Reporte Técnico

Taller de Clustering no supervisado: Información de instituciones de educación superior en U.S.

Integrantes

Julian David Ruiz Herrera

Juan Felipe Usuga Villegas

Jonatan Urrego Zea

Johan Sebastian Cano Garcia

Raul vladimir Gaitan Vaca

1. Introducción

El siguiente reporte técnico trata la agrupación de universidades en Estados Unidos a partir de datos que se proporcionan a través de informes federales de instituciones, datos sobre ayuda financiera federal e información fiscal. Estos datos brindan información sobre el desempeño de las escuelas que reciben dólares de ayuda financiera federal y los resultados de los estudiantes de esas escuelas.

Para realizar agrupaciones se hizo usó una de una documentación para poder encontrar variables de interés que cumplan con los objetivos del proyecto,

los cuales son:

- Desarrollar un agrupamiento de instituciones de educación superior
- Caracterizar cada grupo
- Entender qué hace que un grupo sea una buena opción

Y se usaron métodos de aprendizaje no supervisado para entender qué porcentaje de la varianza en los datos explicaban estas variables.

2. Desarrollo técnico

2.1 Selección de los datos a usar

Para poder implementar cualquier modelo se necesita tener un set de datos completo e imputado con el cual se puedan hacer análisis y operaciones sin inconvenientes. Para poder seleccionar el mejor

grupo de variables, nos basamos en la documentación y mediante un análisis de componentes principales que hicimos en paralelo.

2.1.1 Análisis de componentes principales

Haciendo el análisis de porcentaje de valores faltantes se tiene que 1249 descriptores de los 1725 que tienen los datos, tienen más del 60% de valores como faltantes, por lo tanto, se tiene que hacer uso de la documentación para tratar de eliminar todo este ruido. (Anexo 1)

Después de hacer la intercepción entre las variables de interés y los datos faltantes obtenemos las siguientes variables/columnas:

INSTNM; Nombre de la institución

CONTROL; Identifica si la estructura de gobierno de la institución es:

pública (1),

privada sin ánimo de lucro (2)

privada con ánimo de lucro (3).

HCM2; Es el tipo de HCM que indica problemas financieros o de cumplimiento federal más graves, HCM2 indica un sistema de supervisión mayor dadas irregularidades en la institución y HCM un nivel de monitoreo menor.

(1) tiene problemas (mayor monitoreo)
(0) No tiene problemas, menor monitoreo del dinero.

COSTT4_P; Es el costo de asistencia tomado del programa académico más cursado durante el año.

COSTT4_A; Costo de asistencia anual general para todos los estudiantes.

TUITIONFEE_IN; Costo y tasas de matrícula para estudiantes de dentro del estado

TUITIONFEE_OUT; Costo y tasas de matrícula para estudiantes de fuera del estado.

TUITIONFEE_PROG; Costo y tasas de matrícula general, tanto si es fuera o dentro del estado, esto son aquellas universidades que no cobran alguna diferencia por esto.

STABBR; Estado (ubicación)

Porcentaje de estudiantes dados los ingresos familiares.

INC_PCT_LO, (0-$30,000)

INC_PCT_M1, (30,001-$48,000)

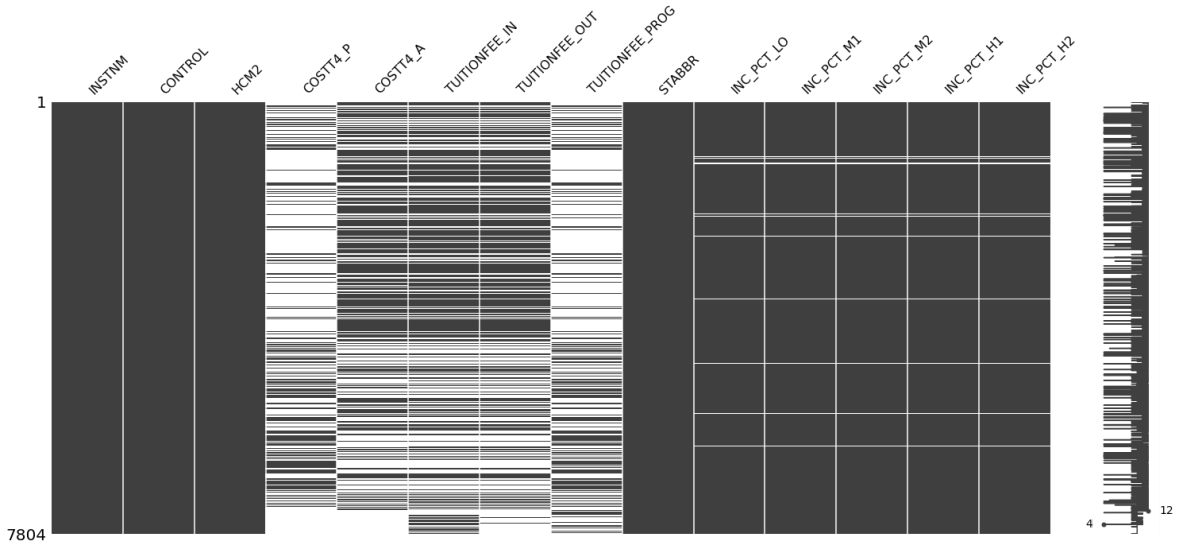
INC_PCT_M2, (48,001-$75,000)

INC_PCT_H1, (75,001-$110,000)

INC_PCT_H2, ($110,001+)

De estas variables se puede agrupar a las instituciones universitarias por el perfil de estudiantes que tienen y sus costos

Se obtiene la siguiente matriz de datos:



En donde blanco=Na, negro= hay dato

2.1.1.1 Metodología

Para realizar la agrupación se utilizaron dos métodos diferentes, la primera fue imputar los datos faltantes como la media de toda la columna y hacer un análisis de componentes principales y por otro lado se usó la función iteractiveimputer() de sklearn con la que se busca hacer imputación utilizando métodos estadísticos más robustos, claramente estos métodos usan medidas estandarizadas para tener una equivalencia entre variables.

2.1.1.2 Agrupación con datos imputados por la media Y PCA

Con los datos imputados por la media obtenemos la siguiente matriz de correlaciones:

	CONTROL	HCM2	ADM_RATE_ALL	COSTT4_P	COSTT4_A	TUITIONFEE_IN	TUITIONFEE_OUT	TUITIONFEE_PROG
CONTROL	1.000000	0.036080	0.201811	0.227326	0.422923	0.426657	0.205345	0.383112
HCM2	0.036080	1.000000	-0.008993	-0.036198	-0.023897	-0.026675	-0.046669	-0.023083
ADM_RATE_ALL	0.201811	-0.008993	1.000000	0.188117	-0.294142	-0.276167	-0.331703	0.163306
COSTT4_P	0.227326	-0.036198	0.188117	1.000000	nan	nan	nan	0.526439
COSTT4_A	0.422923	-0.023897	-0.294142	nan	1.000000	0.962917	0.902968	nan
TUITIONFEE_IN	0.426657	-0.026675	-0.276167	nan	0.962917	1.000000	0.921244	nan
TUITIONFEE_OUT	0.205345	-0.046669	-0.331703	nan	0.902968	0.921244	1.000000	nan
TUITIONFEE_PROG	0.383112	-0.023083	0.163306	0.526439	nan	nan	nan	1.000000

De la cual se puede interpretar que existe una gran correlación entre todas las variables que hablan de costos mientras que para la variable COSTT4_* se tiene que entre la clase P y A no se encuentra correlación por lo que sí se pueden clasificar como variables diferentes e

incorrelacionadas. Por otra parte, las variables TUITTIONFEE_IN y TUITTIONFEE_OUT presentan una gran correlación por lo que se elimina la que tenga menos información que es TUITTIONFEE_OUT

