Clustering

2024-12-21

Contents

kmeans

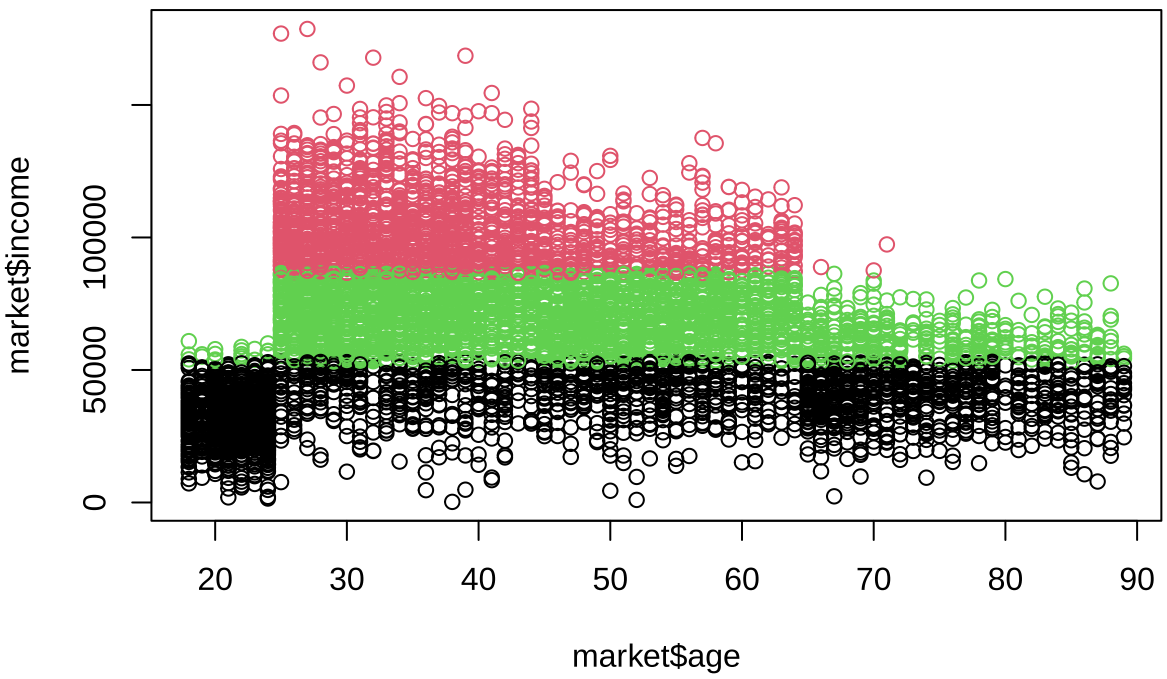
```
stations = read.csv('Ch5_bike_station_locations.csv')
two = kmeans(stations, 2)
two
```

```
## K-means clustering with 2 clusters of sizes 118, 126
##
## Cluster means:
##   latitude longitude
## 1 38.88838 -76.97846
## 2 38.93855 -77.03975
##
## Clustering vector:
##   [1] 2 1 2 1 2 2 1 1 1 1 2 1 1 1 2 2 2 2 2 1 1 2 2 1 2 1 2 2 2 1 1 1 1 2 2 2 2
##  [38] 1 2 2 1 2 2 2 1 2 1 2 1 2 1 2 1 1 1 1 1 1 2 1 2 2 2 2 1 1 1 2 1 2 1 2 2 2
##  [75] 1 2 1 2 1 2 1 2 2 2 1 2 1 2 1 1 2 1 2 1 1 2 2 1 2 2 1 1 2 2 1 2 2 2 2 2 2
## [112] 2 1 2 2 1 2 2 1 1 1 1 2 2 2 1 1 1 1 1 1 2 2 2 2 1 1 1 2 1 2 2 2 2 2 2 2 1
## [149] 2 1 2 2 1 2 2 1 2 1 1 1 1 1 1 1 2 1 2 2 2 1 2 2 2 1 2 1 1 2 2 2 1 1 1 2 1
## [186] 1 1 2 1 2 1 1 2 2 1 1 1 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 1 2 2 1 2 1 1 2 1 2
## [223] 1 1 2 1 1 2 1 1 2 1 2 2 1 1 1 2 1 2 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 0.1754263 0.1575802
##  (between_SS / total_SS =  53.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
market = read.csv('Ch5_age_income_data.csv')
head(market)
```

