Trabajo practico 05-02: Clustering

Ejercicio 1

Cálculo de distancias:

x	y	Euclideana	Manhattan	Minkowski (p = 3)
4	8	4.472135955	6	4.160167646
9	17	14.76482306	20	13.6440899
3	7	3.16227766	4	3.036588972

¿Encuentra diferencias relativas entre las diferentes métricas utilizadas y el resultado obtenido? Explique el comportamiento de cada una utilizando gráficas.

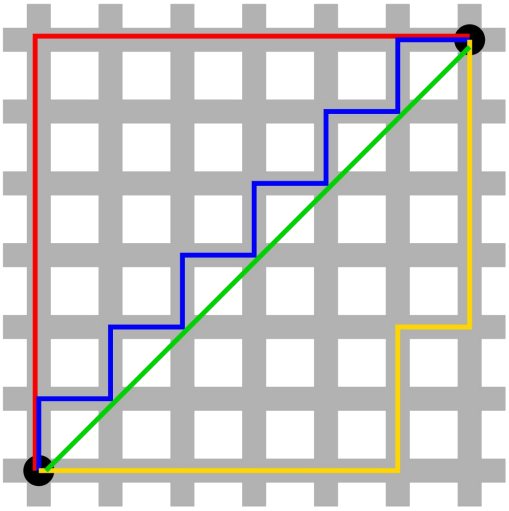
Sí, la diferencia relativa se puede observar en la distancia Manhattan del punto (9,17) respecto del centroide (2,4) y eso está explicado por la forma en que esta se calcula y agravado por encontrarse lejos del mismo.

La distancia Manhattan (o también conocida como “Taxicab geometry”) surge ante la imposibilidad de calcular distancias diagonales a un punto determinado en una ciudad (debido a la existencia de las cuadras). Por lo tanto, lo que propone la distancia Manhattan (siempre en 2D) es reemplazar el cálculo pitagórico de la diagonal y en su lugar sumar la distancia vertical de cuadras junto con la distancia horizontal de cuadras.

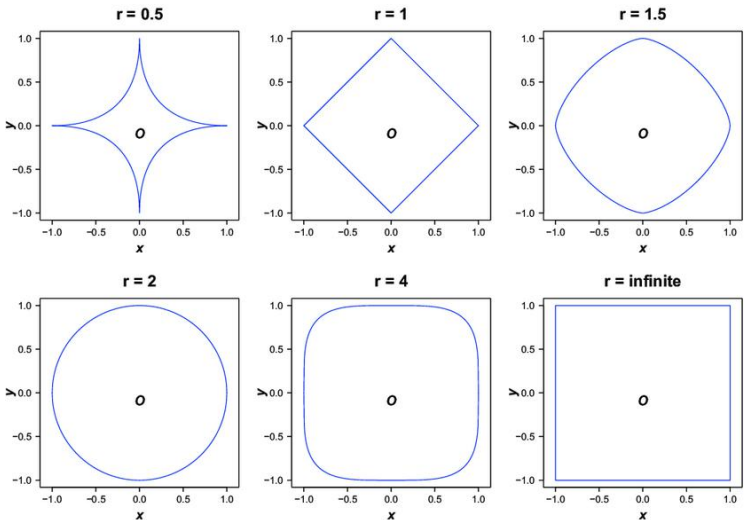
Esto da lugar a que las distancias lejanas resulten en valores más elevados que si se calculara aplicando otro método y es lo que justifica que la distancia del punto (9,17) al centroide (2,4) sea 20.

Respecto a la distancia Minkowski, se trata de un cálculo de distancias vectorizado y generalizado que varía según el valor asignado a p.

Si p vale 1, su cálculo resulta equivalente a la distancia Manhattan. Si p vale 2, su cálculo equivale a la distancia Euclideana por estar elevado al cuadrado (pitágoras).



Euclidea vs. Manhattan



Comportamiento de Minkowski con diferentes valores de *p*

Ejercicio 2

Cálculo de distancias:

#	pronostico	temperatura	humedad	viento	Distancia a #1
1	lluvioso	templado	alta	fuerte	
2	lluvioso	frio	normal	fuerte	0.5
3	nublado	frio	normal	fuerte	0.75
4	soleado	templado	alta	leve	0.75
5	soleado	frio	normal	leve	1
6	lluvioso	templado	normal	leve	0.5
7	soleado	templado	normal	fuerte	0.5
8	nublado	templado	alta	fuerte	0.25
9	nublado	calor	normal	leve	1
10	lluvioso	templado	alta	fuerte	0

¿Cuáles son las instancias más cercanas a la instancia #1?

Se puede observar que la instancia más cerca a #1 es la instancia #10, que tiene una distancia de 0 por ser completamente iguales.

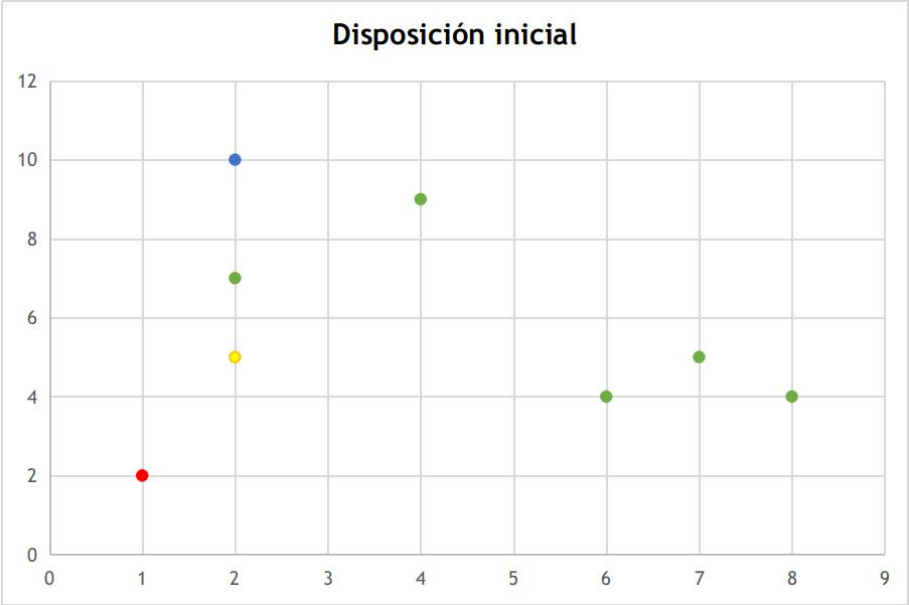
La sigue la instancia #8, a 0.25 de distancia, con un único valor de diferencia (nublado en vez de lluvioso).

Luego vienen las instancias #2, #6 y #7, todas a 0.5 de distancia por tener dos valores diferentes.

