Macrolocalizacion de regiones con K-means

Índice

Carga de datos 1
Normalización de la base de datos 1
Número de clusters óptimo 1
Cómputo de las K-means 2
Flujos migratorios entre clusters 3

Carga de datos

Para comenzar, cargamos los datos de las variables sociales y economicas que definen a las distintas provincias de la Argentina.

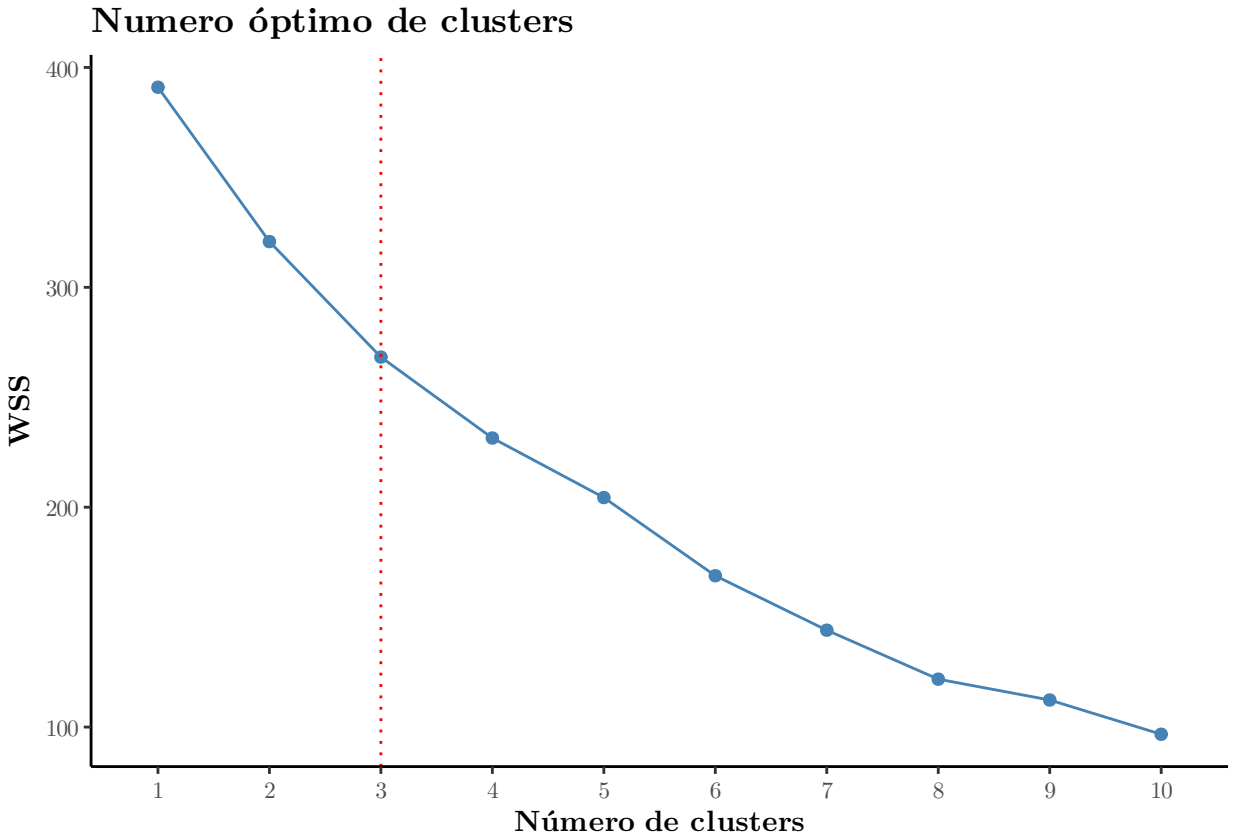
Normalización de la base de datos

Ahora normalizamos la base de datos para lograr coherencia en la aplicación de K-means, debido a quen trabajamos con distintas escalas de datos.

