Análisis de census poblacional

Claudia Quintana Wong

Introducción

El clustering consiste en la agrupación automática de datos y constituye un método de aprendizaje no supervisado. El objetivo de la agrupación es encontrar diferentes grupos dentro de los elementos de los datos. Para ello, los algoritmos de agrupamiento encuentran la estructura en los datos de manera que los elementos del mismo clúster (o grupo) sean más similares entre sí que con los de clústeres diferentesOermite descubrir la estructura intrínseca (y comúnmente oculta) de los datos.

Una cadena de supermercados desea identificar áreas en las que abrir nuevos centros. Este trabajo se centra en la aplicación de un conjunto de técnicas de *clustering* sobre un conjunto de datos que contiene información información censal. El objetivo es agrupar zonas censales cuyos perfiles de clientes se adapten a diferentes tipologías de centros como boutiques, supermercados de presupuesto medio, grandes superficies, entre otros.

Desarrollo

En esta sección, se aplica un modelo de *clustering*. Inicialmente, se realiza un análisis descriptivo sobre el conjunto de datos que permite conocer el formato y naturaleza de los datos. Posteriormente, se desarrolla un algoritmo de agrupamiento que permite organizar los datos en grupos de acuerdo a la similitud entre ellos.

Descripción del conjunto de datos

El conjunto de datos consta 4 variables que se describen a continuación.

- **MeanHHSz**: Tamaño medio de la unidad familiar (HH = Household).
- **MedHHInc**: Nivel de ingresos mediano de la unidad familiar.
- **RegDens**: Densidad de población de la región.
- **RegPop**: Número de habitantes de la región.

```
##     RegDens          RegPop            MedHHInc            MeanHHSz
##  Min.   :  1.0   Length:33178       Length:33178       Min.   :0.000
##  1st Qu.: 26.0   Class :character   Class :character   1st Qu.:2.360
##  Median : 51.0   Mode  :character   Mode  :character   Median :2.550
##  Mean   : 50.5                                         Mean   :2.501
##  3rd Qu.: 75.0                                         3rd Qu.:2.740
##  Max.   :100.0                                         Max.   :8.490
##  NA's   :1013
```

Como se observa en la imagen el conjunto contiene 33.178 instanciass, sin embargo, existen muchos *missing values*. Teniendo en cuenta que la cantidad de filas que contienen NAs es pequeña respecto a la cantidad total de observaciones del *dataset*, con el objetivo de facilitar y garantizar un mejor análisis, se eliminan todas aquellas filas que contienen NA en al menos una de las valriables. A continuación se muestra el resumen del conjunto de datos transformado.

```
##     RegDens          RegPop             MedHHInc            MeanHHSz
##  Min.   :  1.0   Length:32165       Length:32165       Min.   :0.000
##  1st Qu.: 26.0   Class :character   Class :character   1st Qu.:2.380
##  Median : 51.0   Mode  :character   Mode  :character   Median :2.560
##  Mean   : 50.5                                         Mean   :2.579
##  3rd Qu.: 75.0                                         3rd Qu.:2.750
##  Max.   :100.0                                         Max.   :8.490
```

Preprocesado de la información

Las columnas **MedHHInc MedHHInc** y **MeanHHSz** fueron transformadas eliminando el separador decimal y el símbolo $. Asimismo fueron convertidadas a numéricas para su posterior tratamiento.

```
census.scale$MedHHInc <- as.numeric(gsub('[$,]', '', census.scale$MedHHInc))
census.scale$MeanHHSz <- as.numeric(gsub('[$,]', '', census.scale$MeanHHSz))
census.scale$RegPop <- as.numeric(gsub('[,]', '', census.scale$RegPop))
```