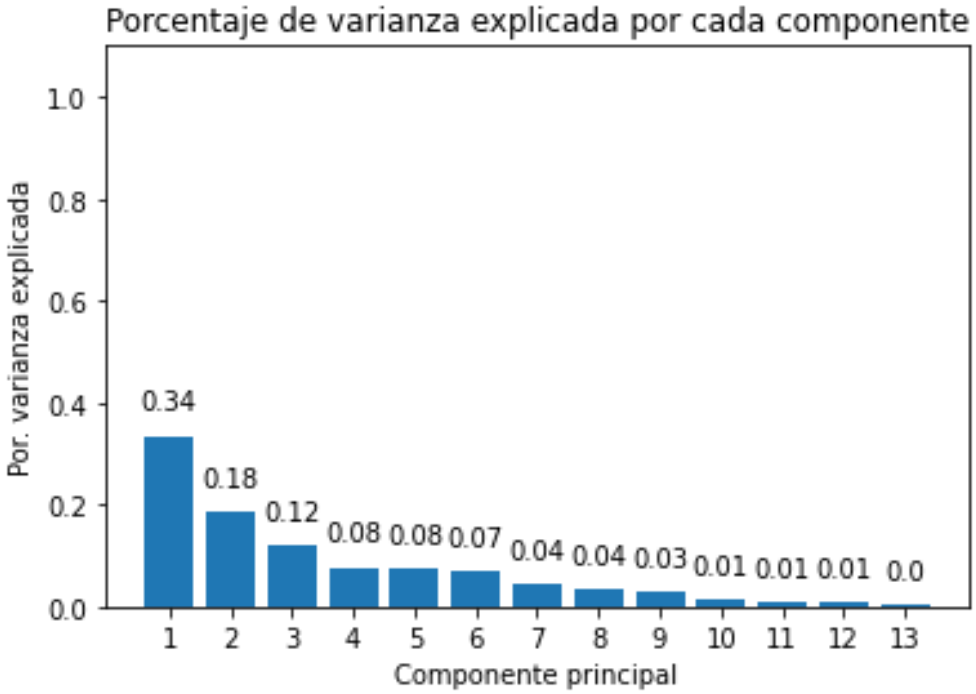
2.1.1.3 Resultados

Cuando se hace el análisis de componentes principales se obtienen los siguientes resultados:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
CONTROL	-0.034787	0.462428	0.145432	0.139237	-0.281206	0.186764	-0.630860	-0.421481	-0.165934	0.040407	0.008056	0.147669	-0.061535
HCM2	-0.023799	0.020314	-0.052340	0.953007	0.186085	-0.223812	0.053611	0.014685	0.006893	-0.012061	0.006951	0.002802	0.000055
ADM_RATE_ALL	-0.125452	-0.058150	0.137959	0.125560	-0.902889	-0.162202	0.290361	0.137560	-0.009761	0.019072	0.027559	-0.004565	-0.009637
COSTT4_P	0.010335	0.064422	0.666329	-0.013731	0.168809	0.010460	0.487222	-0.533148	-0.010918	0.034428	0.002464	0.003284	-0.000676
COSTT4_A	0.363902	0.368320	-0.090060	0.031475	-0.047179	0.143643	0.151654	0.058229	0.074559	-0.022700	-0.096089	-0.598921	-0.547313
TUITIONFEE_IN	0.369222	0.368683	-0.086746	0.034129	-0.068555	0.139938	0.138400	0.043565	0.087132	-0.035878	-0.053256	-0.163469	0.797048
TUITIONFEE_OUT	0.386791	0.302866	-0.092516	-0.007749	0.007025	0.071482	0.255381	0.166665	0.105529	-0.036752	0.127841	0.750299	-0.245174
TUITIONFEE_PROG	0.024261	0.051191	0.682751	0.053662	0.090401	0.128172	-0.270727	0.638080	0.146545	-0.016161	0.002149	-0.028938	0.011959
INC_PCT_LO	-0.394592	0.285704	-0.062560	-0.043510	0.052686	0.007019	0.067467	0.000673	0.278967	-0.378645	0.716742	-0.120016	0.012972
INC_PCT_M1	0.079309	-0.362502	-0.028624	0.199216	-0.065096	0.790163	0.103979	0.004146	-0.325304	-0.158852	0.226379	-0.008081	-0.001436
INC_PCT_M2	0.306322	-0.363271	0.007380	0.072483	-0.118456	0.092665	-0.195785	-0.239567	0.714589	0.300229	0.221524	-0.024405	-0.010340
INC_PCT_H1	0.380315	-0.245496	0.104606	-0.022601	-0.077010	-0.259312	-0.182220	-0.139450	0.002877	-0.806391	-0.085666	0.004398	-0.013535
INC_PCT_H2	0.403692	-0.083744	0.083719	-0.063213	0.021622	-0.355835	-0.099403	0.043578	-0.481820	0.290256	0.591038	-0.117709	0.023849

TUITTIONFEE_OUT	PC1	PC2	PC3
Con	COSTT4_A TUITTIONFEE_IN TUITTIONFEE_OUT INC_PCT_LO INC_PCT_H1 INC_PCT_H2	CONTROL COSTT4_A TUITTIONFEE_IN INC_PCT_M1 INC_PCT_M2	COSTT4_P TUITTIONFEE_PROGRAM
Sin	COSTT4_A TUITTIONFEE_OUT INC_PCT_M2 INC_PCT_H2 INC_PCT_H1 INC_PCT_LO	CONTROL COSTT4_A TUITTIONFEE_OUT INC_PCT_M1	COSTT4_P TUITTIONFEE_PROG

2.1.1.4 Conclusión de la agrupación por medias



Con las tres componentes principales conformadas se obtiene que las agrupaciones estiman el 64% de la varianza, lo cual no es muy eficiente. Pero si se fuera a dejar está agrupación se tendría tres agrupaciones con una base en los costos de las variables valor de universidad(TUITTIONFEE_*), los ingresos de la familia(INC_PCT_*) y en menor medida para la tercera componente el tipo de estructura de gobierno que tiene la institución(CONTROL).

2.1.14 Agrupación con datos imputados por ITERATIVEIMPUTER() Y Cluster

Las únicas variables que no tienen datos faltantes son, INSTNM, CONTROL, HCM2 y STBBR, pero cuando se observan las variables INC_PCT_* notamos que tienen una gran cantidad de filas únicamente con el valor “PrivacySupressed”. Se asume esta notación como un valor faltante.

Se rellenará los datos faltantes de las variables INC_PCT_* usando multiple imputation y basándose en la variable CONTROL.

```
#   Column           Non-Null Count
---  ------           --------------
0   INSTNM           7804 non-null
1   CONTROL          7804 non-null
2   HCM2             7804 non-null
3   COSTT4_P         2541 non-null
4   COSTT4_A         4137 non-null
5   TUITIONFEE_IN    4415 non-null
6   TUITIONFEE_OUT   4196 non-null
7   TUITIONFEE_PROG  2712 non-null
8   STABBR           7804 non-null
9   INC_PCT_LO       7804 non-null
10  INC_PCT_M1       7804 non-null
11  INC_PCT_M2       7804 non-null
12  INC_PCT_H1       7804 non-null
13  INC_PCT_H2       7804 non-null
```

Con las columnas INC_PCT_* completas y sin ningún dato faltante, se puede pasar a imputar COSTT4_* y TUITTIONFEE_*.

Para la variable COSTT4_* leyendo la documentación se observa que aunque parezca que tienen muchos datos nulos, son complementarias, pues cada una se corresponde al costo de matrícula para un tipo específico de universidad entre dos posibles, es decir que para el manejo de datos faltantes en estas dos variables se combinan en una de las columnas los datos de ambas.

Entonces, cambiando los valores faltantes de uno por los complementarios del otro se obtiene que

```
COSTT4_A         6678 non-null
```

Respecto a las variables TUITIONFEE_* dado que TUITIONFEE_IN y TUITIONFEE_OUT son variables que nos hablan del costo de matrícula para estudiantes del mismo estado y fuera el estado, podemos tomar el TUITIONFEE_PROG como un valor para llenar los elementos faltantes de TUITIONFEE_IN tal como en el caso anterior.

Datos a usar:

Dado el análisis realizado de la documentación y de las PCA, decidimos usar en un principio las siguientes variables:

INSTNM; Nombre de la institución

CONTROL; Identifica si la estructura de gobierno de la institución es:

pública (1),

privada sin ánimo de lucro (2)

privada con ánimo de lucro (3).

HCM2; Es el tipo de HCM que indica problemas financieros o de cumplimiento federal más graves, HCM2 indica un sistema de supervisión mayor dadas irregularidades en la institución y HCM un nivel de monitoreo menor.

(1) tiene problemas (mayor monitoreo)
(0) No tiene problemas, menor monitoreo del dinero.

COSTT4_P; Es el costo de asistencia tomado del programa académico más cursado durante el año.

COSTT4_A; Costo de asistencia anual general para todos los estudiantes.

TUITIONFEE_IN; Costo y tasas de matrícula para estudiantes de dentro del estado

TUITIONFEE_OUT; Costo y tasas de matrícula para estudiantes de fuera del estado.

TUITIONFEE_PROG; Costo y tasas de matrícula general, tanto si es fuera o dentro del estado, esto son aquellas universidades que no cobran alguna diferencia por esto.

STABBR; Estado (ubicación)

Porcentaje de estudiantes dados los ingresos familiares.

INC_PCT_LO, (0-$30,000)

INC_PCT_M1, (30,001-$48,000)

INC_PCT_M2, (48,001-$75,000)

INC_PCT_H1, (75,001-$110,000)

INC_PCT_H2, ($110,001+)

2.2 Manejo de los datos faltantes

Para poder realizar nuestro análisis adecuadamente debemos verificar y arreglar si es el caso los datos faltantes.

