Universidad Rafael Landívar

Facultad de Ingeniería

Análisis de Datos

Mgtr. Dan Stanly Bolaños Aguirre

Proyecto - Tablero Auto Partes

Gerardo Acabal # 1152418

Santiago Bocel # 1076818

Kevin Ortíz # 1242018

Guatemala, 23 de noviembre de 2021

Contenido

2. Introducción

El proyecto consiste en realizar un tablero analítico utilizando lenguajes R y Python de modelos estadísticos para obtener diferentes resultados y con esto tener una mayor visión de los resultados que se encuentran en la base de datos.

3. Contenido técnico

- Clustering

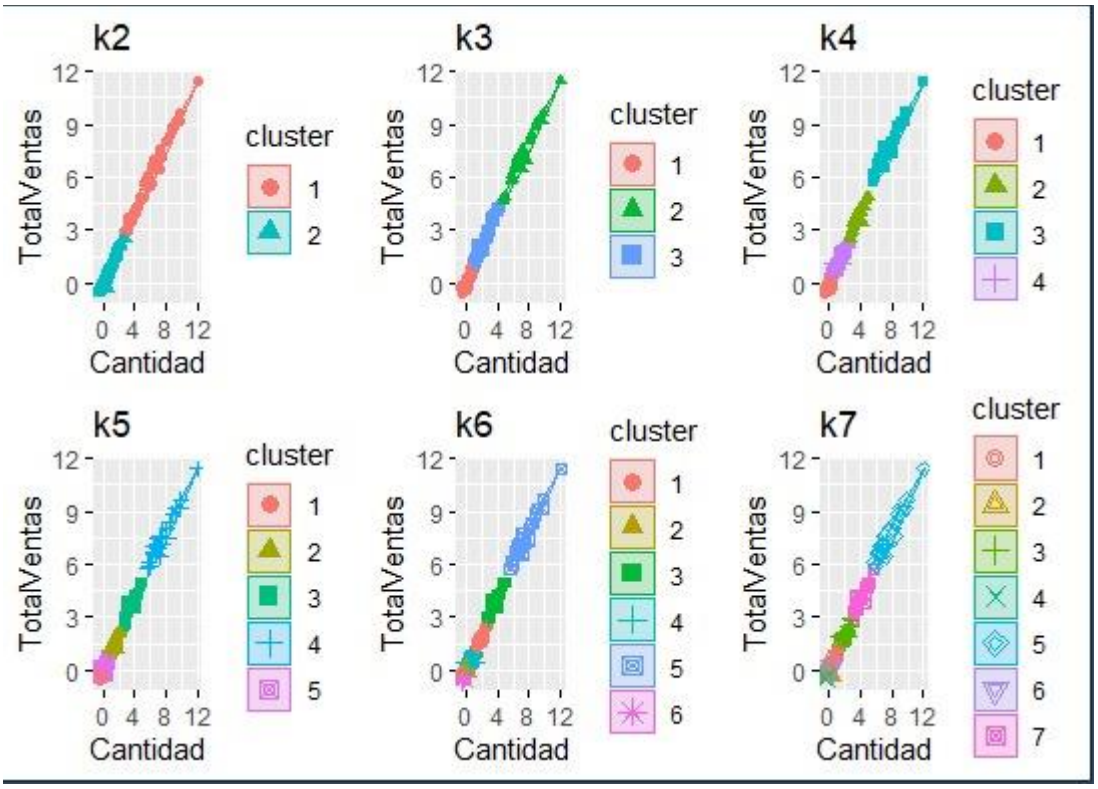
  El término clustering hace referencia a un amplio abanico de técnicas no supervisadas cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones. Las particiones se establecen de forma que, las observaciones que están dentro de un mismo grupo, son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método no supervisado, ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable). Esta característica diferencia al clustering de las técnicas supervisadas, que emplean un set de entrenamiento en el que se conoce la verdadera clasificación.

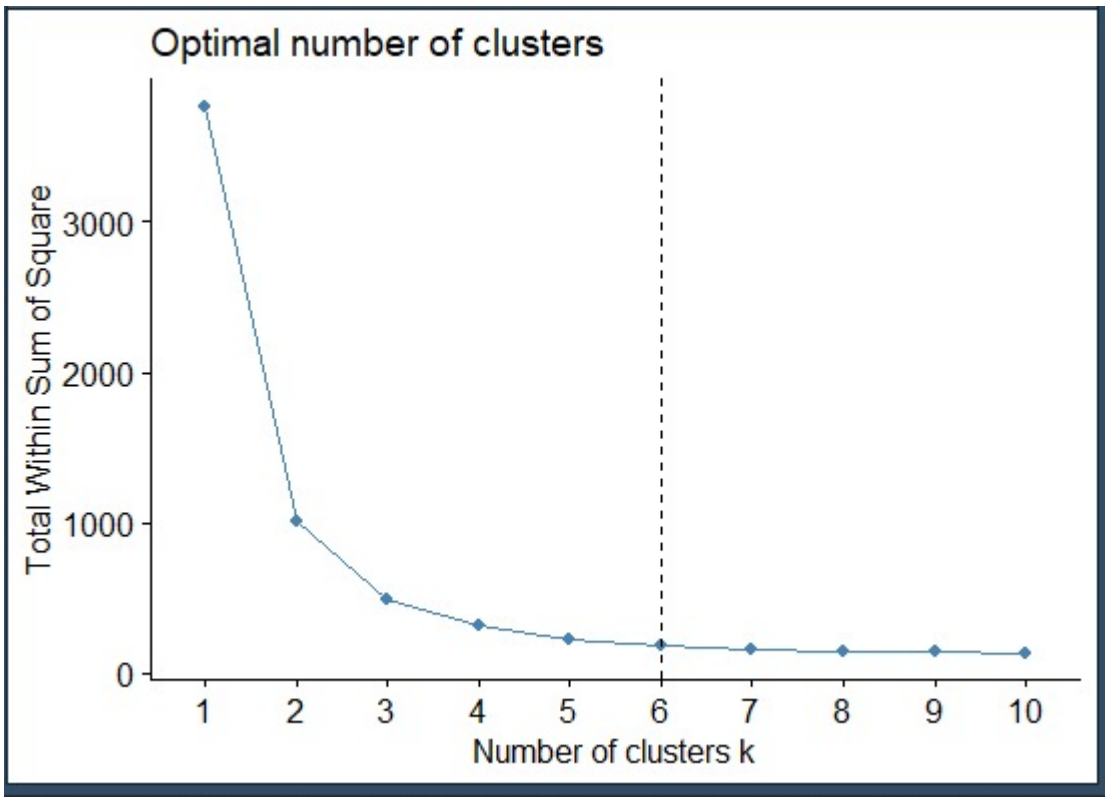
- Regresión

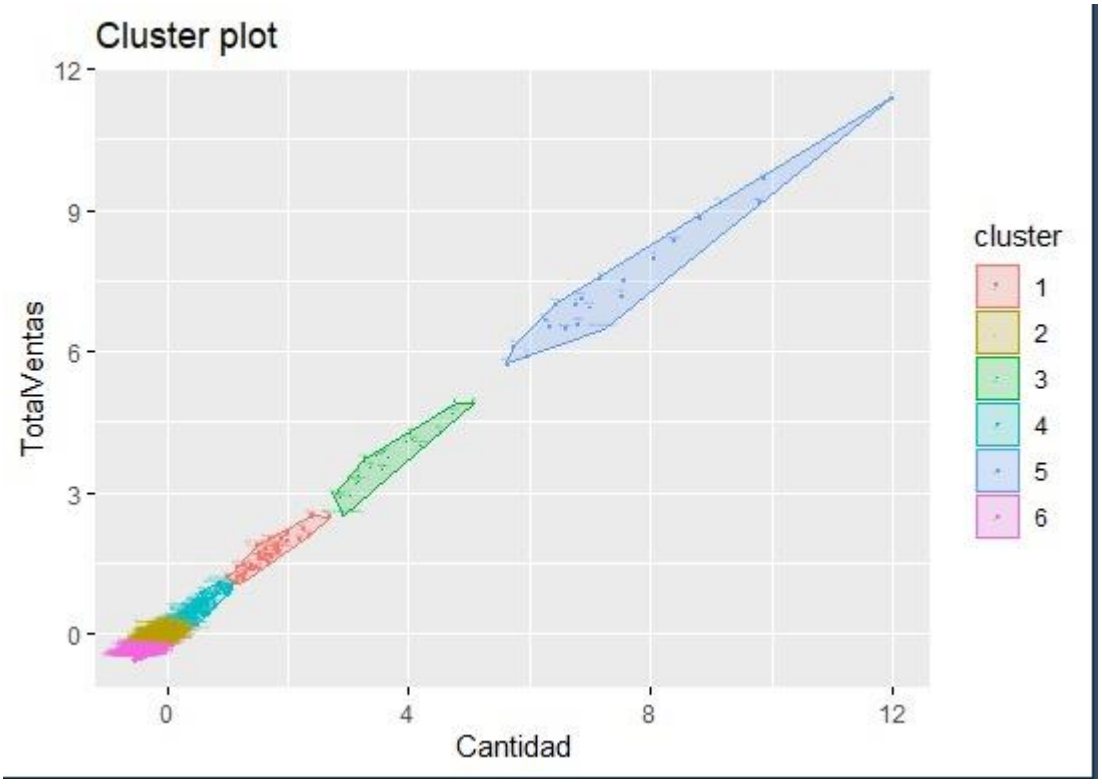
  El análisis de regresión es un enfoque estadístico ampliamente utilizado que busca identificar relaciones entre variables. La idea es agrupar datos relevantes para tomar mejores decisiones. La regresión paso a paso (stepwise) es la construcción iterativa paso a paso de un modelo de regresión que implica la selección automática de variables independientes. La disponibilidad de paquetes de software estadístico hace posible la regresión stepwise, incluso en modelos con cientos de variables.

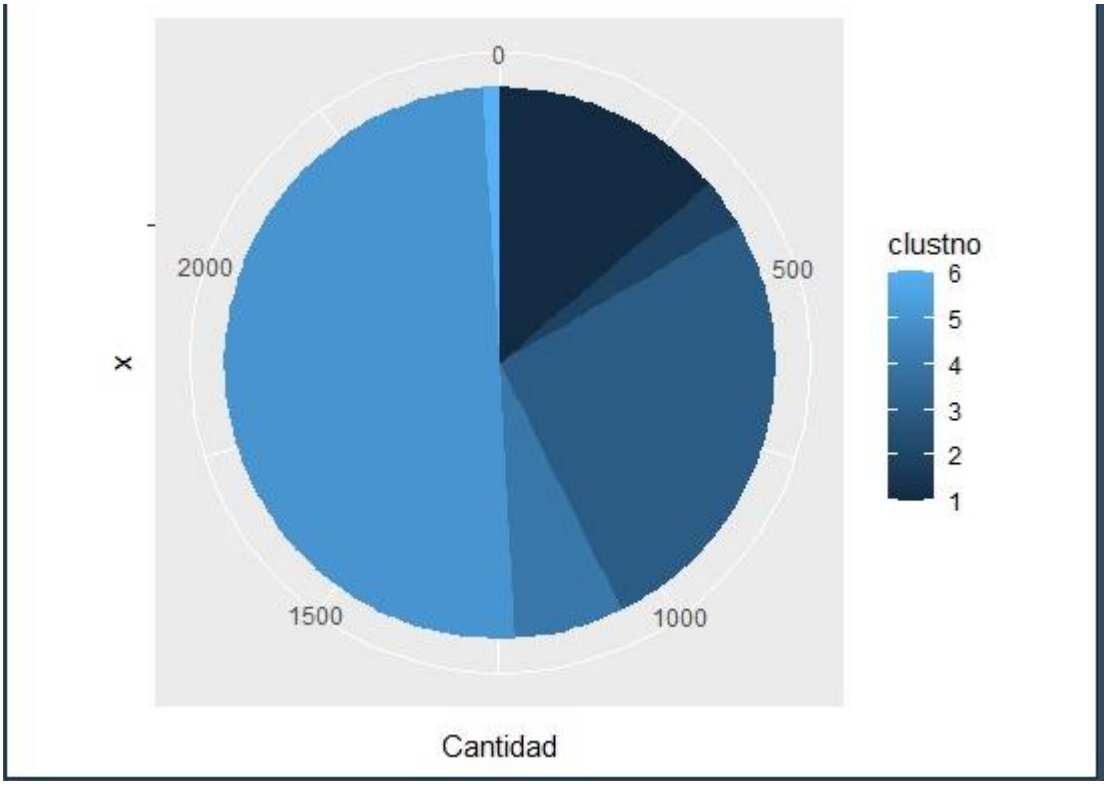
- Series de Tiempo

  En R expresamos una serie de tiempo con dos o más vectores, uno con el punto de tiempo y otro con el valor correspondiente. Es conveniente organizar estos vectores en una estructura data.frame, para asegurarnos que se mantengan alienados. Idealmente el vector índice de tiempo debe tener una fecha -y hora, minutos, segundos si aplica- completa: año, mes, día con separadores claros, respetando el mismo formato a lo largo de toda la serie. Si la marca de tiempo tiene el formato correcto es posible hacer agregar datos con facilidad.


k2
k3
k4
k5
k6
k7
TotalVentas
Cantidad
cluster
0 4 8 12
0 3 6 9 12
1 2 3 4 5 6 7


Optimal number of clusters
Total Within Sum of Square
Number of clusters k
0 1000 2000 3000
1 2 3 4 5 6 7 8 9 10


Cluster plot
TotalVentas
Cantidad
cluster
1
2
3
4
5
6
0
3
6
9
12
0
4
8
12


clustno
6
5
4
3
2
1
0
500
1000
1500
2000
x
Cantidad