ANÁLISIS DE COMPONENTES PRINCIPALES PARA CLASIFICACIÓN DE GRUPOS DE COMIDA CON BASE EN SU INFORMACIÓN NUTRIMENTAL

Elizabeth Rodríguez

Elizabeth Viveros

Leonardo Marín

Ángel Rafael Ortega

Karla Alfaro

Mario Rodríguez

TABLA DE CONTENIDOS

- Introducción
- Datos y contexto
- Análisis exploratorio de los datos
- Teoría del Análisis de Componentes Principales
- Resultados del Análisis de Componentes Principales
- Análisis de conglomerados (clustering)
- Conclusiones
- Referencias

INTRODUCCIÓN

- Métodos numéricos: Análisis de Componentes Principales con los algoritmos:
  - Algoritmo de SVD.
  - Algoritmo QR.
  - Método de la potencia.
  - PCA de sklearn.

- Análisis no supervisado de clasificación de los grupos de alimentos con base en su información nutrimental habiendo eliminado la multicolinealidad de la base al aplicar Análisis de Componentes Principales.

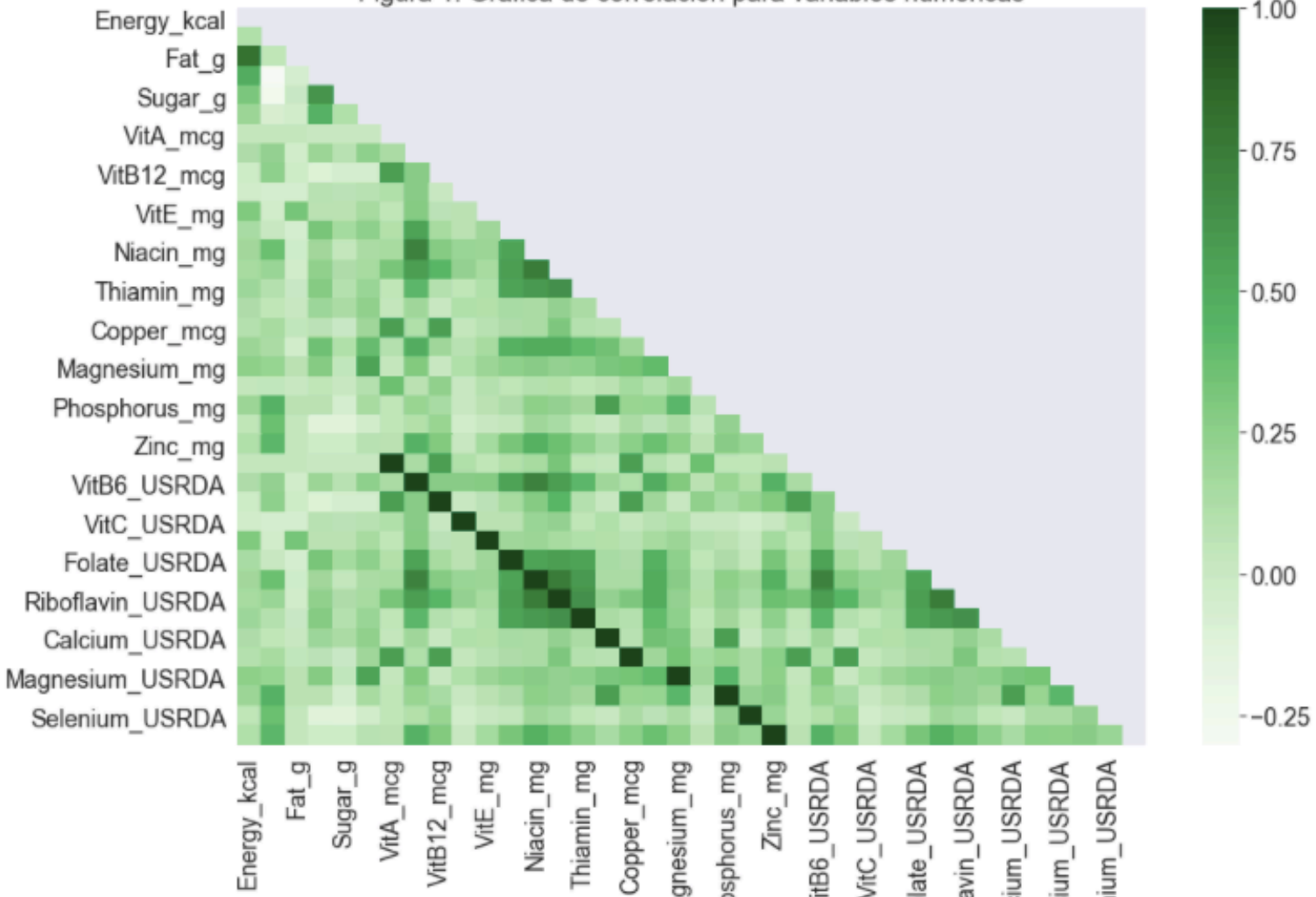
DATOS Y CONTEXTO

- Datos sobre composición alimenticia de la USDA National Nutrient Database for Standard Reference (SR) (2014).
- 8,618 observaciones y 45 variables

ID	FoodGroup	ShortDescrip	Descrip	CommonName	MfgName	ScientificName	Energy_kcal	Protein_g	Fat_g	...	Folate_USRDA	Niacin_USRDA
1001	Dairy and Egg Products	BUTTER,WITH SALT	Butter, salted	NaN	NaN	NaN	717	0.85	81.11	...	0.0075	0.002625
1002	Dairy and Egg Products	BUTTER,WHIPPED,WITH SALT	Butter, whipped, with salt	NaN	NaN	NaN	717	0.85	81.11	...	0.0075	0.002625
1003	Dairy and Egg Products	BUTTER OIL,ANHYDROUS	Butter oil, anhydrous	NaN	NaN	NaN	876	0.28	99.48	...	0.0000	0.000188
1004	Dairy and Egg Products	CHEESE,BLUE	Cheese, blue	NaN	NaN	NaN	353	21.40	28.74	...	0.0900	0.063500
1005	Dairy and Egg Products	CHEESE,BRICK	Cheese, brick	NaN	NaN	NaN	371	23.24	29.68	...	0.0500	0.007375