En nuestro proceso nos encontramos con 3 conjuntos de variables relacionadas a los cuales les faltaban valores y estos eran posibles completar. estos serían: **INC_PCT_*, COSTT4_*, TUITIONFEE_*.**

2.2.1 Para INC_PCT_*

Para estas variables llenaremos los datos faltantes usando multiple imputation y basándose en la variable CONTROL, además aprovechamos y cambiamos el formato.

Antes

```
 9   INC_PCT_LO       5031 non-null   object
 10  INC_PCT_M1       4343 non-null   object
 11  INC_PCT_M2       3927 non-null   object
 12  INC_PCT_H1       3589 non-null   object
 13  INC_PCT_H2       3963 non-null   object
dtypes: float64(5), int64(2), object(7)
memory usage: 853.7+ KB
```

Después:

```
 9   INC_PCT_LO       7804 non-null   float64
 10  INC_PCT_M1       7804 non-null   float64
 11  INC_PCT_M2       7804 non-null   float64
 12  INC_PCT_H1       7804 non-null   float64
 13  INC_PCT_H2       7804 non-null   float64
dtypes: float64(10), int64(2), object(2)
```

2.2.2 Para COSTT4_*

Dado que cuando se tenía un costo anual general para los estudiantes no se suele tener un costo por programa anual de los mismos, pero si se tiene cuando falta, podemos decir que estas 2 variables se auto complementan, por eso decidimos combinarlas para reducir la cantidad de nulos. (Esta combinación la hacemos en la variable **COSTT4_A**).

Antes:

```
 3   COSTT4_P         2541 non-null   float64
 4   COSTT4_A         4137 non-null   float64
```

Después:

```
 3   COSTT4_P         2541 non-null   float64
 4   COSTT4_A         6678 non-null   float64
```

2.2.3 Para TUITIONFEE_*

Para este caso hacemos algo similar al anterior, pero esta vez combinamos en **TUITIONFEE_IN** y **TUITIONFEE_OUT** la variable **TUITION FEE PROG.**

Antes

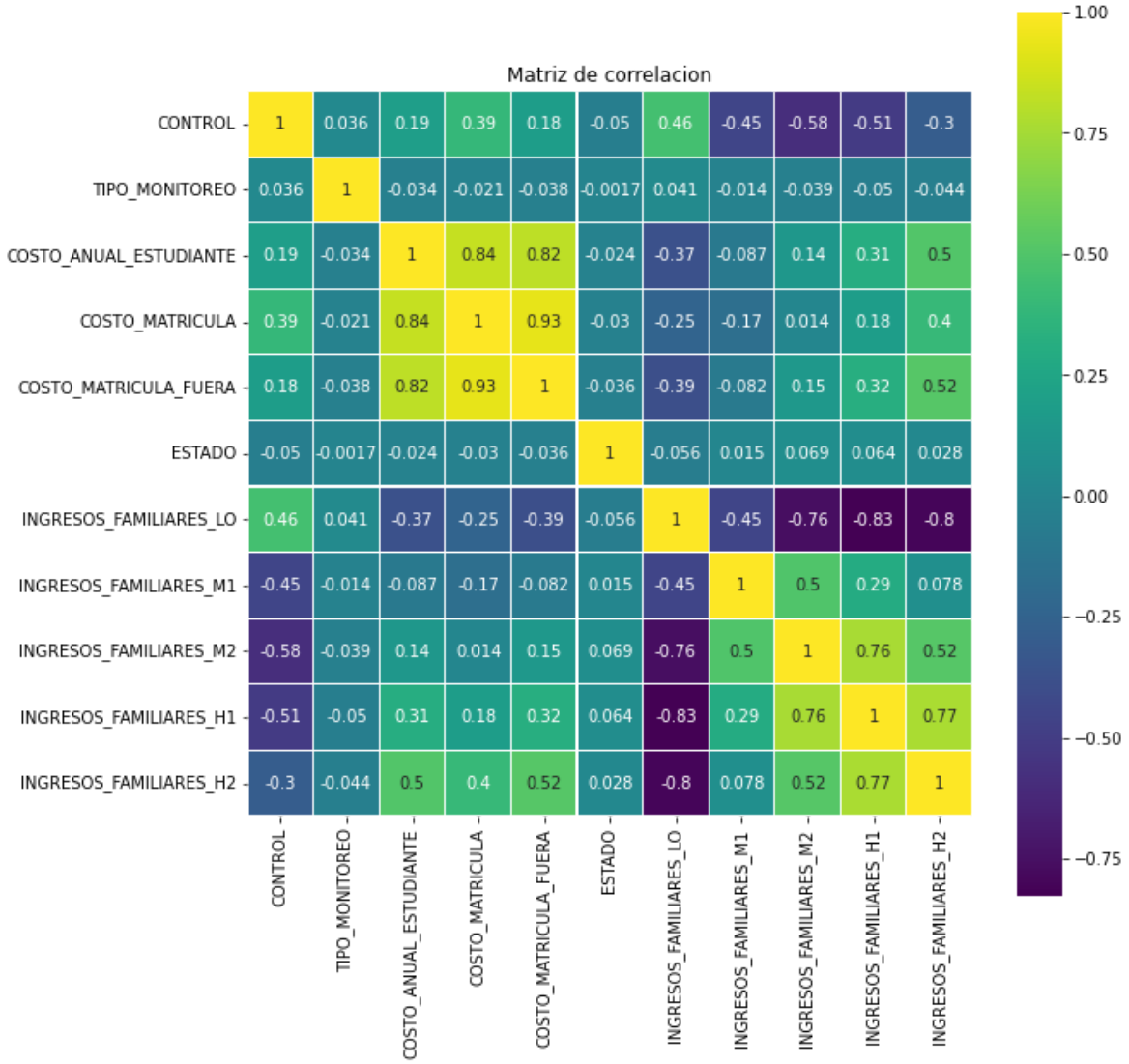
```
 5   TUITIONFEE_IN    4415 non-null   float64
 6   TUITIONFEE_OUT   4196 non-null   float64
 7   TUITIONFEE_PROG  2712 non-null   float64
```

Después

```
5   TUITIONFEE_IN    7127 non-null   float64
6   TUITIONFEE_OUT   6908 non-null   float64
7   TUITIONFEE_PROG  2712 non-null   float64
```