Número de clusters óptimo

Verificamos el numero de clusters optimos a través de el Within-cluster sum of square, el cual es un indicador que mide la suma de las distancias entre las variables dentro de los cluster y sus centroides, la idea es minimizar esta discrepancia dentro de cada grupo, teniendo en cuenta el trade-off que implica cuando este es mínimo, el cual es el caso en que el numero de clusters es igual al número de variables a clusterizar, en donde si tenemos j observaciones, el numero de clusters sería tal que j=k, y no se estaría dando ninguna información relevante a los efectos de poder resumir caracteristicas comunes entre los grupos.

$$WCSS = \sum_{i=1}^{N_C} \sum_{\mathbf{x} \in Ci} distancia(\mathbf{x}, \bar{\mathbf{x}}_{\mathbf{Ci}})$$



Obtenemos que el numero óptimo de clusters se encuentra en torno al **número 3**, debido a que la contribución marginal de aumentar el numero de clusters a 4, no aportaría una reducción muy elevada al **WSS**, y seguiriamos perdiendo generalidad en la clusterización de los grupos sin una ganancia significativa de similitud dentro de los grupos.

Cómputo de las K-means

Setearemos una semilla para el cálculo de los centroides iniciales del algoritmo de K-means y procederemos a realizar el cálculo de los clusters a través de 1000 iteraciones, eligiendo la que de menor numero de WSS.

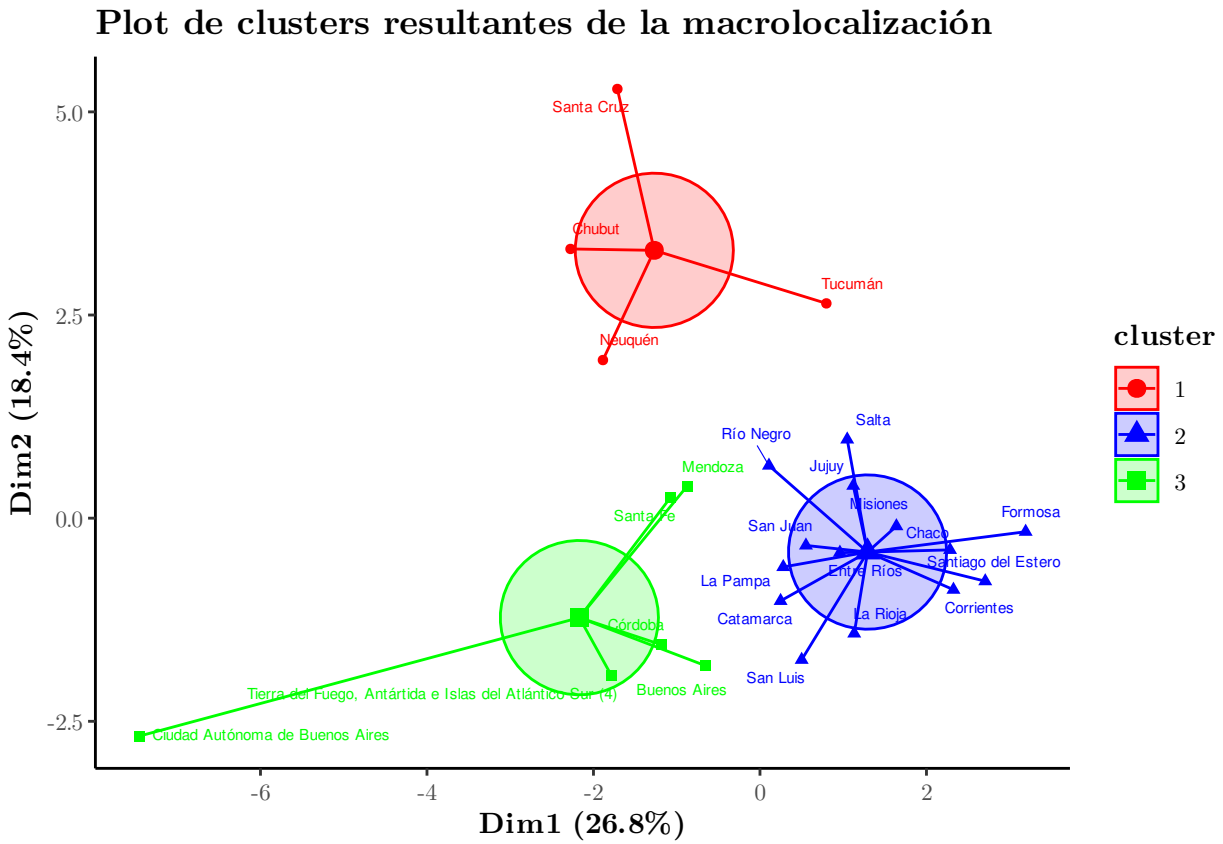
Podemos ver la media de los parámetros para cada uno de los clusters

