Trabajo practico nº3 de Big data acerca de kernels y Cluester

---

Integrantes: Tomas Rodríguez Heredia,Diego Exequiel Coria, Rocco Pimenta
Fecha de entrega: Marte 13 de mayo
Profesora: Romero Noelia

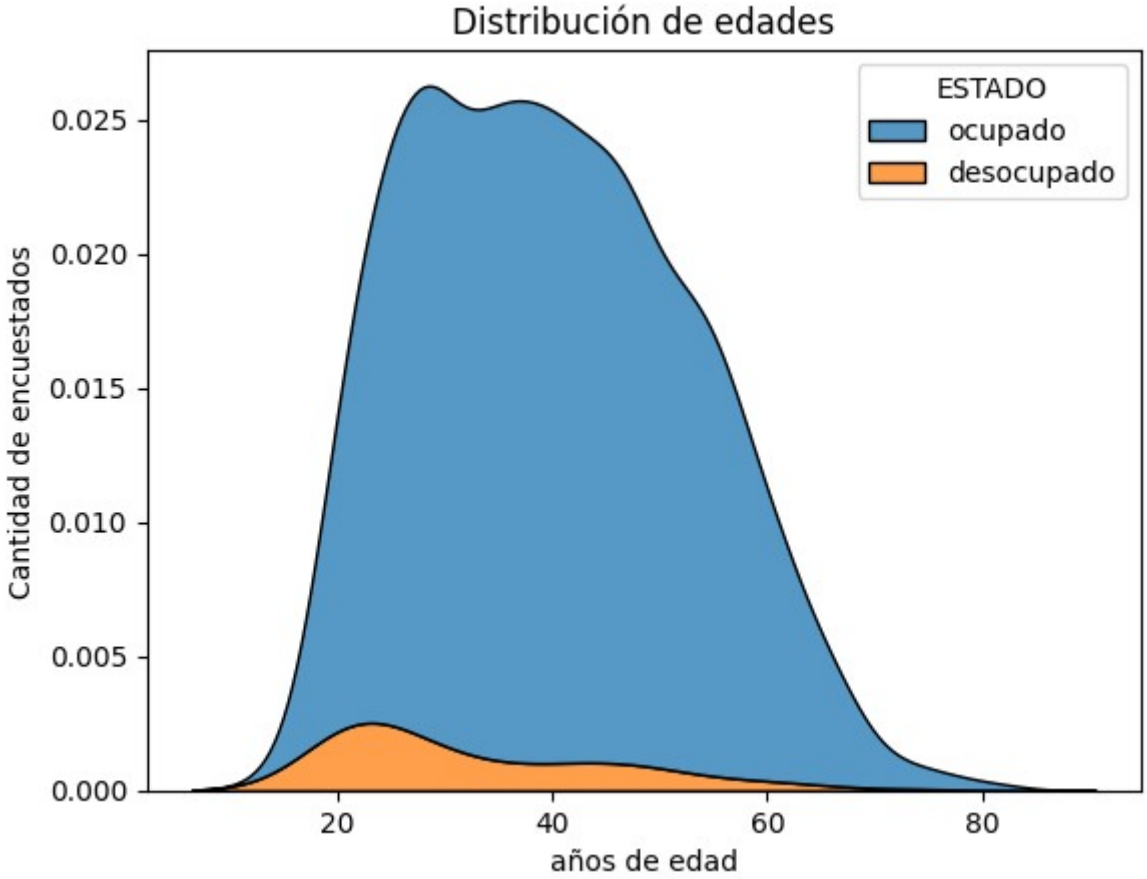
Introducción:

En este tercer trabajo estuvimos estudiando y trabajando la utilización de los kernels y los Clúster en nuestras variables, seguimos usando de base la EPH como indica en el trabajo. Puntualmente utilizamos estos sistemas para representar visualmente nuestras variables y relaciones que queriamos demostrar.