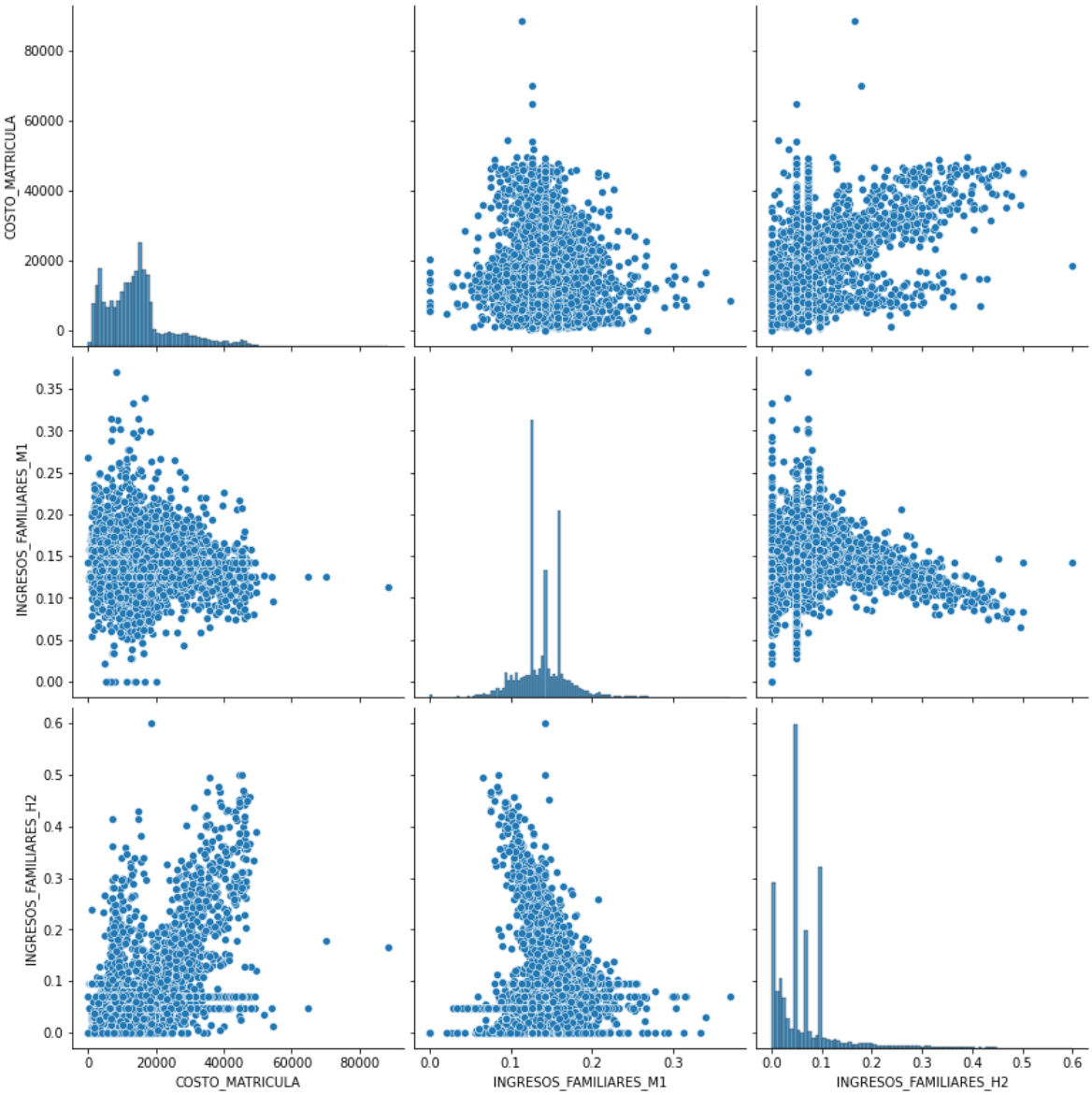
2.3 Análisis de nuestras variables

Antes de continuar analizamos cómo se correlacionan estas variables, para así poder descartar aquellas que no nos sirvan. Además hacemos unos cambios en el nombre de las variables para poder entenderlas más fácilmente.



Descartamos aquellas variables con mucha correlación y dejamos las que consideramos nos pueden ayudar más en nuestro análisis.


Matriz de correlacion
CONTROL
TIPO_MONITOREO
COSTO_MATRICULA
ESTADO
INGRESOS_FAMILIARES_M1
INGRESOS_FAMILIARES_H2
1
0.036
0.39
-0.05
-0.45
-0.3
0.036
1
-0.021
-0.0017
-0.014
-0.044
0.39
-0.021
1
-0.03
-0.17
0.4
-0.05
-0.0017
-0.03
1
0.015
0.028
-0.45
-0.014
-0.17
0.015
1
0.078
-0.3
-0.044
0.4
0.028
0.078
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

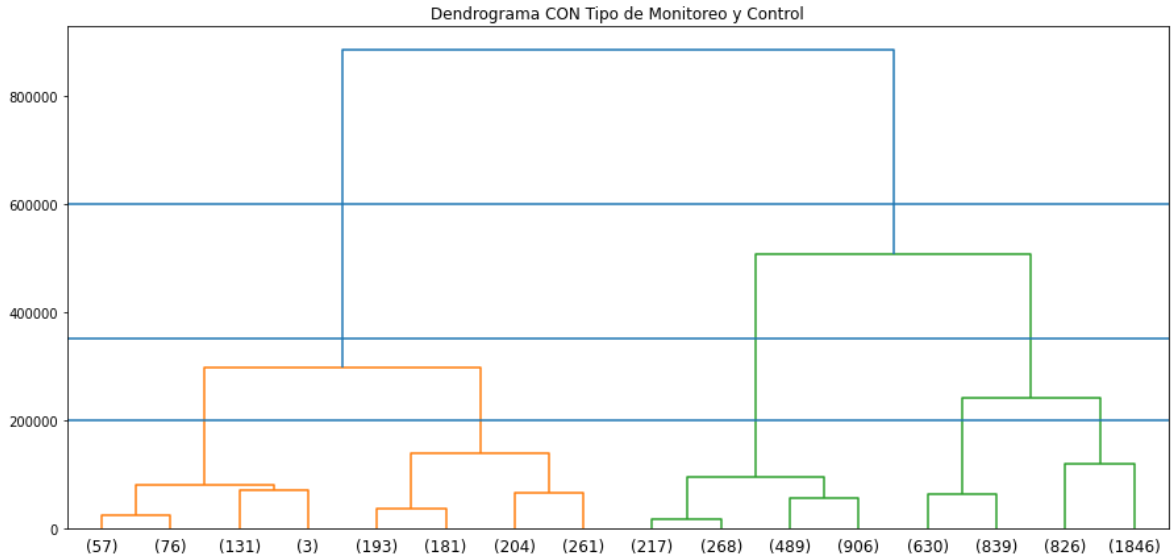


2.4 Cantidad óptima de clusters

Encontramos que el número óptimo de clusters para nuestro conjunto de datos es de 3, este valor se encontró mediante los siguientes análisis.

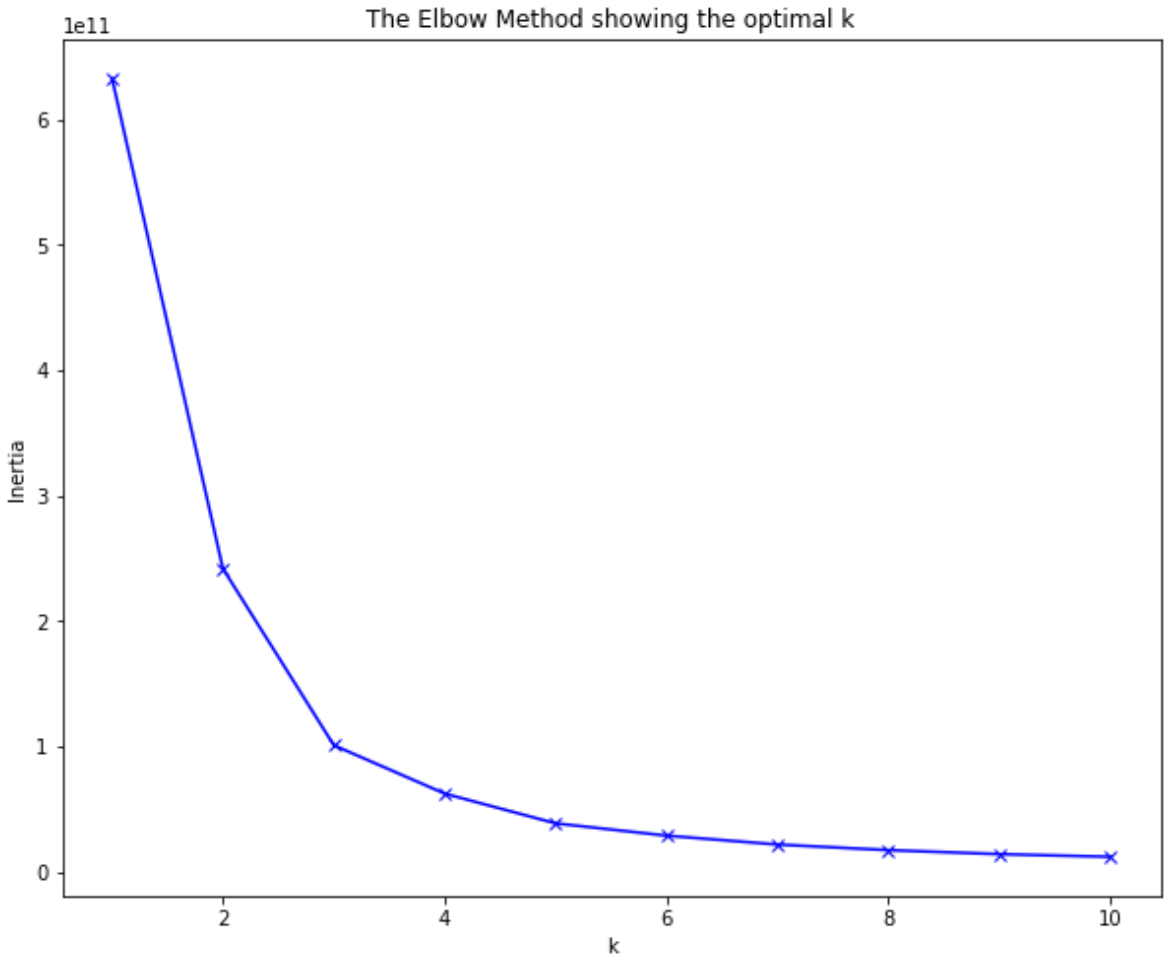
2.4.1 Dendograma

Mediante este método encontramos que el número óptimo de clusters podría ser de 2, 3 y hasta 4 grupos.