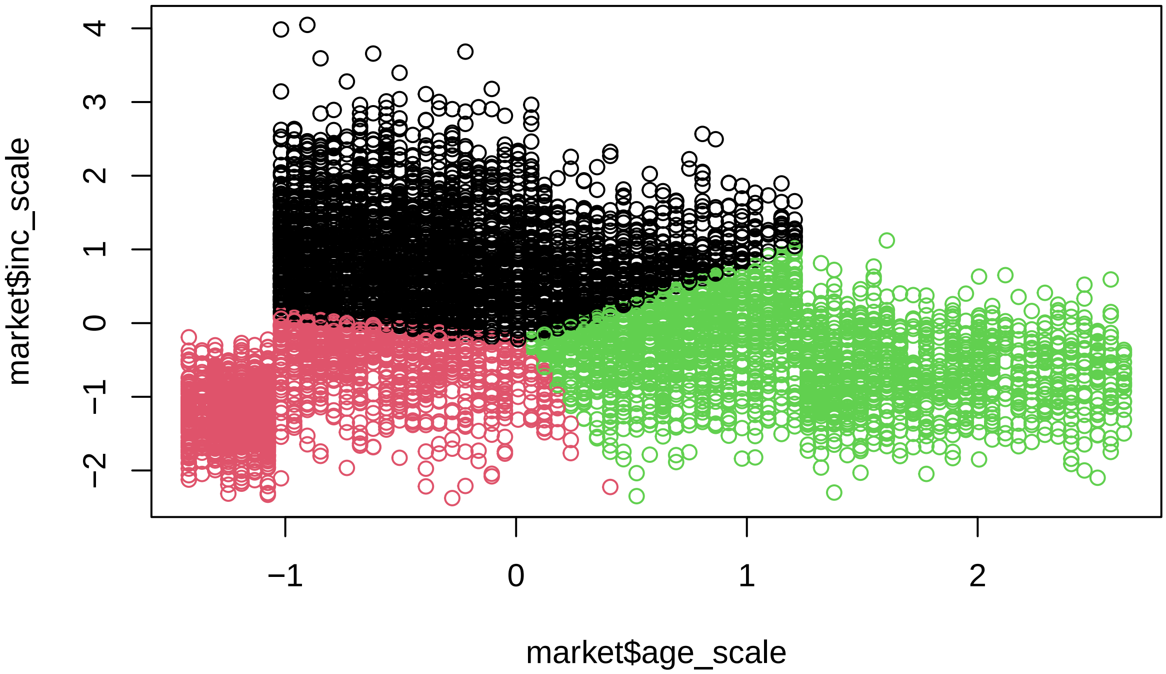
```
##     bin age   income
## 1 60-69  64 87083.24
## 2 30-39  33 76807.82
## 3 20-29  24 12043.60
## 4 30-39  33 61972.00
## 5 70-79  78 60120.32
## 6 60-69  62 40058.42
```

```
three = kmeans(market[,c(2,3)], 3)
plot(market$age, market$income, col=three$cluster)
```



```
market$age_scale = as.numeric(scale(market$age))
market$inc_scale = as.numeric(scale(market$income))
```

```
three_scale = kmeans(market[, c(4,5)],3)
plot(market$age_scale, market$inc_scale, col=three_scale$cluster)
```


market$inc_scale
4
3
2
1
0
−1
−2
−1
0
1
2
market$age_scale