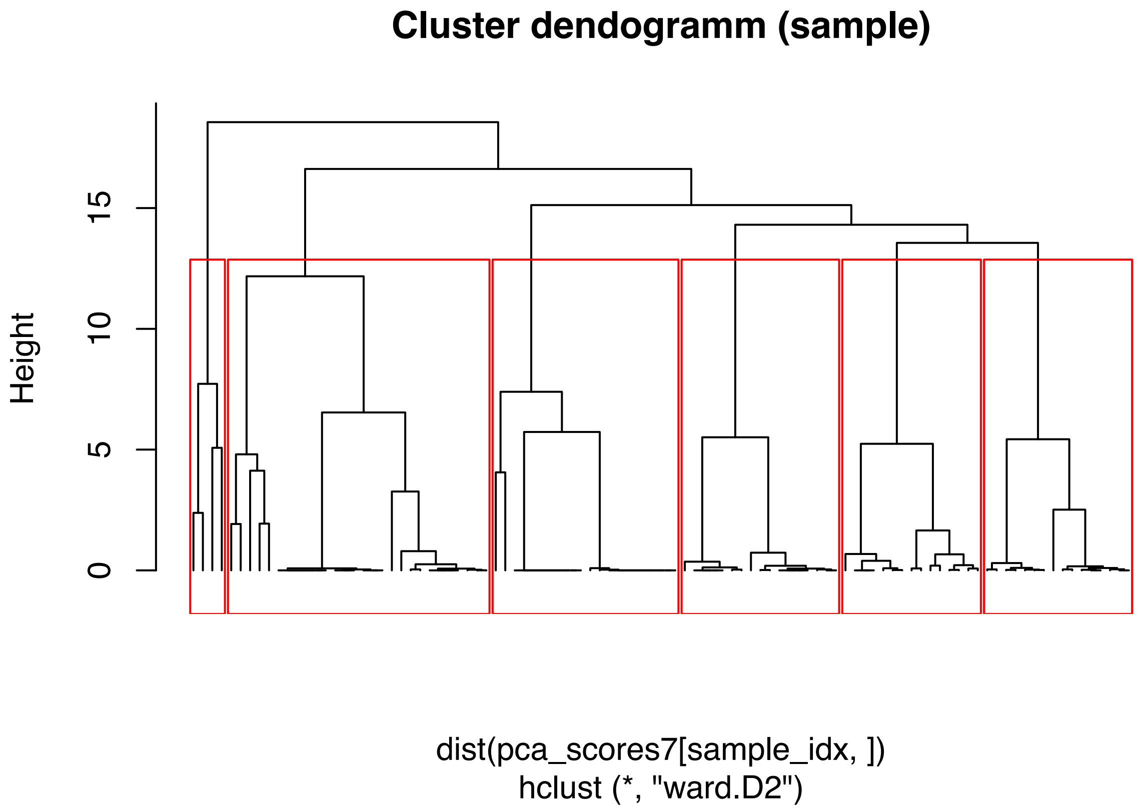
3.3 Hierarchical clustering

```
d <- dist(pca_scores7)
hc <- hclust(d, method = "complete")
plot(hc)
```



It is complicated to see, so we're going to use de ward method.

```
hc <- hclust(d, method="ward.D2")

# sample (max 100 points)
sample_idx <- sample(1:nrow(pca_scores7), size=min(100, nrow(pca_scores7)))
hc_sample <- hclust(dist(pca_scores7[sample_idx,]), method="ward.D2")


plot(hc_sample, labels=FALSE, hang=-1,
     main="Cluster dendogramm (sample)")
rect.hclust(hc_sample, k=6, border="red")  #6 clusters
```