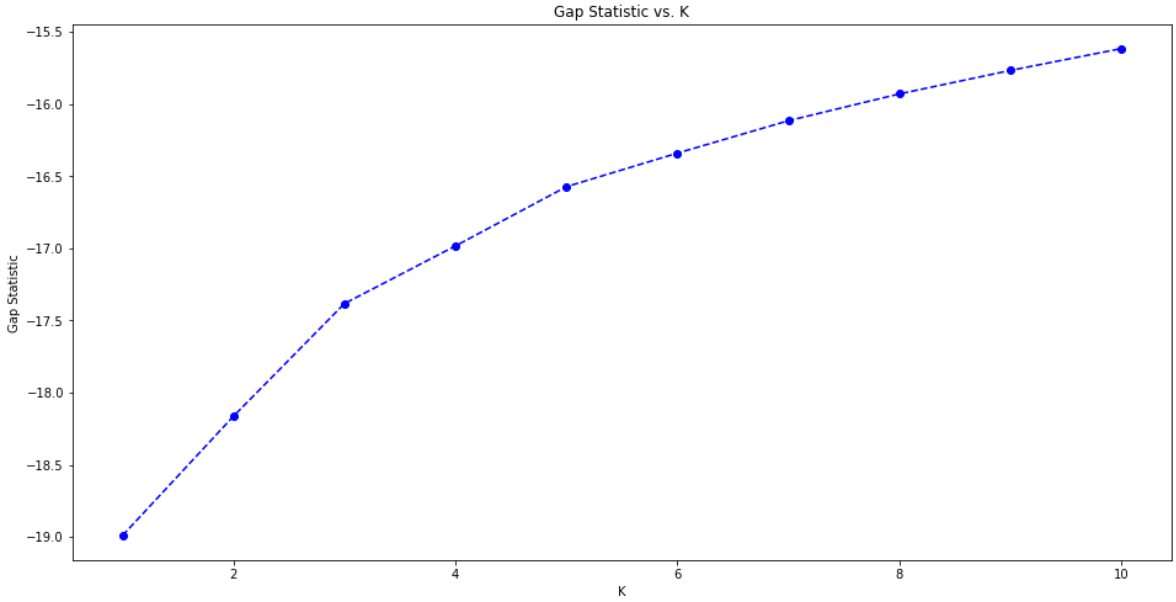
2.4.2 Elbow Curve

De este método podemos concluir que el número óptimo de clusters está entre 3 o 4 clusters.



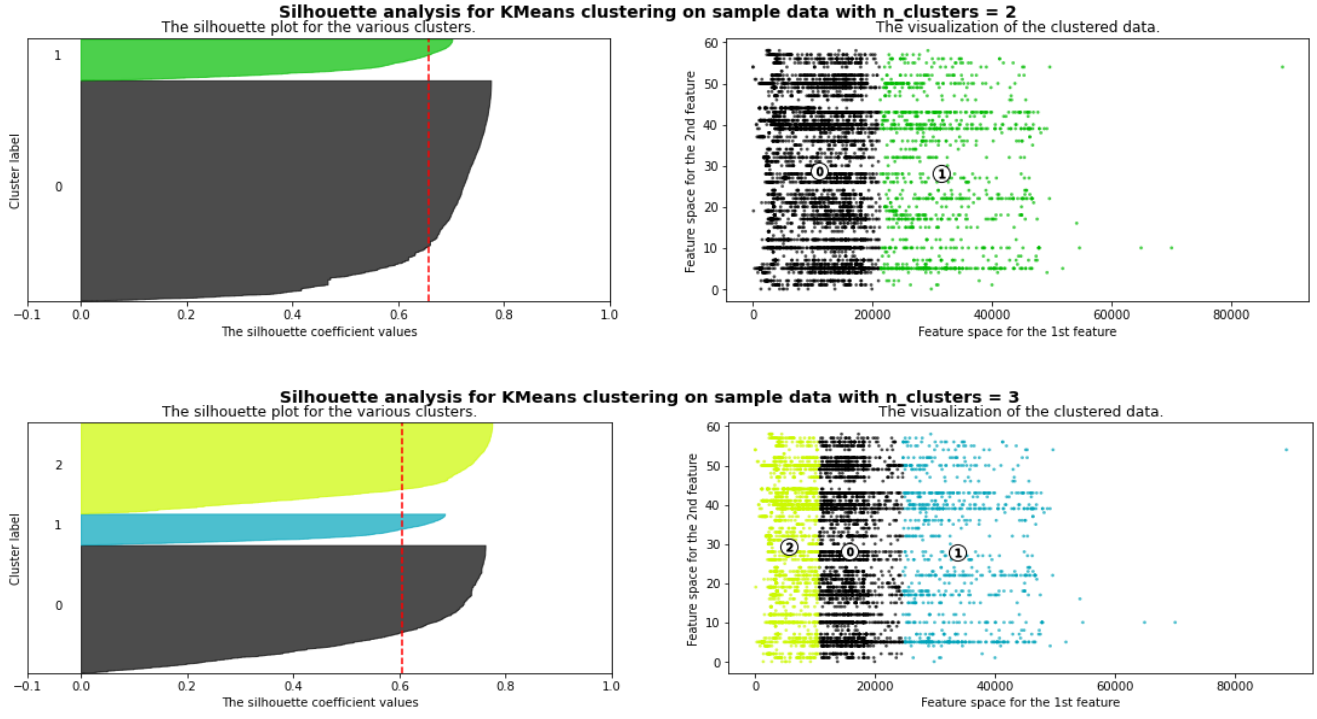
2.4.3 Estadístico de Gap

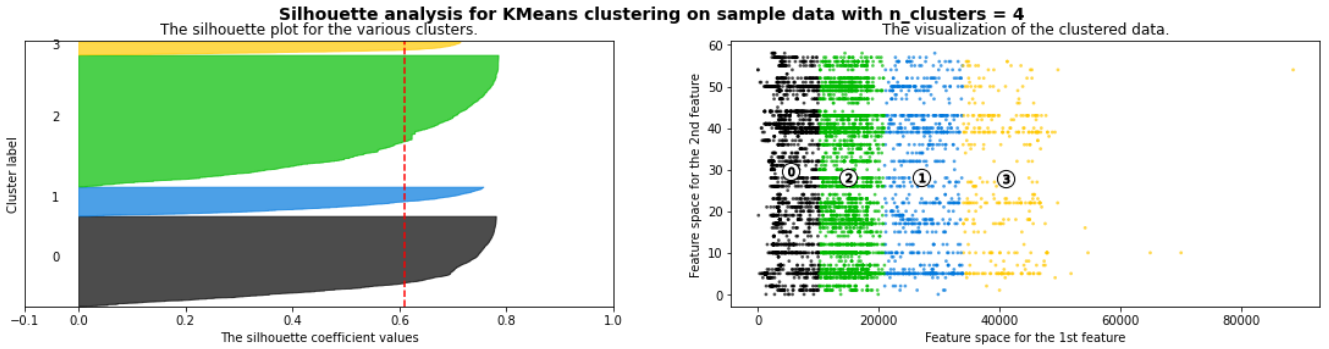
Para este caso el método no nos arroja información clara sobre el número óptimo de clusters, se podría decir que en 3 clusters hay un pequeño cambio que podría ser tomado en cuenta, aun así no lo consideramos como una señal clara.

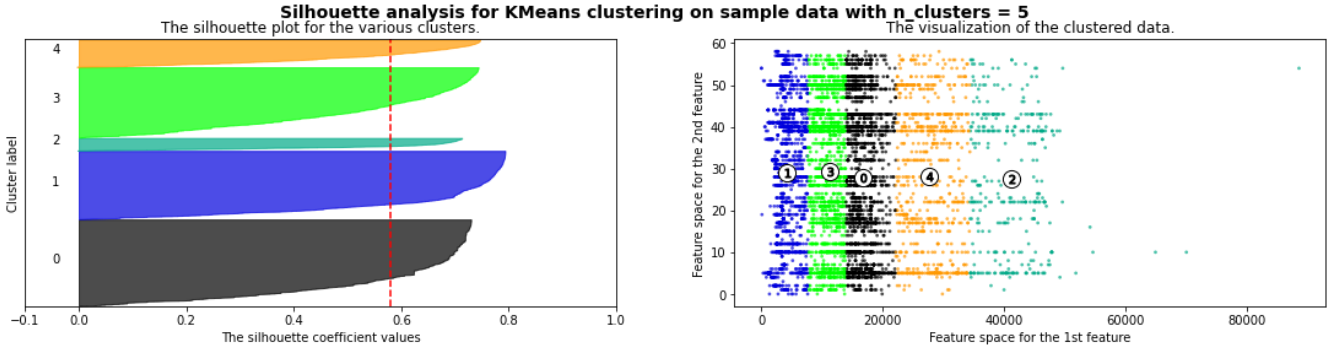


2.4.4 Coeficiente de Silueta

Para este método vemos que 3 es el número óptimo de clusters.