Parte 1:----------

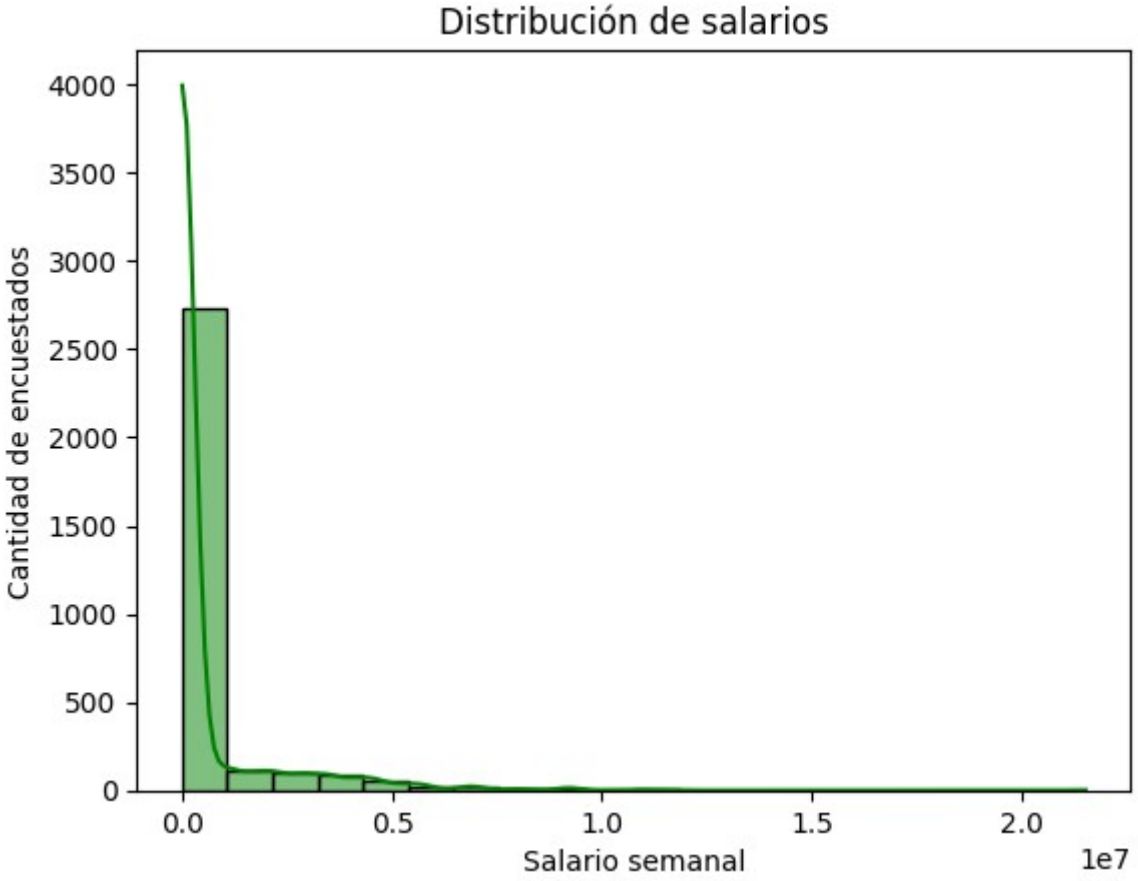
1)

En este punto realizamos gráficos con Matplotlib.pyplot y kernels para medir la cantidad de desocupados y ocupados según la edad. utilizamos la información de CH06 que contiene la edad y “ESTADO” para saber si tienen trabajo.