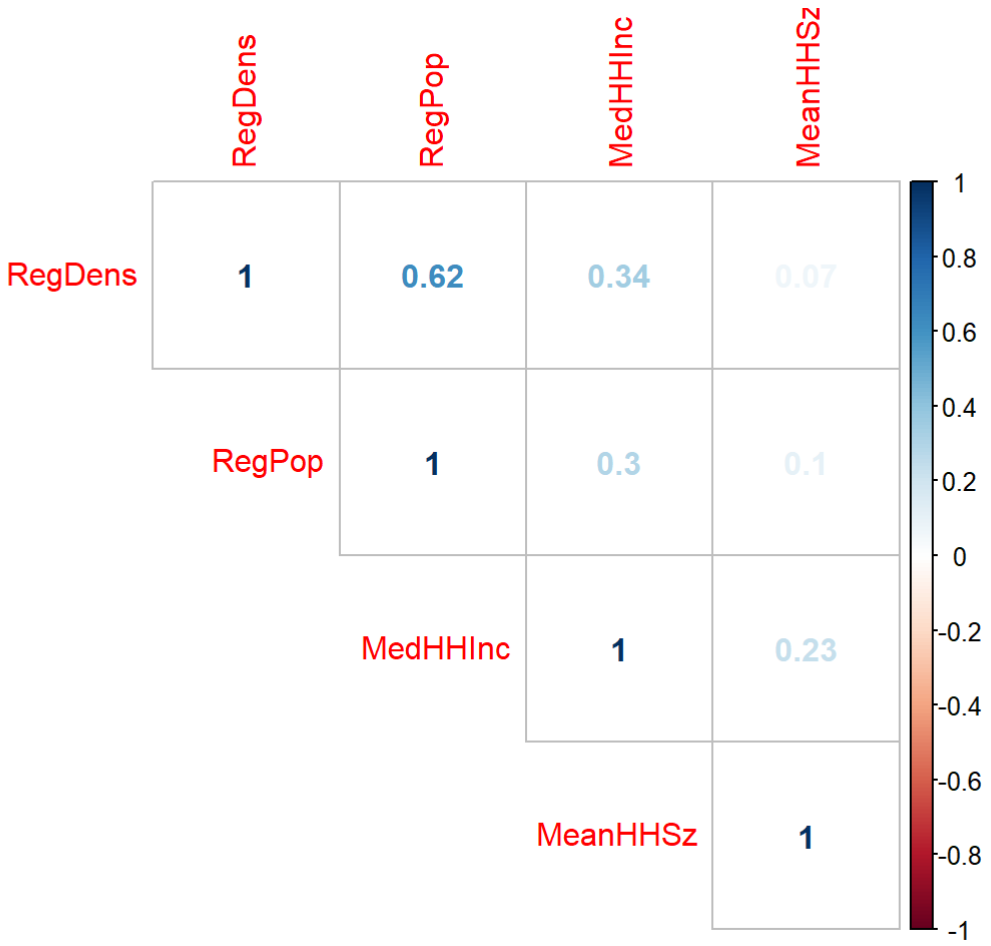
Tratamiento de outliers y estandarización

La estandarización de datos es necesaria en el cálculo de distancias no tengan más peso aquellas variables con mayor variabilidad/rango.

```
census.scale$RegDens <- discrete_by_quantile(census.scale$RegDens)/4
census.scale$RegPop <- discrete_by_quantile(census.scale$RegPop)/4
census.scale$MedHHInc <- discrete_by_quantile(census.scale$MedHHInc)/4
census.scale$MeanHHSz <- discrete_by_quantile(census.scale$MeanHHSz)/4
```

Análisis descriptivo

Para conocer mejor la distribución de los datos, se realiza un breve análisis descriptivo mediante el cálculo de la matriz de correlación para descubrir una posible relación entre las variables.



Se observa que existe una correlación lineal positiva entre la densidad de población y la cantidad de habitantes de la región, expresadas en las variables **RegDens** y **RegPop** respectivamente.

