

# Final Analisis

Luis Tujab 1103920

2023-11-21

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(ggplot2)
library(useful)

## Warning: package 'useful' was built under R version 4.3.2
```

#Estadísticas general

```
data <- read.csv("Mall_Customers.csv")
data
```

	CustomerID	Genre	Age	Annual_Income_.k..	Spending_Score
## 1	1	Male	19	15	39
## 2	2	Male	21	15	81
## 3	3	Female	20	16	6
## 4	4	Female	23	16	77
## 5	5	Female	31	17	40
## 6	6	Female	22	17	76
## 7	7	Female	35	18	6
## 8	8	Female	23	18	94
## 9	9	Male	64	19	3
## 10	10	Female	30	19	72
## 11	11	Male	67	19	14
## 12	12	Female	35	19	99
## 13	13	Female	58	20	15
## 14	14	Female	24	20	77
## 15	15	Male	37	20	13
## 16	16	Male	22	20	79
## 17	17	Female	35	21	35
## 18	18	Male	20	21	66
## 19	19	Male	52	23	29
## 20	20	Female	35	23	98

## 21	21	Male	35	24	35
## 22	22	Male	25	24	73
## 23	23	Female	46	25	5
## 24	24	Male	31	25	73
## 25	25	Female	54	28	14
## 26	26	Male	29	28	82
## 27	27	Female	45	28	32
## 28	28	Male	35	28	61
## 29	29	Female	40	29	31
## 30	30	Female	23	29	87
## 31	31	Male	60	30	4
## 32	32	Female	21	30	73
## 33	33	Male	53	33	4
## 34	34	Male	18	33	92
## 35	35	Female	49	33	14
## 36	36	Female	21	33	81
## 37	37	Female	42	34	17
## 38	38	Female	30	34	73
## 39	39	Female	36	37	26
## 40	40	Female	20	37	75
## 41	41	Female	65	38	35
## 42	42	Male	24	38	92
## 43	43	Male	48	39	36
## 44	44	Female	31	39	61
## 45	45	Female	49	39	28
## 46	46	Female	24	39	65
## 47	47	Female	50	40	55
## 48	48	Female	27	40	47
## 49	49	Female	29	40	42
## 50	50	Female	31	40	42
## 51	51	Female	49	42	52
## 52	52	Male	33	42	60
## 53	53	Female	31	43	54
## 54	54	Male	59	43	60
## 55	55	Female	50	43	45
## 56	56	Male	47	43	41
## 57	57	Female	51	44	50
## 58	58	Male	69	44	46
## 59	59	Female	27	46	51
## 60	60	Male	53	46	46
## 61	61	Male	70	46	56
## 62	62	Male	19	46	55
## 63	63	Female	67	47	52
## 64	64	Female	54	47	59
## 65	65	Male	63	48	51
## 66	66	Male	18	48	59
## 67	67	Female	43	48	50
## 68	68	Female	68	48	48
## 69	69	Male	19	48	59
## 70	70	Female	32	48	47

## 71	71	Male	70	49	55
## 72	72	Female	47	49	42
## 73	73	Female	60	50	49
## 74	74	Female	60	50	56
## 75	75	Male	59	54	47
## 76	76	Male	26	54	54
## 77	77	Female	45	54	53
## 78	78	Male	40	54	48
## 79	79	Female	23	54	52
## 80	80	Female	49	54	42
## 81	81	Male	57	54	51
## 82	82	Male	38	54	55
## 83	83	Male	67	54	41
## 84	84	Female	46	54	44
## 85	85	Female	21	54	57
## 86	86	Male	48	54	46
## 87	87	Female	55	57	58
## 88	88	Female	22	57	55
## 89	89	Female	34	58	60
## 90	90	Female	50	58	46
## 91	91	Female	68	59	55
## 92	92	Male	18	59	41
## 93	93	Male	48	60	49
## 94	94	Female	40	60	40
## 95	95	Female	32	60	42
## 96	96	Male	24	60	52
## 97	97	Female	47	60	47
## 98	98	Female	27	60	50
## 99	99	Male	48	61	42
## 100	100	Male	20	61	49
## 101	101	Female	23	62	41
## 102	102	Female	49	62	48
## 103	103	Male	67	62	59
## 104	104	Male	26	62	55
## 105	105	Male	49	62	56
## 106	106	Female	21	62	42
## 107	107	Female	66	63	50
## 108	108	Male	54	63	46
## 109	109	Male	68	63	43
## 110	110	Male	66	63	48
## 111	111	Male	65	63	52
## 112	112	Female	19	63	54
## 113	113	Female	38	64	42
## 114	114	Male	19	64	46
## 115	115	Female	18	65	48
## 116	116	Female	19	65	50
## 117	117	Female	63	65	43
## 118	118	Female	49	65	59
## 119	119	Female	51	67	43
## 120	120	Female	50	67	57

## 121	121	Male	27	67	56
## 122	122	Female	38	67	40
## 123	123	Female	40	69	58
## 124	124	Male	39	69	91
## 125	125	Female	23	70	29
## 126	126	Female	31	70	77
## 127	127	Male	43	71	35
## 128	128	Male	40	71	95
## 129	129	Male	59	71	11
## 130	130	Male	38	71	75
## 131	131	Male	47	71	9
## 132	132	Male	39	71	75
## 133	133	Female	25	72	34
## 134	134	Female	31	72	71
## 135	135	Male	20	73	5
## 136	136	Female	29	73	88
## 137	137	Female	44	73	7
## 138	138	Male	32	73	73
## 139	139	Male	19	74	10
## 140	140	Female	35	74	72
## 141	141	Female	57	75	5
## 142	142	Male	32	75	93
## 143	143	Female	28	76	40
## 144	144	Female	32	76	87
## 145	145	Male	25	77	12
## 146	146	Male	28	77	97
## 147	147	Male	48	77	36
## 148	148	Female	32	77	74
## 149	149	Female	34	78	22
## 150	150	Male	34	78	90
## 151	151	Male	43	78	17
## 152	152	Male	39	78	88
## 153	153	Female	44	78	20
## 154	154	Female	38	78	76
## 155	155	Female	47	78	16
## 156	156	Female	27	78	89
## 157	157	Male	37	78	1
## 158	158	Female	30	78	78
## 159	159	Male	34	78	1
## 160	160	Female	30	78	73
## 161	161	Female	56	79	35
## 162	162	Female	29	79	83
## 163	163	Male	19	81	5
## 164	164	Female	31	81	93
## 165	165	Male	50	85	26
## 166	166	Female	36	85	75
## 167	167	Male	42	86	20
## 168	168	Female	33	86	95
## 169	169	Female	36	87	27
## 170	170	Male	32	87	63

## 171	171	Male	40	87	13
## 172	172	Male	28	87	75
## 173	173	Male	36	87	10
## 174	174	Male	36	87	92
## 175	175	Female	52	88	13
## 176	176	Female	30	88	86
## 177	177	Male	58	88	15
## 178	178	Male	27	88	69
## 179	179	Male	59	93	14
## 180	180	Male	35	93	90
## 181	181	Female	37	97	32
## 182	182	Female	32	97	86
## 183	183	Male	46	98	15
## 184	184	Female	29	98	88
## 185	185	Female	41	99	39
## 186	186	Male	30	99	97
## 187	187	Female	54	101	24
## 188	188	Male	28	101	68
## 189	189	Female	41	103	17
## 190	190	Female	36	103	85
## 191	191	Female	34	103	23
## 192	192	Female	32	103	69
## 193	193	Male	33	113	8
## 194	194	Female	38	113	91
## 195	195	Female	47	120	16
## 196	196	Female	35	120	79
## 197	197	Female	45	126	28
## 198	198	Male	32	126	74
## 199	199	Male	32	137	18
## 200	200	Male	30	137	83

`summary(data)`

##	CustomerID	Genre	Age	
##	Annual_Income_.k..			
##	Min. : 1.00	Length:200	Min. :18.00	Min. : 15.00
##	1st Qu.: 50.75	Class :character	1st Qu.:28.75	1st Qu.: 41.50
##	Median :100.50	Mode :character	Median :36.00	Median : 61.50
##	Mean :100.50		Mean :38.85	Mean : 60.56
##	3rd Qu.:150.25		3rd Qu.:49.00	3rd Qu.: 78.00
##	Max. :200.00		Max. :70.00	Max. :137.00
##	Spending_Score			
##	Min. : 1.00			
##	1st Qu.:34.75			
##	Median :50.00			
##	Mean :50.20			
##	3rd Qu.:73.00			
##	Max. :99.00			

#Limpieza de los datos eliminar ID

R// Se elimino el ID ya que considero que es un valor que no aporta nada al analisis de los datos y podria generar algun tipo de error.

```
dataLim <- data[, -c(1)]  
summary(dataLim)
```

```
##      Genre      Age      Annual_Income_.k.. Spending_Score  
## Length:200      Min.   :18.00      Min.   : 15.00      Min.   : 1.00  
## Class :character 1st Qu.:28.75      1st Qu.: 41.50      1st Qu.:34.75  
## Mode  :character Median :36.00      Median : 61.50      Median :50.00  
##              Mean  :38.85      Mean   : 60.56      Mean   :50.20  
##              3rd Qu.:49.00      3rd Qu.: 78.00      3rd Qu.:73.00  
##              Max.   :70.00      Max.   :137.00      Max.   :99.00
```

#Covertir el genero en un 0 o 1 -> 0 para hombre y 1 para mujer

R// Esta proceso de convertir a un valor binario es mayor ya que mantener los valores de hombre y mujer van en contra de la estandarizacion para mantener los datos de una manera homogenea

```
dataLim$Genre <- ifelse(dataLim$Genre == "Female",1,0)  
dataLim$Genre <- as.integer(dataLim$Genre)  
dataLim
```

```
##      Genre Age Annual_Income_.k.. Spending_Score  
## 1      0  19              15              39  
## 2      0  21              15              81  
## 3      1  20              16               6  
## 4      1  23              16              77  
## 5      1  31              17              40  
## 6      1  22              17              76  
## 7      1  35              18               6  
## 8      1  23              18              94  
## 9      0  64              19               3  
## 10     1  30              19              72  
## 11     0  67              19              14  
## 12     1  35              19              99  
## 13     1  58              20              15  
## 14     1  24              20              77  
## 15     0  37              20              13  
## 16     0  22              20              79  
## 17     1  35              21              35  
## 18     0  20              21              66  
## 19     0  52              23              29  
## 20     1  35              23              98  
## 21     0  35              24              35  
## 22     0  25              24              73  
## 23     1  46              25               5  
## 24     0  31              25              73  
## 25     1  54              28              14  
## 26     0  29              28              82
```

## 27	1	45	28	32
## 28	0	35	28	61
## 29	1	40	29	31
## 30	1	23	29	87
## 31	0	60	30	4
## 32	1	21	30	73
## 33	0	53	33	4
## 34	0	18	33	92
## 35	1	49	33	14
## 36	1	21	33	81
## 37	1	42	34	17
## 38	1	30	34	73
## 39	1	36	37	26
## 40	1	20	37	75
## 41	1	65	38	35
## 42	0	24	38	92
## 43	0	48	39	36
## 44	1	31	39	61
## 45	1	49	39	28
## 46	1	24	39	65
## 47	1	50	40	55
## 48	1	27	40	47
## 49	1	29	40	42
## 50	1	31	40	42
## 51	1	49	42	52
## 52	0	33	42	60
## 53	1	31	43	54
## 54	0	59	43	60
## 55	1	50	43	45
## 56	0	47	43	41
## 57	1	51	44	50
## 58	0	69	44	46
## 59	1	27	46	51
## 60	0	53	46	46
## 61	0	70	46	56
## 62	0	19	46	55
## 63	1	67	47	52
## 64	1	54	47	59
## 65	0	63	48	51
## 66	0	18	48	59
## 67	1	43	48	50
## 68	1	68	48	48
## 69	0	19	48	59
## 70	1	32	48	47
## 71	0	70	49	55
## 72	1	47	49	42
## 73	1	60	50	49
## 74	1	60	50	56
## 75	0	59	54	47
## 76	0	26	54	54

## 77	1	45	54	53
## 78	0	40	54	48
## 79	1	23	54	52
## 80	1	49	54	42
## 81	0	57	54	51
## 82	0	38	54	55
## 83	0	67	54	41
## 84	1	46	54	44
## 85	1	21	54	57
## 86	0	48	54	46
## 87	1	55	57	58
## 88	1	22	57	55
## 89	1	34	58	60
## 90	1	50	58	46
## 91	1	68	59	55
## 92	0	18	59	41
## 93	0	48	60	49
## 94	1	40	60	40
## 95	1	32	60	42
## 96	0	24	60	52
## 97	1	47	60	47
## 98	1	27	60	50
## 99	0	48	61	42
## 100	0	20	61	49
## 101	1	23	62	41
## 102	1	49	62	48
## 103	0	67	62	59
## 104	0	26	62	55
## 105	0	49	62	56
## 106	1	21	62	42
## 107	1	66	63	50
## 108	0	54	63	46
## 109	0	68	63	43
## 110	0	66	63	48
## 111	0	65	63	52
## 112	1	19	63	54
## 113	1	38	64	42
## 114	0	19	64	46
## 115	1	18	65	48
## 116	1	19	65	50
## 117	1	63	65	43
## 118	1	49	65	59
## 119	1	51	67	43
## 120	1	50	67	57
## 121	0	27	67	56
## 122	1	38	67	40
## 123	1	40	69	58
## 124	0	39	69	91
## 125	1	23	70	29
## 126	1	31	70	77



## 127	0	43	71	35
## 128	0	40	71	95
## 129	0	59	71	11
## 130	0	38	71	75
## 131	0	47	71	9
## 132	0	39	71	75
## 133	1	25	72	34
## 134	1	31	72	71
## 135	0	20	73	5
## 136	1	29	73	88
## 137	1	44	73	7
## 138	0	32	73	73
## 139	0	19	74	10
## 140	1	35	74	72
## 141	1	57	75	5
## 142	0	32	75	93
## 143	1	28	76	40
## 144	1	32	76	87
## 145	0	25	77	12
## 146	0	28	77	97
## 147	0	48	77	36
## 148	1	32	77	74
## 149	1	34	78	22
## 150	0	34	78	90
## 151	0	43	78	17
## 152	0	39	78	88
## 153	1	44	78	20
## 154	1	38	78	76
## 155	1	47	78	16
## 156	1	27	78	89
## 157	0	37	78	1
## 158	1	30	78	78
## 159	0	34	78	1
## 160	1	30	78	73
## 161	1	56	79	35
## 162	1	29	79	83
## 163	0	19	81	5
## 164	1	31	81	93
## 165	0	50	85	26
## 166	1	36	85	75
## 167	0	42	86	20
## 168	1	33	86	95
## 169	1	36	87	27
## 170	0	32	87	63
## 171	0	40	87	13
## 172	0	28	87	75
## 173	0	36	87	10
## 174	0	36	87	92
## 175	1	52	88	13
## 176	1	30	88	86

## 177	0	58	88	15
## 178	0	27	88	69
## 179	0	59	93	14
## 180	0	35	93	90
## 181	1	37	97	32
## 182	1	32	97	86
## 183	0	46	98	15
## 184	1	29	98	88
## 185	1	41	99	39
## 186	0	30	99	97
## 187	1	54	101	24
## 188	0	28	101	68
## 189	1	41	103	17
## 190	1	36	103	85
## 191	1	34	103	23
## 192	1	32	103	69
## 193	0	33	113	8
## 194	1	38	113	91
## 195	1	47	120	16
## 196	1	35	120	79
## 197	1	45	126	28
## 198	0	32	126	74
## 199	0	32	137	18
## 200	0	30	137	83

#Nulos por columna

R/ Se verifica que la cantidad de nulos no sea considerable para eliminarlos

```
Nulos <- colSums(is.na(dataLim))
```

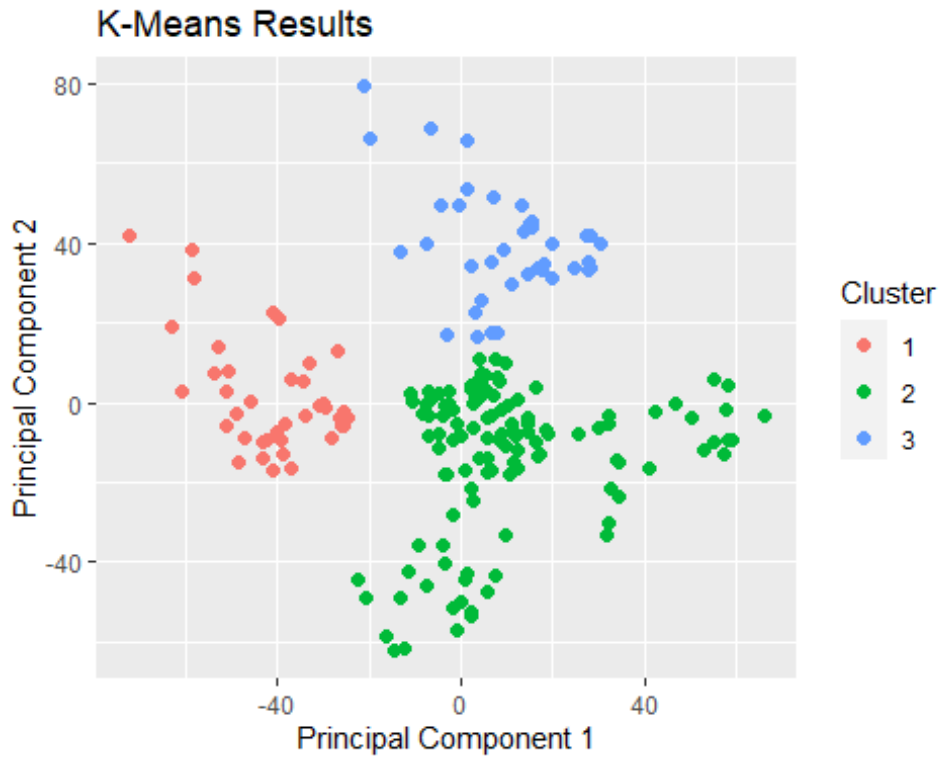
Nulos

##	Genre	Age	Annual_Income_.k..
Spending_Score			
##	0	0	0
0			

#Metodo del kmeans 3 centros como prueba

```
datTrain <- kmeans(dataLim, centers = 3)
```

```
plot(datTrain, data = dataLim)
```



#Cuadro de

clusters y error para la grafica de codo

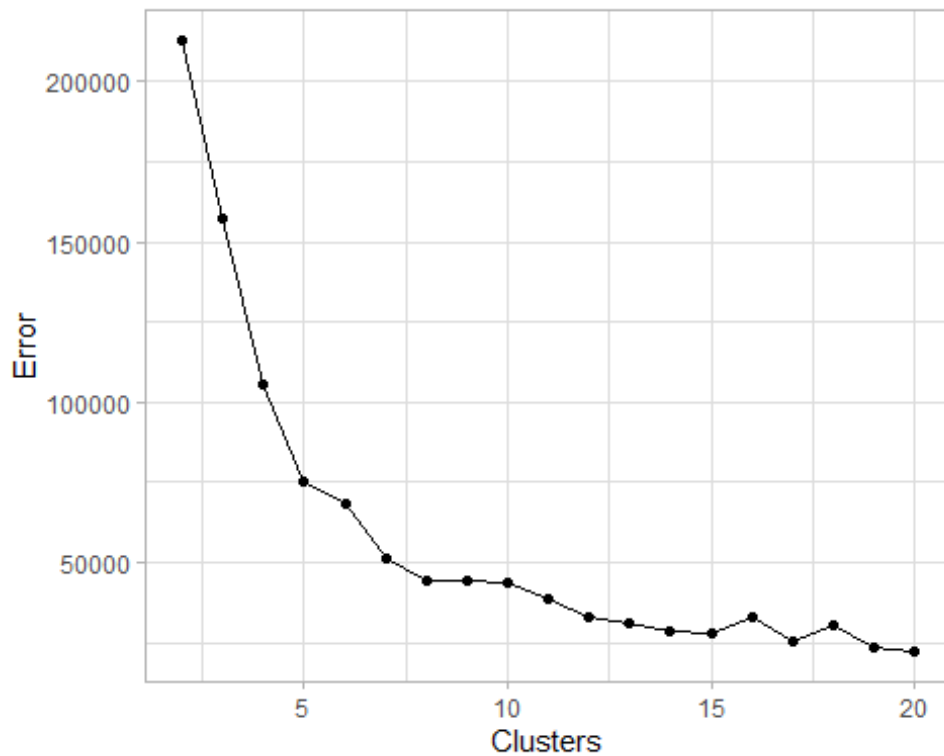
```
Gew <- data.frame(matrix(ncol = 2, nrow = 0))
colnames(Gew) <- c("Clusters", "Error")
for (i in 1:20)
{
  datTrain <- kmeans(x= dataLim, centers = i)
  Gew[i-1,] <- c(i, datTrain$tot.withinss)
}
Gew
```

##	Clusters	Error
## 1	2	212889.44
## 2	3	157200.67
## 3	4	105299.99
## 4	5	75399.62
## 5	6	68331.80
## 6	7	51130.69
## 7	8	44355.31
## 8	9	44346.95
## 9	10	43585.09
## 10	11	38573.41
## 11	12	32920.48
## 12	13	31048.72
## 13	14	28290.76
## 14	15	28001.65
## 15	16	32676.82
## 16	17	25569.35