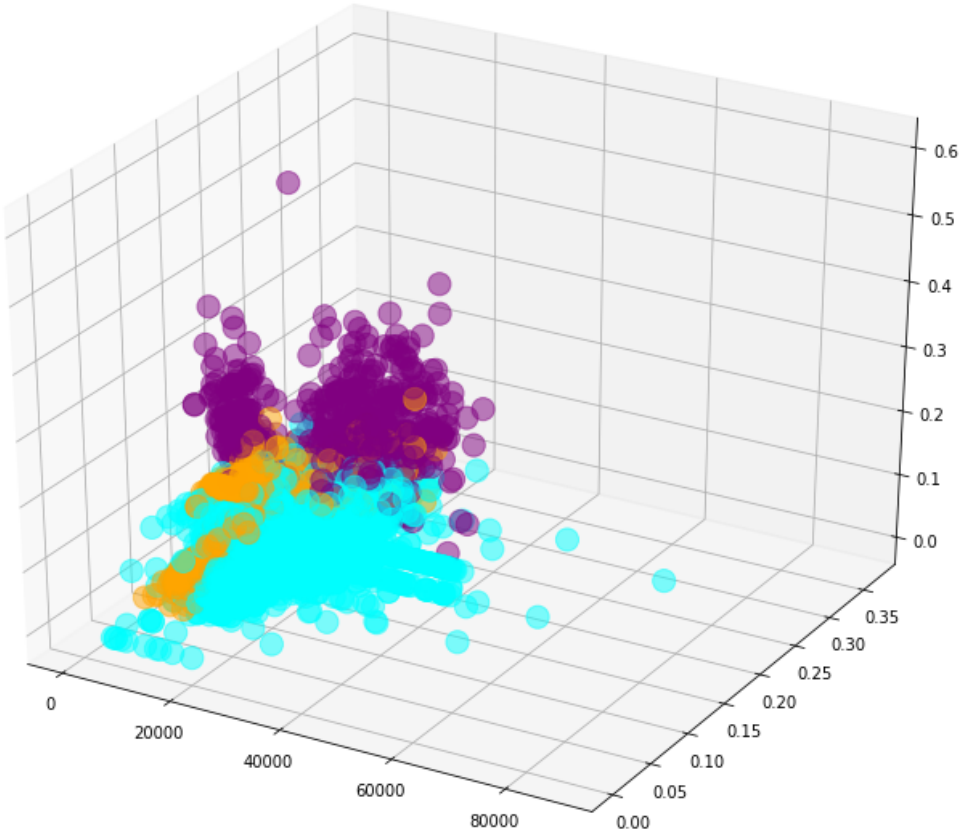


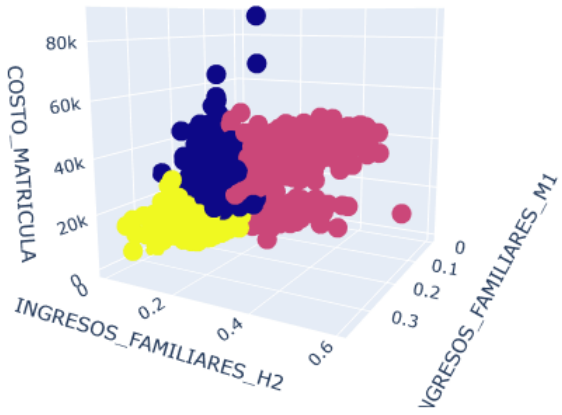


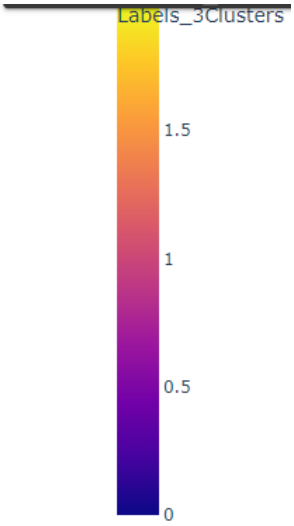


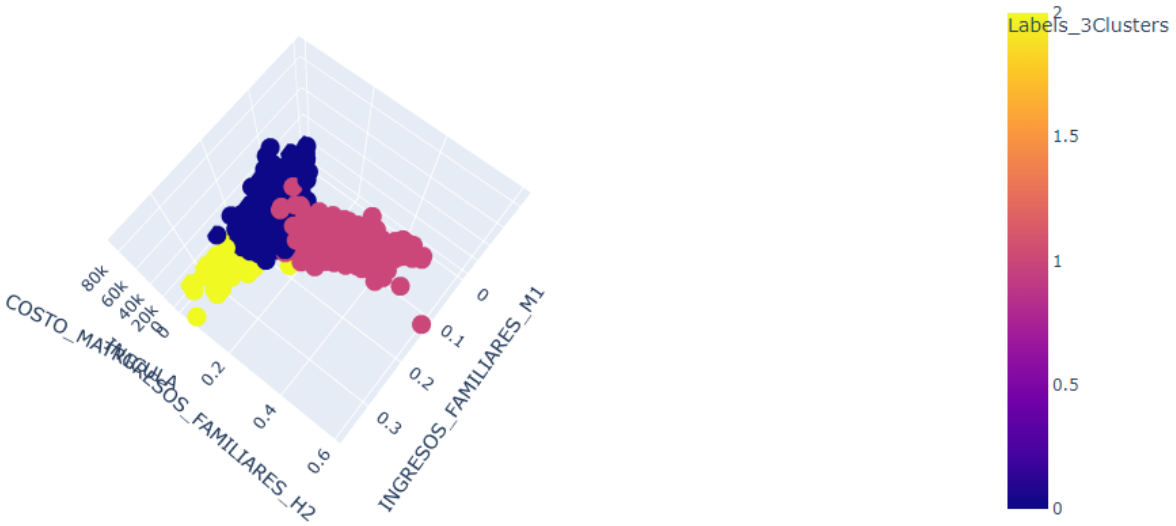
2.5 Clustering - Jerárquico

Ya teniendo el número de clusters más óptimo, realizamos el agrupamiento.