Modelo de *custering*

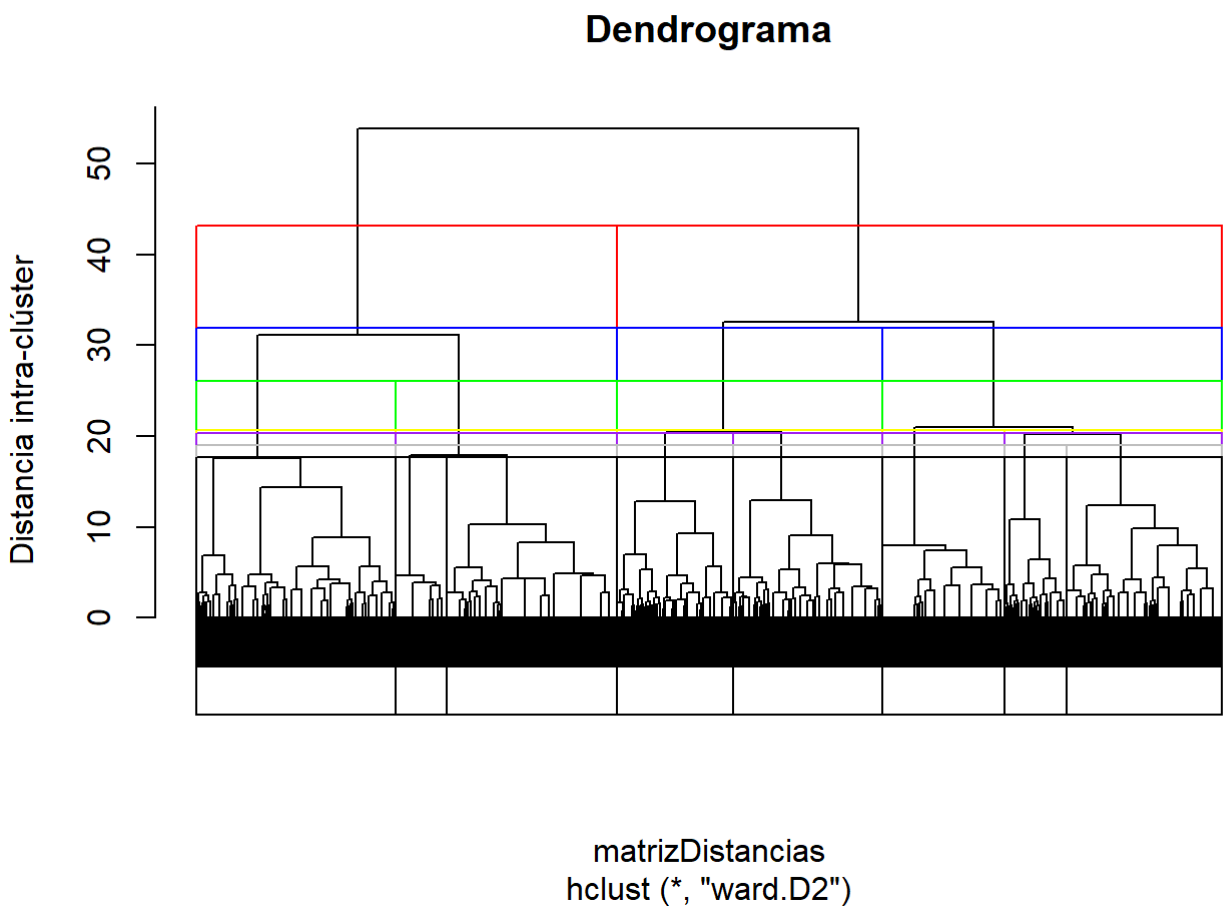
El objetivo de los algoritmos de agrupamiento es minimizar la varianza intergrupo y maximizar la distancia intra-cluster. Uno de los algoritmos de agrupamiento más sencillo y extendido, tanto en la literatura como en la práctica, es el algoritmo de *K-means*, que es aplicable en los casos en que se tenga una representación de los datos como elementos en un espacio métrico, como en este caso. Sin embargo, este algoritmo tiene una desventaja y es que la cantidad de clústers debe conocerse de antemano y pasarse como parámetro al método.

Clustering jerárquico

El análisis hecho hasta el momento no ofrece conocimiento suficiente sobre los datos por lo que no se puede establecer un número de grupos a priori. Por esta razón, es necesario aplicar un algoritmo de agrupamiento jerárquico que nos dará como respuesta una cantidad aproximada de grupos.

La medida que se utiliza para computar la distancia entre dos puntos es la distancia euclidiana. El método jerárquico es muy costoso computacionalmente porque se basa en el cálculo de la matriz de distancias, por lo que es necesario trabajar sobre una muestra aleatoria del conjunto de datos original y luego con la cantidad de grupos escogida, aplicar un algoritmo menos costoso sobre todo el conjunto de datos.

A partir de la salida del método jerárquico es posible representar el dendograma. Un dendograma permite visualizar los clústers escogidos por el método en cada una de las iteracione por lo que puede ser útil para escoger el número de clústers de manera visual.



Teniendo en cuenta que en el eje Y se representa la métrica varianza intergrupo y en el eje X se puede visualizar los clústers seleccionados en cada iteración, se puede notar que la cantidad óptima puede estar entre 4(representado por la línea verde) y 6 (línea morada) clústers porque la ganacia comienza a disminuir en menor medida. En este caso, se decide seleccionar K=4 para lograr una mejor explicabilidad de los grupos obtenidos posteriormente.

El método de Calinsky se utiliza para reafirmar la decisión sobre el número de clústers, indicando que la variación entre 4 y 6 comienza a disminuir.

```
##          2 groups 3 groups 4 groups 5 groups 6 groups 7 groups 8 groups
## SSE      3257.767 2678.279 2233.822 1968.520 1749.526 1576.055 1441.007
## calinski 8676.975 7016.332 6674.216 6221.616 6002.426 5847.117 5696.376
##          9 groups 10 groups
## SSE      1306.057  1215.670
## calinski 5706.594  5582.112
```

Es importante notar que en ambos métodos mientras más clústers más se reduce la distancia intra-grupo, sin embargo, en la práctica la elección de un número de clústers muy grande dificulta la interpretación.