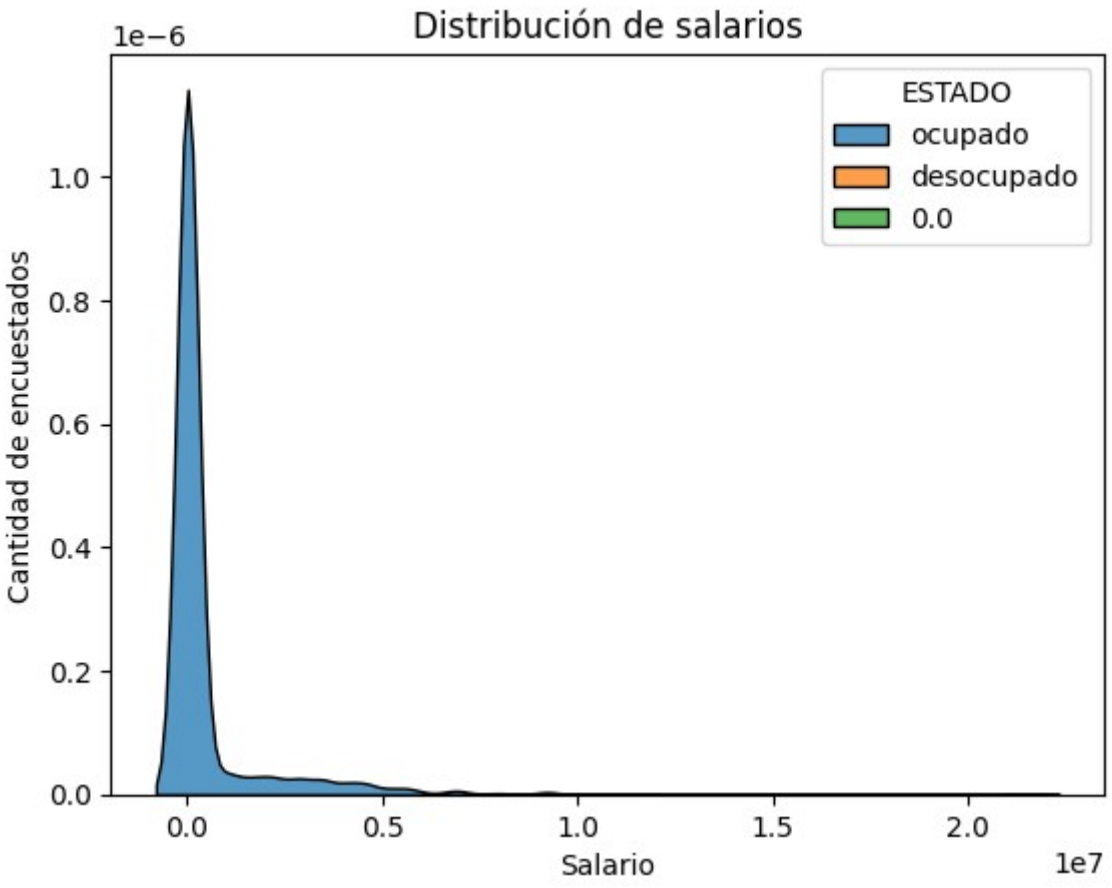


Como conclusión lo que se puede observar en el primer grafico es como mediante el crecimiento de la población estudiada, estos van obteniendo trabajo hasta que llega la edad cercana a la jubilación que es donde caen los datos. Lo que se ve en el segundo grafico es que muchas de las

personas desempleadas rondan entre los 25-30 años de edad y esto va decreciendo según son mas viejos.
2)
para la creación de la variable referida a los años de educación construimos una formula la cual tenía en cuenta los años del ultimo ciclo educativo cursado y en caso de no haber sido terminado le restábamos los años faltantes. Esto teniendo en cuenta que cada periodo, con excepciones, corresponde a aproximadamente 6 años.
3)
Para la creación de la variable de salario_semanal tomamos la columna los datos de los encuestados y los dividimos por 40; en el caso particular del años 2004 también debemos tener en cuenta la inflación acumulada del periodo.



Con los gráficos expuestos podemos notar como la mayoría de los encuestados se concentran dentro de un ingreso menor a al promedio dentro de la Patagonia. Para el mismo tuvimos que descartar los ingresos nulos en el caso particular del histograma.



4)
En forma muy parecida al punto 2 se puede observar que entre la dos ocupaciones, tanto la principal como la secundaria, termina trabajando 8 horas en promedio.

5)