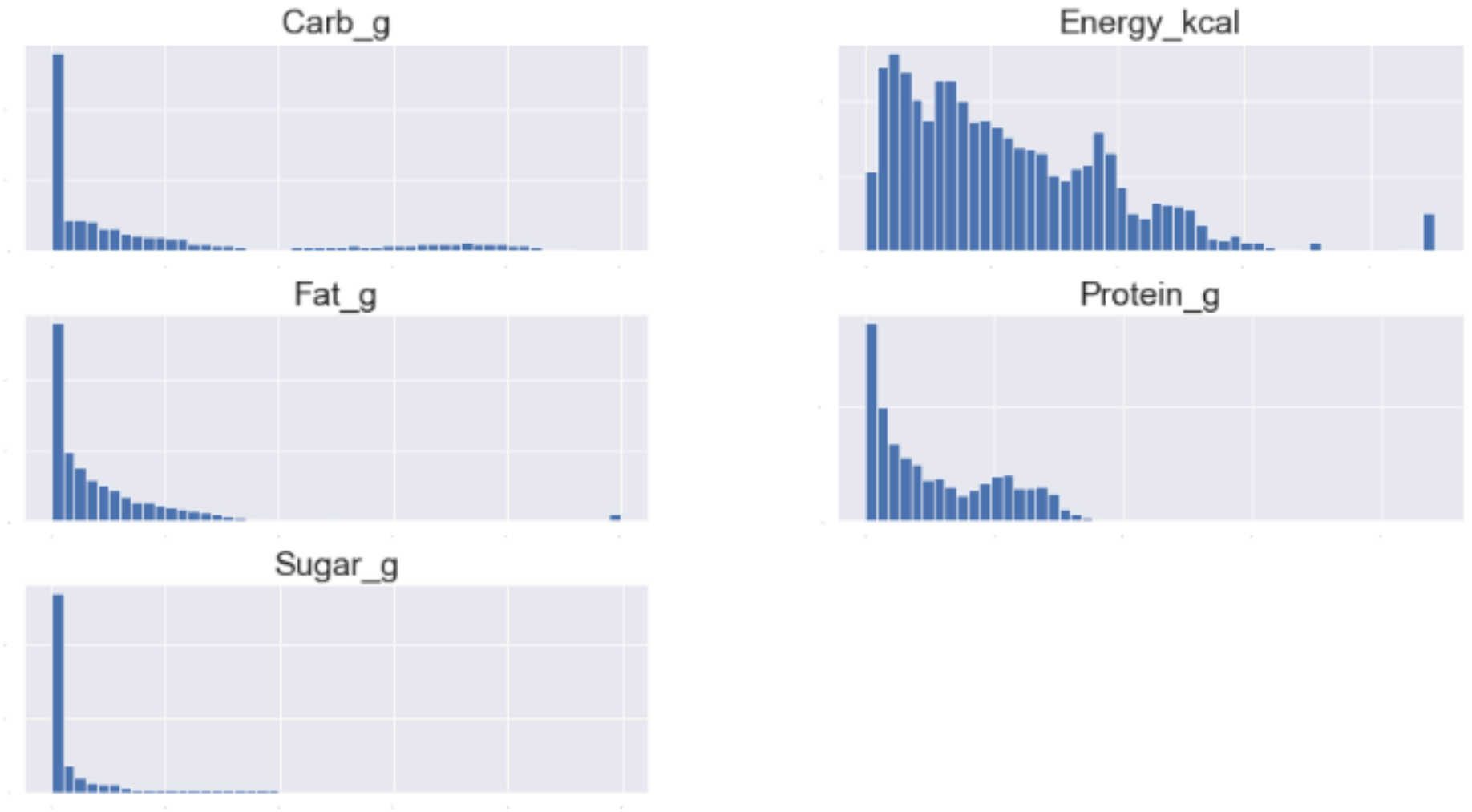
DIAGNÓSTICO DE MULTICOLINEALIDAD

Figura 1: Gráfica de correlación para variables numéricas



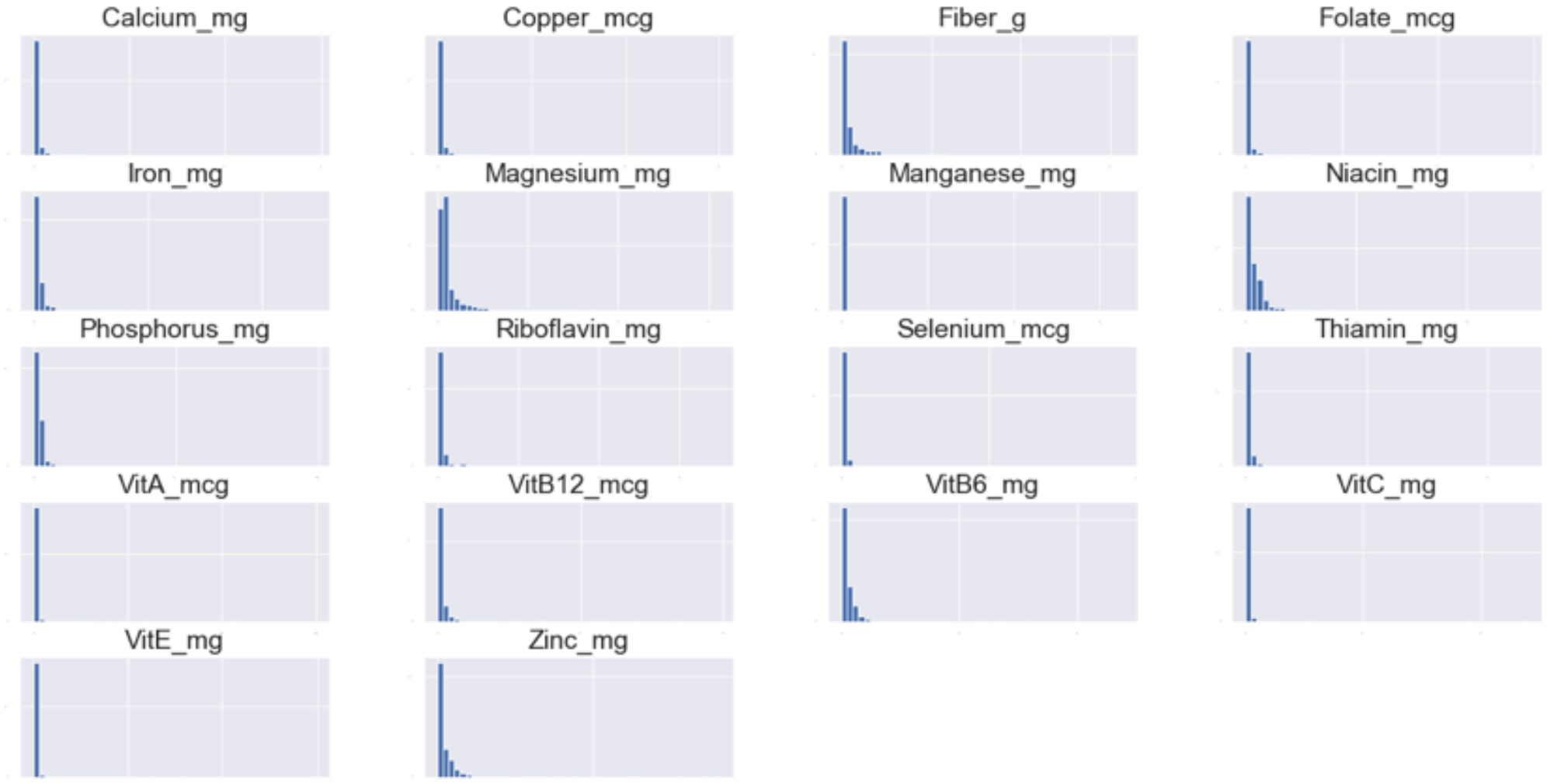
DISTRIBUCIÓN DE LOS MACRONUTRIENTES

Figura 2: Distribución de los macronutrientes

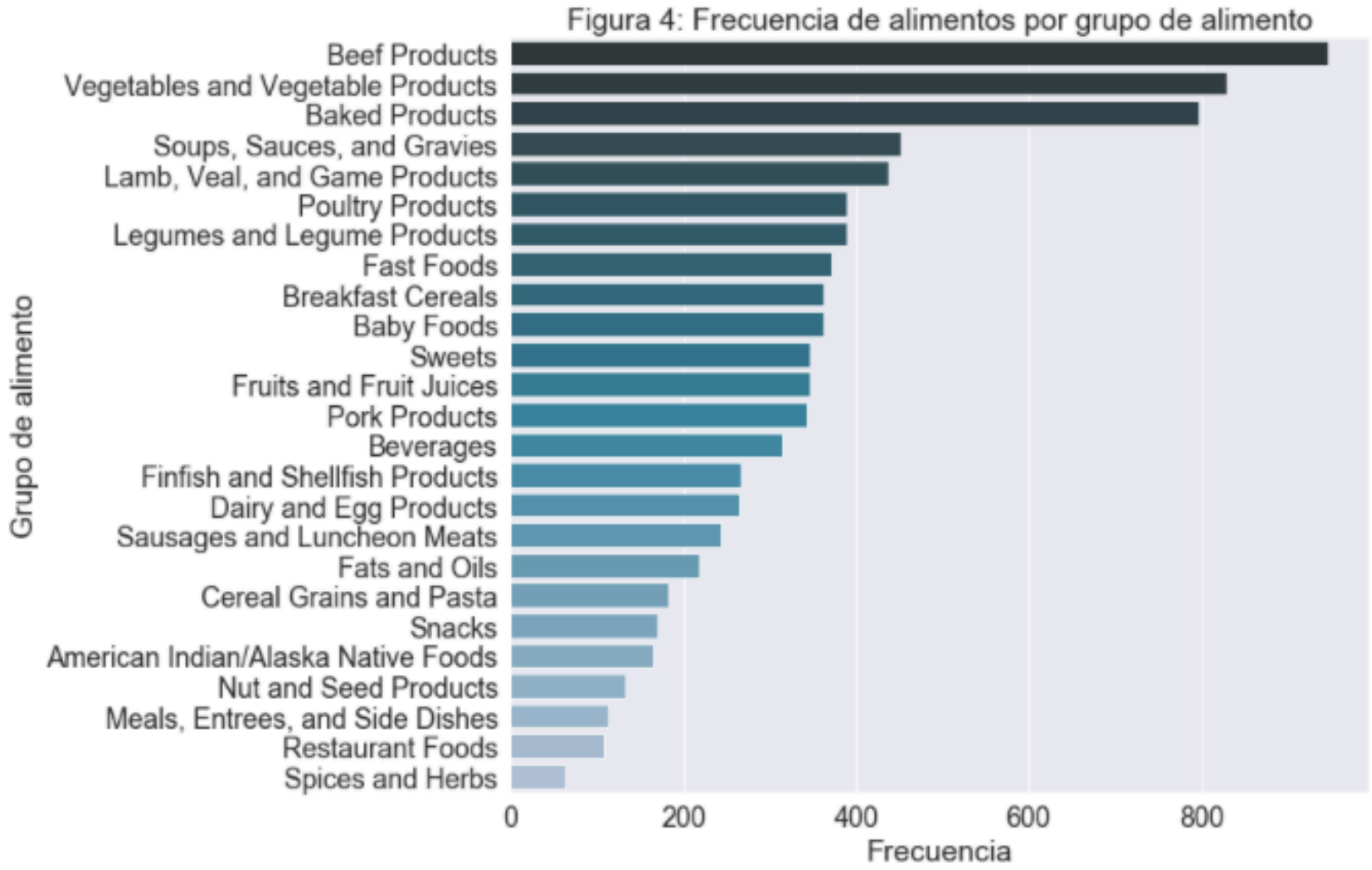


DISTRIBUCIÓN DE LOS MICRONUTRIENTES

Figura 3: Distribución de los micronutrientes

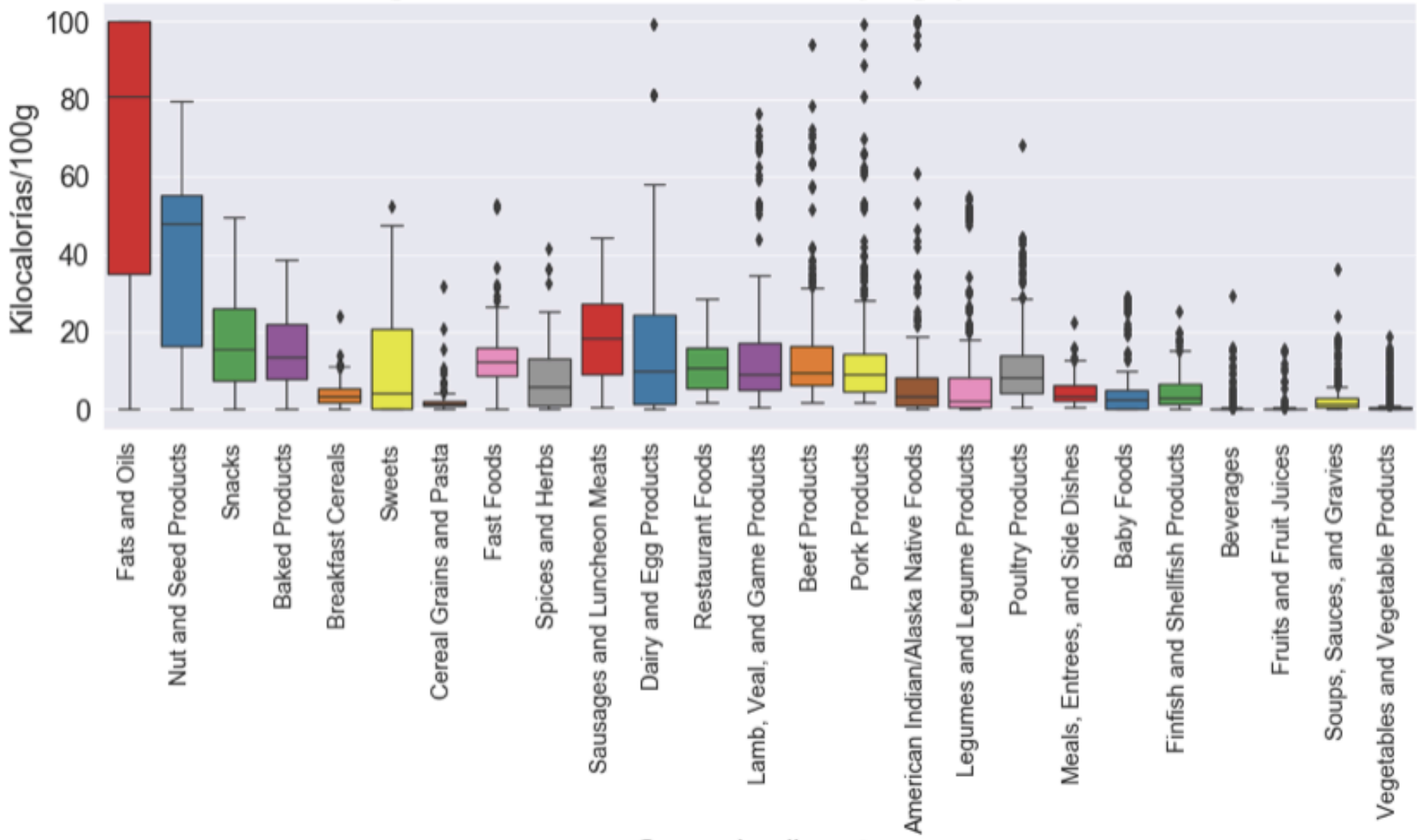


