

# LABORATORIO KMEAN

EDDIE ALEJANDRO GIRÓN CARRANZA

1307419

16-ABRIL-2023

## ELIMINACIÓN DE COLUMNA "ID"

Description: df [2,000 x 7]							
	Sex <int>	Marital.status <int>	Age <int>	Education <int>	Income <int>	Occupation <int>	Settlementsize <int>
1	0	0	67	2	124670	1	2
2	1	1	22	1	150773	1	2
3	0	0	49	1	89210	0	0
4	0	0	45	1	171565	1	1
5	0	0	53	1	149031	1	1
6	0	0	35	1	144848	0	0
7	0	0	53	1	156495	1	1
8	0	0	35	1	193621	2	1
9	0	1	61	2	151591	0	0
10	0	1	28	1	174646	2	0

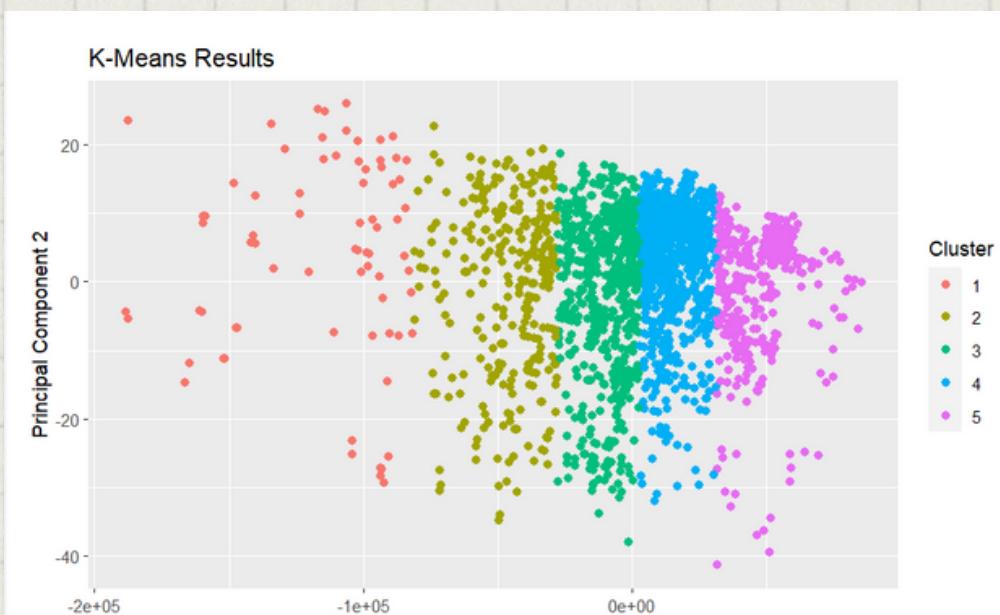


## VERIFICACIÓN DE REGISTROS NO NULOS

```
##(r)
sum(is.na(data))
```

Warning: is.na() applied to non-(list or vector) of type 'closure' [1] 0

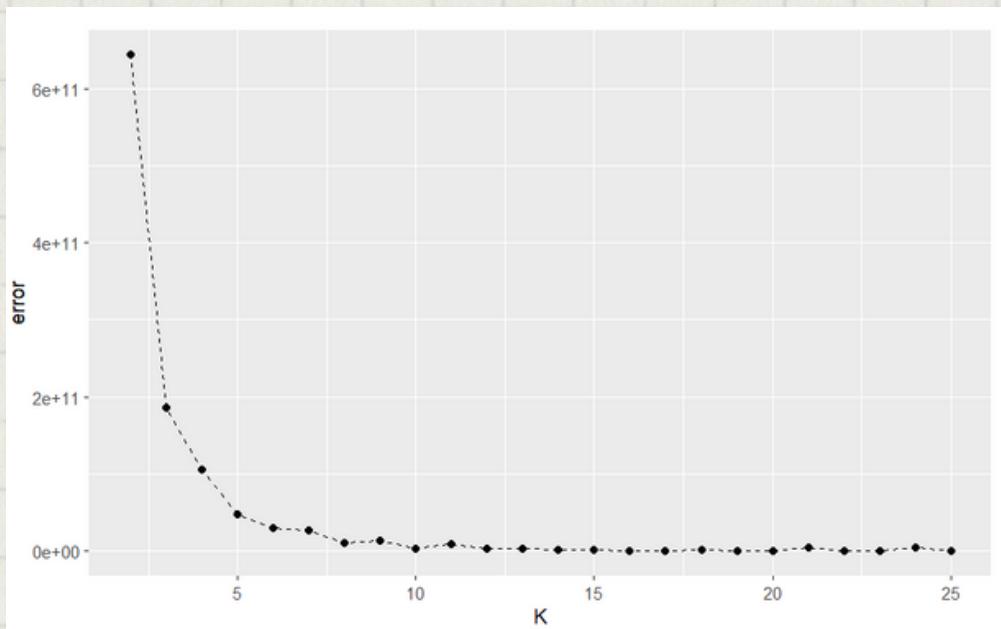
## KMEANS CON 5 CENTROS



## ERROR CON DISTINTA CANTIDAD DE CLUSTERS

	K <dbl>	error <dbl>
1	2	643786252591
2	3	185759595463
3	4	104871861338
4	5	47767115181
5	6	29523611739
6	7	26358245786
7	8	10649722290
8	9	13961439810
