Pasamos a responder lo planteado:

1. Al procesar los datos, una forma de saber cuáles son las variables más correlacionadas es mediante el coeficiente de correlación. Indique los 2 pares más correlacionadas y comente si le parecen razonables estos pares.

R. Los dos grupos más relacionados corresponden a “OtraDeuda vs DeudaTarjeta” con coeficiente 0,644955 e “IngresoMensual vs AnhosEmpleo” con coeficiente 0,625093. Notamos en ambos casos una relación entre las variables sean estas por aumento de deuda en el primer par, y en el segundo caso por el aumento de ingreso a medida que se avanza en los años de empleo.

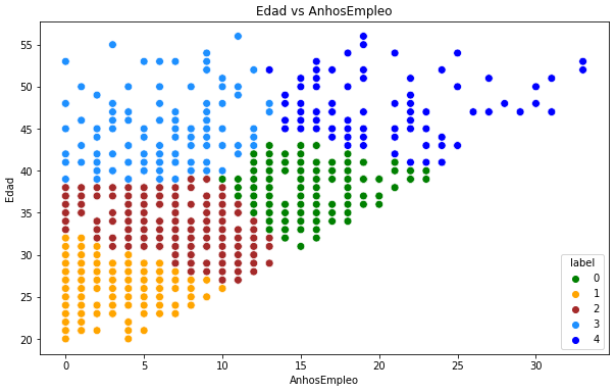
2. Indique subjetivamente los 5 pares de variables (ejm: AnhosEmpleo e Ingreso Mensual) que usted cree que son las más relevantes para definir patrones de los clientes. Puede incluir el resultado anterior. ¿Cuál es el número de clusters de cada par? ¿Cómo lo eligió?

R. Tomamos los siguientes pares, elegidos por los coeficientes más altos.

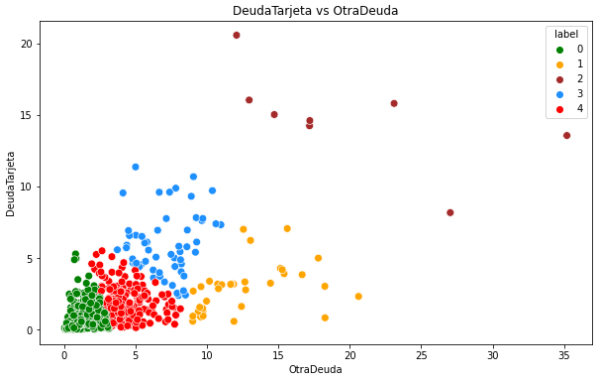
Par	Coeficiente	Nr. Cluster
OtraDeuda vs DeudaTarjeta	0,644955	5
IngresoMensual vs AnhosEmpleo	0,625093	4
IngresoMensual vs OtraDeuda	0,603356	4
RatioIngresoDeuda vs OtraDeuda	0,572545	5
AnhosEmpleo vs Edad	0,554241	5

3. Al aplicar K-Means con el número de clusters elegidos: ¿Cuál par tiene los patrones más definidos? ¿Cuál par tiene los patrones menos definidos? Comente sobre los resultados.

R. El par que encontramos mejor definido corresponde a “AnhosEmpleo vs Edad”, si bien los elementos están muy cohesionados, no se visualizan datos lejanos o muy alejados con esta combinación, se muestra el grafico:

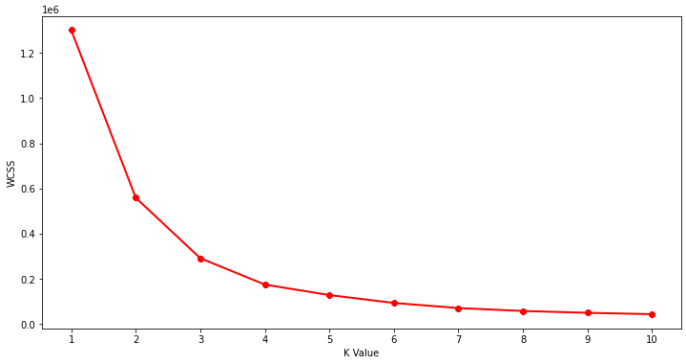


El par menos definido corresponde a “OtraDeuda vs DeudaTarjeta”, al ver el grafo se aprecia que el grupo 2 está muy lejos y disperso.

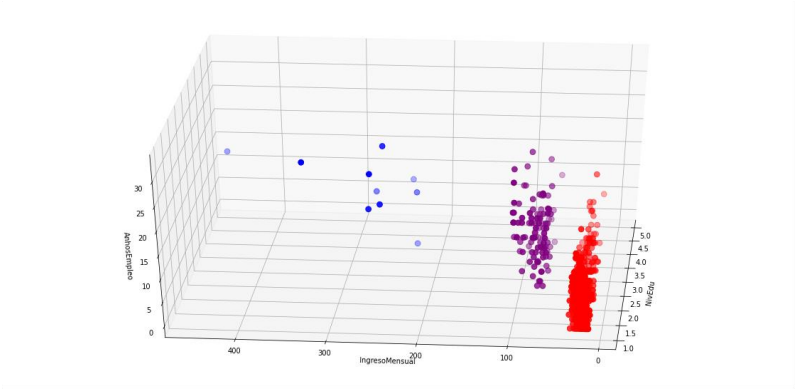


4. Al mejor par de variables, ahora agregue una variable de forma subjetiva. ¿Cuál es el número de clusters de cada par? ¿Cómo lo eligió? ¿Qué opina de los clusters obtenidos? (2 puntos)

R. Tomamos los siguientes campos "IngresoMensual","AnhosEmpleo" y "NivEdu", por análisis el número de cluster se fija en 5, obtenido del grafo que se muestra a continuación.



El único problema al graficarlo en 3D, corresponde a la variable “NivEdu” que al ser discreta posee solo 3 valores, como se aprecia el grafico



5. Si un compañero le indicara que Id es una variable útil para ser agregada al algoritmo K-Means. ¿Le parece correcto? Justifique su respuesta.

R. Concluimos que al ser un índice (ID), este no afecta en producir un impacto en los datos, debido a que solo va incrementándose 1 en 1.

Integrantes

1. Eric Hutchinson
2. Cristian Leon
3. Luis Reyes