A continuación, se muestran los centroides asociados a cada uno de los grupos jerárquicos y la distribución porcentual de los grupos en la muestra.

cluster <int>	cluster_size <int>	RegDens <dbl>	RegPop <dbl>	MedHHInc <dbl>	MeanHHSz <dbl>	Dist. <dbl>
1	3479	0.8858149	0.9024863	0.9275654	0.7676056	0.2163288
2	3121	0.8032682	0.8802467	0.4648350	0.4951137	0.1940679
3	5323	0.4324629	0.4112812	0.3669453	0.5495491	0.3309912

cluster <int>	cluster_size <int>	RegDens <dbl>	RegPop <dbl>	MedHHInc <dbl>	MeanHHSz <dbl>	Dist. <dbl>
4	4159	0.5047487	0.4776389	0.8165424	0.6752224	0.2586121

4 rows

Se puede notar que no existe un gran desbalanceo entre la cantidad de observaciones presentes en los grupos seleccionados.

K-means

Una vez seleccionada la cantidad de grupos k, se puede ejecutar el algoritmo de K-means como método de optimización sobre todo el conjunto de datos. Los centroides de los grupos resultantes por el jerárquico pueden ser utilizados para inicializar los grupos de K-means puesto que la inicialización de los grupos es determinante en los resultados del algoritmo.

Como resultado de la aplicación de K-means se obtienen los siguientes centroides que representan las características de las regiones agrupadas.

```
##     RegDens    RegPop  MedHHInc  MeanHHSz
## 1 0.8961453 0.8980168 0.9236275 0.7551550
## 2 0.8263386 0.8528702 0.4922415 0.4603232
## 3 0.3792637 0.3998804 0.3662906 0.4726757
## 4 0.4971317 0.4720342 0.7441200 0.7750688
```

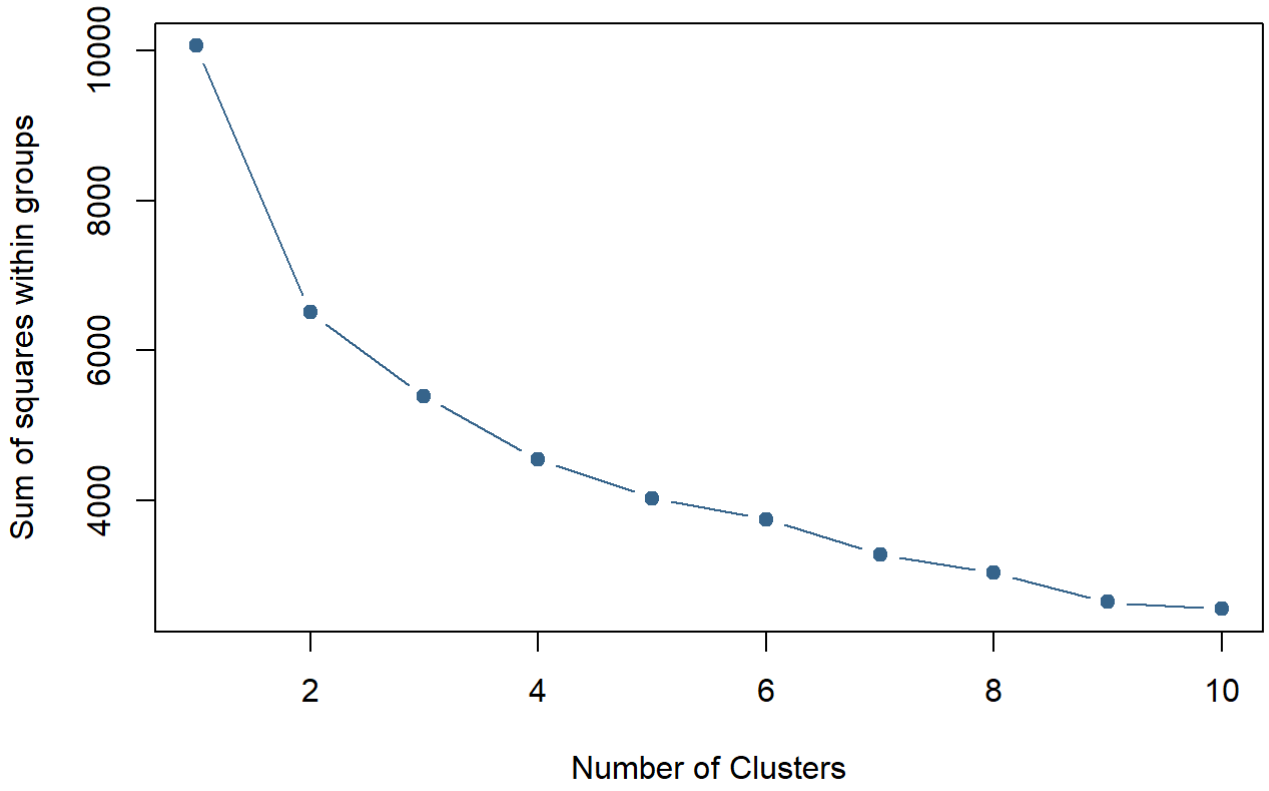
A continuación se muestra la frecuencia y distribución porcentual de los grupos en k-means en el total de observaciones.

```
## [1] 0.2367169 0.1933468 0.2989585 0.2709778
```