FRECUENCIA POR GRUPO DE ALIMENTO



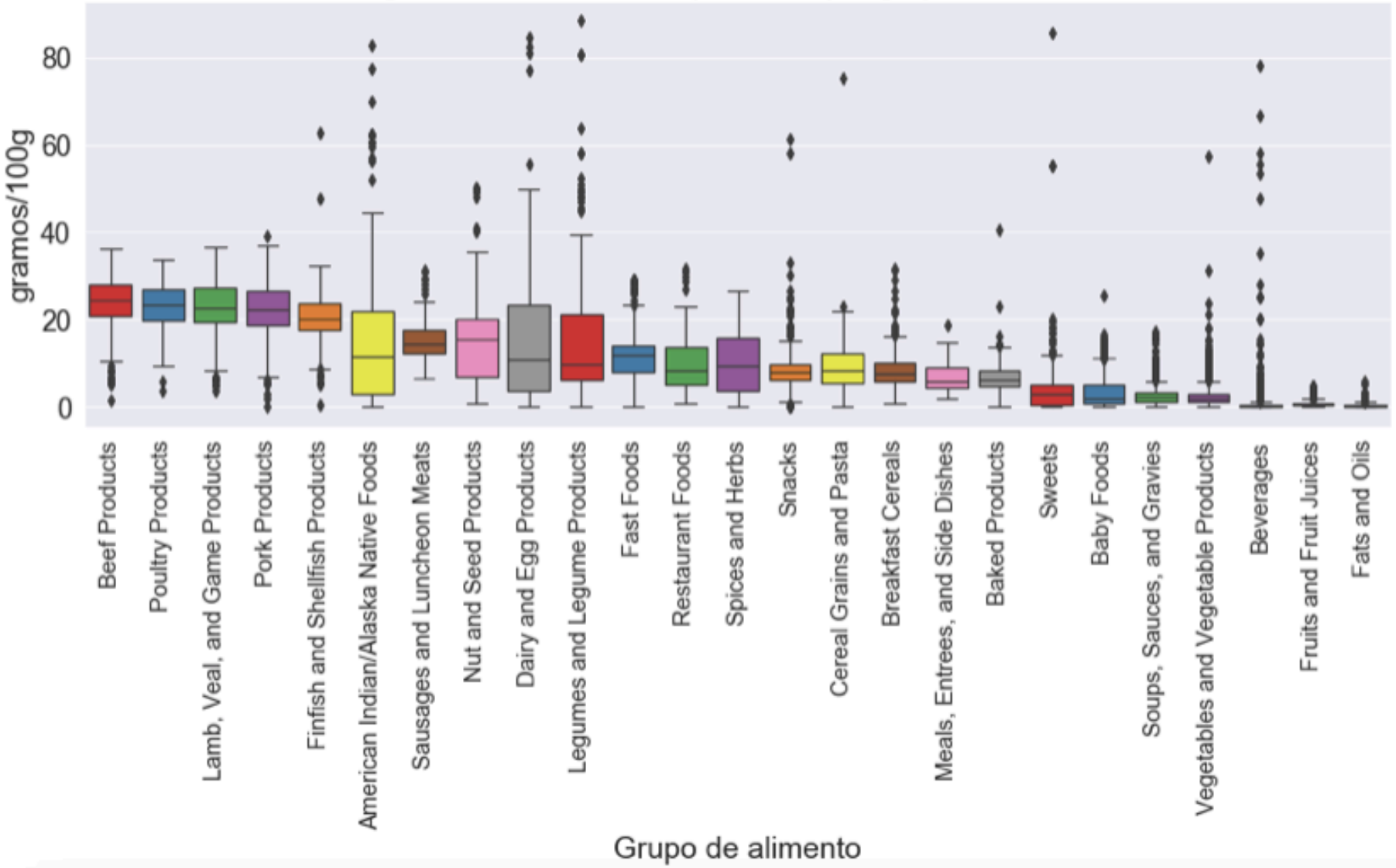
KILOCALORÍAS POR GRUPO DE ALIMENTO

Figura 5: Distribución de kilocalorías por grupo de alimento



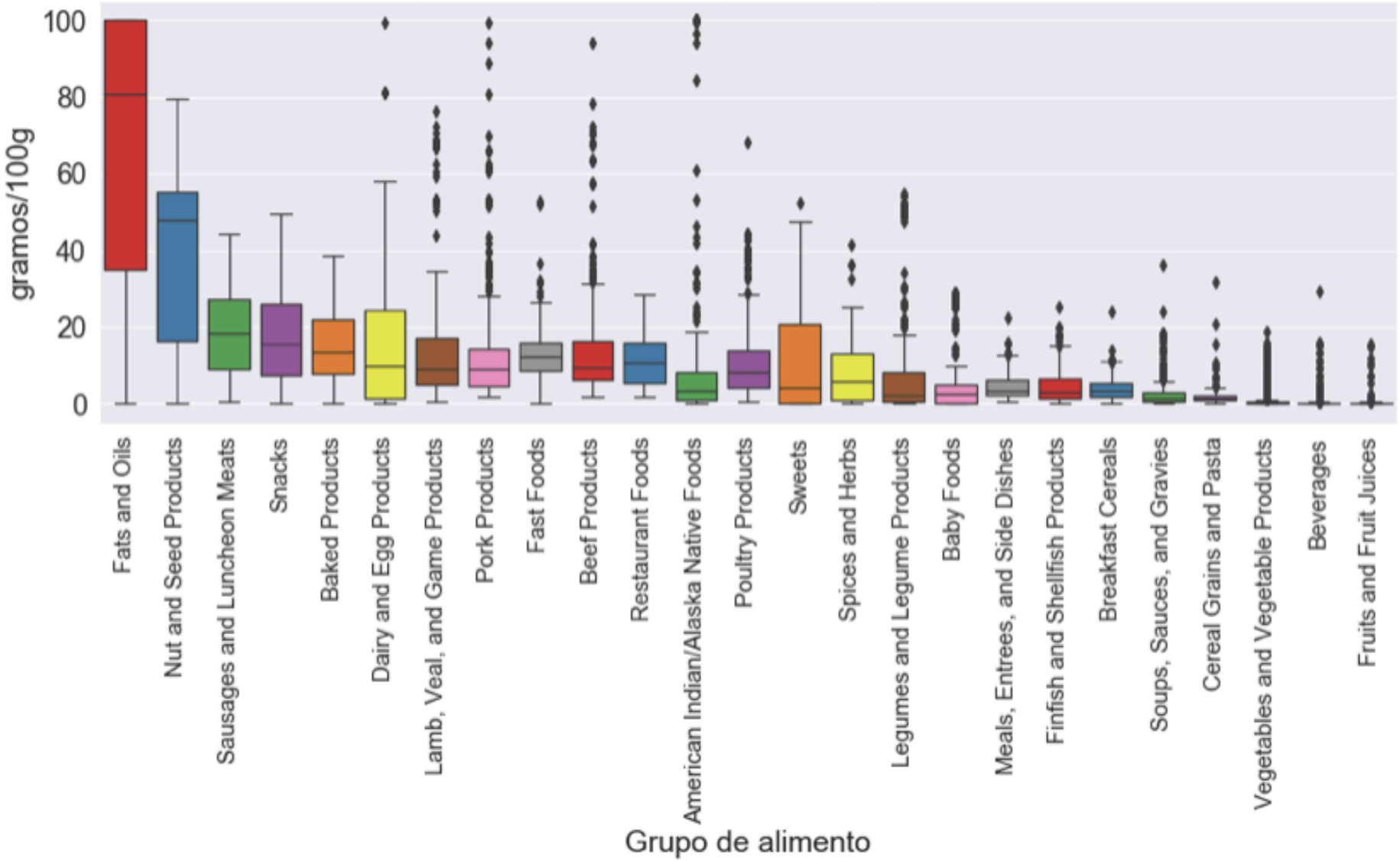
PROTEÍNAS POR GRUPO DE ALIMENTO

Figura 6: Distribución de proteínas por grupo de alimento



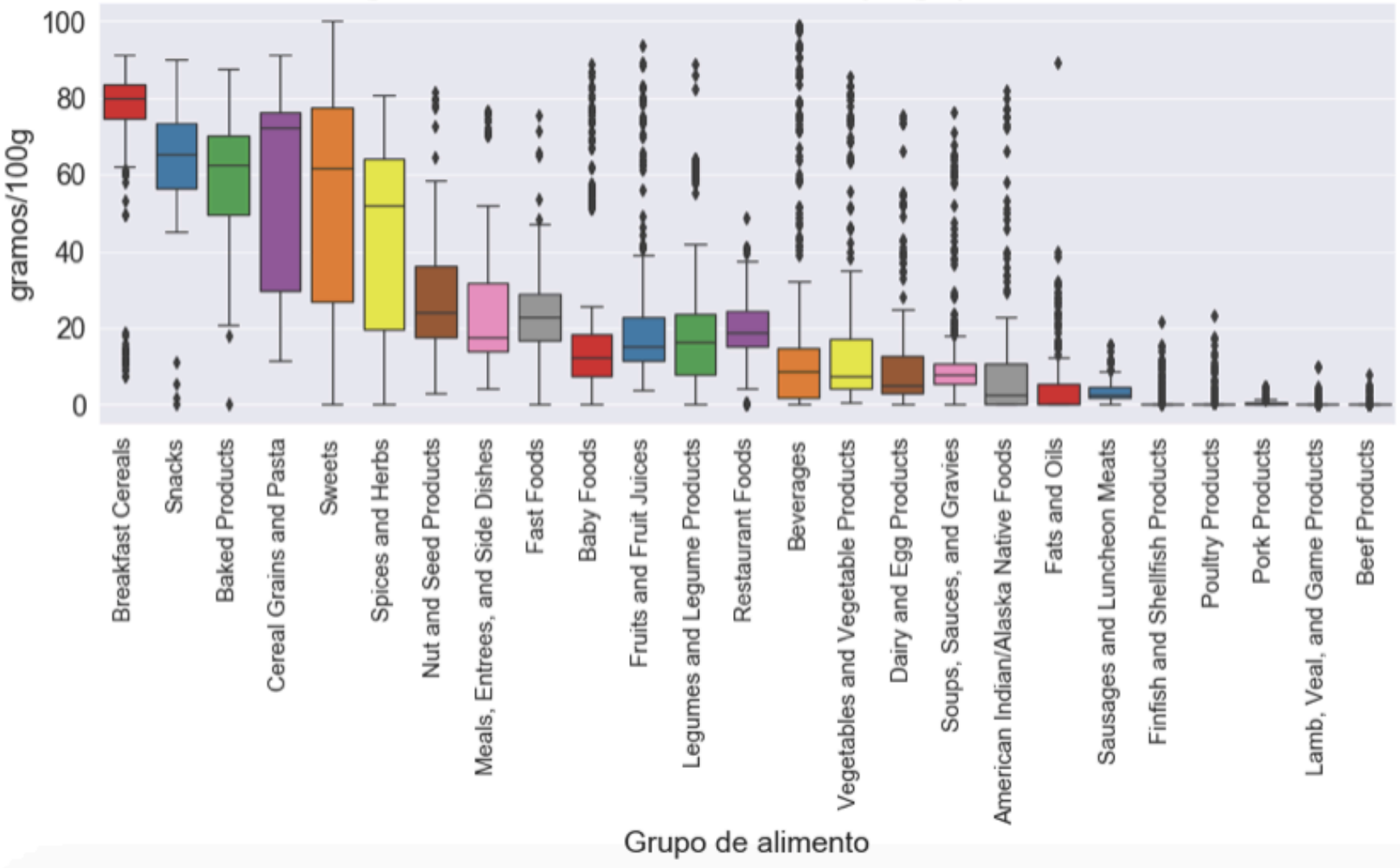
GRASAS POR GRUPO DE ALIMENTO

Figura 7: Distribución de grasas por grupo de alimento