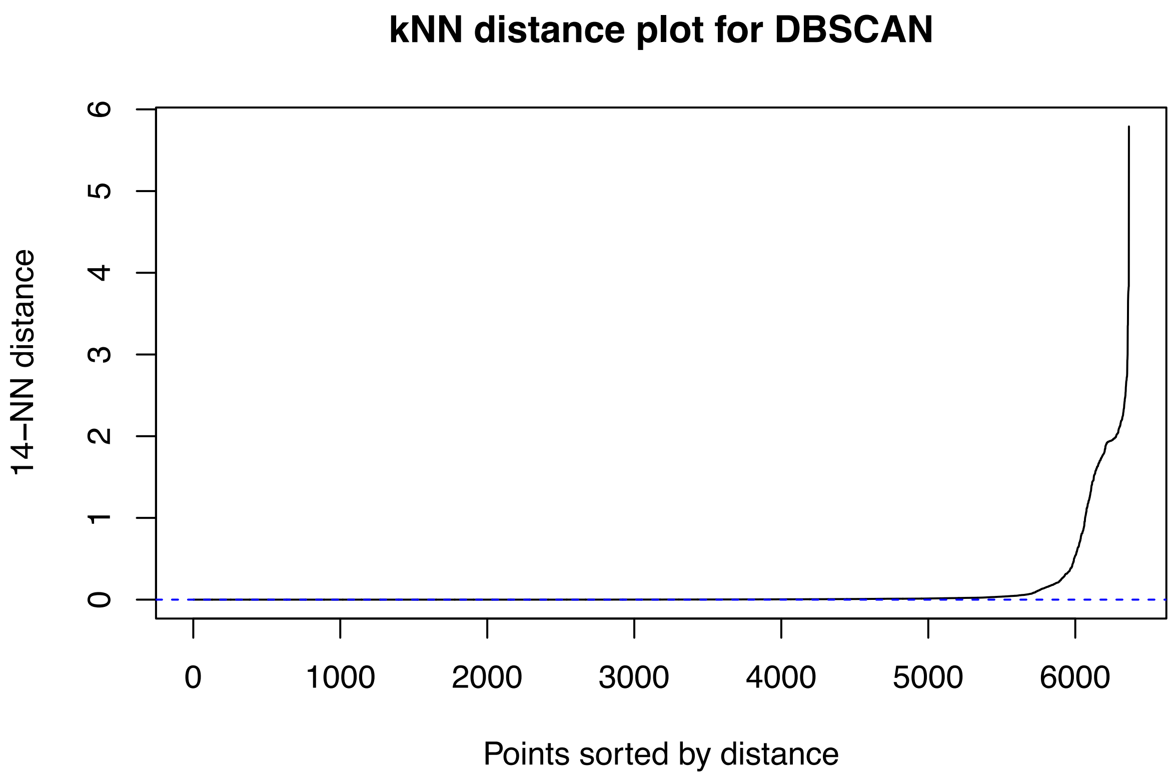


3.4 DBSCAN

kNN distance plot

We do this in order to choose the right epsilon (eps)

```
minPts <- 2 * ncol(pca_scores7)

kNNdistplot(pca_scores7, k = minPts)
abline(h=0, col="blue", lty=2)
title(main="kNN distance plot for DBSCAN")
```



We may choose eps = 2.

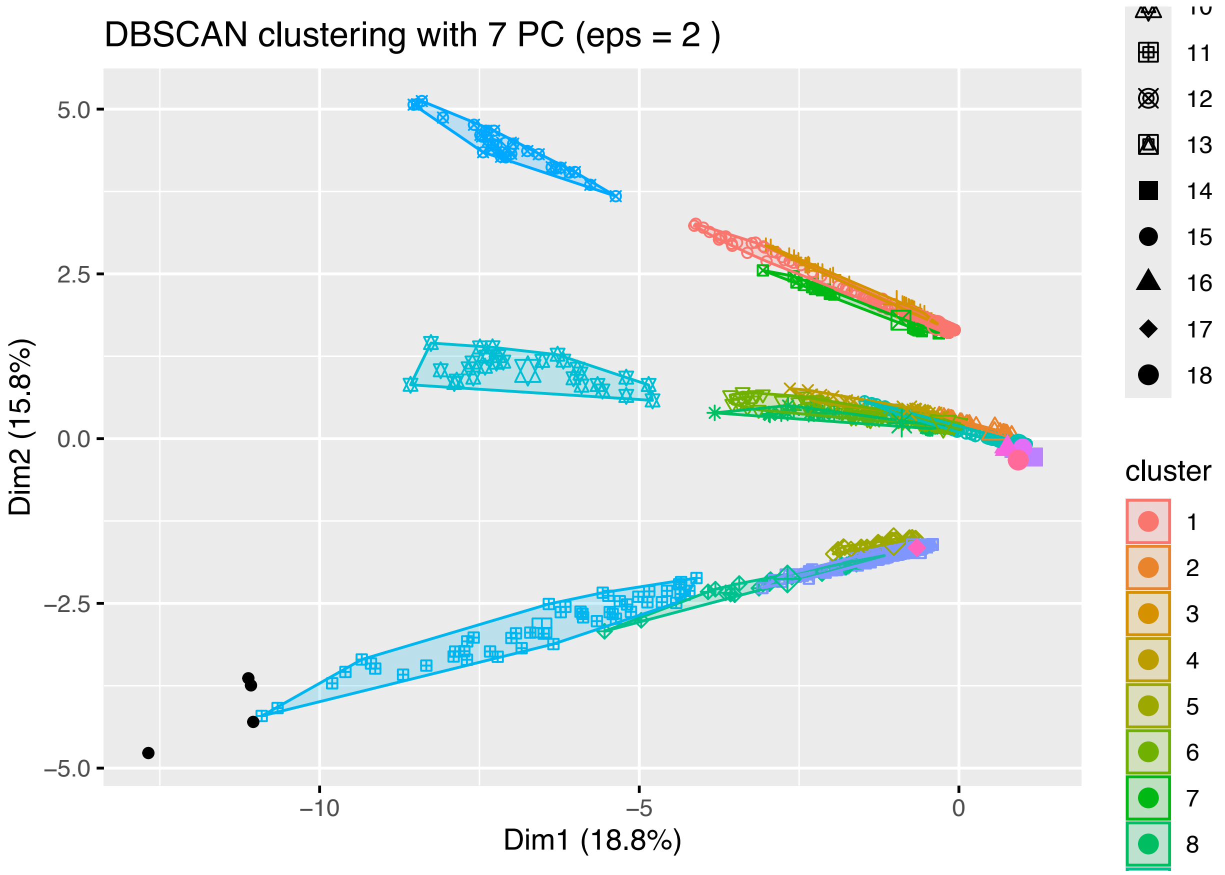
```
eps_val <- 2

dbscan_result <- dbscan(pca_scores7, eps = eps_val, minPts = minPts)

# clusters attributed
table(dbscan_result$cluster)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##    4 1160   54   54   54   54 1160   54   54   27 1188   28   50   28 1134 1083
##   16   17   18   19
##   49   54   27   49
```