cluster <int>	cluster_size <int>
1	7614
2	6219
3	9616
4	8716

4 rows

Se puede notar que la distribución porcentual de K-means y la cantidad de observaciones en cada grupo no presenta grandes diferencias respecto al método jerárquico. Para comprobar si existe correspondencia entre entre el resultado del método jerárquico sobre una muestra y k-means sobre todo el conjunto de datos se ejecuta el método de k-means con diferentes ks. Asimismo, para cada k se aplica el metodo k-means con 3 inicializaciones distintas de centroides para que no sea tan sensible a los centroides de partida.

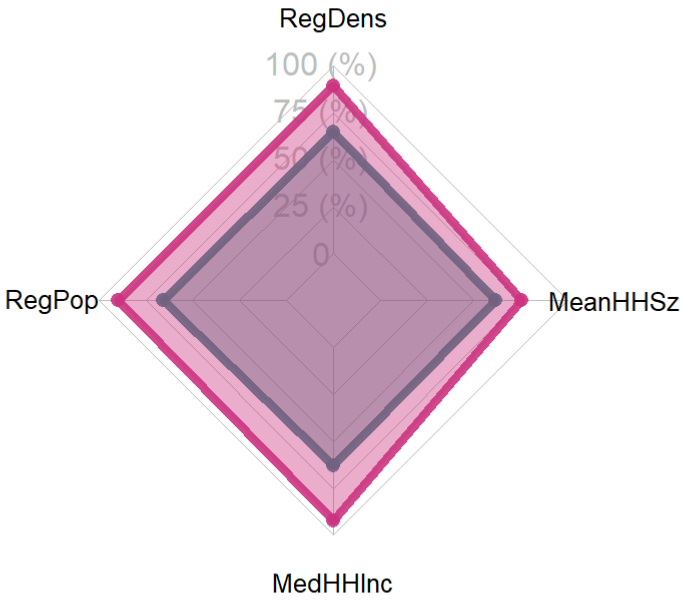


Sobre está gráfica podemos aplicar el Método del Elbow que nos confirma que el número de clústers adecuado puede estar entre 4 y 6. Por lo tanto y teniendo en cuenta lo explicado anteriormente, se mantiene la decisión de establecer K=4.

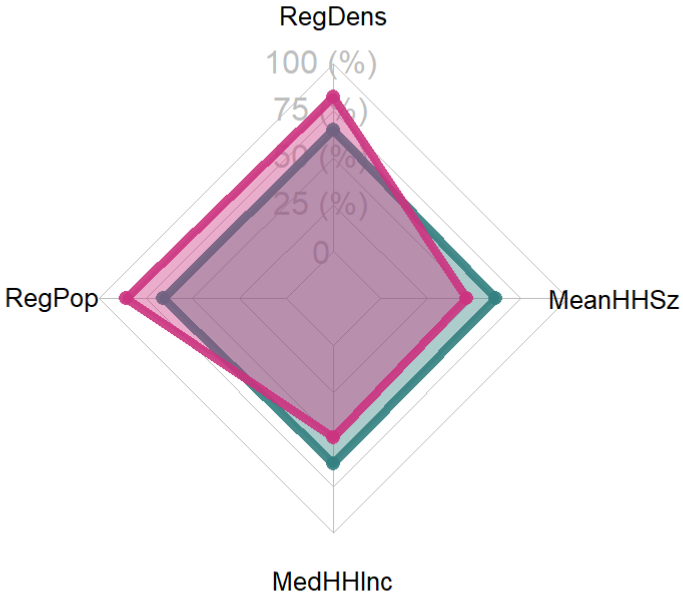
Validación e interpretación de resultados

Al representar los centroides es posible reconocer grupos diferentes. Para un mejor entendimiento visualizamos los centroides a través de gráficos de radar para cada uno de los grupos obtenidos.