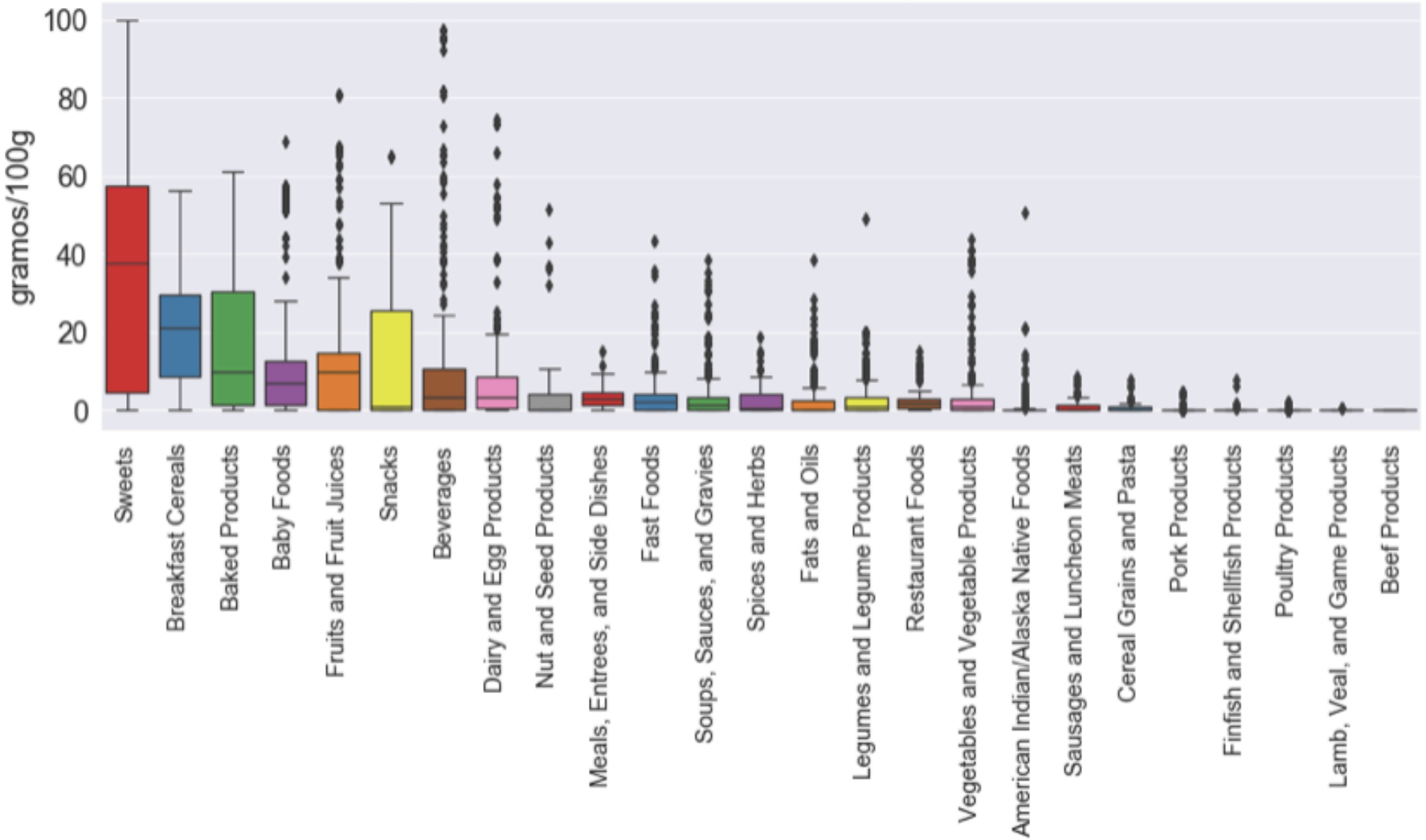
CARBOHIDRATOS POR GRUPO DE ALIMENTO

Figura 8: Distribución de carbohidratos por grupo de alimento



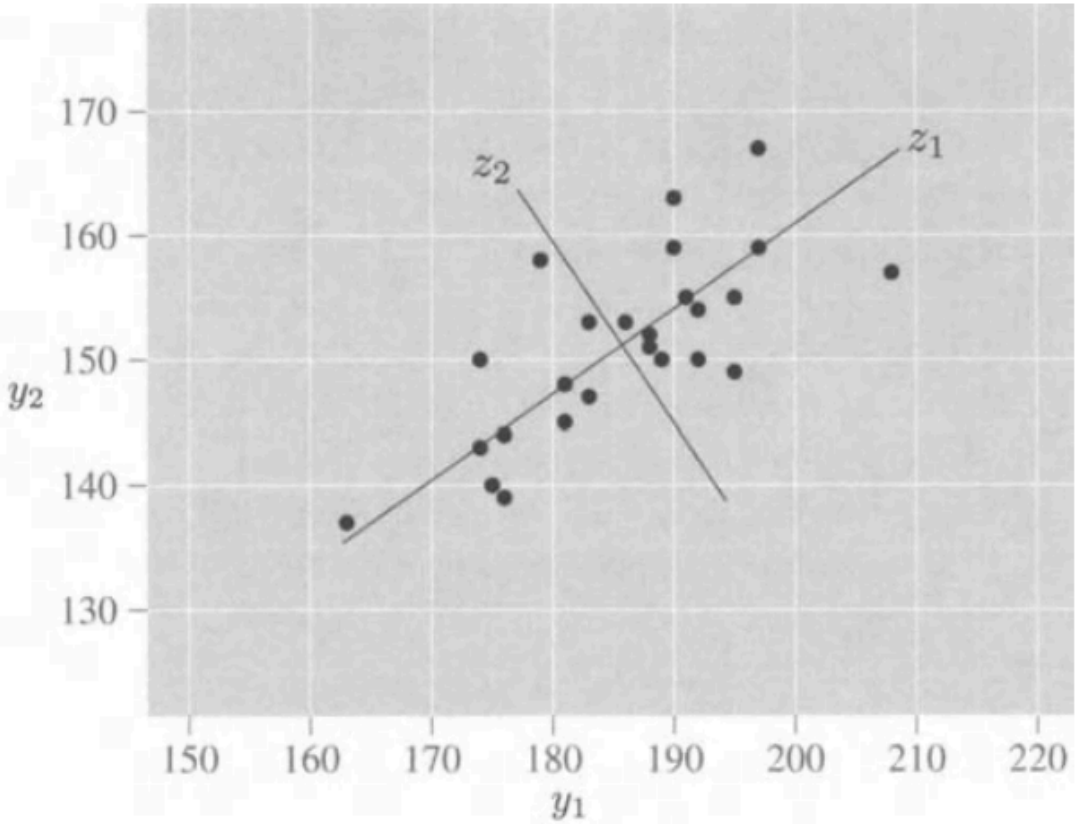
AZÚCARES POR GRUPO DE ALIMENTO

Figura 9: Distribución de azúcares por grupo de alimento



ANÁLISIS DE COMPONENTES PRINCIPALES

- Enfoque geométrico



- Enfoque algebraico

$$z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$
$$z_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$
$$\vdots$$
$$z_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p$$

$$s_{z_i}^2 = \lambda_i$$

1. ALGORITMO SVD

- Se aplica SVD a la matriz de datos (centrada) A:

$$A = USV^T$$

- Las columnas de V son las direcciones de los componentes.
- Las columnas de US nos darán los componentes principales.

2. ALGORITMO QR

- Procedimiento para calcular los eigenvalores y eigenvectores de una matriz A.