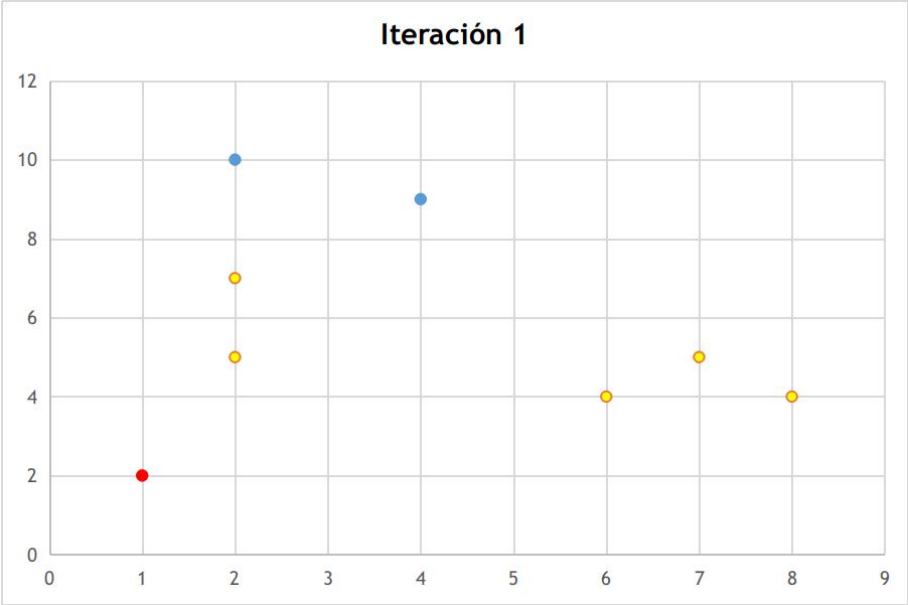
Ejercicio 3

Se muestran los cálculos y los gráficos correspondientes a cada tabla:

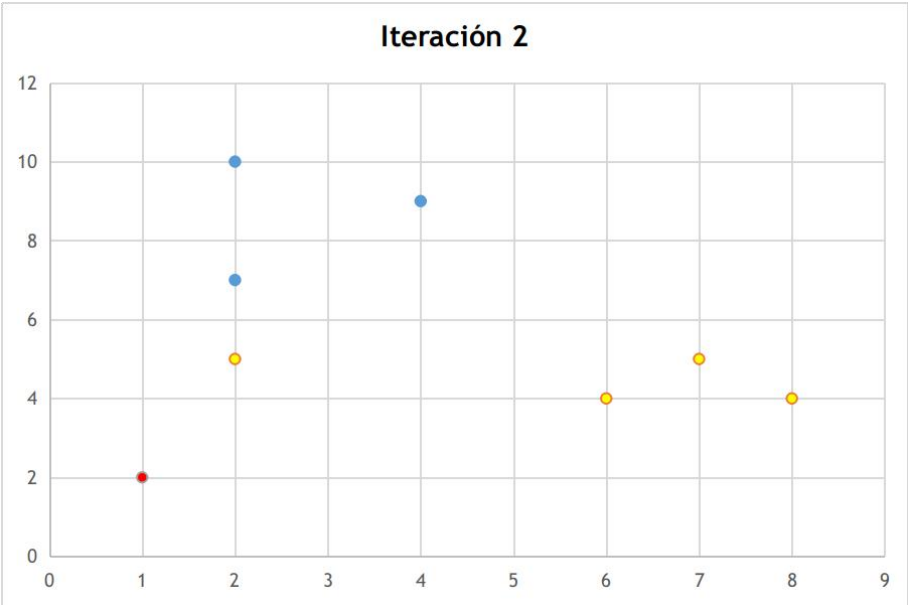
Disposición inicial		
PUNTO	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	2	7
A5	7	5
A6	6	4
A7	1	2
A8	4	9



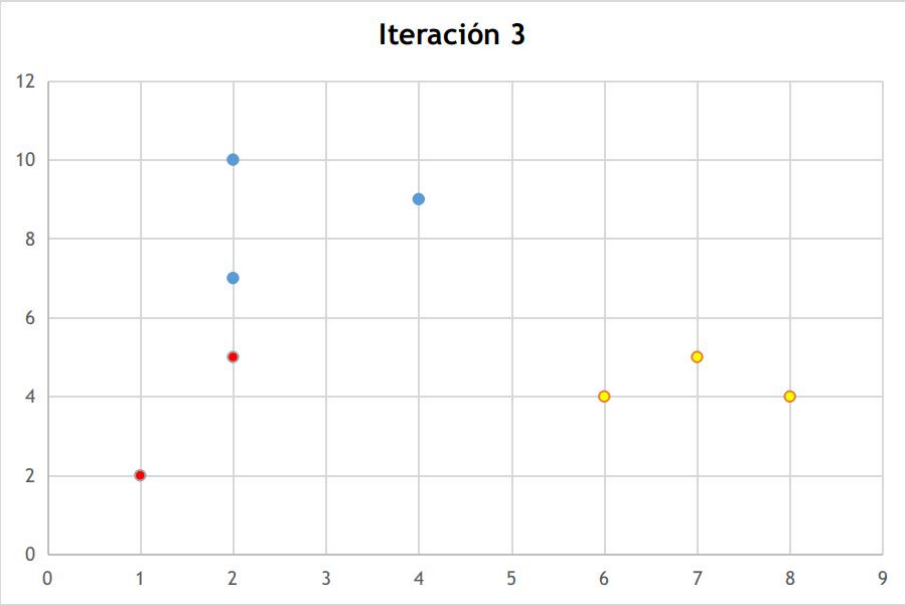
			Iteracion 1		
Distancia a A1	Distancia a A2	Distancia a A7	PUNTO	X	Y
-	-	-	A1	2	10
-	-	-	A2	2	5
8.485281374	6.08276253	7.280109889	A3	8	4
3	2	5.099019514	A4	2	7
7.071067812	5	6.708203932	A5	7	5
7.211102551	4.123105626	5.385164807	A6	6	4
-	-	-	A7	1	2
2.236067977	4.472135955	7.615773106	A8	4	9



						Iteracion 2		
PUNTO	X	Y	Distancia azul	Distancia amarillo	Distancia rojo	PUNTO	X	Y
A1	2	10	1.118033989	5.830951895	8.062257748	A1	2	10
A2	2	5	4.609772229	3	3.16227766	A2	2	5
A3	8	4	7.433034374	3.16227766	7.280109889	A3	8	4
A4	2	7	2.692582404	3.605551275	5.099019514	A4	2	7
A5	7	5	6.020797289	2	6.708203932	A5	7	5
A6	6	4	6.264982043	1.414213562	5.385164807	A6	6	4
A7	1	2	7.762087348	5	0	A7	1	2
A8	4	9	1.118033989	4.123105626	7.615773106	A8	4	9
Centroide azul	3	9.5						
Centroide amarillo	5	5						
Centroide rojo	1	2						



						Iteración 3		
PUNTO	X	Y	Distancia azul	Distancia amarillo	Distancia rojo	PUNTO	X	Y
A1	2	10	1.490711985	6.656763478	8.062257748	A1	2	10
A2	2	5	3.726779962	3.783186488	3.16227766	A2	2	5
A3	8	4	7.086763875	2.304886114	7.280109889	A3	8	4
A4	2	7	1.795054936	4.506939094	5.099019514	A4	2	7
A5	7	5	5.676462122	1.346291202	6.708203932	A5	7	5
A6	6	4	5.734883511	0.559016994	5.385164807	A6	6	4
A7	1	2	6.871842709	5.367727638	0	A7	1	2
A8	4	9	1.374368542	4.828301979	7.615773106	A8	4	9
Centroide azul	2.666666667	8.666666667						
Centroide amarillo	5.75	4.5						
Centroide rojo	1	2						



Ejercicio 4

Luego de aplicar el pre-procesamiento necesario, se obtuvo el siguiente agrupamiento al ejecutar K-Means con 3 clusters. Posteriormente se ejecutó el método Elbow para verificar si la decisión fué acertada.