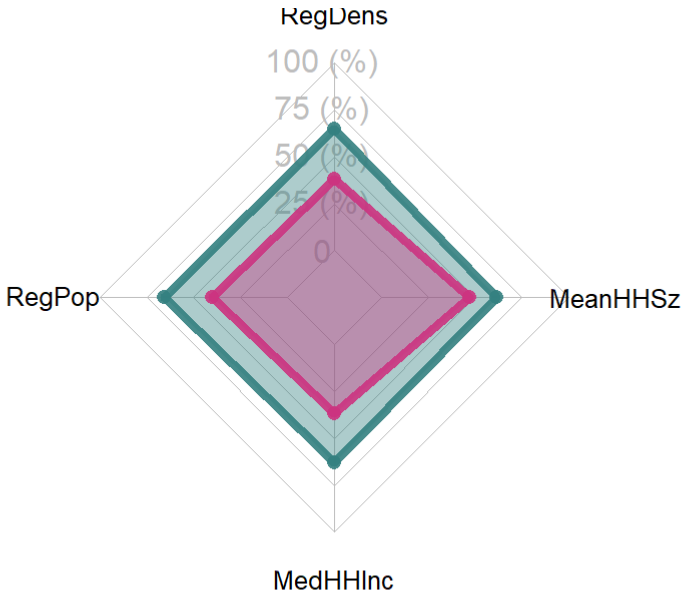
Cluster:1



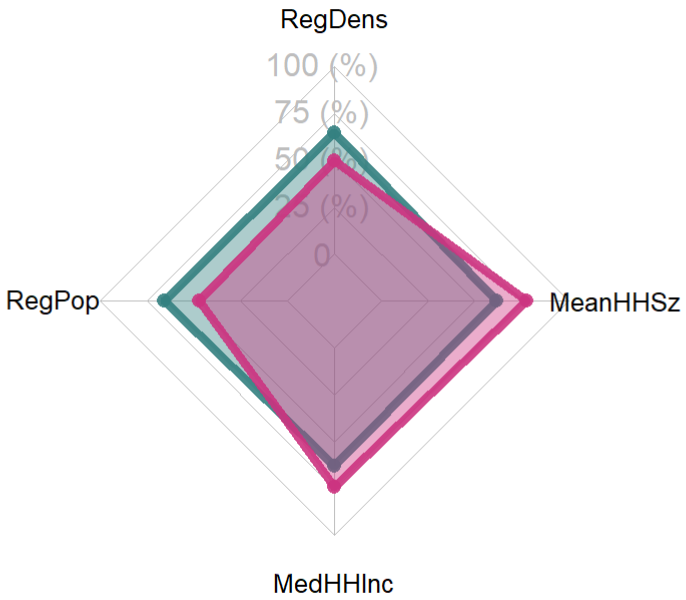
Cluster:2



Cluster:3



Cluster:4



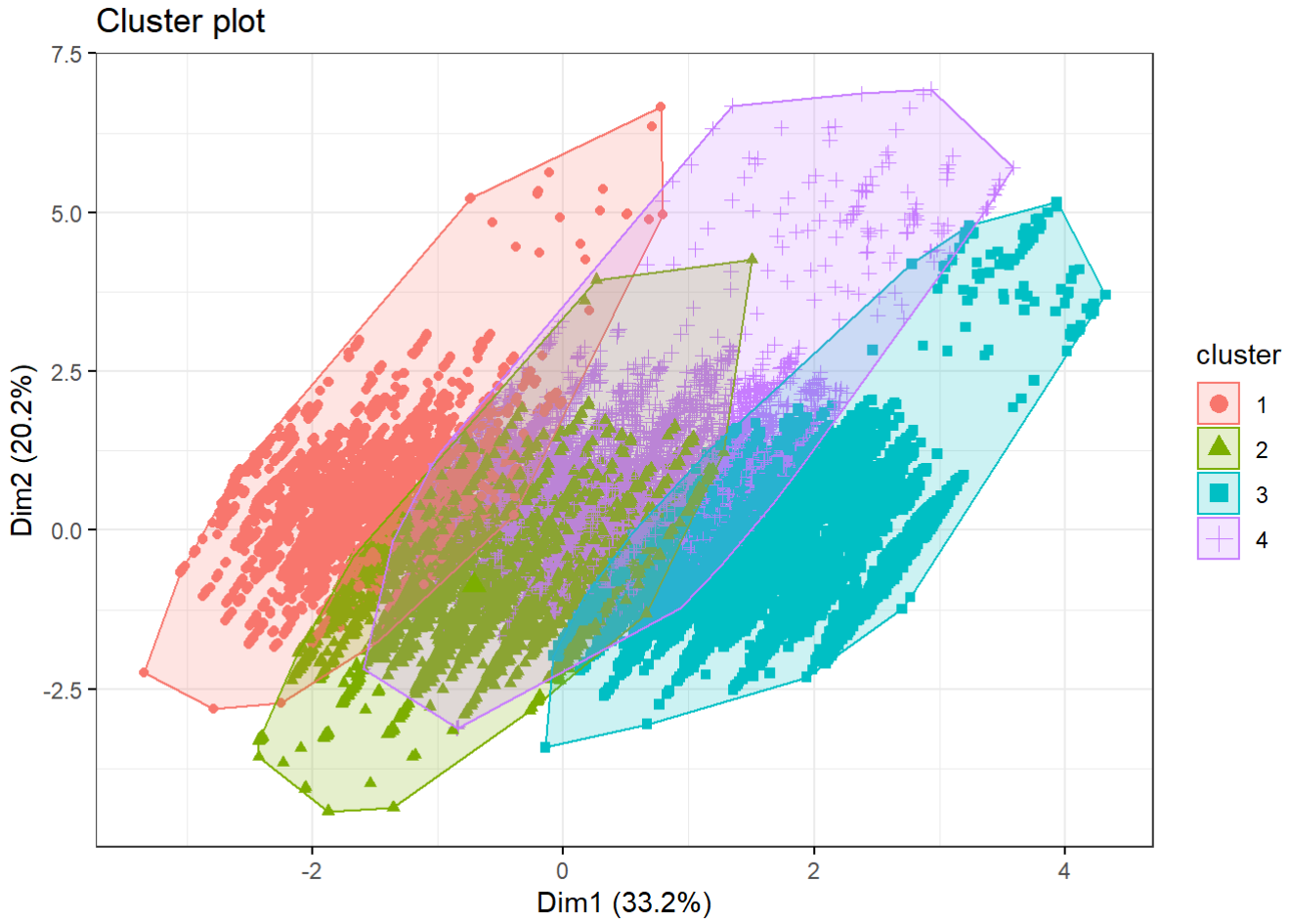
En este gráifco, la silueta azul representa el comportamiento medio de los individuos del dataset, mientras que la silueta rosa representa el comportamiento de los individuos del clúster en cuestión.