- Basado en la descomposición QR:
  $A = QR$ (con Q matriz ortogonal y R triangular superior)

Algoritmo

- Iniciar con $A_0 = A$
- Iterar hasta convergencia:
  - $A_k = Q_k R_k$
  - $A_{k+1} = R_k Q_k$

Para A simétrica:

- eigenvalores en la diagonal de A_k
- eigenvectores correspondientes en las columnas de la composición de las transformaciones Q_k

3. MÉTODO DE LA POTENCIA

- Para una matriz A diagonalizable, genera
  - el eigenvalor más grande (en valor absoluto) de A , λ
  - un vector no nulo v , que es el eigenvector correspondiente a λ $(Av = \lambda v)$
- Para asegurar la convergencia se debe cumplir:
  - A tiene un eigenvalor estrictamente mayor en magnitud respecto a sus otros eigenvalores
  - El vector inicial b_0 tiene una componente distinta de cero en la dirección de un eigenvector asociado con el eigenvalor dominante

- Iniciar con un vector b_0 aleatorio
- Iterar
  - $b_{k+1} = \frac{Ab_k}{||Ab_k||}$
  - $\mu_k = \frac{{b_k}^T Ab_k}{b_k^T b_k}$

➢ μ_k converge al eigenvalor dominante

➢ b_k converge al eigenvector correspondiente

Deflation: volver a aplicar el método a una matriz actualizada:

$$A_{k+1} = A_k - b_k b_k^T A b_k k_k^T$$

PCA A PARTIR DE EIGENVECTORES DE LA MATRIZ DE COVARIANZAS

- Los componentes principales se obtienen de la siguiente forma

- $Z_i = A_k * y_i$

- Donde Z_i es el componente principal i, A_k la matriz de datos original y y_i el i-ésimo eigenvector de la matriz de covarianzas

- La varianza explicada se obtiene a partir de los eigenvalores.

- Para saber que porcentaje de varianza explicada aporta cada componente principal se hace

$$E_i / \Sigma E_i$$

Donde Ei es el i-ésimo eigenvalor de la matriz de covarianzas

COMPARACIÓN ENTRE ALGORITMOS

- Aplicamos PCA utilizando, por una parte, la función PCA del paquete scikit learn, y por otro, la implementación de varios algoritmos programados por el equipo:
  - Utilizando la SVD con el paquete numpy
  - Programando el algoritmo QR
  - Implementando el método de la potencia

1. ALGORITMO SVD

- El mayor error relativo, que se presenta en las componentes principales fue del orden de 10^{-10}.

	elemento	Igualdad (en valor absoluto)	Max error relativo (con valor absoluto)	Error relativo (con valor absoluto)
0	varianza explicada	True	1.935215e-16	[0.0, 1.2189963581231018e-16, 1.57089176848327...
1	valores singulares	True	0.000000e+00	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2	coeficientes	True	0.000000e+00	[[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,...
3	componentes principales	True	1.128716e-10	[[7.521729907317927e-15, 4.695376192208106e-15...

2. ALGORITMO QR

- El mayor error relativo, que se presenta en las componentes principales fue del orden de 10^{-5}.

	elemento	Igualdad (en valor absoluto)	Max error relativo (con valor absoluto)	Error relativo (con valor absoluto)
0	varianza explicada	True	1.912542e-14	[7.848981292888267e-15, 1.682214974209909e-14,...
1	eigenvalores	True	2.441949e-14	[3.259419297349506e-16, 2.4419493161683765e-14...
2	coeficientes	True	5.478970e-08	[[2.866773180195502e-14, 4.697647283337805e-14...
3	componentes principales	True	3.882674e-05	[[1.9991966332607923e-14, 2.3476880961041193e-...

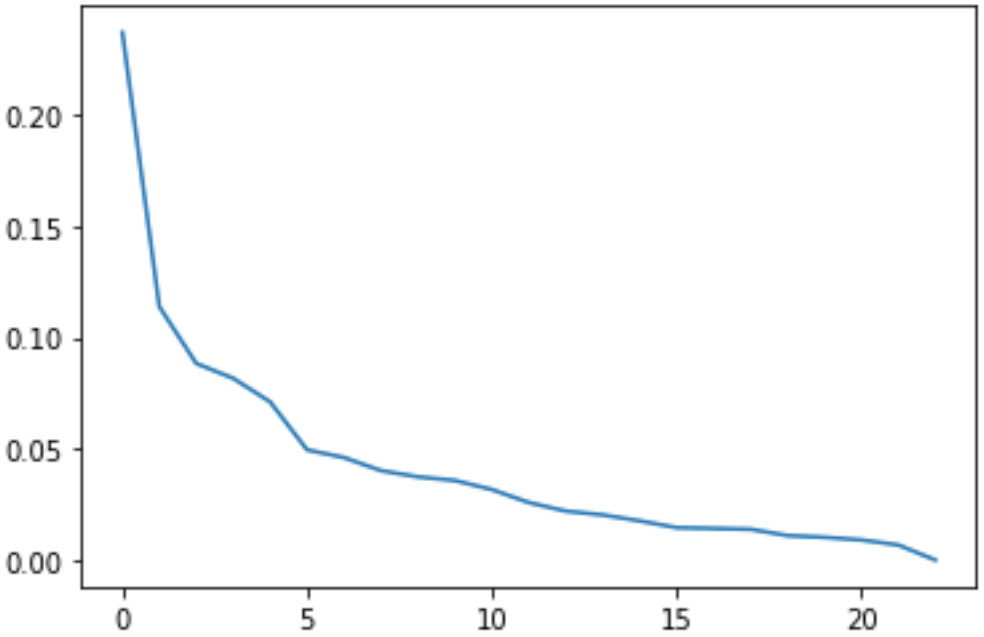
3. MÉTODO DE LA POTENCIA

- El error relativo fue del orden de 10^{-15}.

	elemento	error relativo
0	primer eigenvalor	1.955879e-15
1	primer eigenvector	1.275342e-15

VARIANZA EXPLICADA Y EIGENVALORES

- Criterio de varianza explicada (80%):
  - 10 componentes



- Eigenvalores (>1):
  - 7 componentes

- C1: 5.44991815
- C2: 2.61876222
- C3: 2.03213339
- C4: 1.87934951
- C5: 1.63586122
- C6: 1.140503
- C7: 1.06099038
- C8: 0.92636129
- C9: 0.8621182
- C10: 0.82478215

INTERPRETACIÓN DE COMPONENTES

- Componente 1

Riboflavin_mg	0.341325
Niacin_mg	0.337779
VitB6_mg	0.315663
Iron_mg	0.299857
Folate_mcg	0.284102
Thiamin_mg	0.272453
Zinc_mg	0.243551
Magnesium_mg	0.241348
Phosphorus_mg	0.199403
Fiber_g	0.181570
Copper_mcg	0.180806
VitB12_mcg	0.177985
Carb_g	0.169685
Calcium_mg	0.168112
Energy_kcal	0.157814
Protein_g	0.140620
VitE_mg	0.137122
VitA_mcg	0.133519
Manganese_mg	0.093567
Selenium_mcg	0.092319
VitC_mg	0.087639
Sugar_g	0.076323
Fat_g	0.033008

- Componente 2

Carb_g	0.443416
Sugar_g	0.358769
VitB12_mcg	0.355045
Protein_g	0.343397
Energy_kcal	0.273449
Fiber_g	0.257733
Selenium_mcg	0.239322
VitA_mcg	0.236470
Copper_mcg	0.212669
Zinc_mg	0.177798
Fat_g	0.111670
VitE_mg	0.106372
Calcium_mg	0.105173
Magnesium_mg	0.103361
Folate_mcg	0.097093
Iron_mg	0.093812
Manganese_mg	0.088783
Phosphorus_mg	0.087448
Niacin_mg	0.084801
Thiamin_mg	0.075150
Riboflavin_mg	0.073471
VitC_mg	0.038525
VitB6_mg	0.021129

- Componente 3

Fat_g	0.534051
Energy_kcal	0.462006
Phosphorus_mg	0.274814
Folate_mcg	0.230985
Protein_g	0.213567
VitE_mg	0.207331
Magnesium_mg	0.201225
Riboflavin_mg	0.192098
Thiamin_mg	0.184351
VitB6_mg	0.174648
Niacin_mg	0.164885
Selenium_mcg	0.163361
VitC_mg	0.162303
Copper_mcg	0.152263
Calcium_mg	0.128139
Iron_mg	0.087109
Manganese_mg	0.072630
Sugar_g	0.055247
Carb_g	0.049822
Fiber_g	0.040397
Zinc_mg	0.038639
VitA_mcg	0.021929
VitB12_mcg	0.012760