9	10	3509708121
10	11	8757357425
1-10 of 24 rows		
11	12	2897000797
12	13	2942750291
13	14	1389453893
14	15	792684886
15	16	392061266
16	17	583980250
17	18	979062167
18	19	584704121
19	20	188451275
20	21	3722997406
21	22	43744526
22	23	344026024
23	24	3722997406
24	25	183679447

## GRÁFICA %ERROR X NO.CLUSTERS



GRÁCIAS A LA GRÁFICA, Y A LA LEY DEL CODO, ES POSIBLE DETERMINAR QUE LA MEJOR CANTIDAD DE CLUSTERS ES DE 5 EN RELACIÓN CON LA CANTIDAD DE ERROR

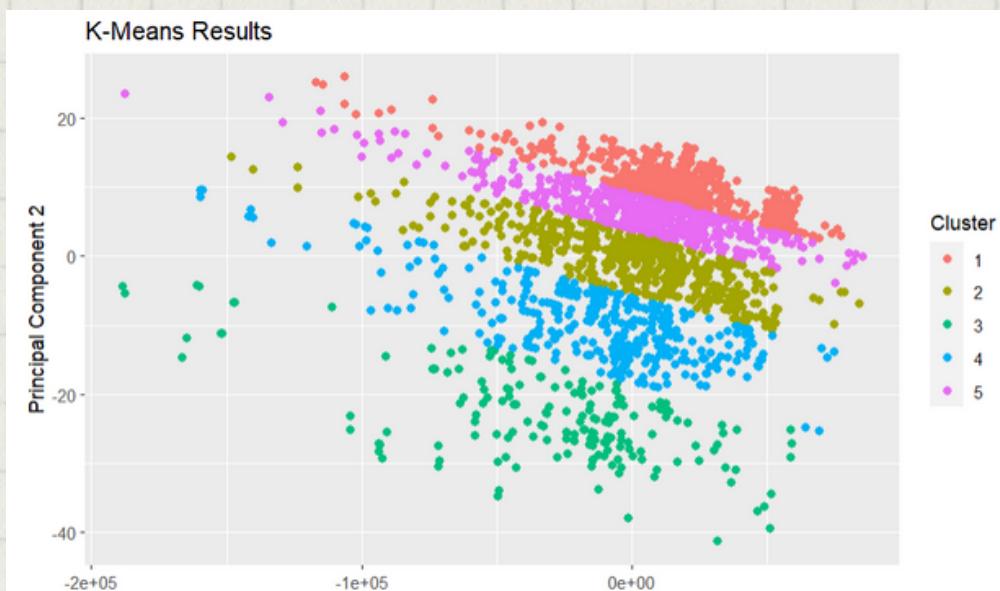
## MODIFICACIÓN DE COLUMNA INCOME

	Sex <int>	Marital.status <int>	Age <int>	Education <int>	Income <dbl>	Occupation <int>	Settlement.size <int>
1	0	0	67	2	0.097499228	1	2
2	1	1	22	1	0.782458689	1	2
3	0	0	49	1	-0.832993913	0	0
4	0	0	45	1	1.328054104	1	1
5	0	0	53	1	0.736747492	1	1
6	0	0	35	1	0.626982889	0	0
7	0	0	53	1	0.932607639	1	1
8	0	0	35	1	1.906817689	2	1
9	0	1	61	2	0.803923534	0	0
10	0	1	28	1	1.408901520	2	0