Al analizar los gráficos presentados se puede concluir que:

- El grupo 1 está conformado por los individuos que viven en zonas de gran densidad y población y que a la vez tienen altos ingresos y casas de mayor tamaño que la media. Este grupo se puede asociar con familias trabajadoras de altos ingresos que viven en la ciudad y pueden permitirse casas de mayor tamaño que la media, por lo que pueden recibir el nombre de **Familias trabajadoras de altos ingresos**

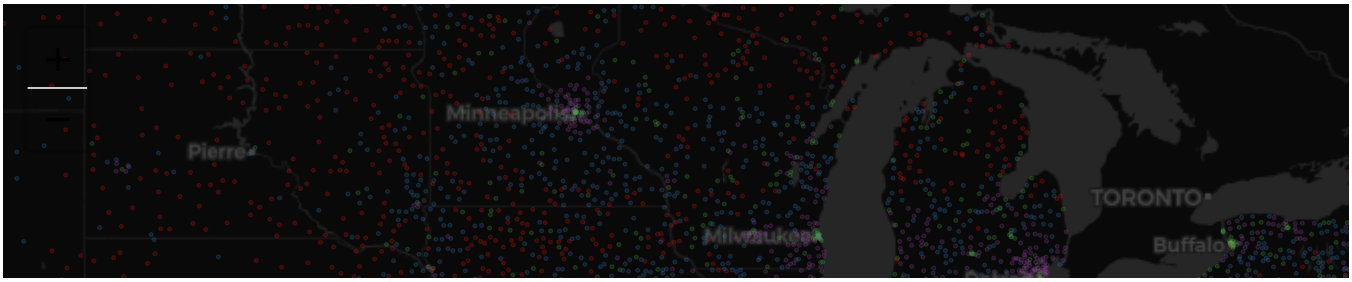
- El grupo de 2 está formado por las personas que vicen en zonas de gran población y densidad, pero sus ingresos y el tamaño de sus casas está por debajo de la media. A este grupo se le podría llamar **Familias trabajadoras** y pueden estara sociados a familias trabajadoras que necesitan vivir cerca de la ciudad pero no pueden permitirse grandes casas.
- El grupo 3 está formado por los individuos que viven en zonas de poca población y densidad, con bajos ingresos y en casas más pequeñas. Este grupo se puede asociar a personas que viven alejadas de la ciudad porque no pueden permitirse una casa en el centro y se puede denominar **Familias de bajos ingresos**.
- El grupo 4 está conformado por familias de altos ingresos y que viven en casas de mayor tamaño que la media, sin embargo viven en zonas no muy pobladas ni muy densas. Este grupo se puede asociar con las personas que tienen un status económico medio alto y viven en las afueras de las ciudades, por lo tanto reciben el nombre de **Familias de medios altos ingresos**