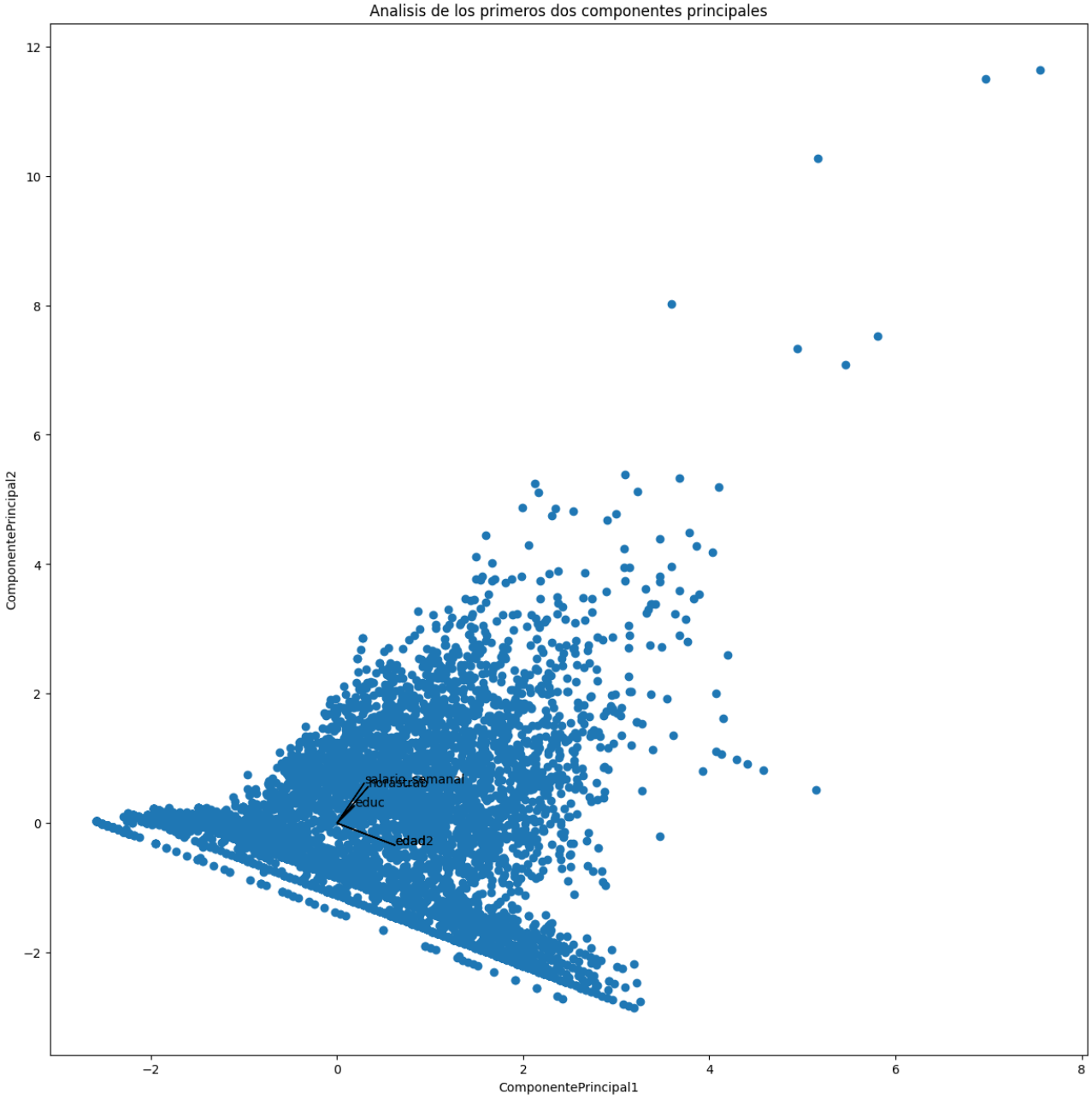
	2004	**2024**	**Total**
Cantidad de observaciones	3264	4513	7777
Cantidad de nas	0	0	0
Cantidad de ocupados	1238	2018	3256
Cantidad de desocupados	600	100	700
Variables limpias y homogenizadas	3264	4513	7777

Parte 2:----------

1) En este punto realizamos una matriz de correlación entre las 5 variables que estuvimos creando. Para eso creamos un .csv que contuviera el dataframe con las variables.

	edad	educ	salario_semanal	horastrab	edad2
edad	1.00	0.12	0.16	0.21	1.00
educ	0.12	1.00	0.12	0.13	0.12
salario_semanal	0.16	0.12	1.00	0.41	0.16
horastrab	0.21	0.13	0.41	1.00	0.21
edad2	1.00	0.12	0.16	0.21	1.00

2) Para esta parte aplicamos PCA en nuestra variables, tuvimos que transformarla utilizando .fit_transform y aplicando PCA conseguimos los “scores” y “loading vectors”. Nos van a servir para visualizar mediante una grafica de dispersión.