0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20000
40000
60000
80000

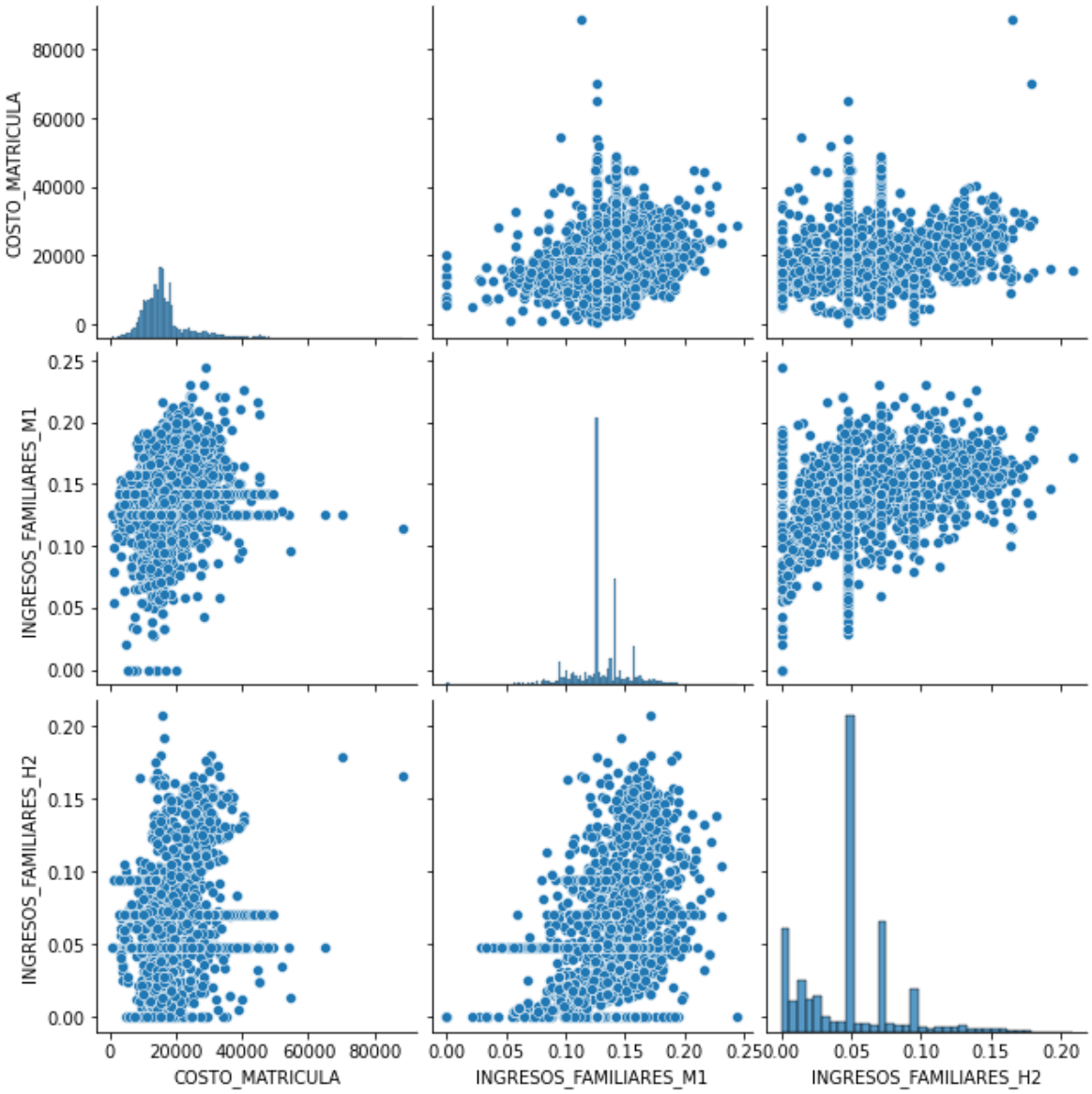

80k
60k
40k
20k
COSTO_MATRICULA
0
0.2
0.4
0.6
INGRESOS_FAMILIARES_H2
0
0.1
0.2
0.3
VGRESOS_FAMILIARES_M1


Labels_3Clusters
2
1.5
1
0.5
0



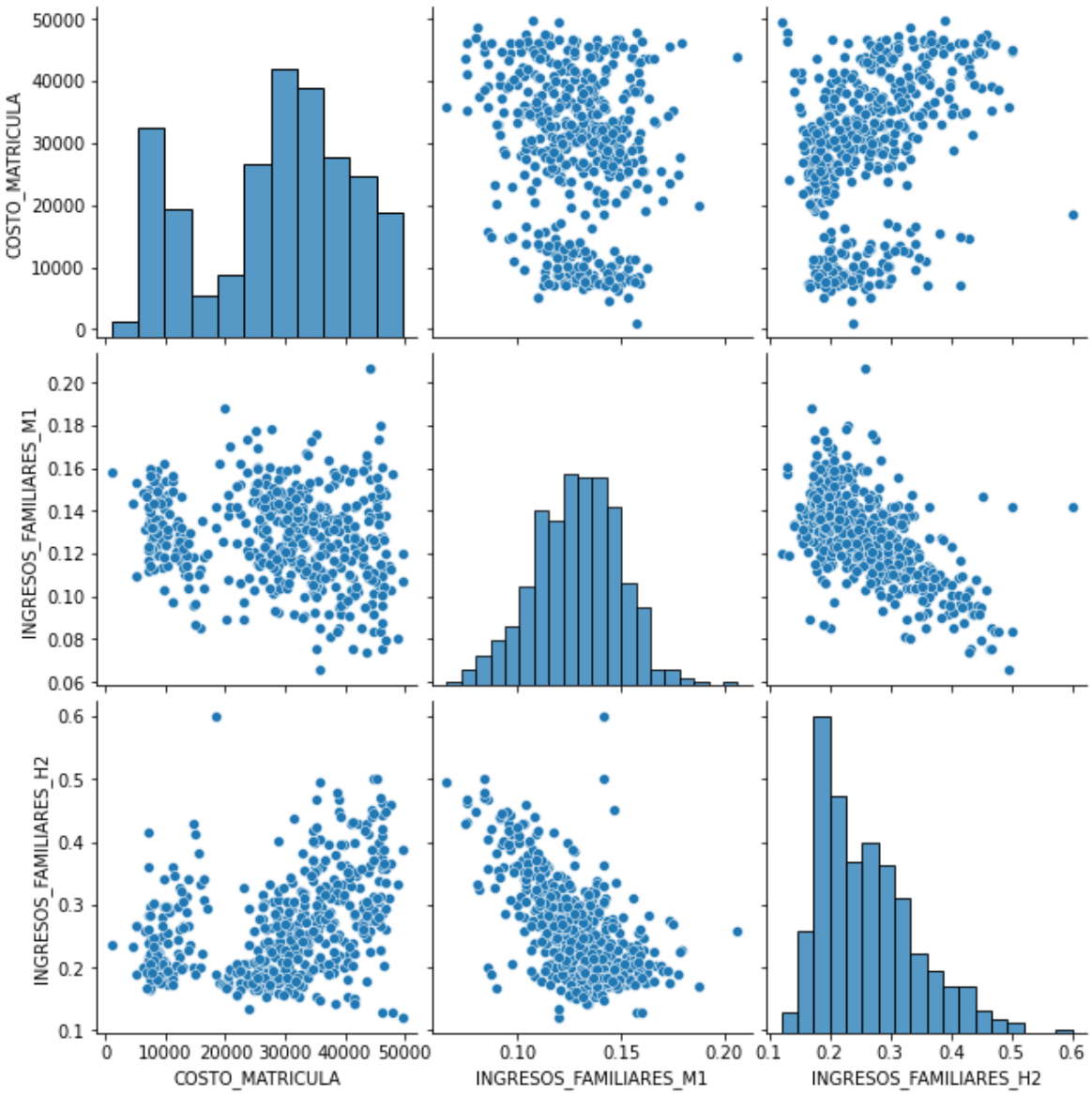
3. Caracterización

Grupo 0:


COSTO_MATRICULA
INGRESOS_FAMILIARES_M1
INGRESOS_FAMILIARES_H2
0
20000
40000
60000
80000
0.00
0.05
0.10
0.15
0.20
0.25

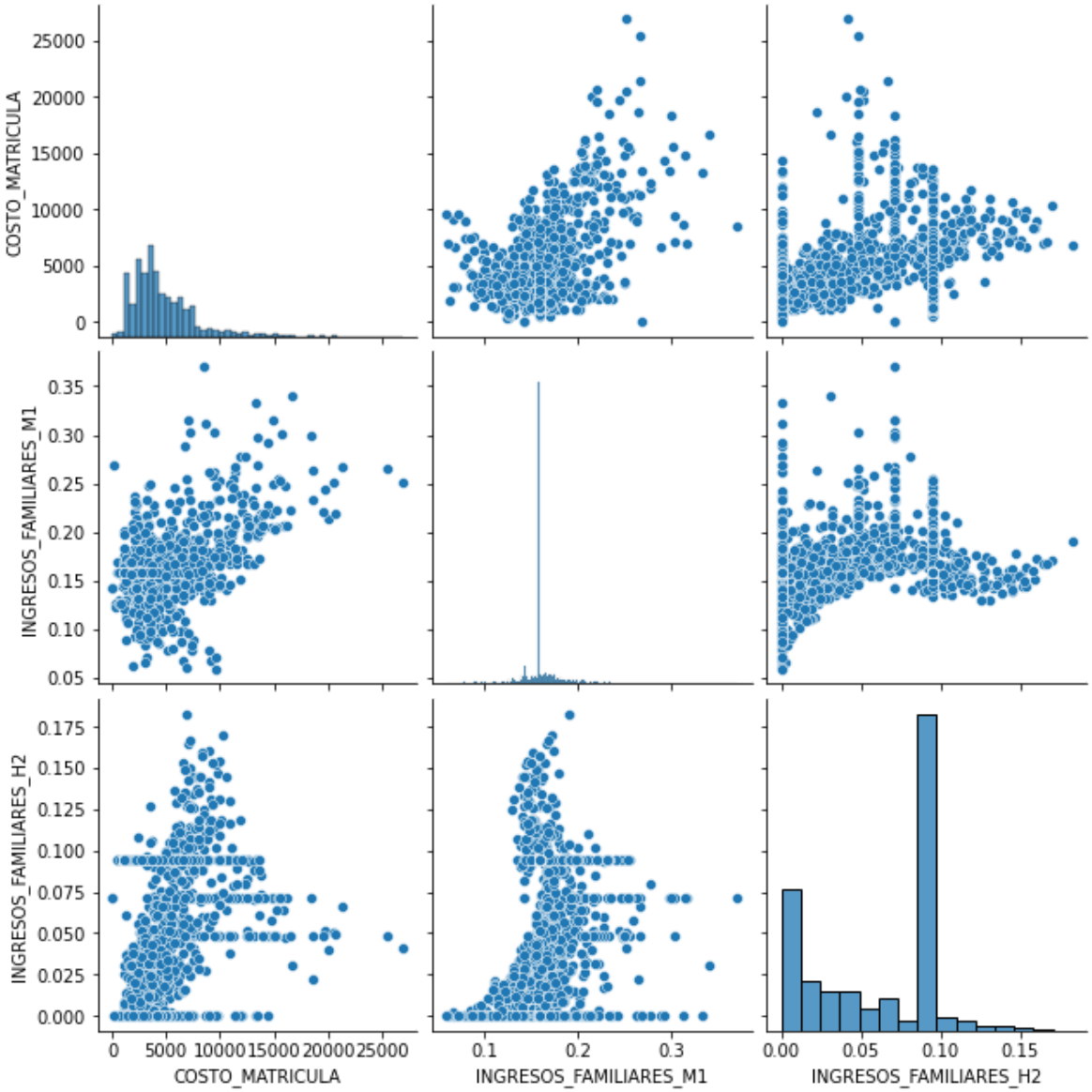
	COSTO_MATRICULA	INGRESOS_FAMILIARES_M1	INGRESOS_FAMILIARES_H2	Labels_3Clusters
count	4836.000000	4836.000000	4836.000000	4836.0
mean	16349.470017	0.130382	0.047558	0.0
std	7281.412394	0.025572	0.032905	0.0
min	501.000000	0.000000	0.000000	0.0
25%	11964.000000	0.125536	0.021767	0.0
50%	15227.500000	0.125962	0.047940	0.0
75%	18048.000000	0.141866	0.071160	0.0
max	88550.000000	0.244444	0.208247	0.0