1-10 of 2,000 rows

Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) ... [100](#) Next

## KMEANS CON 5 CENTROS INCOME MODIFICADO

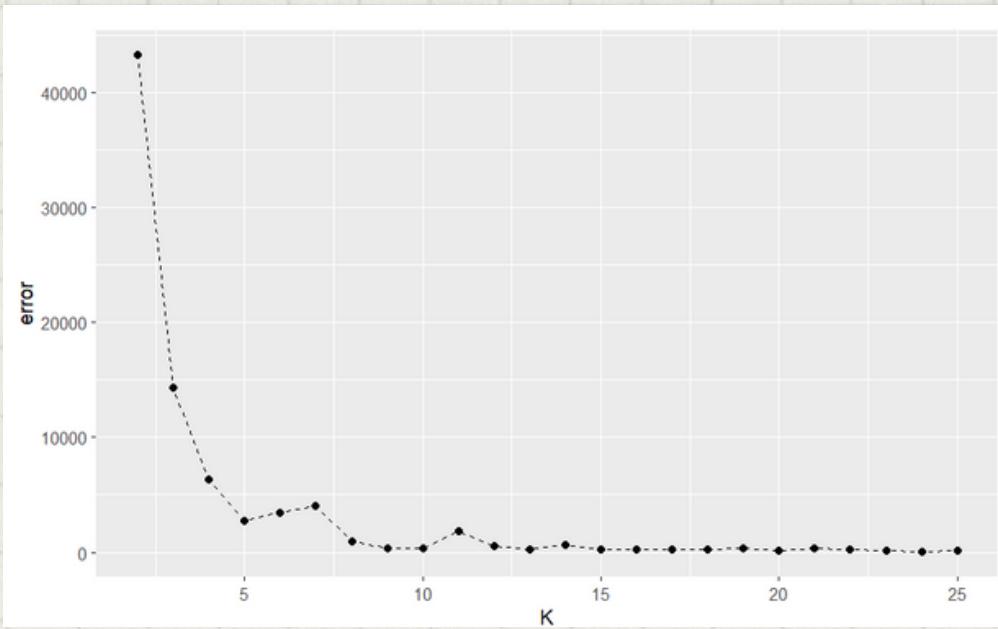


DEBIDO A LA MODIFICACIÓN DE DATOS EN LA COLUMNA INCOME, LOS CENTROS DENTRO DE LA GRÁFICA CAMBIARON DE POSICIÓN, TORNANDOSE ASÍ MAS HORIZONTALES

## ERROR CON DISTINTA CANTIDAD DE CLUSTERS INCOME MODIFICADO

	K <dbl>	error <dbl>
1	2	43236.48368
2	3	14340.17305
3	4	6387.58784
4	5	2733.06271
5	6	3417.15540
6	7	4084.10774
7	8	954.23865
8	9	340.96663
9	10	321.51020
10	11	1881.49937
11	12	582.58745
12	13	286.07606
13	14	632.20896
14	15	225.67495
15	16	246.22418
16	17	212.92668
17	18	274.04396
18	19	318.56030
19	20	110.44937
20	21	360.21628
21	22	221.74649
22	23	110.71762
23	24	55.94189
24	25	182.87170

## GRÁFICA %ERROR X NO.CLUSTERS INCOME MODIFICADO



SEGÚN LA GRÁFICA Y LA LEY DEL CODO,  
ES POSIBLE DECIR QUE AÚN HABIENDO  
MODIFICADO LA COLUMNA INCOME, LA  
CANTIDAD IDEAL DE CLUSTERS ES 5

## PREGUNTAS

1) ¿EXISTE DIFERENCIA EN EL NIVEL DE ERRORES ENTRE EL DATASET CON EL INCOME ORIGINAL Y EL ESTANDARIZADO?

R/ SÍ, EXISTE UNA DIFERENCIA BASTANTE NOTORIA. ESTO DEBIDO A QUE LOS DATOS QUE POSEE EL INCOME ORIGINAL SON VALORES MAS GRANDES Y DIFIEREN MAS UNO DE OTRO, Y EL EL ESTANDARIZADO SE TIENEN VALORES MAS PEQUEÑOS Y CON MENOR DIFERENCIA ENTRE ELLOS

2) ¿EXISTE DIFERENCIA EN EL NUMERO DE CLUSTERS ÓPTIMOS ENTRE EL DATASET CON EL INCOME ORIGINAL Y EL ESTANDARIZADO?

R/ NO, EL NÚMERO DE CLUSTERS ÓPTIMOS PARA AMBOS CASOS ES 5. ESTO DEBIDO A QUE, AUNQUE LOS DATOS SE ENCUENTRAN ESTANDARIZADOS, LA CANTIDAD DE DATOS ES LA MISMA, POR LO QUE AÚN EXISTEN DATOS QUE SE ALEJAN DE LOS CENTROS