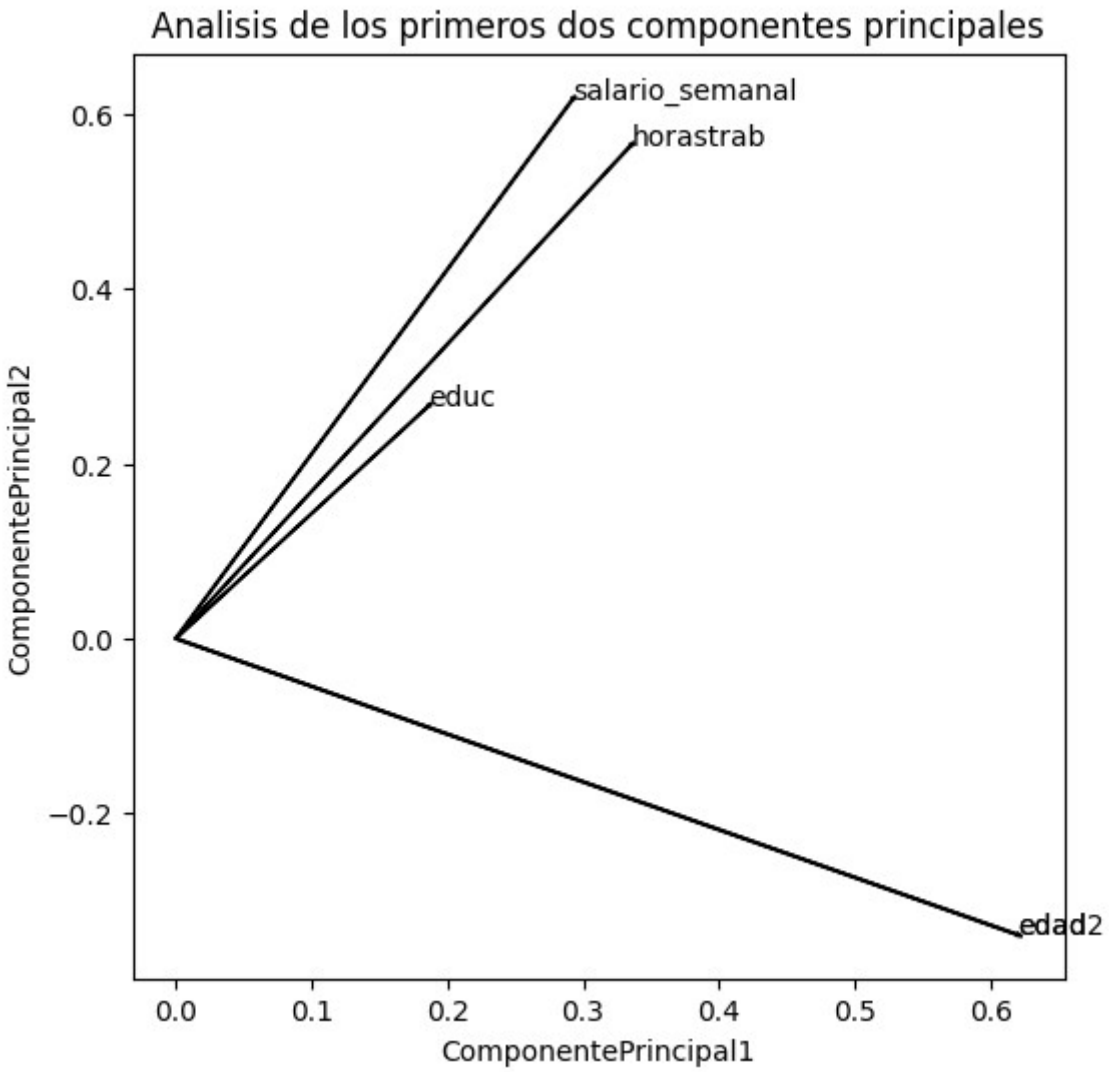


Datos interesante que se pueden ver es como las horas totales del trabajo y el salario semanal casi se superponen, poniendo en claro la relación que tienen estas dos variables.

3)
Construimos el siguiente cuadro con el peso de cada variable:

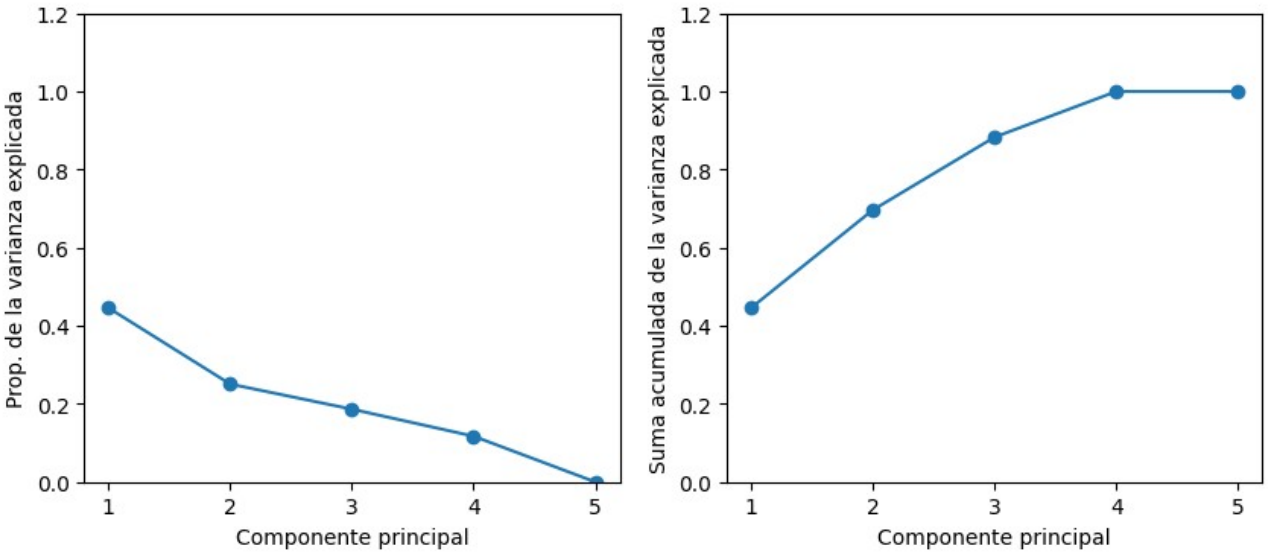
	EDAD	EDUC.	SALARIO.	HORA TRAB.
CP1	0,619	0,186	0,291	0,335
CP2	0,338	0,261	0,615	0,568

Y a partir de esto graficamos.



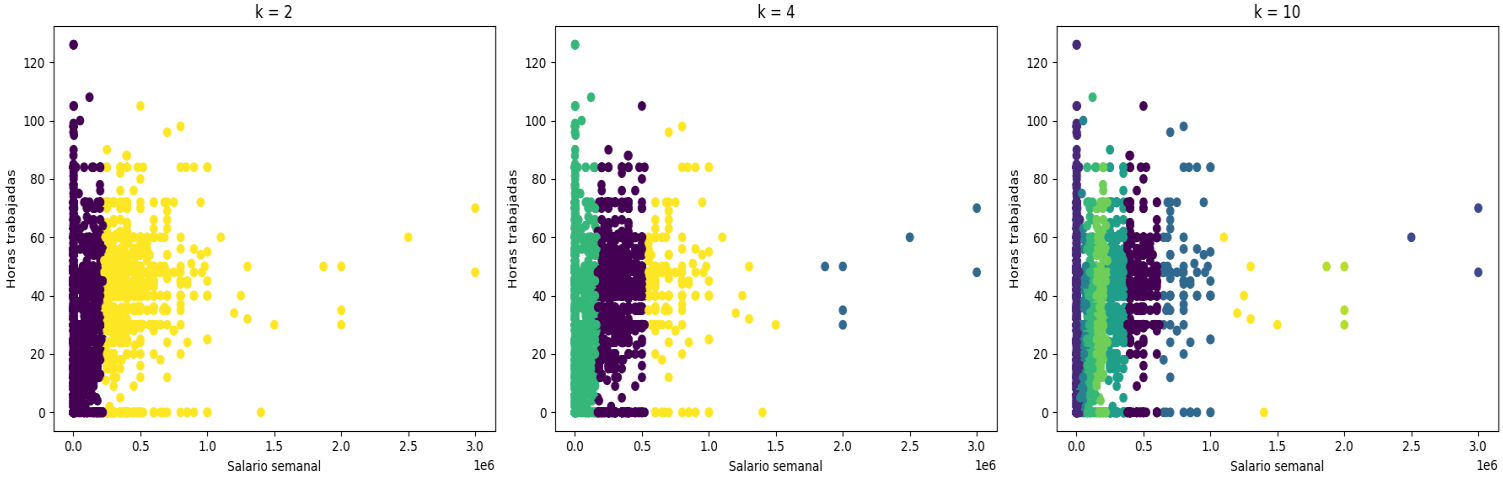
4)

Con los 3 primeros componentes ya logramos explicar más del 80% de la varianza original.
El primer componente explica más del 40% de la matriz original, hay bastante variabilidad no explicada; un segundo ponderador explicaría un poco más de un 20% al igual que un tercero y, por último, un cuarto componente principales explicarían menos del 20%.