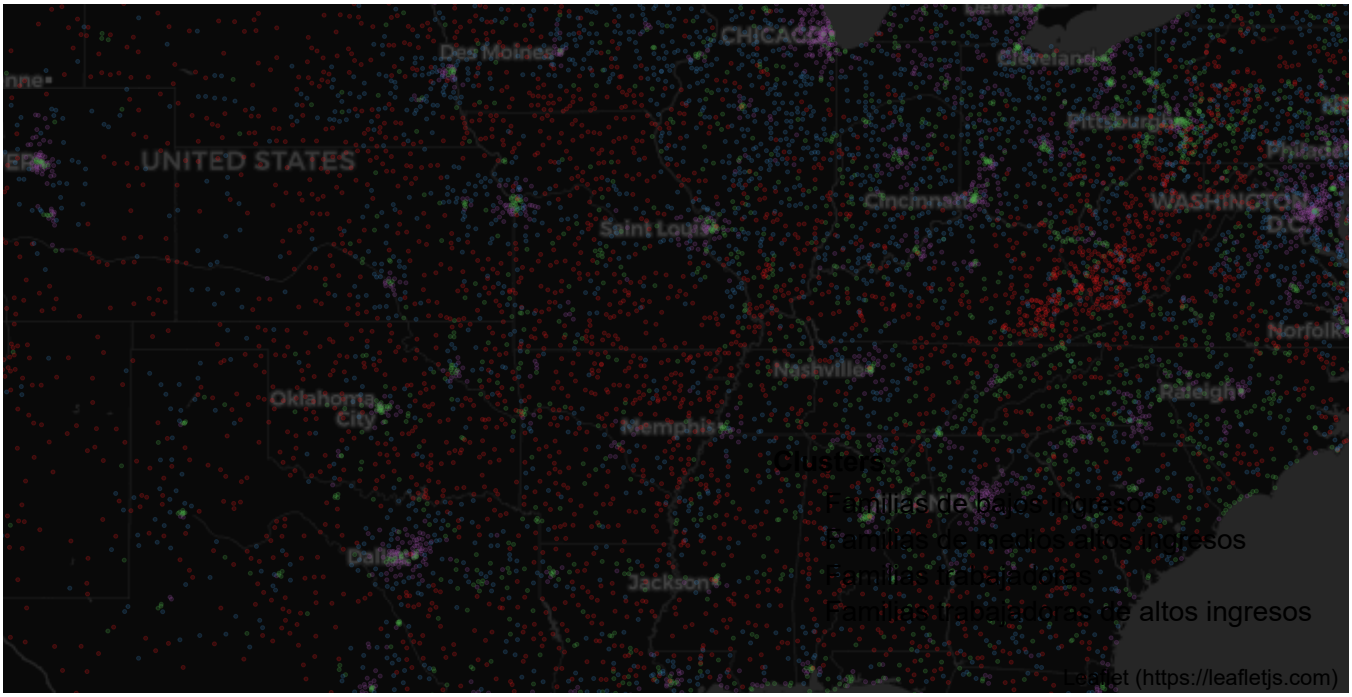
El siguiente gráfico nos permite visualizar los datos y los respectivos grupos en dos dimensiones. Esta representación bidimensional se logra al aplicar PCA sobre las cuatro variables originales.



Teniendo en cuenta la varianza representada en cada componente, se puede afirmar que en las dos primeras componentes se preserva el 50% de la varianza de los datos. Este datos no es concluyente para la toma de decisiones, pero facilita la visualización de los grupos.

En la siguiente figura se representan una muestra de los individuos del dataset teniendo en cuenta la latitud y longitud y el grupo asignado.





Al observar estos dos gráficos se puede notar que no se logra una evidencia clara a los hojos humanos sobre los criterios de separación tenidos en cuenta por el algoritmo, sino que se puede notar que los grupos están bastante cercanos y solapados entre sí.

Conclusiones

En los métodos de agrupamiento, al ser técnicas de aprendizaje no-supervisado, no hay una respuesta correcta. Esto hace que la evaluación de los grupos identificados sea un poco subjetiva. En este trabajo, se han aplicado varios algoritmos de *clustering* y se ha dado una interpretación a cada uno de los grupos, respondiendo a los requerimientos del problema. La validez del análisis y la elección de los diferentes parámetros es subjetiva y en estos casos, depende del escenario a aplicar.

Bibliografía

- https://medium.com/datos-y-ciencia/introducci%C3%B3n-a-los-modelos-de-agrupamiento-en-r-72739633e8f3 (https://medium.com/datos-y-ciencia/introducci%C3%B3n-a-los-modelos-de-agrupamiento-en-r-72739633e8f3)
- https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/ (https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/)
- https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc (https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc)