METODOS MULTIVARIADOS

andres cerquera mejia

Analisis de clusters

Acontinuacion se emplea otra base de datos para hacer el analisis de clusters de una manera en que se pudieran emplear los distintos analisis vistos en clase.

```
## # A tibble: 27 x 8
##    CULTIVOS `SUPERFICIE, HA~ `SUPERFICIE, HA~ `PRODUCCION TON` `PRECIO AL PROD~
##    <chr>                <dbl>            <dbl>            <dbl>            <dbl>
##  1 YUCA SO~             5058.            3060.           33829.            1004.
##  2 YUCA AS~             7695.            5444            35092.             828.
##  3 YUCA IN~             2635             2324.           35988.             239.
##  4 TABACO ~              866.             724.            1578.            3433.
##  5 TABACO ~              953.             516.            1549.            5586.
##  6 COL                     4                4              12             1800
##  7 BERENJE~              120               98             396             1050
##  8 BATATA                 10               10             120             1500
##  9 PLATANO              1535.            1364.            6707.             675
## 10 COCO                  680              640             3336              750
## # ... with 17 more rows, and 3 more variables: `COSTO DE PRODUCCION HA` <dbl>,
## #   `INGRESO BRUTO PRODUCCION` <dbl>, `COSTO TOTAL PRODUCCION` <dbl>
```

```
##  [1] "YUCA SOLA"             "YUCA ASOCIADA"         "YUCA INDUSTRIAL"
##  [4] "TABACO NEGRO"          "TABACO RUBIO"          "COL"
##  [7] "BERENJENA"             "BATATA"                "PLATANO"
## [10] "COCO"                  "AGUACATE"              "NARANJA DULCE"
## [13] "CACAO"                 "MARACUYA"              "PALMA AFRICANA"
## [16] "GUAYABA DULCE"         "MANGO"                 "PAPAYA"
## [19] "ZAPOTE"                "CAÑA PANELERA"         "PIÑA"
## [22] "AJÍ DULCE"             "ÑAME CRIOLLO ASOCIADO" "ÑAME CRIOLLO SOLO"
## [25] "ÑAME ESPINO"           "LIMÓN CRIOLLO"         "AHUYAMA"
```

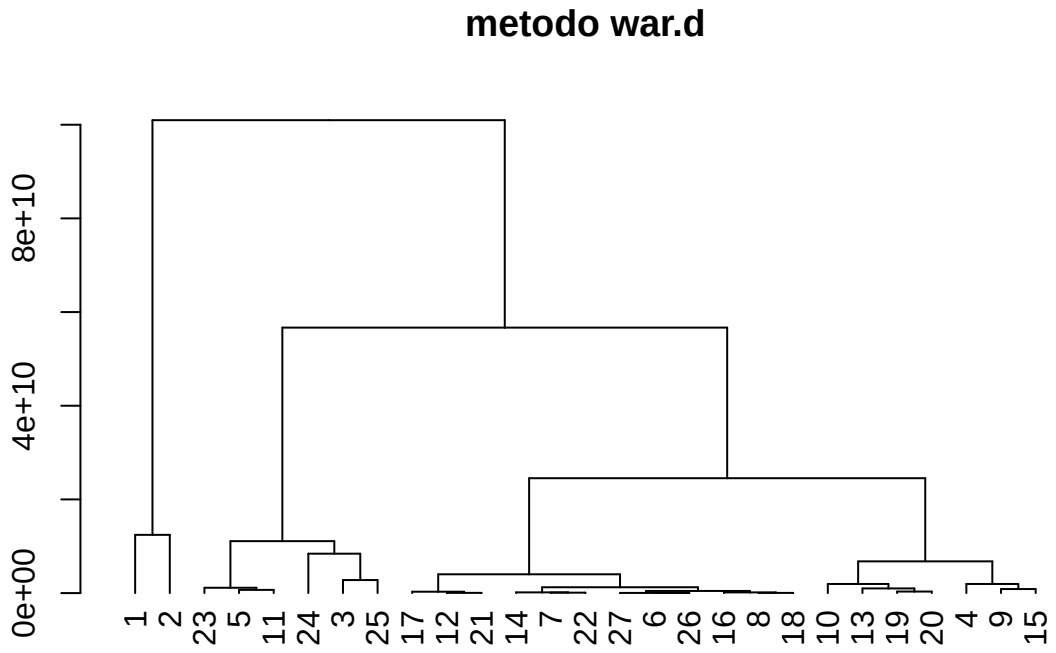
Metodo de ward

```
library(cluster.datasets)
hp= hclust(dist(datas),method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
## Warning in dist(datas): NAs introducidos por coerción
```

```
dend2 <- as.dendrogram(hp)
plot(dend2)
title(main='metodo war.d')
```



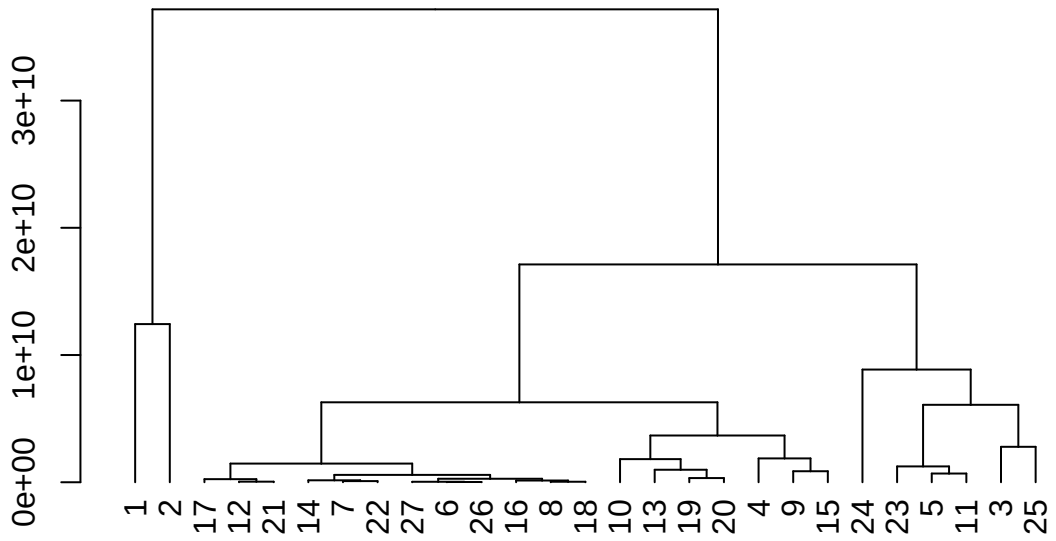
en este metodo jerarquico de varianza minima se parte del supuesto de que los datos se provienen de poblaciones normales para este supuesto se tienen cultivos del departamento de sucre, en este caso parecieran que los cultivos se forman en 4 grupos donde el cultivo 1 y 2 forman un cluster el segundo grupo incluye aquellos 23 a 25 y un tercer grupo desde el 17 hasta el 18, y un cuarto grupo desde el cultivo 10 hasta el 15. aparte del resto de los cultivos que se ve claramente que forman una relacion jerarquica entre ellos.

Metodo de Encadenamiento completo

```
hc= hclust(dist(datas),"com")
```

```
## Warning in dist(datas): NAs introducidos por coerción
```

```
dend1 <- as.dendrogram(hc)
plot(dend1)
title(main='Encadenamiento Completo')
```



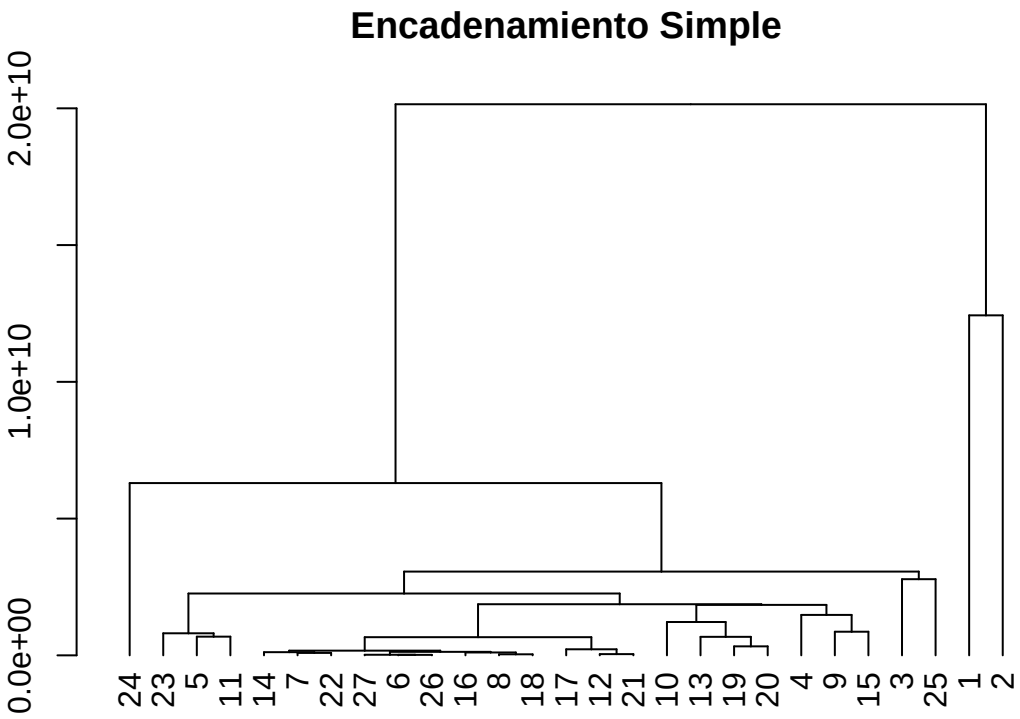
este metodo toma la maxima distancia entre un par de grupos, es decir la distancia entre 2 conglomerados se calcula como la distancia entre sus dos elementos mas alejados. se observan 4 grupos definidos donde uno esta conformado por cultivo 1 y 2 mientras el segundo va desde el 17 hasta el cultivo 18 y un tercer grupo donde se incluyen los cultivos 10 hasta el cultivo 15 y el cuarto grupo desde el 24 hasta el 25.

metodo encadenamiento simple

```
hm <- hclust(dist(datas), "sing")
```

```
## Warning in dist(datas): NAs introducidos por coerción
```

```
dend3 <- as.dendrogram(hm)
op <- par(mfrow= c(1,1), mar = c(4,4,3,3))
plot(dend3)
title(main='Encadenamiento Simple')
```

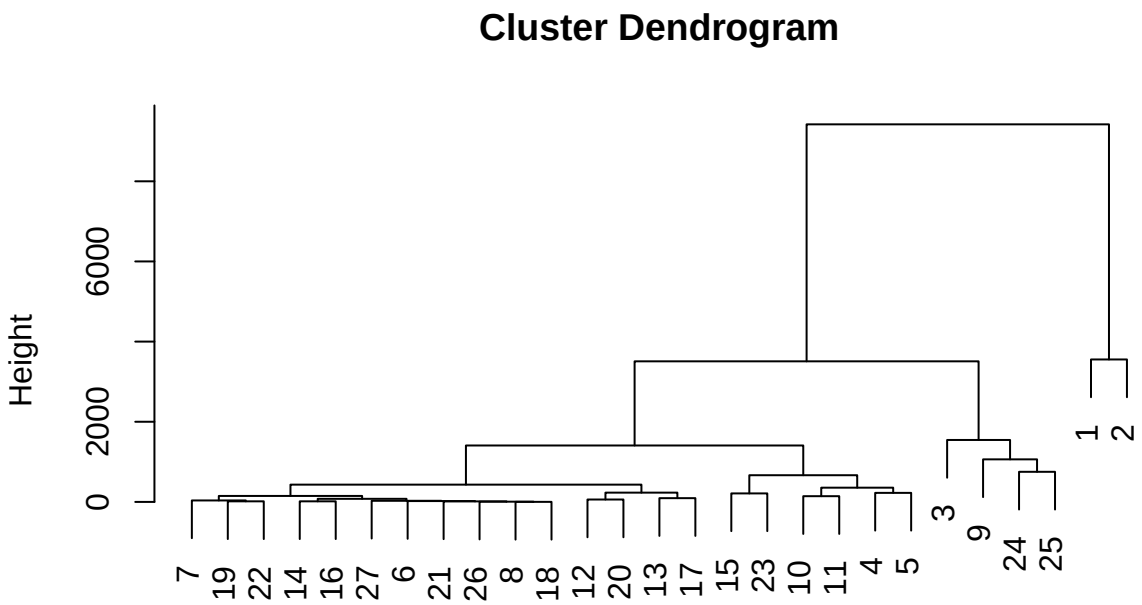


este metodo agrupa con la minima distancia de los vecinos mas cercano es decir la distancia minima entre dos clusters y se fusionan los dos clusteres mas cercanos, como se puede observar de nuevo el cultivo 1 y 2 forma un grupo aspecto que coincide con los otros dos metodos.

DENDOGRAMA DE LOS CULTIVOS

variables: superficie_has_sembrada,superficies_has_cosechada se construye un dendograma donde se tienen en cuenta las dos variables anteriores y de esta manera saber cuales cultivos estan relacionados acorde a su supeficie. se tiene el cultivo 1 y 2 yuca sola y yuca asociada formando un cluster mientras que la mayoria de los demas clusters se encuentran del valor 2000 hacia abajo formando 2 clusters en el valor 2000.

```
superficies_cluster=hclust(dist(datas[,2:3]))
plot(superficies_cluster)
```



ANALISIS POR K_MEANS

el metodo k-means se emplea en grupos muy grandes >10000 cabe mencionar que se debe establacer una semilla set.seed, pero en este caso esta no se empleo ya que es una poblacion muy pequeña ademas este metodo trata de buscar en las observaciones grupos con caracteristicas similares para este analisis es necesario determinar el numero de clusters tambien.

```
cluste=kmeans(datas[,-1],3)#realizamos clustering con k=3
names(cluste) #asignacion de observaciones
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
cluste$cluster #asinacion de observaciones a cluster, se clasifican los 27 datos en los 3 grupos(cluste
```

```
##  [1] 3 3 2 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1
```

```
cluste$totss #inercia total, inercia de grupos con respecto al centroide.
```

```
## [1] 2.223597e+21
```

```
cluste$betweenss #inercia intergrupos,
```

```
## [1] 2.015858e+21
```

```
cluste$withinss #inercia intragrupos,inercia individual de cada grupo, una por cada cluster.
```

```
## [1] 5.686206e+19 8.326102e+19 6.761564e+19
```

```
cluste$tot.withinss#inercia intra grupos (total) suma de las inercias
```

```
## [1] 2.077387e+20
```

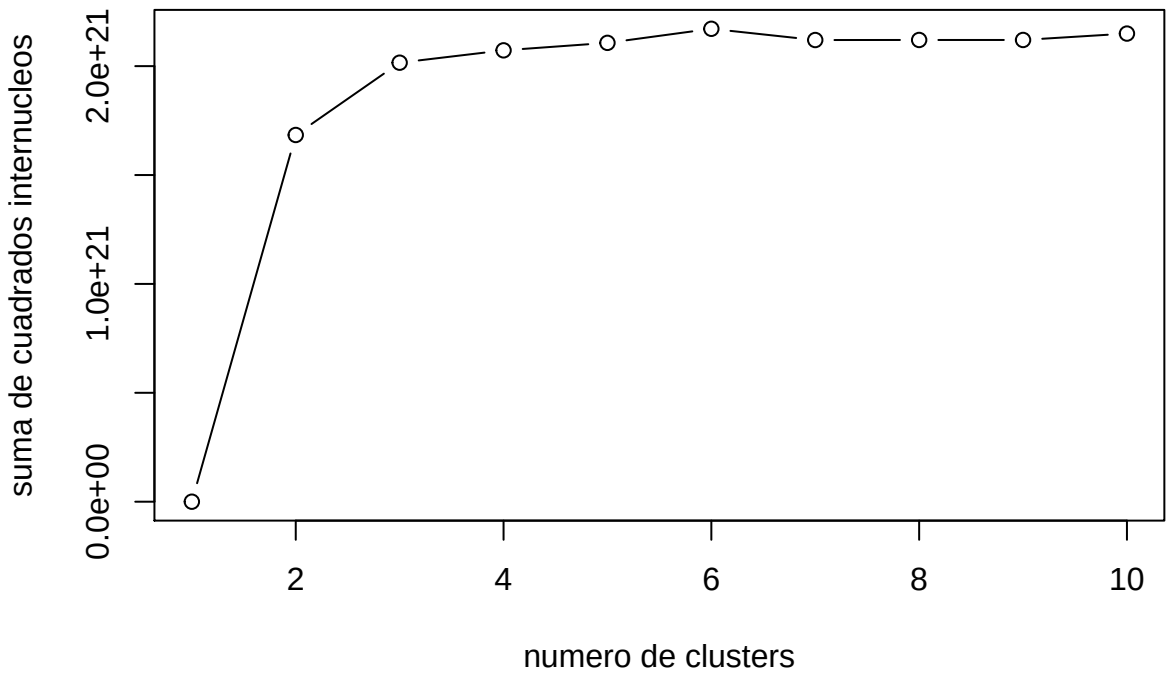
-Determinar un numero de clusters optimo, se determina que el numero de clusters optimos para trabajar es 1 sin embargo se emplearon 3 clusters.

```
sumbt=kmeans(datas[,-1],1)$betweenss
for(i in 2:10) sumbt[i] = kmeans(datas[,-1],centers = i )$betweenss
plot(1:10,sumbt,type= "b",xlab = "numero de clusters",ylab = "suma de cuadrados internucleos")
library(factoextra)
```

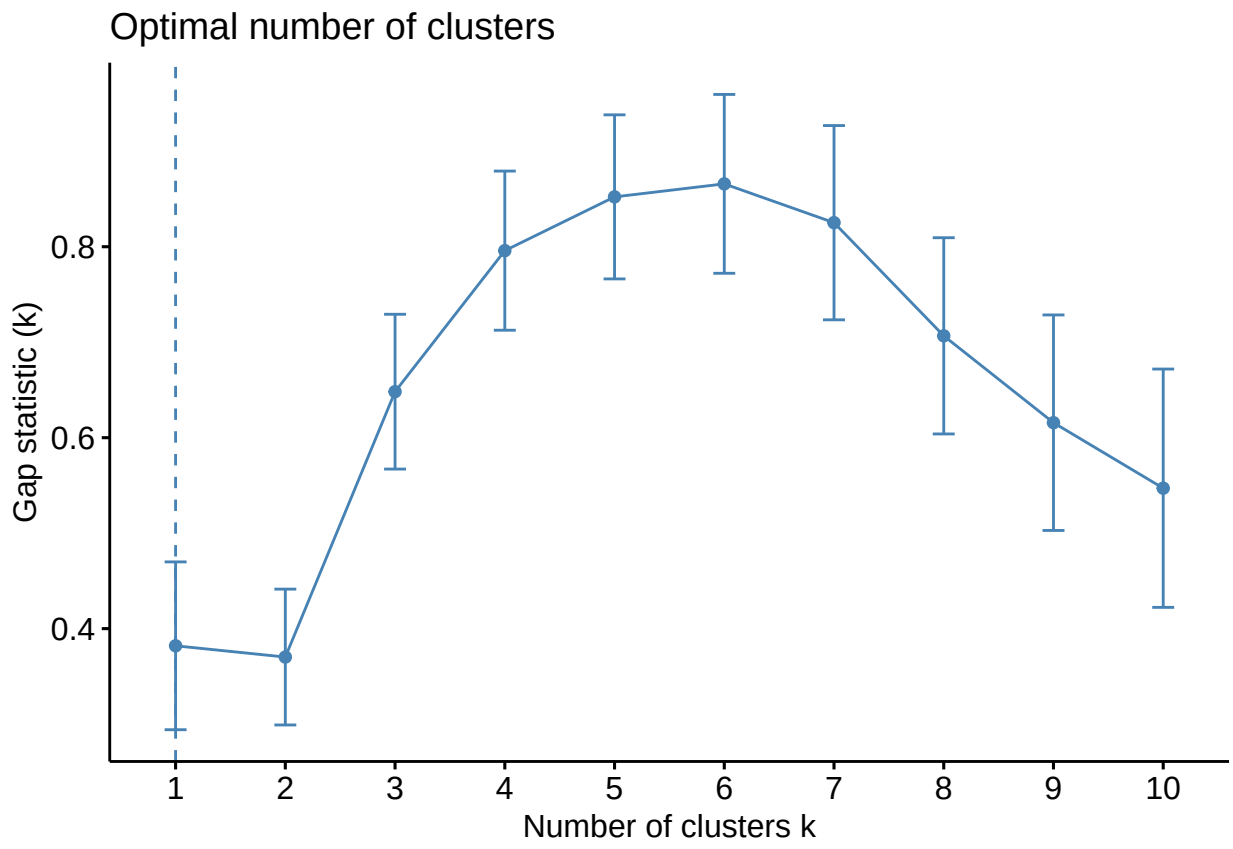
```
## Warning: package 'factoextra' was built under R version 4.0.2
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

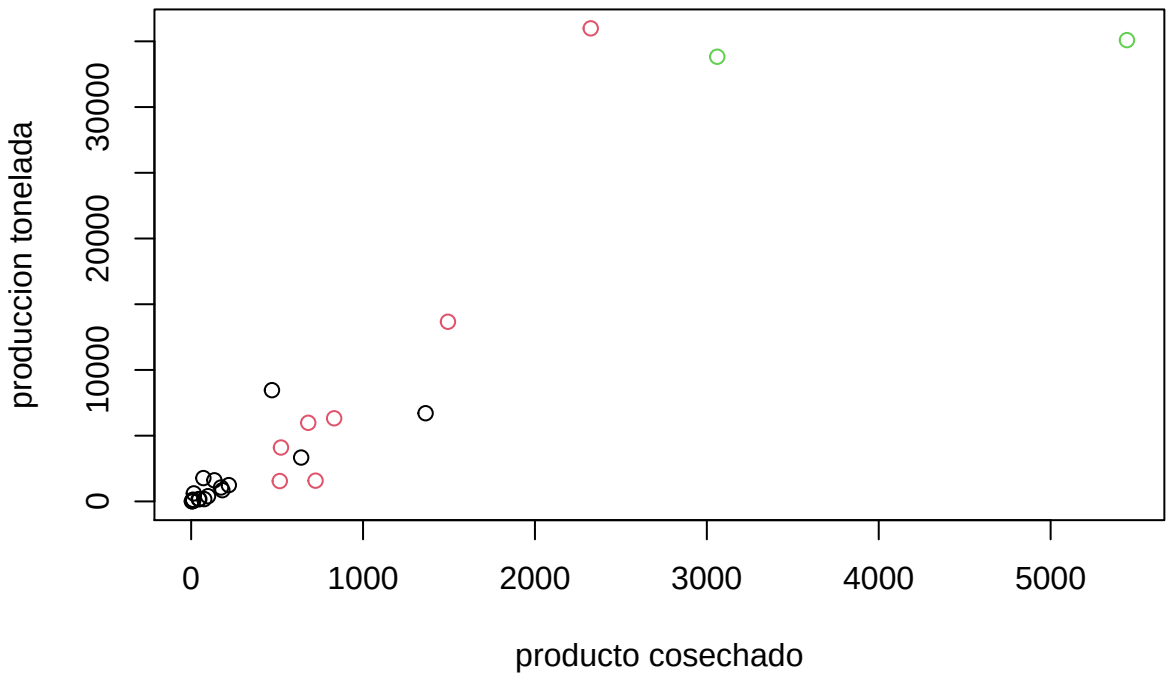


```
fviz_nbclust(datas[,-1],kmeans,method = "gap_stat")
```



inspeccionando los resultados

```
plot(datas$`SUPERFICIE, HAS COSECHADA`,datas$`PRODUCCION TON`,col=cluste$cluster,xlab = "producto cosecl
```



se observan los 3 diferentes grupos establecidos por k_means, n donde la mayoria de los datos estan en el segundo y primer grupo y son aquellos con menor producto cosechado y menor produccion tonelada

ahora vamos a inspeccionar las medias de cada cluster

se calcula la media de cada uno de los clusters con las respectivas variables, como en este caso se tienen 3 clusters entonces se saca el promedio para cada variable respectiva.

```
aggregate(datas[,-1],by = list(cluste$cluster),mean )
```

```
##   Group.1 SUPERFICIE, HAS SEMBRADA SUPERFICIE, HAS COSECHADA PRODUCCION TON
## 1       1                287.5056                198.7472       1491.311
## 2       2               1587.6214               1013.3143       9883.211
## 3       3               6376.8750               4252.2500      34460.750
##   PRECIO AL PRODUCTOR $/KG COSTO DE PRODUCCION HA INGRESO BRUTO PRODUCCION
## 1                 1462.176               5935542               1407460564
## 2                 2160.656               3634523               9016864996
## 3                  915.675               2893022              31499313552
##   COSTO TOTAL PRODUCCION
## 1              733856123
## 2             3558291091
## 3            12813341519
```

‘