Visualization

```
fviz_cluster(dbscan_result, data = pca_scores7,
             stand = FALSE,
             geom = "point",
             main = paste("DBSCAN clustering with 7 PC (eps =", eps_val, ")"))
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



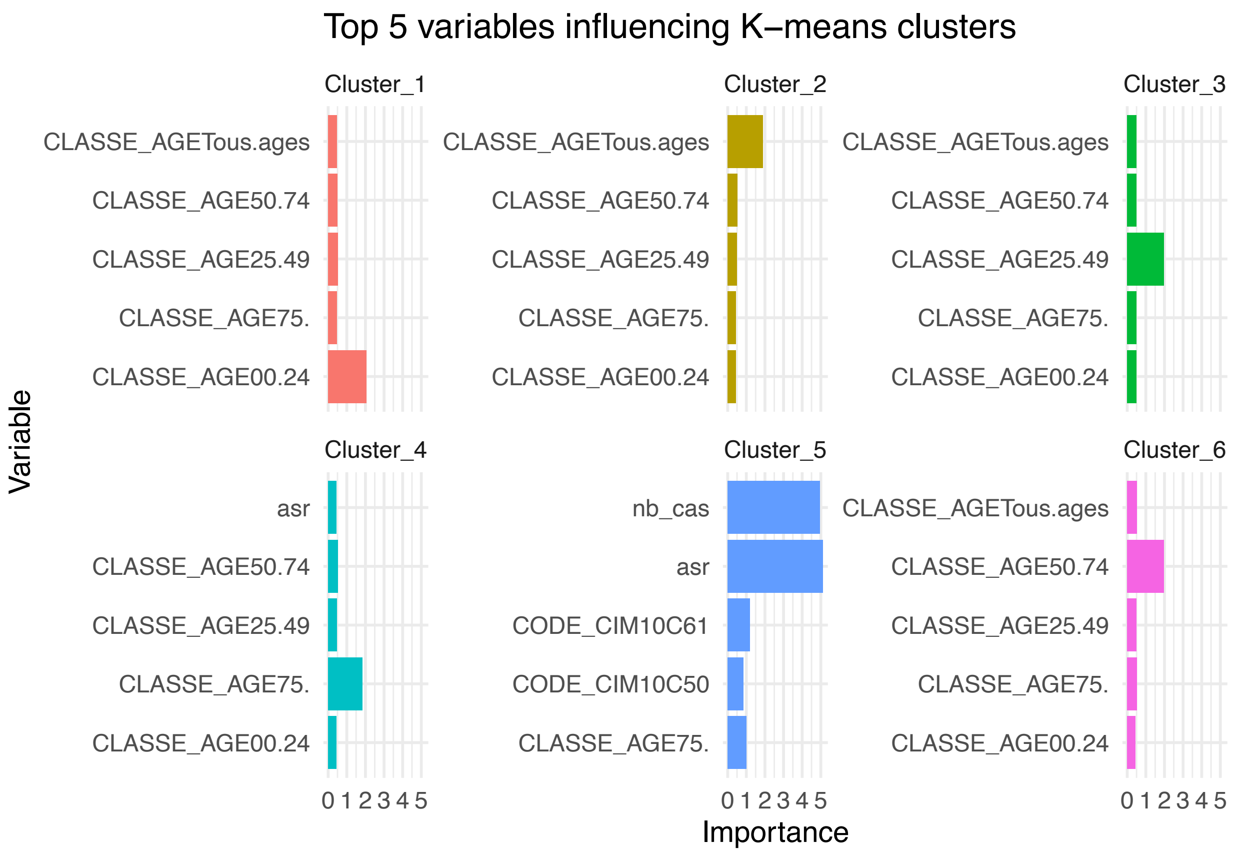
# 4. Interpretation of clustering for the dataset reduced to breast and prostate cancers

```
compute_and_plot_variable_importance <- function(pca_scores, cluster_labels, pca_loadings,
                                                 top_n = 5, plot_title = "Variable importance per clust
  library(dplyr)
  library(ggplot2)

  clusters_unique <- sort(unique(cluster_labels))

  cluster_centers <- sapply(clusters_unique, function(cl) {
    colMeans(pca_scores[cluster_labels == cl, , drop = FALSE])
  })
  cluster_centers <- t(cluster_centers)

  get_variable_importance <- function(cluster_center, pca_loadings){
    cluster_center <- as.matrix(cluster_center, ncol = 1)
    importance <- abs(pca_loadings %*% cluster_center)
    importance <- as.vector(importance)
    names(importance) <- rownames(pca_loadings)
```

```r
    sort(importance, decreasing = TRUE)
  }

  all_importance <- data.frame()
  for (i in 1:nrow(cluster_centers)) {
    imp <- get_variable_importance(cluster_centers[i, ], pca_loadings)
    df <- data.frame(
      Variable = names(imp),
      Importance = imp,
      Cluster = paste0("Cluster_", clusters_unique[i])
    )
    all_importance <- rbind(all_importance, df)
  }

  top_df <- all_importance %>%
    group_by(Cluster) %>%
    slice_max(order_by = Importance, n = top_n)

  p <- ggplot(top_df, aes(x = reorder(Variable, Importance), y = Importance, fill = Cluster)) +
    geom_bar(stat = "identity") +
    coord_flip() +
    facet_wrap(~Cluster, scales = "free_y") +
    labs(title = plot_title, x = "Variable", y = "Importance") +
    theme_minimal() +
    theme(legend.position = "none")

  print(p)
  return(top_df)
}
```

4.1 K-means

```r
top_kmeans <- compute_and_plot_variable_importance(
  pca_scores = pca_scores7,
  cluster_labels = km_result$cluster,
  pca_loadings = pca_result$rotation[, 1:7],
  plot_title = "Top 5 variables influencing K-means clusters"
)
```

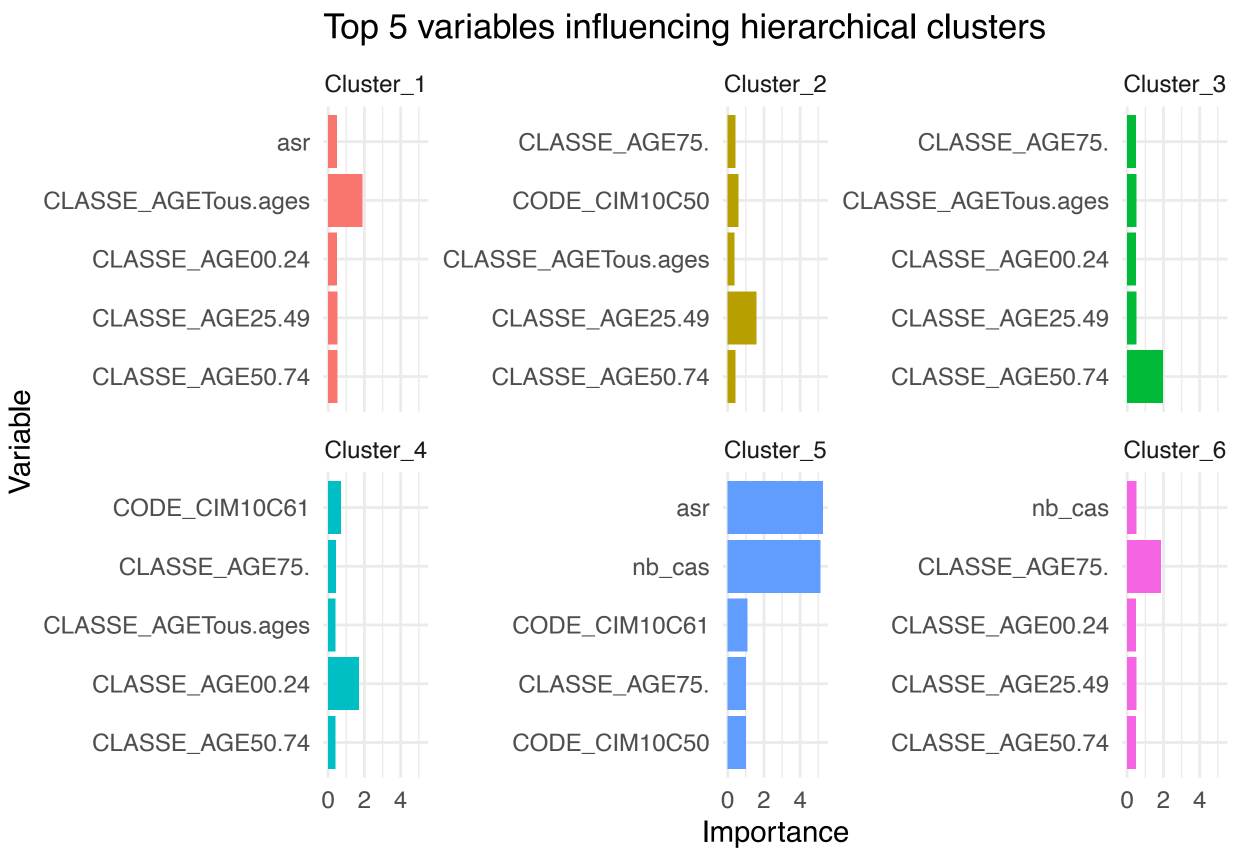


```
top_kmeans
```

```
## # A tibble: 30 x 3
## # Groups:   Cluster [6]
##    Variable            Importance Cluster
##    <chr>                    <dbl> <chr>
##  1 CLASSE_AGE00.24          2.05  Cluster_1
##  2 CLASSE_AGE25.49          0.534 Cluster_1
##  3 CLASSE_AGE50.74          0.487 Cluster_1
##  4 CLASSE_AGETous.ages      0.479 Cluster_1
##  5 CLASSE_AGE75.            0.475 Cluster_1
##  6 CLASSE_AGETous.ages      1.88  Cluster_2
##  7 CLASSE_AGE50.74          0.526 Cluster_2
##  8 CLASSE_AGE25.49          0.492 Cluster_2
##  9 CLASSE_AGE00.24          0.442 Cluster_2
## 10 CLASSE_AGE75.            0.437 Cluster_2
## # i 20 more rows
```

4.2 Hierarchical clustering

```
hc_clusters <- cutree(hc, k = 6)

top_hierarchical <- compute_and_plot_variable_importance(
  pca_scores = pca_scores7,
  cluster_labels = hc_clusters,
  pca_loadings = pca_result$rotation[, 1:7],
  plot_title = "Top 5 variables influencing hierarchical clusters"
)
```



```
top_hierarchical
```

```
## # A tibble: 30 x 3
## # Groups:   Cluster [6]
##    Variable            Importance Cluster
##    <chr>                    <dbl> <chr>
##  1 CLASSE_AGETous.ages      1.88  Cluster_1
##  2 CLASSE_AGE25.49          0.515 Cluster_1
##  3 CLASSE_AGE50.74          0.514 Cluster_1
##  4 asr                      0.491 Cluster_1
##  5 CLASSE_AGE00.24          0.480 Cluster_1
##  6 CLASSE_AGE25.49          1.59  Cluster_2
##  7 CODE_CIM10C50            0.598 Cluster_2
##  8 CLASSE_AGE75.            0.434 Cluster_2
##  9 CLASSE_AGE50.74          0.423 Cluster_2
## 10 CLASSE_AGETous.ages      0.380 Cluster_2
## # i 20 more rows
```