Adicionalmente, se calculó la moda del tipo de trigo en cada cluster resultante para poder tener una suerte de predictor. En las conclusiones se habla de esto.

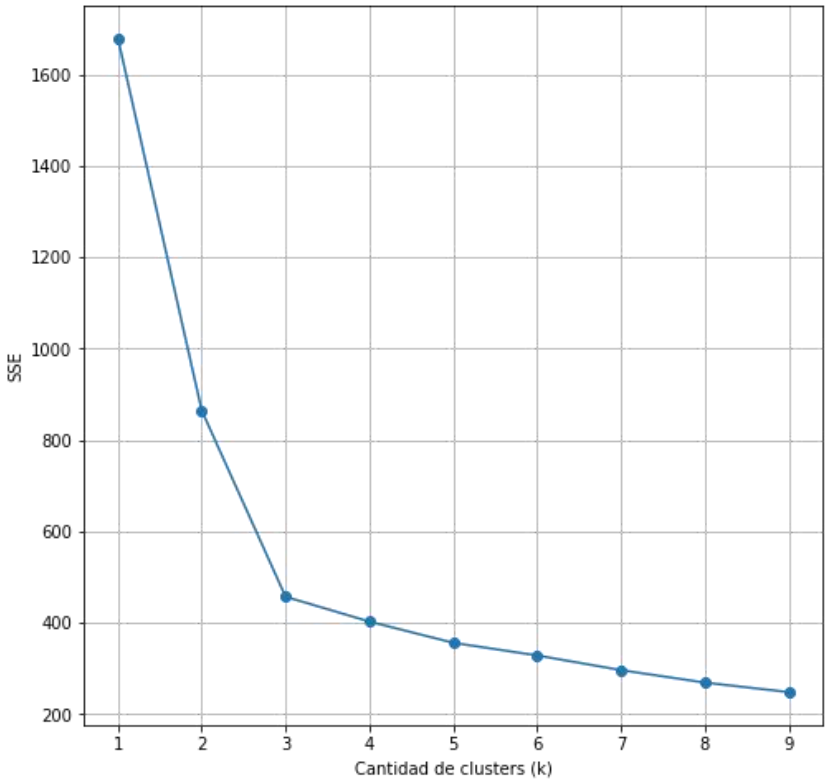
Para facilitar la interpretación, se muestran primero el gráfico del método Elbow y luego el agrupamiento resultante con las conclusiones al pie del mismo.

Observaciones:

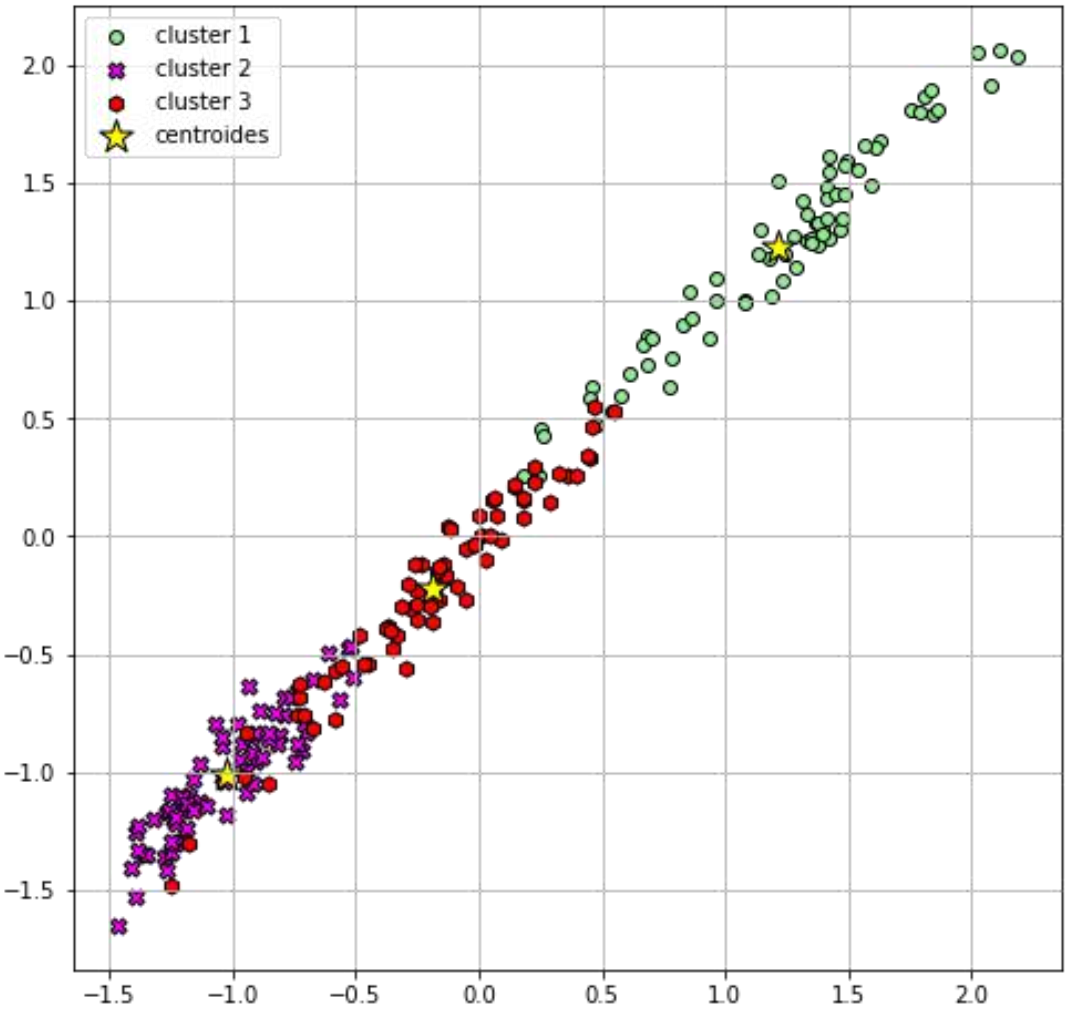
- Se aplicó K-Means utilizando 3 clusters y como resultado se observa una agrupación de datos bastante correcta.