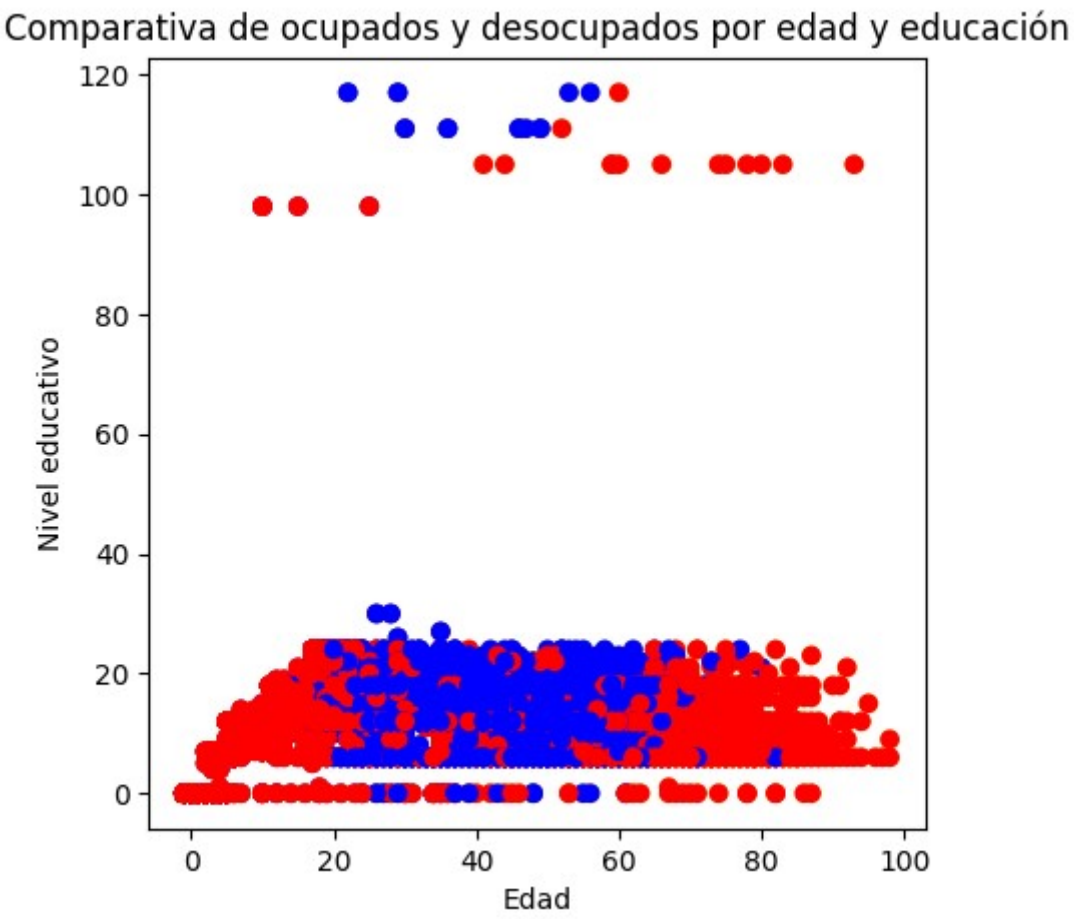
Cluester k-medias:

5)
A)Para esta consigna nos pide elegir predictores, nosotros elegimos el salario semanal y las horas totales trabajadas. También nos pide que hagamos un grafico para distintos números de Clúster, a la hora de graficar nos quedaría así:

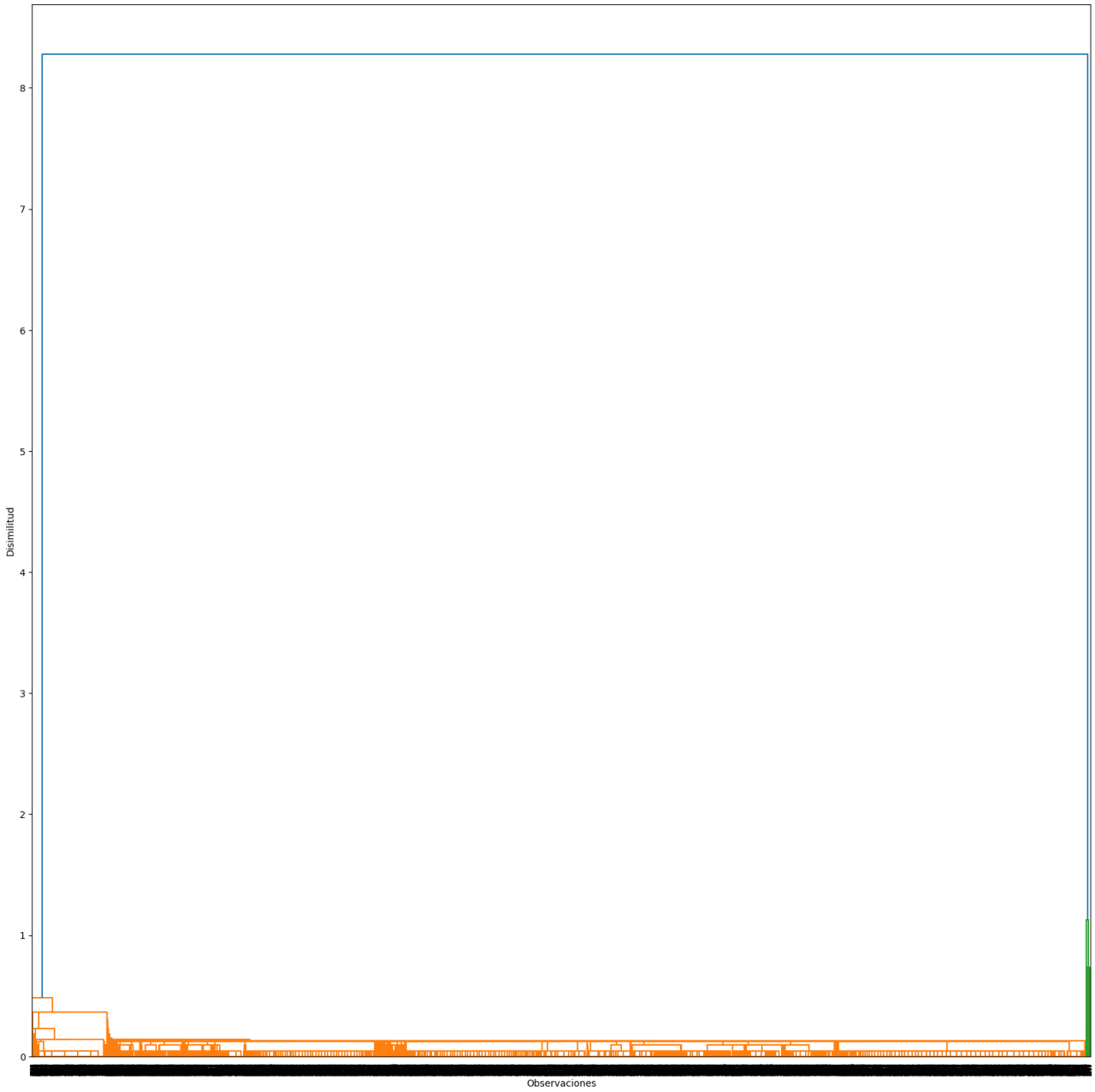


Cada punto de un color diferente representa un clúster distinto. Entre mas clúster hay mas agrupaciones.

B)
Para esta parte usamos nuestra variable educ que contiene todo sobre el nivel educativo que llegan a cursar los encuestados de la EPH y nuestra variable edad que contiene lógicamente la edad. Con estos datos vamos a hacer un grafico que compare ocupados y desocupados según su nivel educativo y edad.


Comparativa de ocupados y desocupados por edad y educación
Nivel educativo
Edad
0
20
40
60
80
100
120

El color azul es para Ocupados y el rojo para Desocupados.
Se puede observar como las personas con nivel educativo alto suelen estar empleados hasta su jubilación. Y contraposición en el nivel educativo que la gente empieza a trabajar se ven mas baches rojos osea desempleados.
6)
Para este ultimo punto del trabajo lo que vamos a hacer es un dendrograma de las mismas variables que trabajamos arriba educ y edad. ¿Que es un dendrograma? En síntesis es un grafico para mostrar agrupaciones jerárquicas.



en este caso por la cantidad de datos es difícil sacar una conjetura del mismo, mas o menos nos muestra como jerárquicamente la variable educ es modificada por edad.