Grupo 1:


COSTO_MATRICULA
INGRESOS_FAMILIARES_M1
INGRESOS_FAMILIARES_H2
0
10000
20000
30000
40000
50000
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.1
0.2
0.3
0.4
0.5
0.6
0.15

	COSTO_MATRICULA	INGRESOS_FAMILIARES_M1	INGRESOS_FAMILIARES_H2	Labels_3Clusters
count	493.000000	493.000000	493.000000	493.0
mean	28450.799189	0.128369	0.261723	1.0
std	12379.088393	0.020925	0.079477	0.0
min	1032.000000	0.065450	0.120000	1.0
25%	19790.000000	0.114094	0.197253	1.0
50%	30470.000000	0.129288	0.246377	1.0
75%	38069.000000	0.142857	0.307886	1.0
max	49630.000000	0.206463	0.600000	1.0

Grupo 2:



	COSTO_MATRICULA	INGRESOS_FAMILIARES_M1	INGRESOS_FAMILIARES_H2	Labels_3Clusters
count	1798.000000	1798.000000	1798.000000	1798.0
mean	4713.516129	0.161703	0.062045	2.0
std	3061.785974	0.029052	0.041032	0.0
min	0.000000	0.059269	0.000000	2.0
25%	2700.000000	0.157250	0.019417	2.0
50%	3922.000000	0.157771	0.091783	2.0
75%	6054.750000	0.165918	0.094380	2.0
max	26900.000000	0.370370	0.182879	2.0

4. Conclusiones del agrupamiento

1. Conclusión

Grupo 0 (Azul): En este grupo se encuentran universidades con un costo promedio de matrícula de (16000 dólares), pero que puede llegar a variar mucho. En este grupo se encuentra un mayor porcentaje de estudiantes M1 en comparación con los de H2.

Grupo 1 (Rosado): En este grupo se encuentran universidades con un costo promedio de matrícula de (28000 dólares), siendo este el grupo con el mayor costo promedio de matrícula y el mayor porcentaje de estudiantes H2.

Grupo 2 (Amarillo): En este grupo se encuentran universidades con un costo promedio de matricula de (4000 dolares), siendo este el grupo con el menor costo promedio de matricula y el mayor porcentajes de estudiantes M1.

- Dado todo el análisis anterior podríamos resumir los 3 grupos en los siguiente:

Grupo 0: Universidades de costo medio y con estudiantes principalmente de clase media

Grupo 1: Universidades de costo alto y estudiantes principalmente de clase alta

Grupo 2: Universidades de costo bajo y estudiantes principalmente de clase media

2. Conclusión

También encontramos que existe una relación entre los grupos que encontramos, con la variable *CONTROL*. Recordemos que esta variable nos identifica si la estructura de gobierno de la institución es:

(1) Pública.

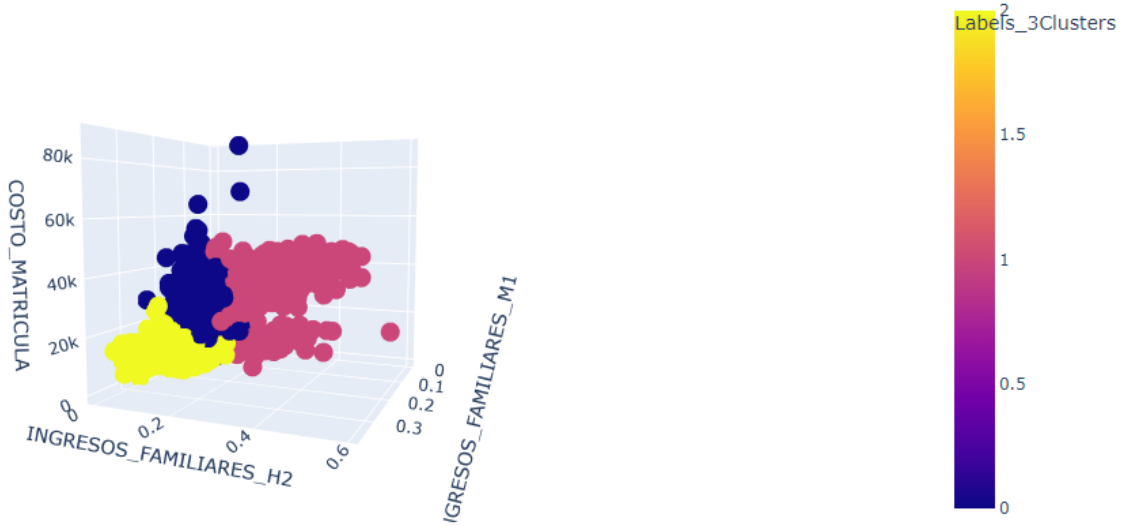
(2) Privada sin ánimo de lucro.

(3) Privada con ánimo de lucro.

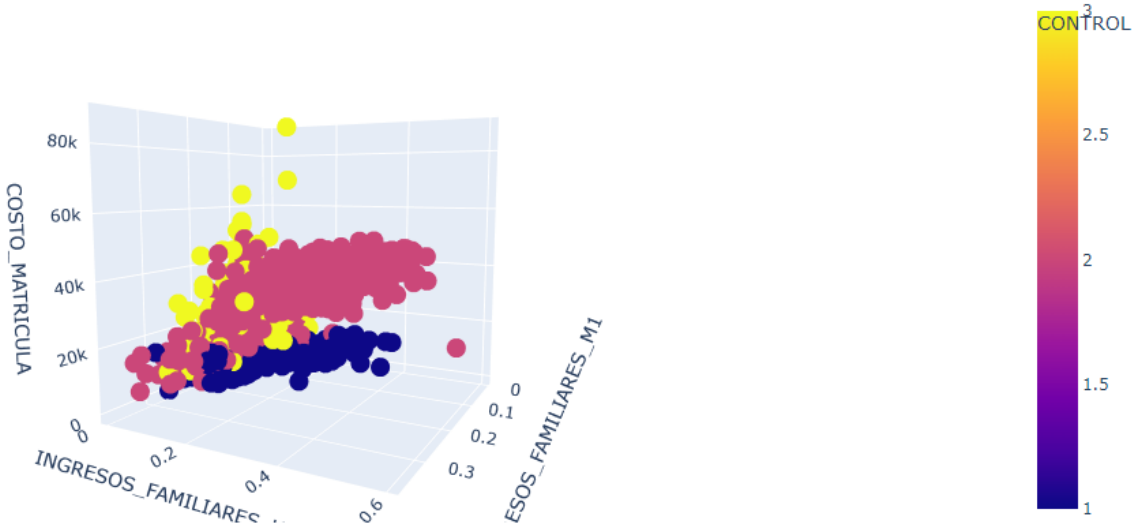
Vemos que nuestro grupo 2 conformado por las universidades con menor costo promedio de matrícula está relacionado con aquellas universidades que son públicas.

Vemos que nuestro grupo 1 conformado por las universidades con mayor costo promedio de matrícula está relacionado con aquellas universidades que son privadas sin ánimo de lucro.

Finalmente también vemos que para nuestro grupo 0 con los costos medios se relaciona con aquellas universidades que son privadas con ánimo de lucro.



Dado el control.



5. Cómo implementar esto en Colombia

Para generar un conjunto de datos que nos permita hacer esto en Colombia necesitamos al menos recopilar variables que funjan como equivalentes a las variables que utilizamos para el contexto colombiano. En general los valores de matrícula y los costos anuales o semestrales están aproximadamente bien cubiertos considerando una clasificación de las universidades por ciudad o

por departamento. Solo quedaría evaluar un equivalente a las variables HCM_PCT_*Para el contexto colombiano el dato más descriptivo y de fácil acceso es el porcentaje de estudiantes por estrato socioeconómico por universidad. En general no es bien conocido el porcentaje actualizado de estudiantes por estrato socioeconómico en todas las universidades del país y se tendría que empezar a hacer encuestas demográficas para obtener esta información. Los demás datos necesarios en general son conocidos o de acceso relativamente fácil, por lo que una vez obtenido el porcentaje de estudiantes por estrato socioeconómico este análisis realizado se puede replicar para Colombia.

Es posible hacer un análisis similar sin clasificar las universidades geográficamente dado que en general en todo el territorio nacional los costos son similares en comparación con las diferencias que hay entre estados de Estados Unidos, de manera que se consideren todas las universidades del país y se haga todo el análisis directamente.

Referencias:

planteamiento para el enfoque de los datos:
https://www.cbsnews.com/news/the-biggest-problems-with-americas-colleges/