- En el caso de que haya superposiciones de puntos pertenecientes a diferentes clusters, es importante recordar que se puede deber a:
  - Mala selección de clusters.
  - Se están graficando solo dos variables de un dataset multidimensional, por lo tanto puede que una variable clave (que determina a cuál cluster pertenece un punto) no haya sido graficada.

- De todas maneras, para determinar la cantidad de clusters que mejor agrupa los datos, se pueden aplicar dos técnicas:
  - Método de Elbow (se realizará a continuación).
  - Coeficiente silueta.



- El método Elbow nos confirma que la cantidad ideal de clusters es 3 y que la superposición de puntos de diferentes clusters se debe a la maldición de la dimensionalidad explicada anteriormente.



Conclusiones:

1. En el scatterplot se graficaron las variables *área* y *perímetro* del dataset y con cada color y forma se indicó a cuál cluster pertenecía cada tupla del dataframe.

2. A simple vista, y sin importar a qué cluster pertenece cada punto, se puede decir que hay una relación de dependencia lineal entre las variables graficadas (y que está matemáticamente respaldada también, pues son área y perímetro).

3. El cluster #3 representa a los tipos de trigo con mayor área y perímetro de todo el dataset, los cuales pertenecen al trigo de tipo 1.

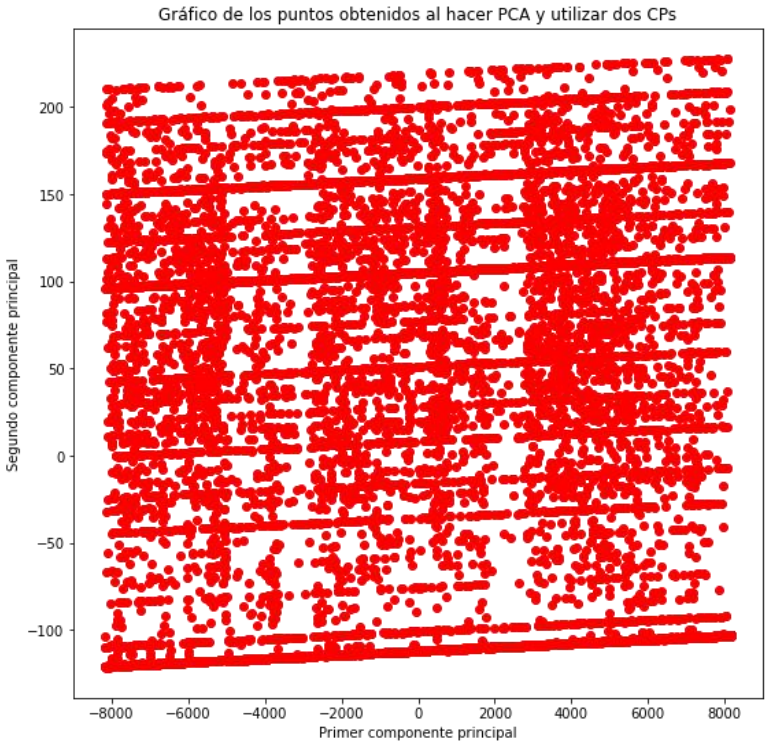
4. El cluster #2 posee en su mayoría trigo tipo 3. Este tipo de trigo posee un área y un perímetro con valores promedio respecto a los otros dos presentes en el dataset, pues se ubica en el centro del scatter.

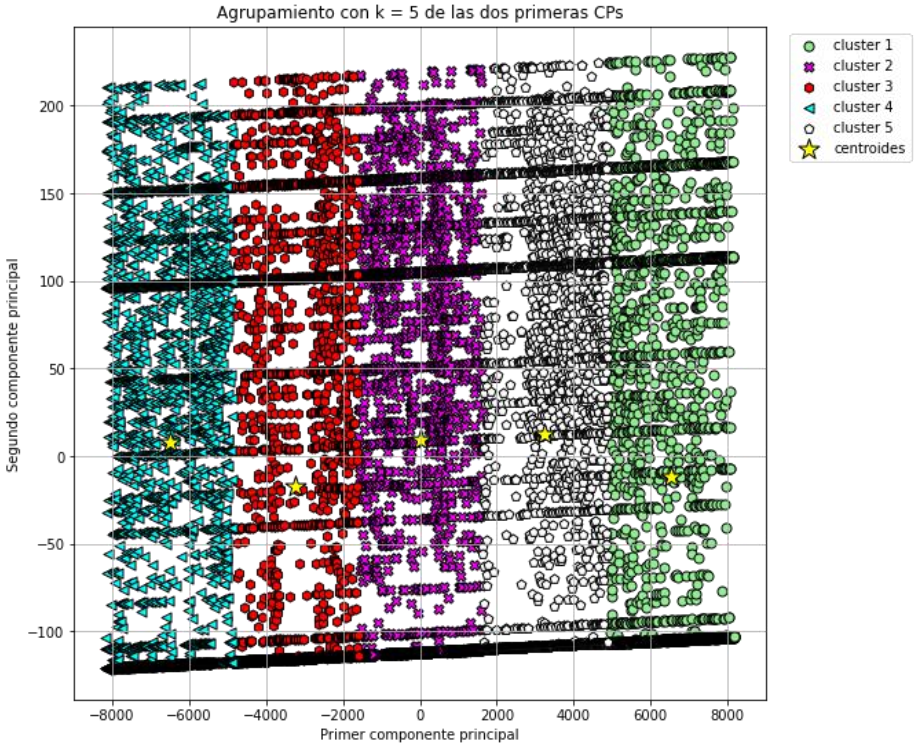
5. El cluster #1 posee una moda de trigo de tipo 2, cuyos valores de área y perímetro son los de menor tamaño de todo el dataset

Ejercicio 5

Para poder definir cuales son las features que permiten el mejor agrupamiento se realizarán las siguientes acciones:

- Imputar por hot-deck los 8 valores faltantes de anios_en_unlu.
- Convertir la columna promedios_1er_anio a numerico.
- Pre-procesar las columnas String para discretizarlas.
- Escalar los valores.
- Realizar un análisis por componentes principales (PCA) sobre *todas* las columnas del dataset.
- Aplicar K-Means con k = 5.