- Componente 4

VitA_mcg	0.530395
Copper_mcg	0.389929
VitB12_mcg	0.346550
Manganese_mg	0.311369
Protein_g	0.311112
Sugar_g	0.217373
Phosphorus_mg	0.207871
Carb_g	0.174108
Zinc_mg	0.166322
Selenium_mcg	0.161623
Niacin_mg	0.156394
VitB6_mg	0.114372
Thiamin_mg	0.103518
Calcium_mg	0.099340
Magnesium_mg	0.071805
Iron_mg	0.059443
Energy_kcal	0.052280
VitC_mg	0.047584
Riboflavin_mg	0.046996
Fiber_g	0.042130
Folate_mcg	0.032481
Fat_g	0.026520
VitE_mg	0.026174

INTERPRETACIÓN DE COMPONENTES

- Componente 5

Fat_g	0.394450
Calcium_mg	0.388692
Magnesium_mg	0.352742
Phosphorus_mg	0.344929
Fiber_g	0.332161
Energy_kcal	0.293596
VitE_mg	0.238019
Niacin_mg	0.203433
Thiamin_mg	0.161520
Copper_mcg	0.161241
Riboflavin_mg	0.153460
Folate_mcg	0.137697
VitB6_mg	0.133747
Manganese_mg	0.125600
Iron_mg	0.097110
Carb_g	0.082723
Zinc_mg	0.064319
VitB12_mcg	0.058857
Sugar_g	0.048421
VitC_mg	0.024713
Protein_g	0.013176
VitA_mcg	0.008137
Selenium_mcg	0.005049

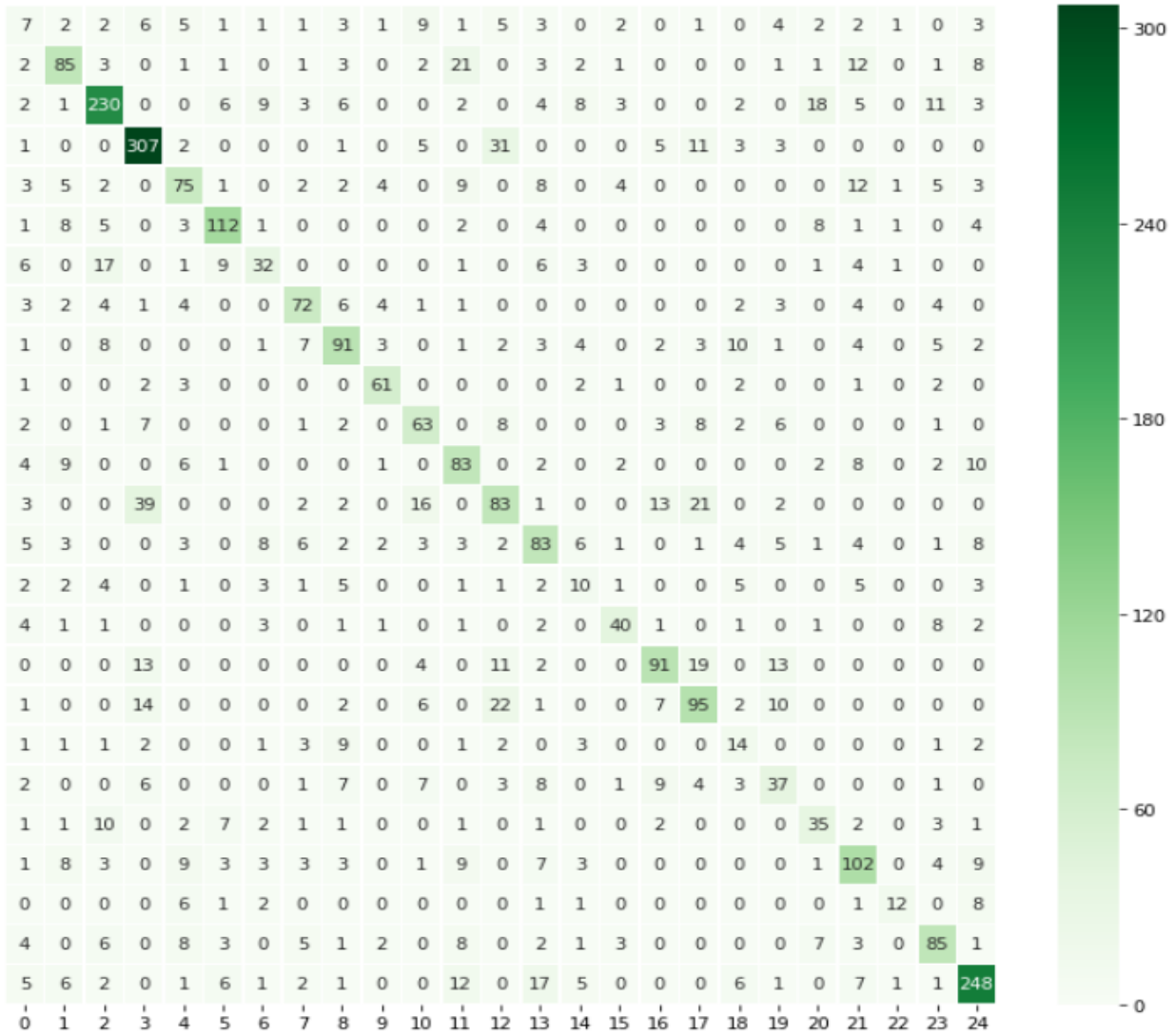
- Componente 6

VitC_mg	0.545219
VitE_mg	0.475300
Sugar_g	0.337022
Carb_g	0.279369
Selenium_mcg	0.243773
VitB6_mg	0.200525
Manganese_mg	0.182527
Protein_g	0.141469
Magnesium_mg	0.140120
Fiber_g	0.134870
Energy_kcal	0.132865
Thiamin_mg	0.130000
VitB12_mcg	0.123150
Fat_g	0.116492
Copper_mcg	0.086214
Phosphorus_mg	0.083792
Calcium_mg	0.070532
Folate_mcg	0.054688
Zinc_mg	0.054538
Riboflavin_mg	0.050758
VitA_mcg	0.039028
Iron_mg	0.030539
Niacin_mg	0.026277

- Componente 7

Calcium_mg	0.462817
VitC_mg	0.451276
Fiber_g	0.414733
Phosphorus_mg	0.323521
Sugar_g	0.298613
Magnesium_mg	0.269604
Riboflavin_mg	0.141794
Manganese_mg	0.136507
Iron_mg	0.133322
Folate_mcg	0.130939
VitE_mg	0.125438
Copper_mcg	0.119212
Zinc_mg	0.105925
VitA_mcg	0.101341
Thiamin_mg	0.071844
Energy_kcal	0.066823
VitB12_mcg	0.056077
Niacin_mg	0.045698
Fat_g	0.045318
VitB6_mg	0.024691
Carb_g	0.022079
Selenium_mcg	0.021893
Protein_g	0.006299

ÁRBOLES DE DECISIÓN PARA CLASIFICACIÓN



- Matriz de confusión
- Precisión score:
  - 0.5673
- Recall score:
  - 0.5558

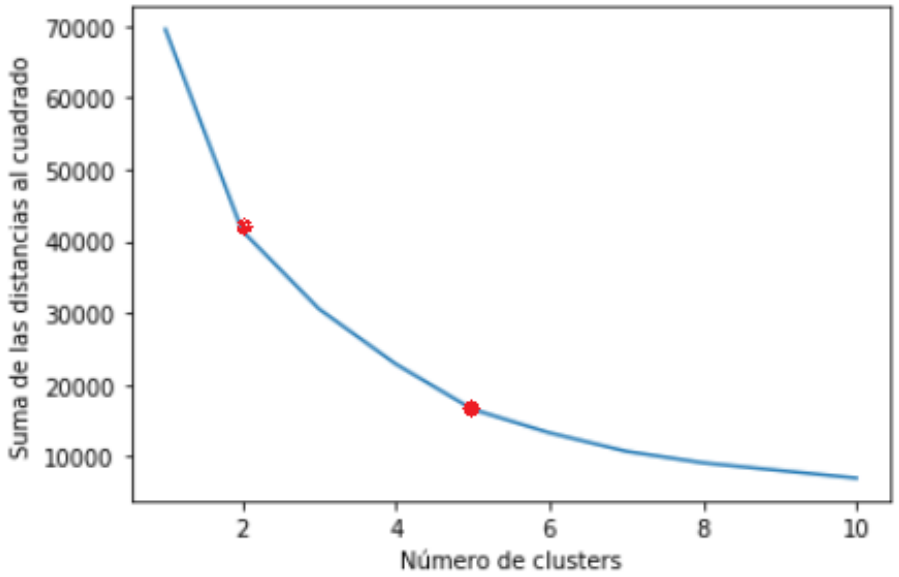
PARALELIZACIÓN (DASK)