Cuadro 1: Resumen de indicadores por cluster

Indicadores	Cluster 1	Cluster 2	Cluster 3
Habitantes (2010)	530,187.00	677,181.00	3,042,344.00
tasa promocion efectiva secundaria (2017)	76.25	78.60	82.05
promedio mortalidad infantil 2016-2019	8.20	9.50	8.25
Homicidios dolosos promedio (2016-2019	7.25	3.75	5.30
Muertes en Accidentes Viales promedio (2016-2019	12.40	14.05	7.50
Robos (excluye los agravados por el resultado de lesiones y/o muertes) promedio (2016-2019	973.40	654.85	1,566.95
Robos agravados por el resultado de lesiones y/o muertes promedio (2016-2019	27.80	1.70	9.50
Violaciones promedio (2016-2019	15.45	11.15	8.75
exportaciones promedio (2016-2019) en USD	2,857.20	794.20	1,420.55
Demanda de MWH energía electrica (2016)	3.80	2.20	3.35

Indicadores	Cluster 1	Cluster 2	Cluster 3
Pobreza Promedio2017-2019	26.05	32.55	29.35
Tasa actividad Promedio 2017-2019	44.55	42.50	46.00
cantidad empresas 2016-2017	1,569.62	846.96	1,771.95
remuneracion real trab registrados priv (2016-2019)	29,995.27	14,997.52	17,149.25
Porcentaje empleados en Agricultura, ganadería, pesca y actividades extractivas (2016-2019)	0.21	0.13	0.05
Porcentaje empleados en Comercio, servicios, electricidad, gas, agua y construccion (2016-2019)	0.67	0.69	0.71
Porcentaje empleados en Industria (2016-2019)	0.10	0.15	0.22

agregamos a la base de datos original la denominación de cada cluster



Se puede ver claramente la diferencia en la similitud de los tres grupos, siendo solamente la Capital Federal la única que presenta mayor disimilitud con respecto a su cluster, lo cual indica que esta capital lleva una dinámica socio-económica muy peculiar con respecto al promedio de las provincias argentinas, inclusive considerablemente distinta que las provincias con la cual ostenta mayor similitud.

Flujos migratorios entre clusters

Dentro de los cluster, existen provincias de las cuales salen gran cantidad de inmigrantes, al igual que existen provincias que son receptoras de estos mismos, en primer lugar caracterizaremos cual es el cluster con mayor nivel de *expulsion* de inmigrantes interprovinciales.

Cuadro 2: Porcentaje de migrantes por cluster

Cluster	Porcentaje
1	10.41 %
2	61.67 %
3	27.92 %

Se puede notar que las provincias del *cluster Nº3* son las que mayor nivel de expulsión poseen, seguidas por las provincias pertenecientes al cluster Nº2, y por úlitmo se encuentran las pertenecientes al cluster Nº3.

Tables	Are	Cool
col 1 is	left-aligned	$1600
col 2 is	centered	$12
col 3 is	right-aligned	$1