```
## 17      18  30345.33
## 18      19  23644.26
## 19      20  22275.92
```

#Grafica de codo

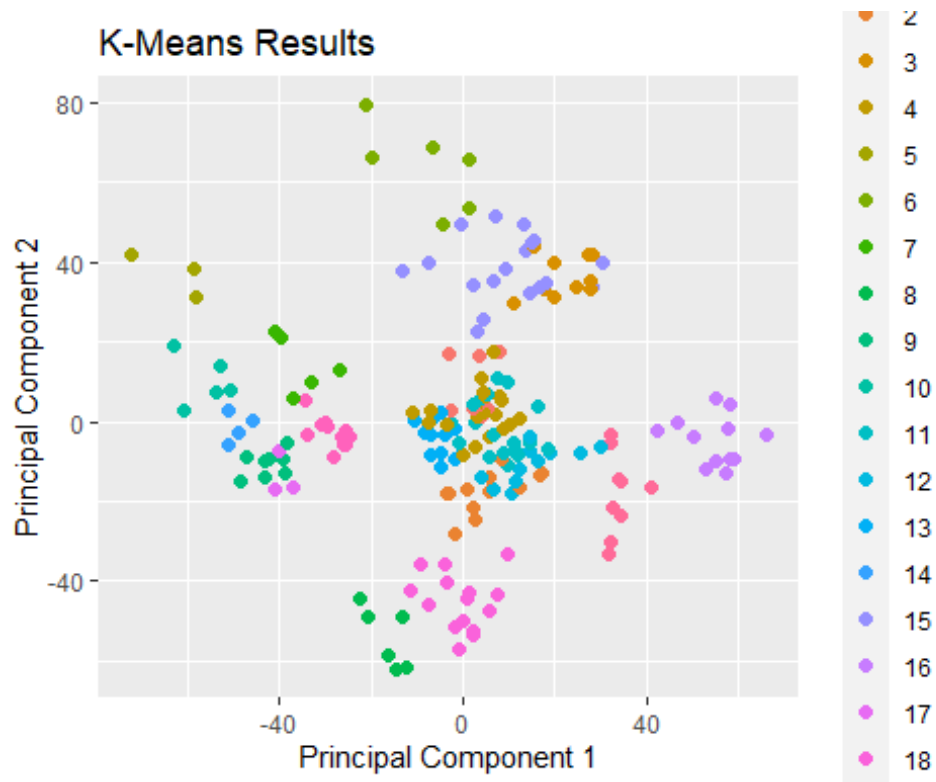
```
ggplot(data = Gew, aes(x = Clusters, y = Error)) +  
  geom_line() +  
  geom_point() +  
  theme_light()
```



#Grafica con todos los centros solo de demostración

R/ Demostracion para los 18 centros creados no es el valor final a tomar

```
plot(datTrain, data = dataLim)
```



#Grafica con el numero de clusters, este si es el valor oficial a tomar ya que es el valor donde la grafica se comienza a estabilizar

```
datTrain <- kmeans(dataLim, centers = 7)  
plot(datTrain, data = dataLim)
```