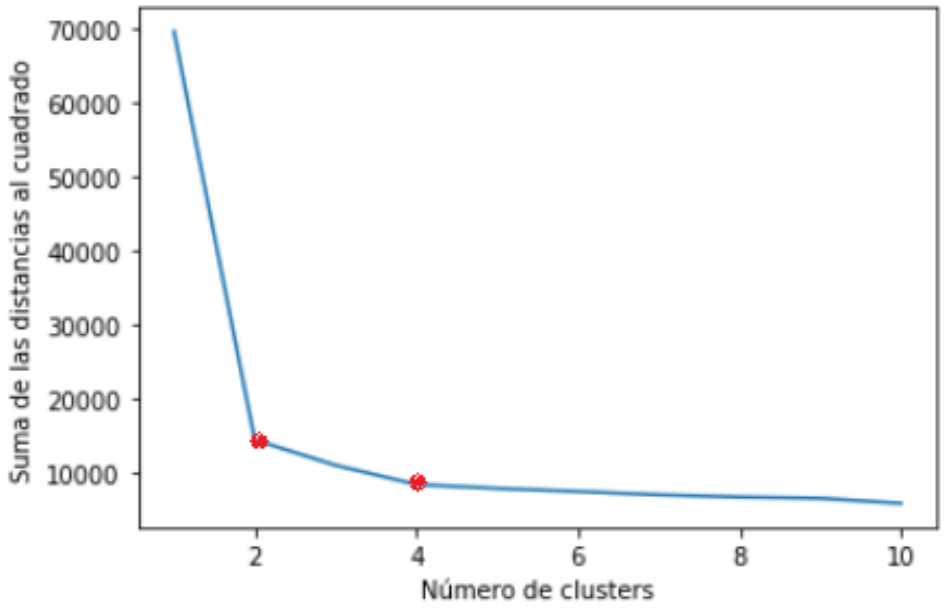
Client

Scheduler: tcp://127.0.0.1:61276
Dashboard: http://127.0.0.1:61279/status

Cluster

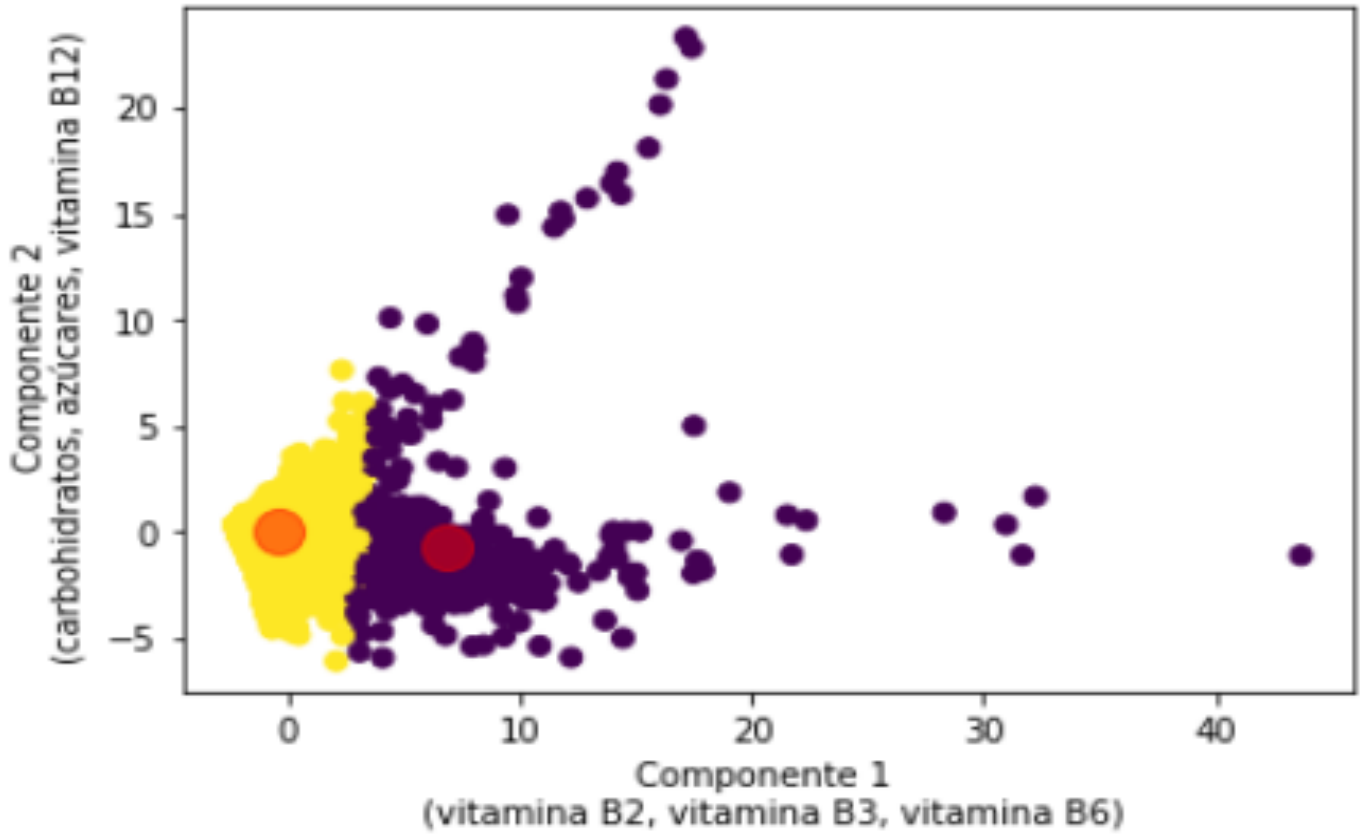
Workers: 4
Cores: 12
Memory: 17.02 GB





ANÁLISIS DE CONGLOMERADOS

- Se eligió un parámetro de k=2. Este es el número de grupos en los que se agrupan los datos.



CONCLUSIÓN

- Se logró reducir la dimensionalidad y eliminar la multicolinealidad de los datos al hacer PCA.
  - Se redujo a 10 componentes principales con el criterio de varianza explicada al 80%.
  - Se redujo a 7 componentes principales con el criterio de eigenvalor >1.

- La programación de los 4 métodos comparados dieron resultados muy similares.

- El análisis de clasificación de los tipos de comida fue bueno:
  - Precisión: 0.5673
  - Recall: 0.5558

REFERENCIAS

- Codesansar. (s.f). Power Method Algorithm for Finding Dominant Eigen Value and Eigen Vector
- Dan, D. J. (2014). Power-Method-PCA
- Data.world. (2017). USDA National Nutrient DB
- Equipo SVD. (2020). Examen de cómputo matricial equipo SVD
- Equipo QR. (2020). Examen de cómputo matricial equipo QR
- Fox, J., Chalmers, P., Monette, G., & Sanchez, G. (2020). PowerMethod: Power Method for Eigenvectors in matlib: Matrix Functions for Teaching and Learning Linear Algebra and Multivariate Statistics
- Palacios M. Erick. (2020). Notas de MNO 2020. SVD
- Palacios M. Erick. (2020). Notas de MNO 2020. Componentes principales
- Palacios M. Erick. (2020). Notas de MNO 2020. Cómputo en paralelo - Dask
- Ramos, Irene. (2020). "Tipología_manejo_agrícola"
- Rencher, Alvin C & William F. Christensen. (2012). Methods of Multivariate Analysis. Department of Statistics, Brigham Young University, Provo, UT.- Third Edition. Ch 12.
- Sharma Subhash. (1996). Applied Multivariate Techniques. University of South Carolina. Ch4.
- U.S. Department of Agriculture, Agricultural Research Service. 2014. USDA National Nutrient Database for Standard Reference, Release 27. Methods and Application of Food Composition Laboratory Home Page, http://www.ars.usda.gov/nea/bhnrc/mafcl
- Wikipedia. QR algorithm