Observaciones:

Aparentemente el PCA arrojó buenos resultados, con tan solo dos componentes principales se obtuvo casi el 100% de la varianza y esto despierta ciertas dudas:

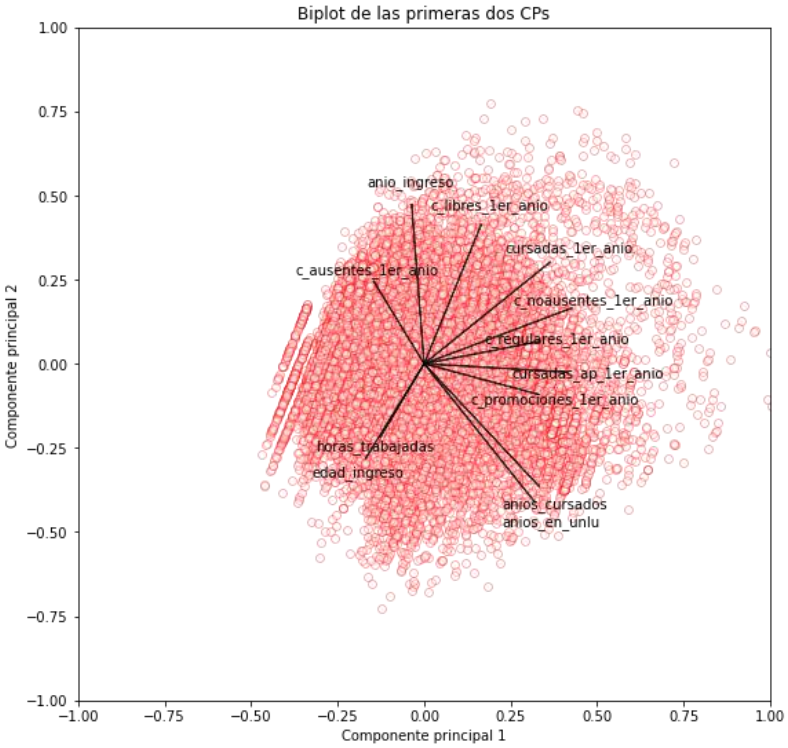
- Lo primero que se me ocurre al realizar K-Means sobre un dataset con tipos de datos mixtos (Strings, numericos y Bools) es qué medida de distancia se puede aplicar de igual manera para todos (¿minkowski?) y que a la vez pueda ser compatible con K-Means (que usa distancia euclideana).

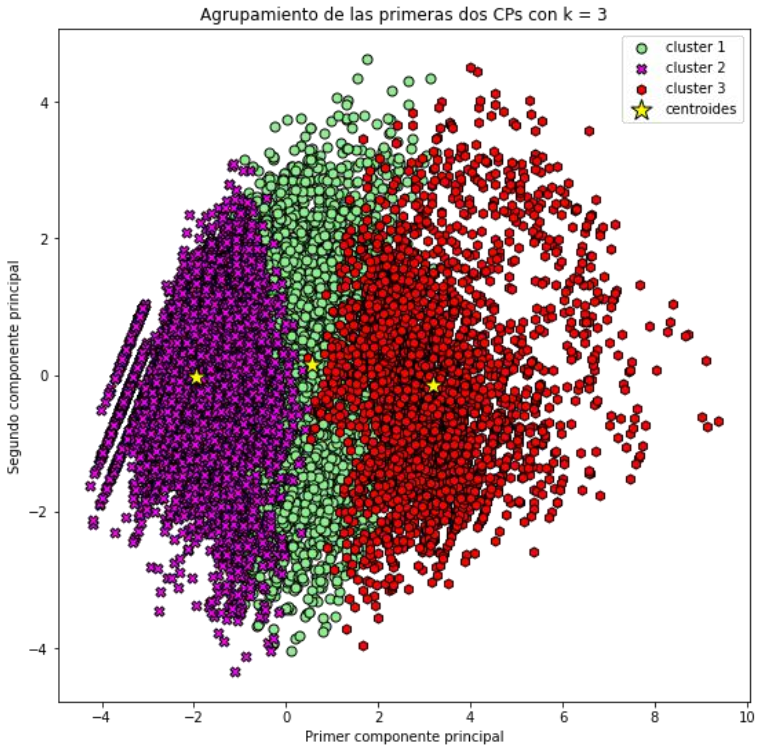
- Este ejercicio me deja bastante confundido, principalmente porque es bastante dificil medir la distancia entre valores numéricos y valores no numéricos como un String. **¿Qué tan representativo sería eso? ¿Tiene sentido?**

- Se puede utilizar el coeficiente de Jaccard pero solo tendría validez si todo el dataset estuviera compuesto por valores de tipo String, cosa que aquí no sucede, y luego así hacer algo similar a lo que requería el ejercicio 2.

Debido a lo mencionado, se decidió elegir TODAS las *features numéricas* para luego realizar PCA sobre ellas, algo que a mí criterio tiene un poco más de sentido.

Con el fin evitar concluir erróneamente, se realizó el gráfico biplot previo al scatterplot. Para facilitar la interpretación, se mostrarán ambos gráficos juntos en la siguiente página.





Observaciones:

- Varianza capturada por cada CP:
  - CP1: 0.41771296066644065
  - CP2: 0.13153433332430756
  - CP3: 0.10728291451551471
  - CP4: 0.09264840004781584
  - CP5: 0.08071581588454496

- Se pudo capturar un 82.98% de la varianza utilizando todas las variables numéricas.

- Se necesitaron 5 componentes principales para lograr dicho porcentaje.

Características de cada cluster:

1. El cluster más cercano a la CP2 es el cluster 2 (color magenta), y la mayor parte de la varianza de dicha CP está compuesta por las features *horas_trabajadas*, *edad_ingreso* y *c_ausentes_1er_anio.*

2. El cluster 3 (rojo) es el que está más a la derecha en el gráfico y eso indica que está altamente influenciado (ver biplot más arriba) por la varianza de las features relacionadas a cantidad de asignaturas cursadas, regularizadas o desaprobadas (*cursadas_1er_anio, c_noausentes_1er_anio, c_regulares_1er_anio, cursadas_ap_1er_anio, c_promociones_1er_anio*)

3. El cluster 1 (color verde claro) se aloja entremedio de los clusters mencionados antes pero con la diferencia de que posee varios puntos ubicados en ambos extremos de la CP2 y en el medio de la CP1 del scatterplot. En la región superior de la CP2 se encuentran las features *anio_ingreso* y *c_libres_1er_anio* mientras que en la región inferior de la CP2 se ubican entre las features *horas_trabajadas-edad_ingreso* y las features *anios_cursados-anios_en_unlu.*

Conclusiones:

Alumnos pertenecientes al Cluster 1:

1. Se dividen dos grandes grupos:
2. Alumnos que ingresaron trabajando pero a la vez llevan al menos un año cursando en la universidad.
3. Alumnos que ingresaron trabajando, no lo hicieron inmediatamente después de finalizar la secundaria, y que no completaron el primer año de cursada o quedaron libres en varias materias de primer año.

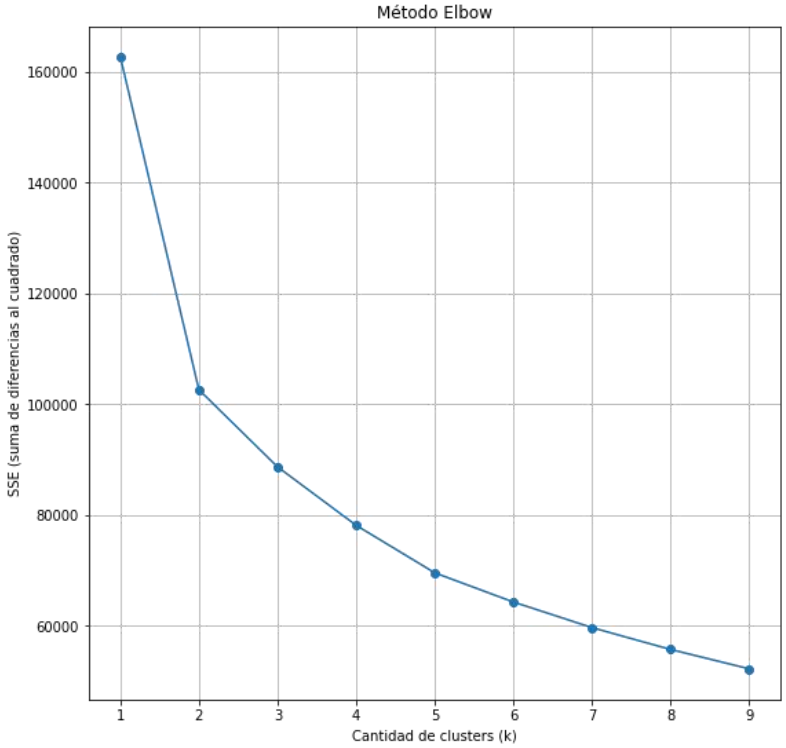
Alumnos pertenecientes al Cluster 2:

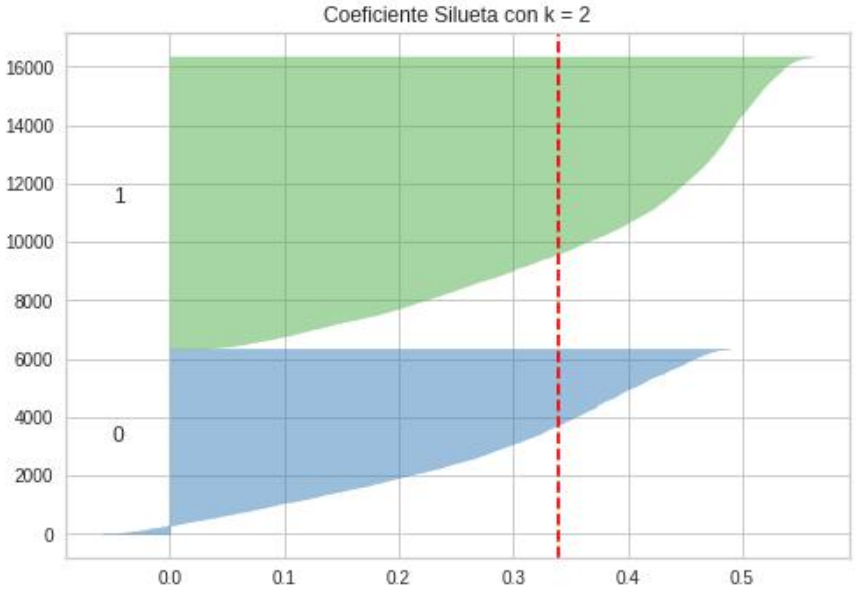
1. Han faltado bastante durante el primer año de su carrera.
2. Tienen que trabajar.
3. No ingresaron luego de finalizar el secundario, sino algunos años después.
4. Casualmente, los últimos dos factores son justificantes del primero y tiene sentido que estén relacionados.
5. Por lo tanto, se puede decir que los alumnos que pertenezcan a este cluster han dejado la carrera en el primer año, pues se encuentran en el extremo opuesto al cluster 3.

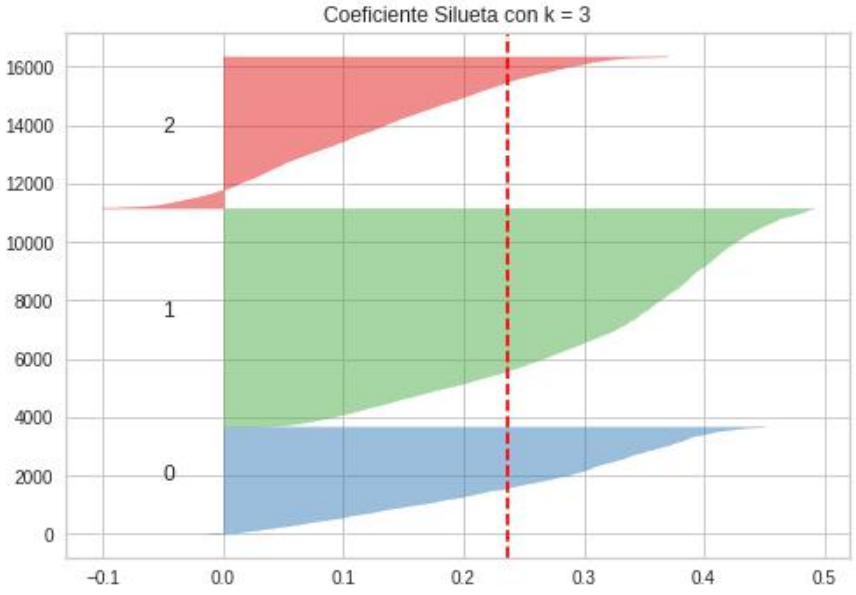
Alumnos pertenecientes al Cluster 3:

1. Como se explicó arriba, este cluster agrupa a los alumnos que al menos completaron primer año.
2. Sin embargo, y a pesar de que la feature *activo_2017* fué descartada por ser binaria, este cluster se ubica cerca del extremo inferior derecho de la CP1 donde se encuentran las features *anios_cursados* y *anios_en_unlu* lo que permite estimar que al menos una parte de este grupo continúa cursando su carrera.

Luego de llegar a las previas conclusiones, se aplicaron el método de Elbow y el coeficiente Silueta con el fin de obtener el *k* ideal para la agrupación. Se adjuntan ambos gráficos en la siguiente página.


Método Elbow
SSE (suma de diferencias al cuadrado)
Cantidad de clusters (k)
160000
140000
120000
100000
80000
60000
1
2
3
4
5
6
7
8
9


Coeficiente Silueta con k = 2
1
0
16000
14000
12000
10000
8000
6000
4000
2000
0
0.0
0.1
0.2
0.3
0.4
0.5


Coeficiente Silueta con k = 3
2
1
0
16000
14000
12000
10000
8000
6000
4000
2000
0
-0.1
0.0
0.1
0.2
0.3
0.4
0.5

Observaciones:

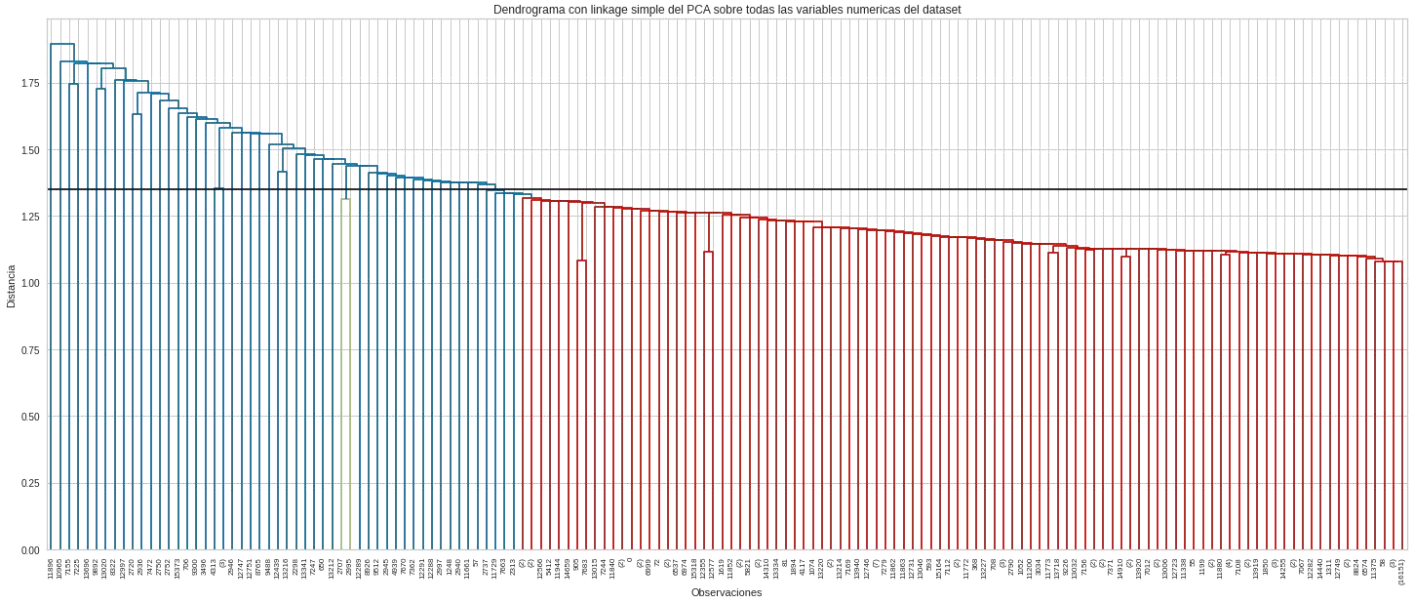
- El método de Elbow indica que los mejores valores para asignar a K son 2 y 4.
- El coeficiente silueta indica que el mejor valor para K es 2 (a pesar de tener algunos errores) con un score del 33%.
  - Con 5 clusters el coeficiente silueta es: 0.213476055207649.
  - Con 6 clusters el coeficiente silueta es: 0.1977395533491375.
  - Con 7 clusters el coeficiente silueta es: 0.20211103837086009.
  - Con 8 clusters el coeficiente silueta es: 0.19842425790886384.
  - Con 9 clusters el coeficiente silueta es: 0.20575752630560057.
  - Con 10 clusters el coeficiente silueta es: 0.20947190356790438.
- El coeficiente silueta para K = 3 (valor elegido por mi interpretación) determina que hay algunos errores de agrupación para los puntos pertenecientes al cluster 3 (rojo), pues su score es de 23% (un detrimento importante comparado con K = 2).

Conclusiones:

1. El coeficiente silueta nos indica que las conclusiones apuntadas anteriormente acerca del scatterplot pueden no ser del todo ciertas debido a que la cantidad inicial de clusters no es la óptima.

2. En el caso de que sí lo fuera, con K = 2, aun así estaríamos ante un score bajo, síntoma de que persistirían los errores de clasificación.

Para finalizar, se procede con la aplicación de un algoritmo jerárquico con linkage simple sobre el mismo dataset y utilizando el mismo análsisis de PCA sobre todas las features numéricas con la intención de encontrar una mejor manera de agrupar los datos, tal como lo pide el enunciado.



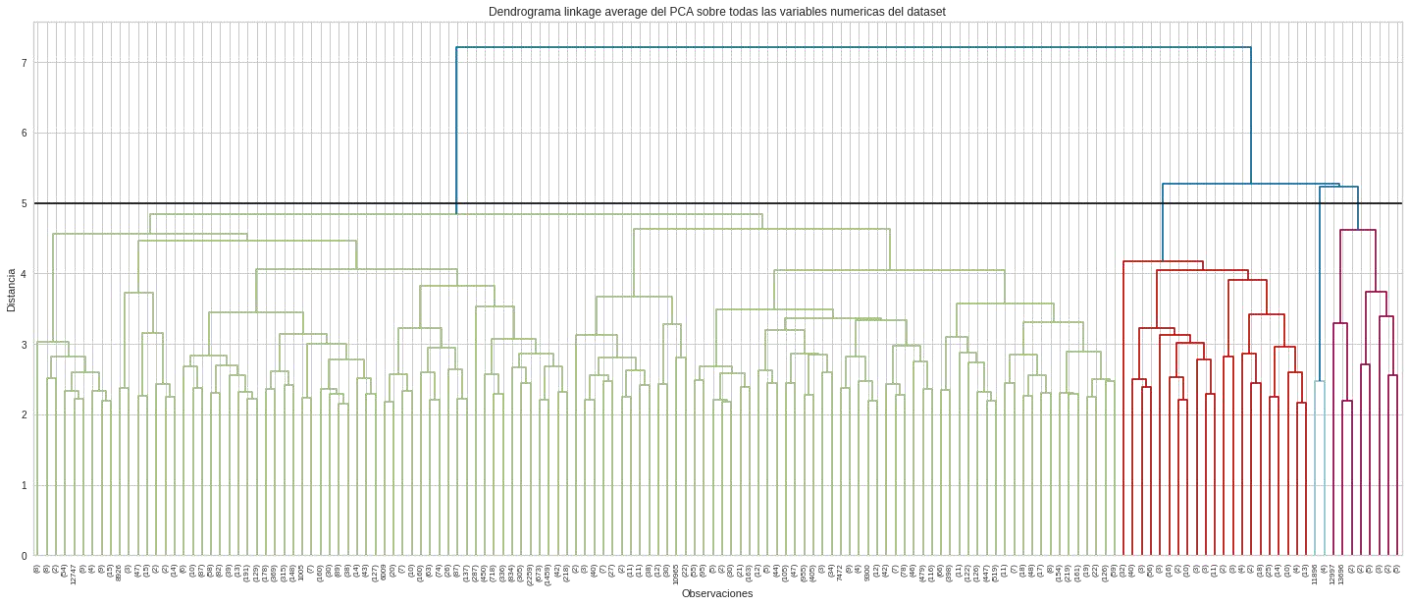
Observaciones:

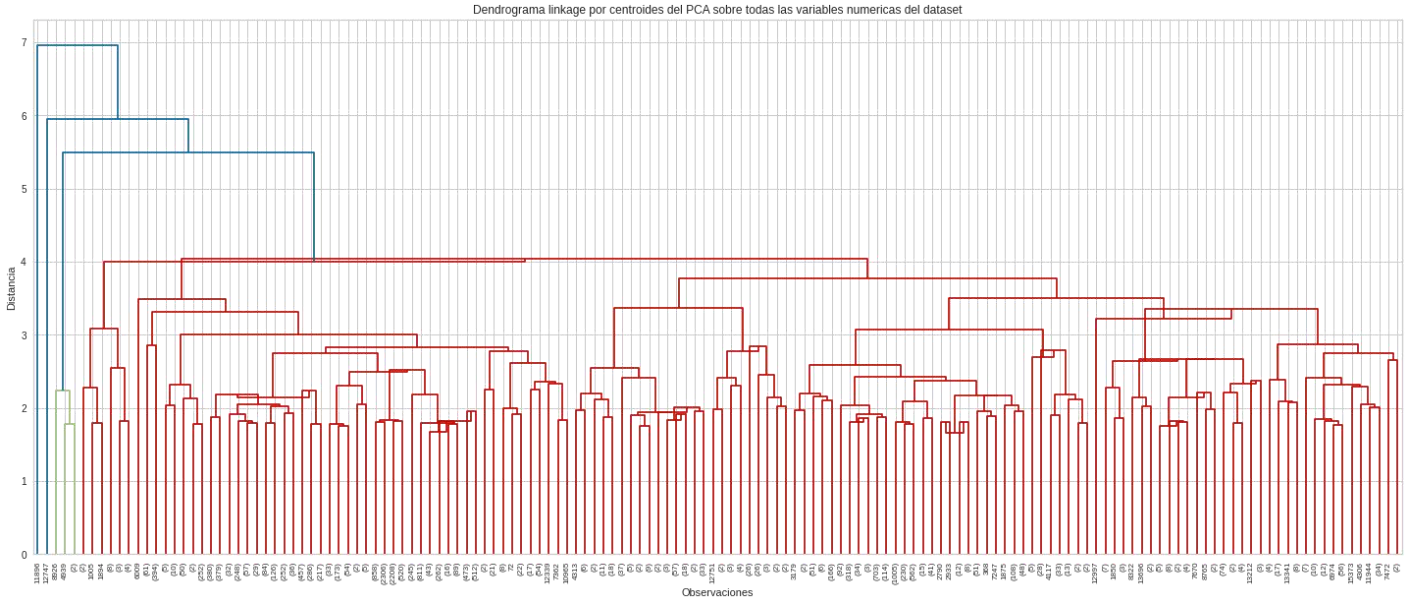
- Debido a que este agrupamiento es realizado por linkage simple no permite encontrar una mejor forma de agrupar los datos, pues se obtienen también tres clusters utilizando el valor 1.35 como distancia máxima entre cada observación.

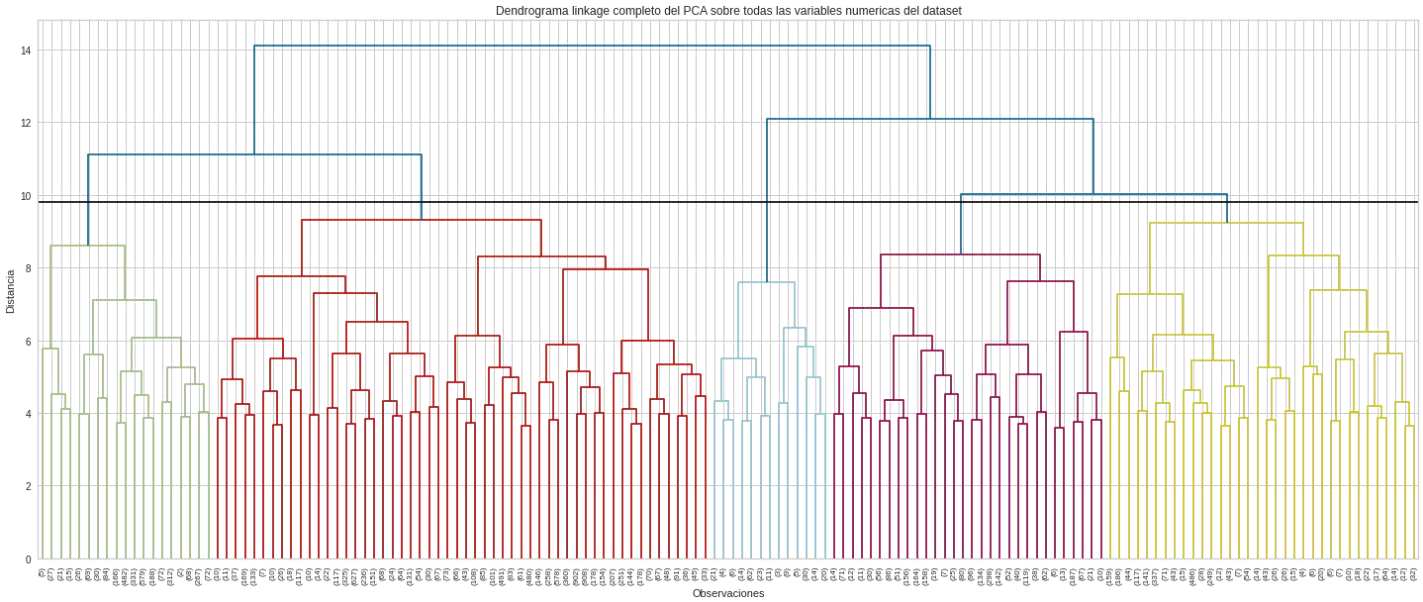
Ejercicio 6

Al utilizar el mismo dataset, se re-utilizó el análisis por PCA sobre todas las features numéricas tal como en el punto anterior. También se obvio la ejecución de la aglomeración utilizando linkage simple, pues fué realizada en el punto anterior y sus resultados fueron desfavorables.

En resumen, se muestran a continuación los dendrogramas de las algomeraciones realizadas por linkage average, linkage por centroides y linkage completa.







Conclusiones:

1. Sin dudas el mejor método para agrupar este dataset utilizando un algoritmo jerárquico y aglomerativo es por linkage completo.

2. Respecto al valor de corte se pueden utilizar varios criterios. Sin embargo, los colores del dendrograma dicen mucho y es preferible utilizar una altura que evite que se fusionen clusters de diferente color (lo que se traduciría en una mezcla de grupos con diferentes características).

3. Es por esto que se decidió que la altura de corte sea 9.8 sobre el eje Y, un valor que evita la fusión de clusters.

Referencias:

- Imputación por la media - https://scikit-learn.org/stable/modules/impute.html
- PCA con python - https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- PCA con K-Means - https://365datascience.com/pca-k-means/
- Recuperación de CPs luego de hacer PCA - https://stackoverflow.com/questions/22984335/recovering-features-names-of-explained-variance-ratio-in-pca-with-sklearn