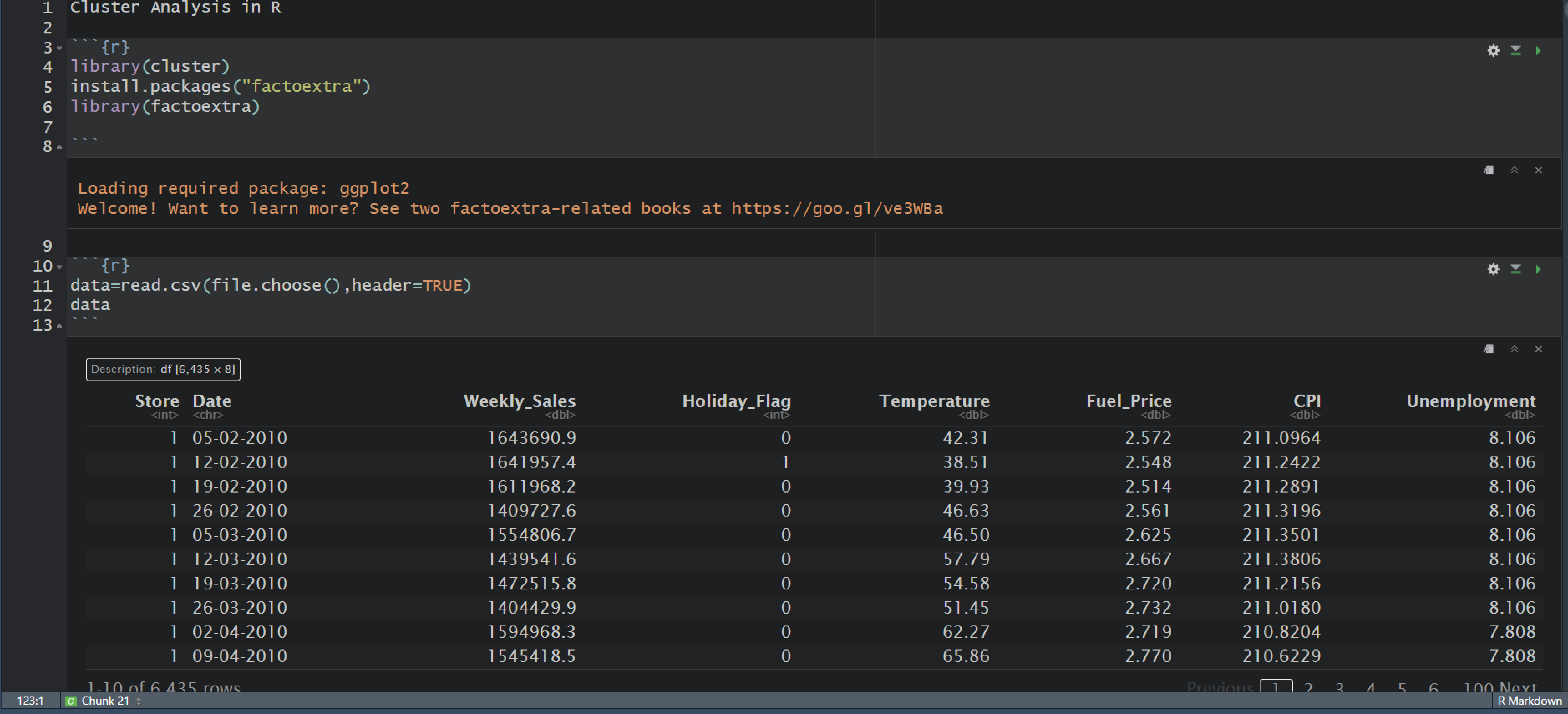


~~~
Cluster Analysis in R

```{r}
library(cluster)
install.packages("factoextra")
library(factoextra)

```
~~~

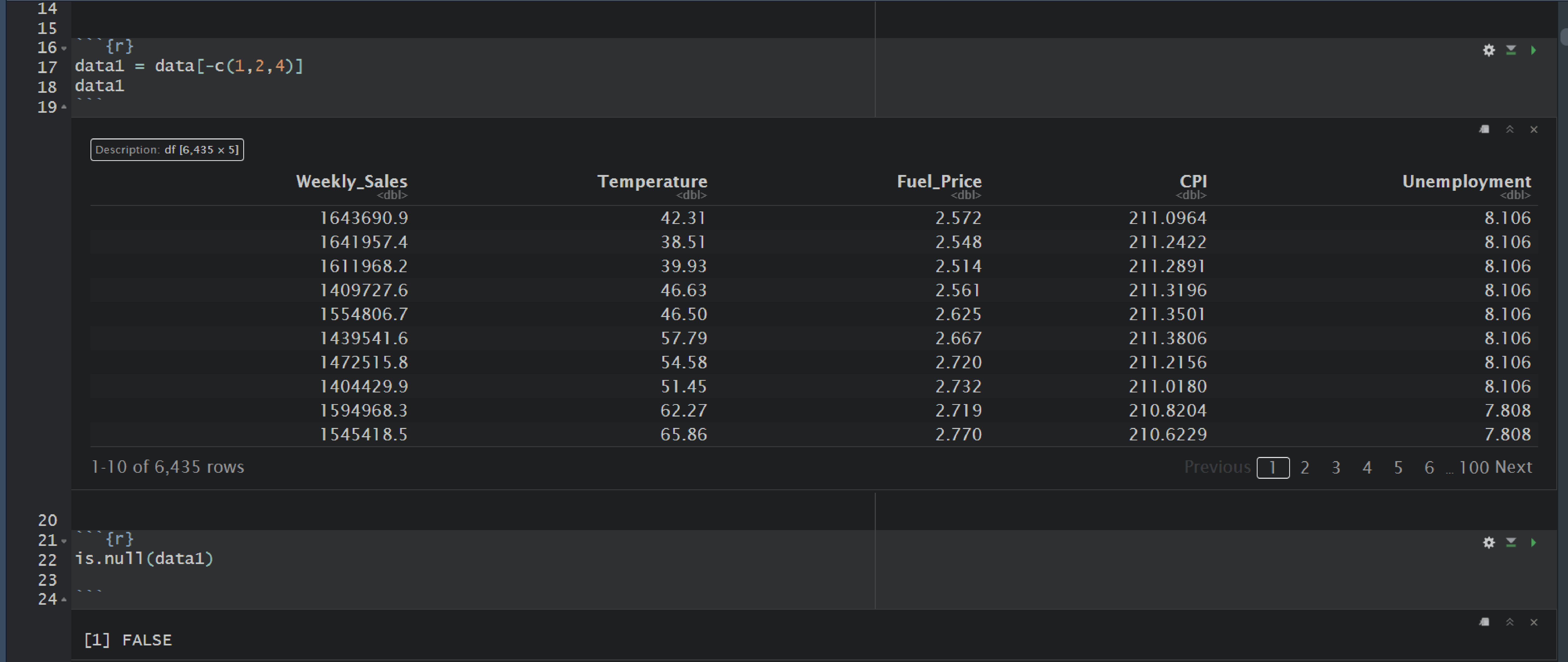
```
Loading required package: ggplot2
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

~~~
```{r}
data=read.csv(file.choose(),header=TRUE)
data
```
~~~

Description: df [6,435 x 8]

Store <int>	Date <chr>	Weekly_Sales <dbl>	Holiday_Flag <int>	Temperature <dbl>	Fuel_Price <dbl>	CPI <dbl>	Unemployment <dbl>
1	05-02-2010	1643690.9	0	42.31	2.572	211.0964	8.106
1	12-02-2010	1641957.4	1	38.51	2.548	211.2422	8.106
1	19-02-2010	1611968.2	0	39.93	2.514	211.2891	8.106
1	26-02-2010	1409727.6	0	46.63	2.561	211.3196	8.106
1	05-03-2010	1554806.7	0	46.50	2.625	211.3501	8.106
1	12-03-2010	1439541.6	0	57.79	2.667	211.3806	8.106
1	19-03-2010	1472515.8	0	54.58	2.720	211.2156	8.106
1	26-03-2010	1404429.9	0	51.45	2.732	211.0180	8.106
1	02-04-2010	1594968.3	0	62.27	2.719	210.8204	7.808
1	09-04-2010	1545418.5	0	65.86	2.770	210.6229	7.808

1-10 of 6,435 rows



~~~
```{r}
data1 = data[-c(1,2,4)]
data1
```
~~~

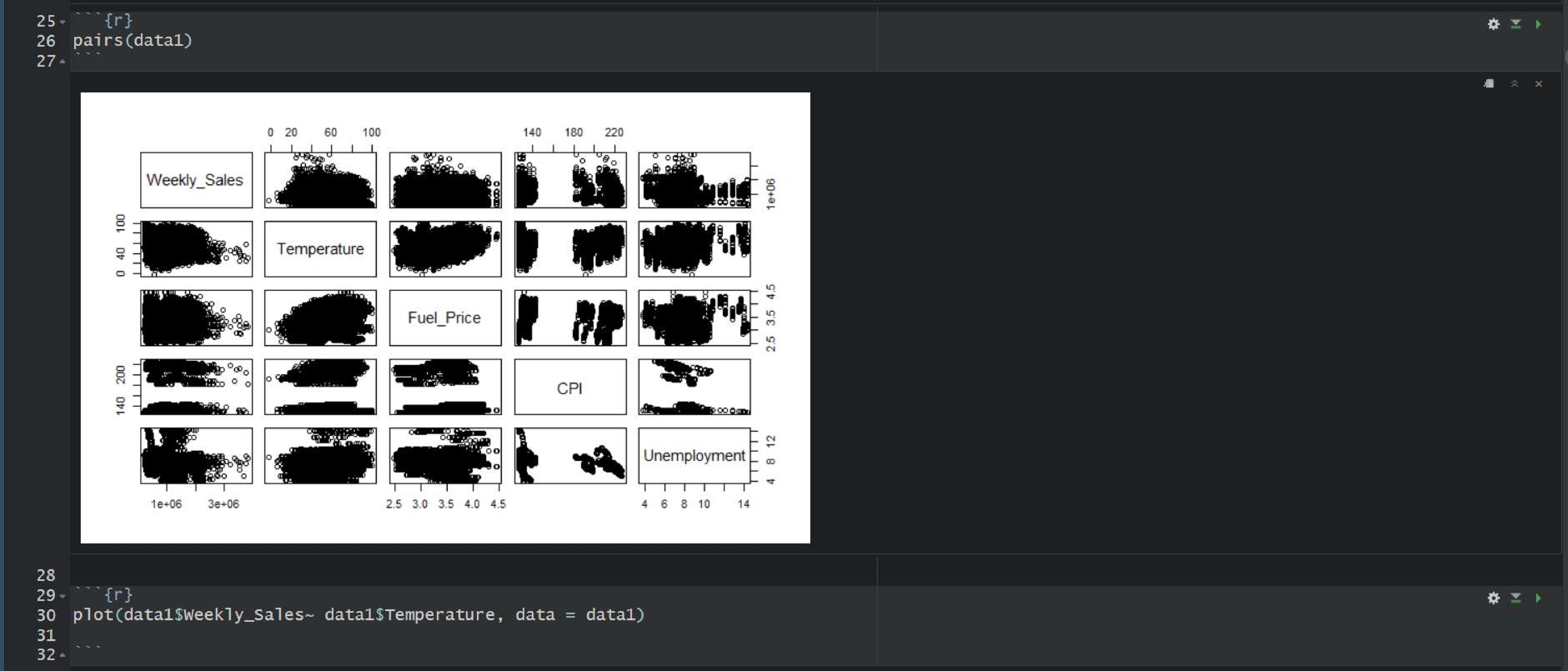
Description: df [6,435 × 5]

Weekly_Sales <dbl>	Temperature <dbl>	Fuel_Price <dbl>	CPI <dbl>	Unemployment <dbl>
1643690.9	42.31	2.572	211.0964	8.106
1641957.4	38.51	2.548	211.2422	8.106
1611968.2	39.93	2.514	211.2891	8.106
1409727.6	46.63	2.561	211.3196	8.106
1554806.7	46.50	2.625	211.3501	8.106
1439541.6	57.79	2.667	211.3806	8.106
1472515.8	54.58	2.720	211.2156	8.106
1404429.9	51.45	2.732	211.0180	8.106
1594968.3	62.27	2.719	210.8204	7.808
1545418.5	65.86	2.770	210.6229	7.808

1-10 of 6,435 rows

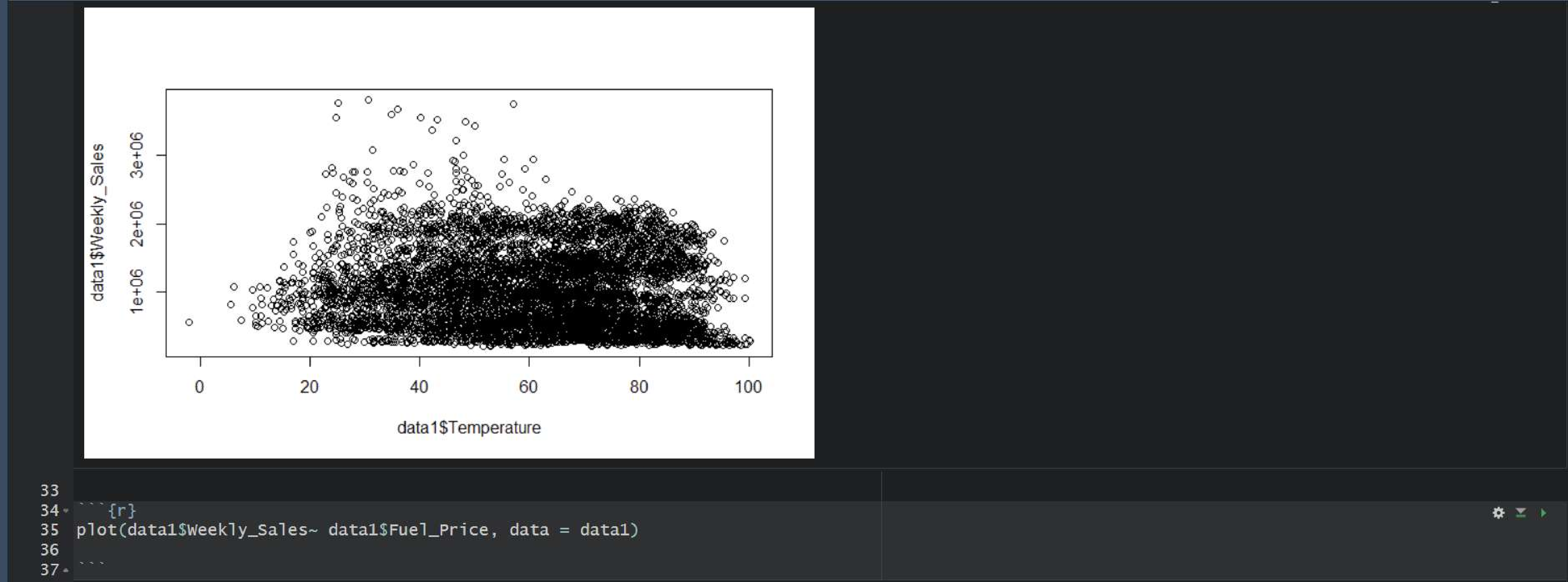
~~~
```{r}
is.null(data1)

```
~~~

```
[1] FALSE
```

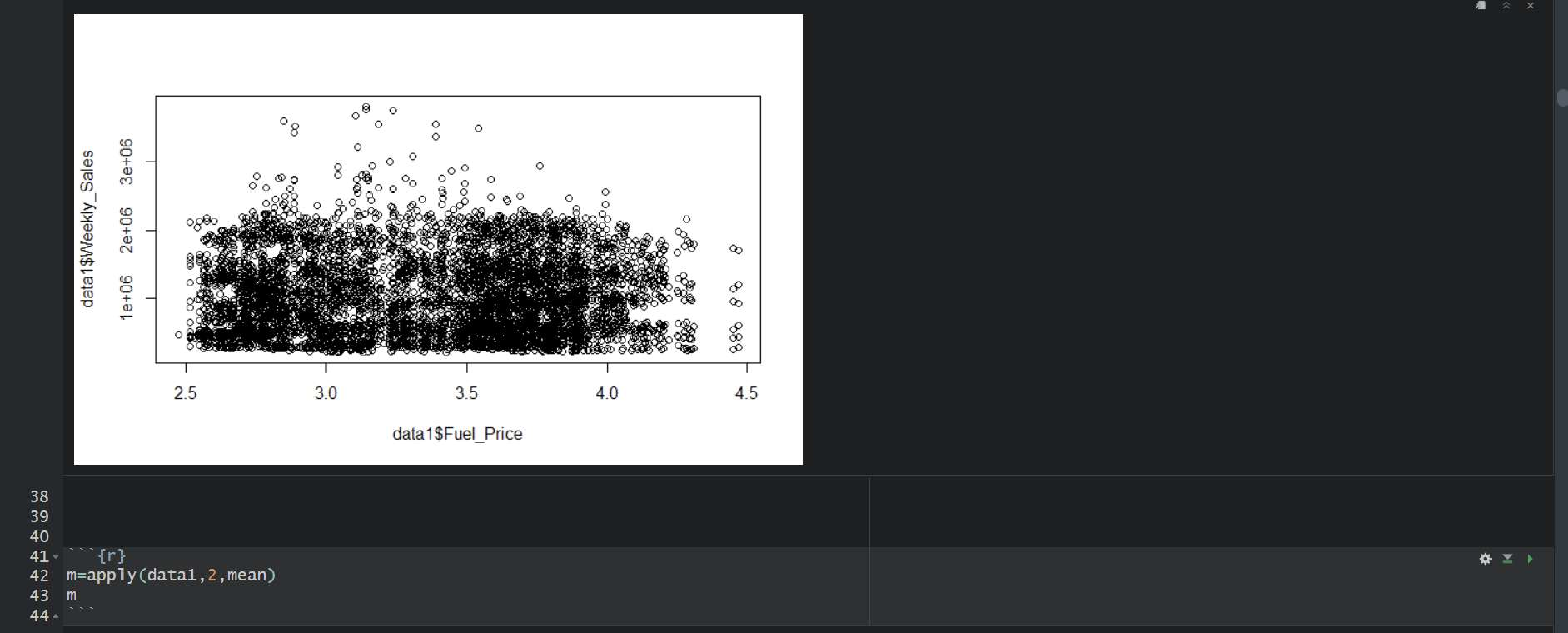


```{r}
pairs(data1)
```

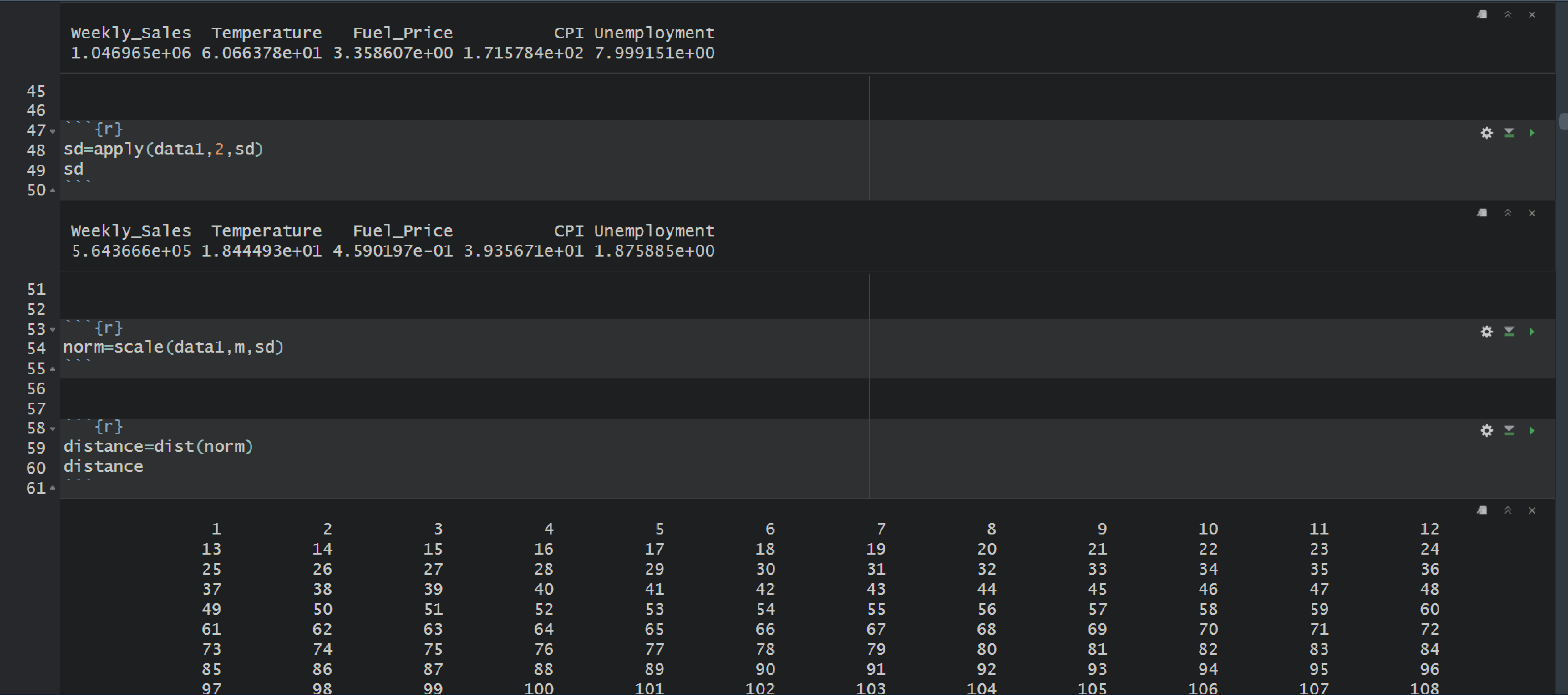
```{r}
plot(data1$Weekly_Sales~ data1$Temperature, data = data1)

```



~~~
```{r}
plot(data1$Weekly_Sales~ data1$Fuel_Price, data = data1)

```
~~~



~~~
```{r}
m=apply(data1,2,mean)
m
```
~~~



```
Weekly_Sales  Temperature   Fuel_Price          CPI Unemployment 
1.046965e+06 6.066378e+01 3.358607e+00 1.715784e+02 7.999151e+00 
```

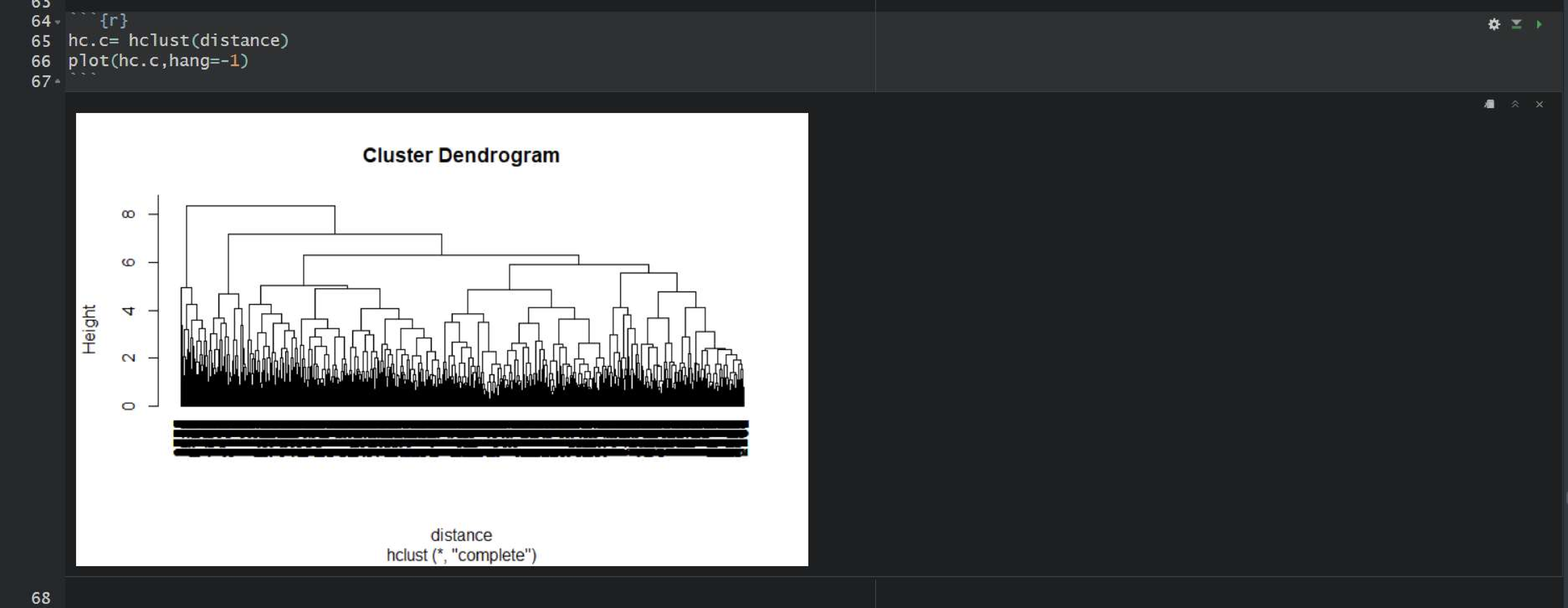
~~~
```{r}
sd=apply(data1,2,sd)
sd
```
~~~

```
Weekly_Sales  Temperature   Fuel_Price          CPI Unemployment 
5.643666e+05 1.844493e+01 4.590197e-01 3.935671e+01 1.875885e+00 
```

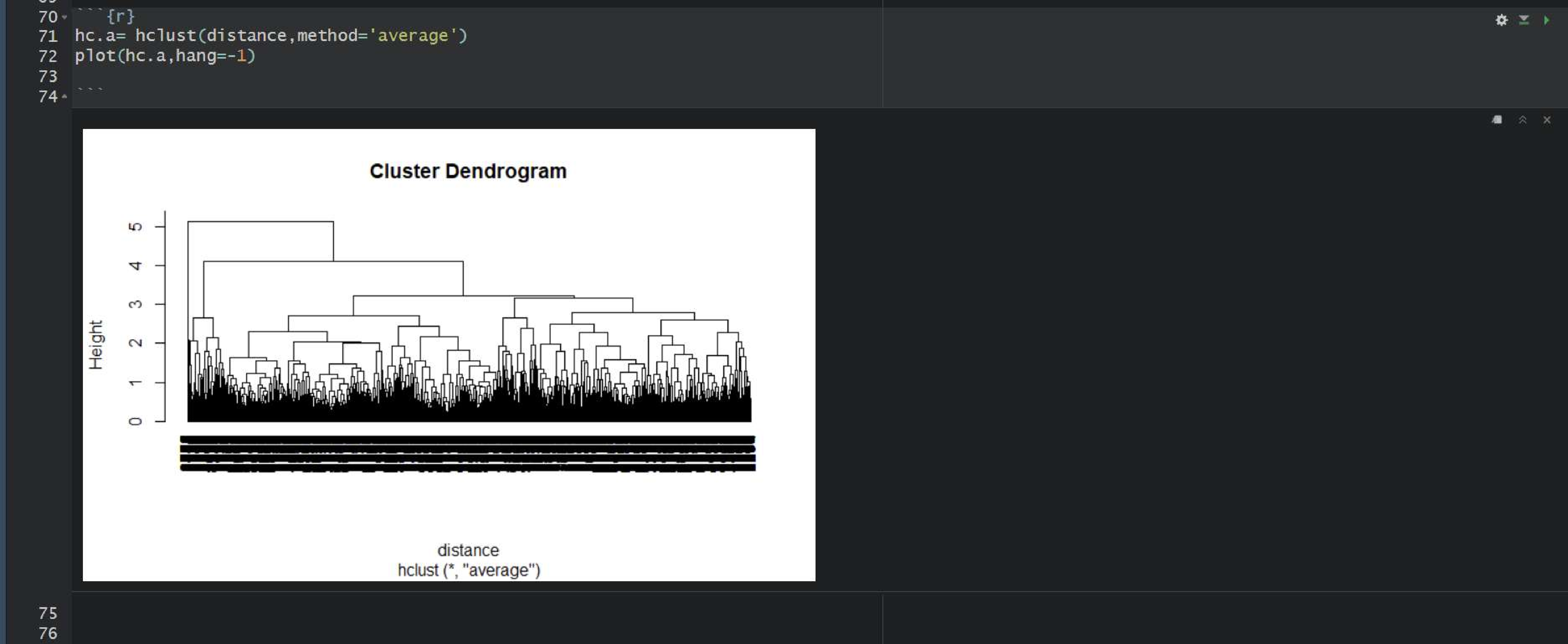
~~~
```{r}
norm=scale(data1,m,sd)
```
~~~

~~~
```{r}
distance=dist(norm)
distance
```
~~~

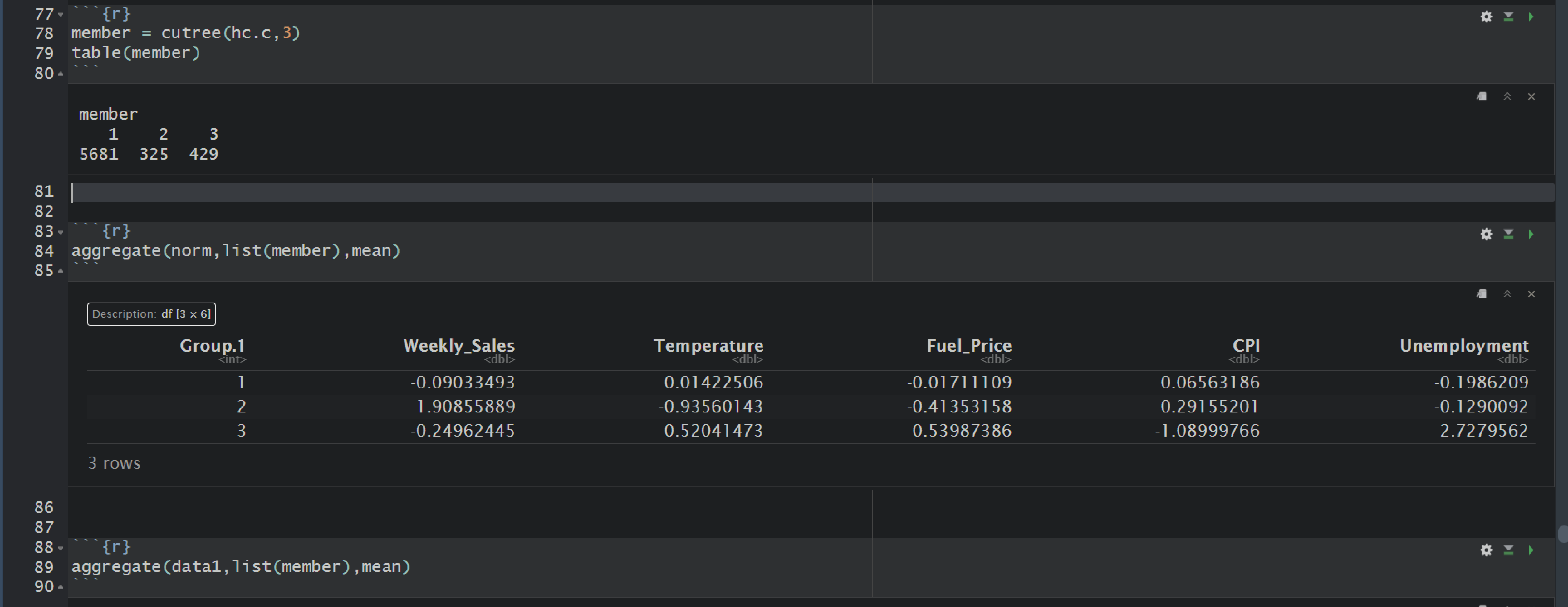
```
              1          2          3          4          5          6          7          8          9         10         11         12
             13         14         15         16         17         18         19         20         21         22         23         24
             25         26         27         28         29         30         31         32         33         34         35         36
             37         38         39         40         41         42         43         44         45         46         47         48
             49         50         51         52         53         54         55         56         57         58         59         60
             61         62         63         64         65         66         67         68         69         70         71         72
             73         74         75         76         77         78         79         80         81         82         83         84
             85         86         87         88         89         90         91         92         93         94         95         96
             97         98         99        100        101        102        103        104        105        106        107        108
```



```{r}
hc.c= hclust(distance)
plot(hc.c,hang=-1)
```



~~~
```{r}
hc.a= hclust(distance,method='average')
plot(hc.a,hang=-1)

```
~~~



```{r}
member = cutree(hc.c,3)
table(member)
```

```
member
   1    2    3 
5681  325  429 
```

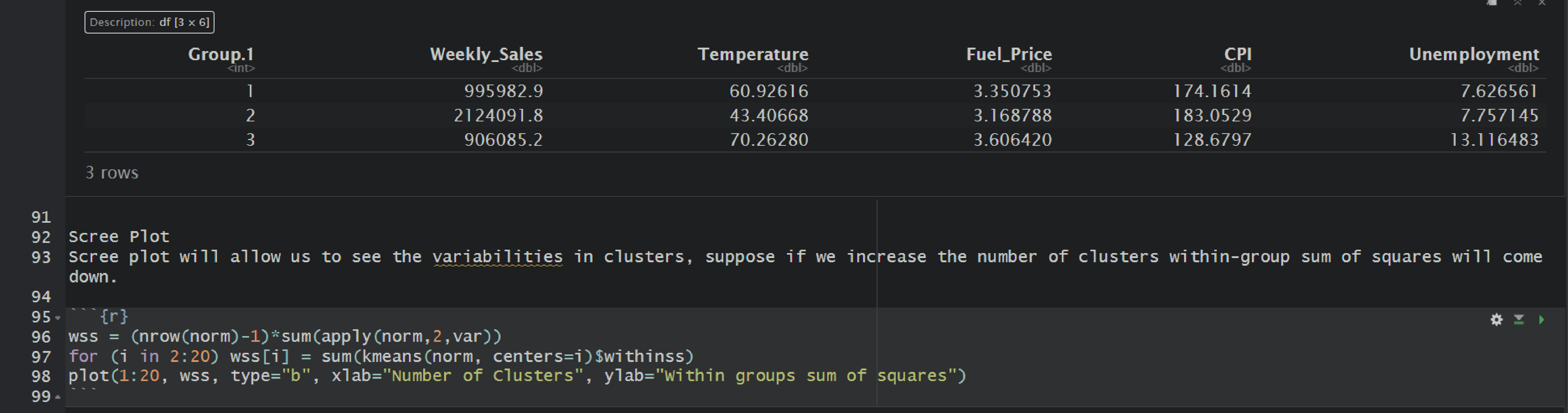
```{r}
aggregate(norm,list(member),mean)
```

Description: df [3 x 6]

Group.1 <int>	Weekly_Sales <dbl>	Temperature <dbl>	Fuel_Price <dbl>	CPI <dbl>	Unemployment <dbl>
1	-0.09033493	0.01422506	-0.01711109	0.06563186	-0.1986209
2	1.90855889	-0.93560143	-0.41353158	0.29155201	-0.1290092
3	-0.24962445	0.52041473	0.53987386	-1.08999766	2.7279562

3 rows

```{r}
aggregate(data1,list(member),mean)
```



Description: df [3 × 6]

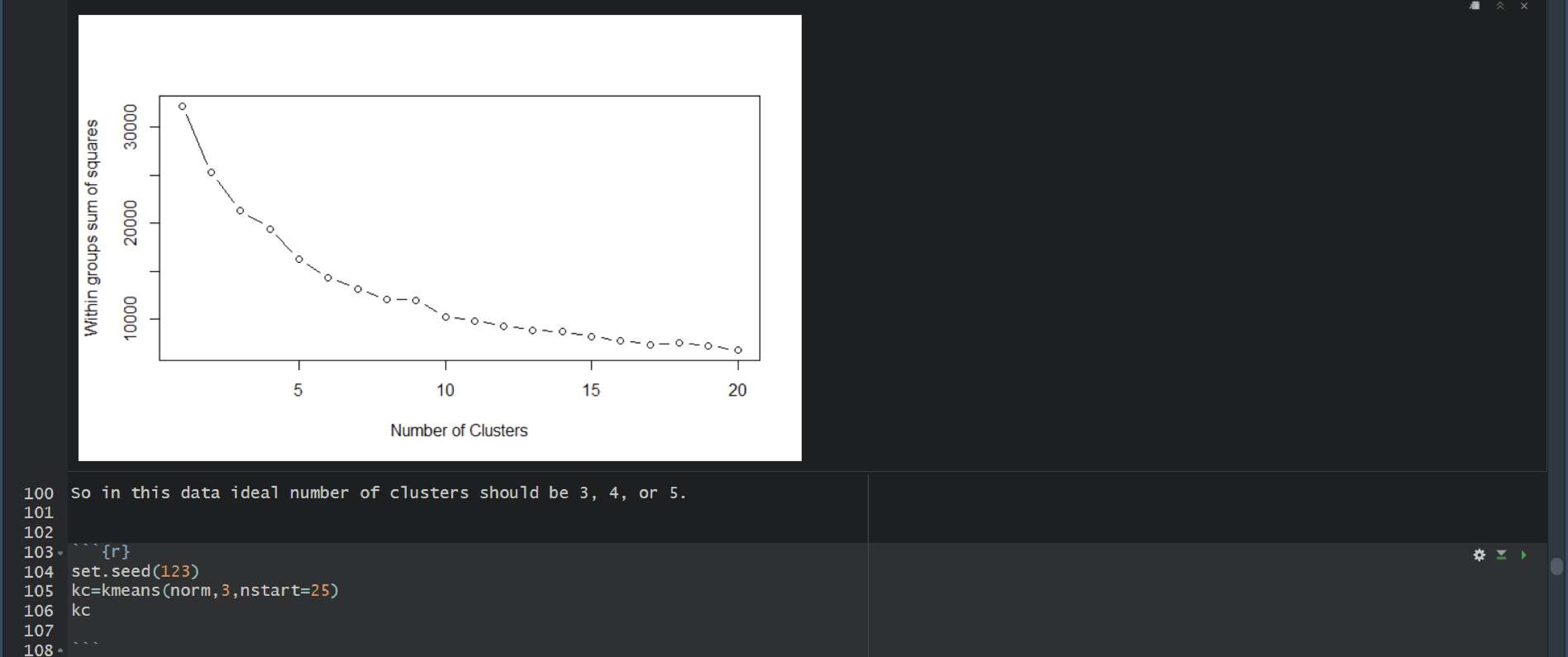
Group.1 <int>	Weekly_Sales <dbl>	Temperature <dbl>	Fuel_Price <dbl>	CPI <dbl>	Unemployment <dbl>
1	995982.9	60.92616	3.350753	174.1614	7.626561
2	2124091.8	43.40668	3.168788	183.0529	7.757145
3	906085.2	70.26280	3.606420	128.6797	13.116483

3 rows

Scree Plot

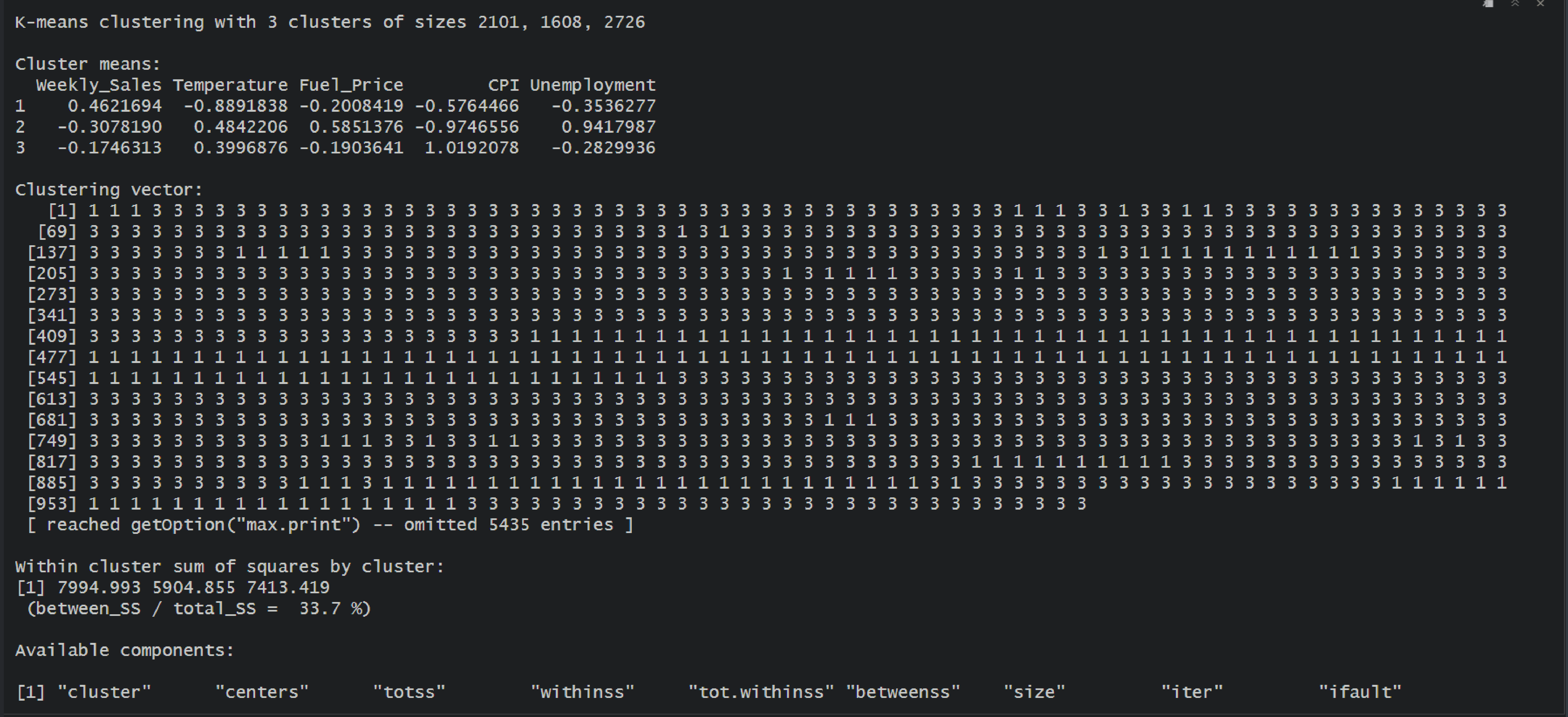
Scree plot will allow us to see the variabilities in clusters, suppose if we increase the number of clusters within-group sum of squares will come down.

~~~
```{r}
wss = (nrow(norm)-1)*sum(apply(norm,2,var))
for (i in 2:20) wss[i] = sum(kmeans(norm, centers=i)$withinss)
plot(1:20, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```
~~~

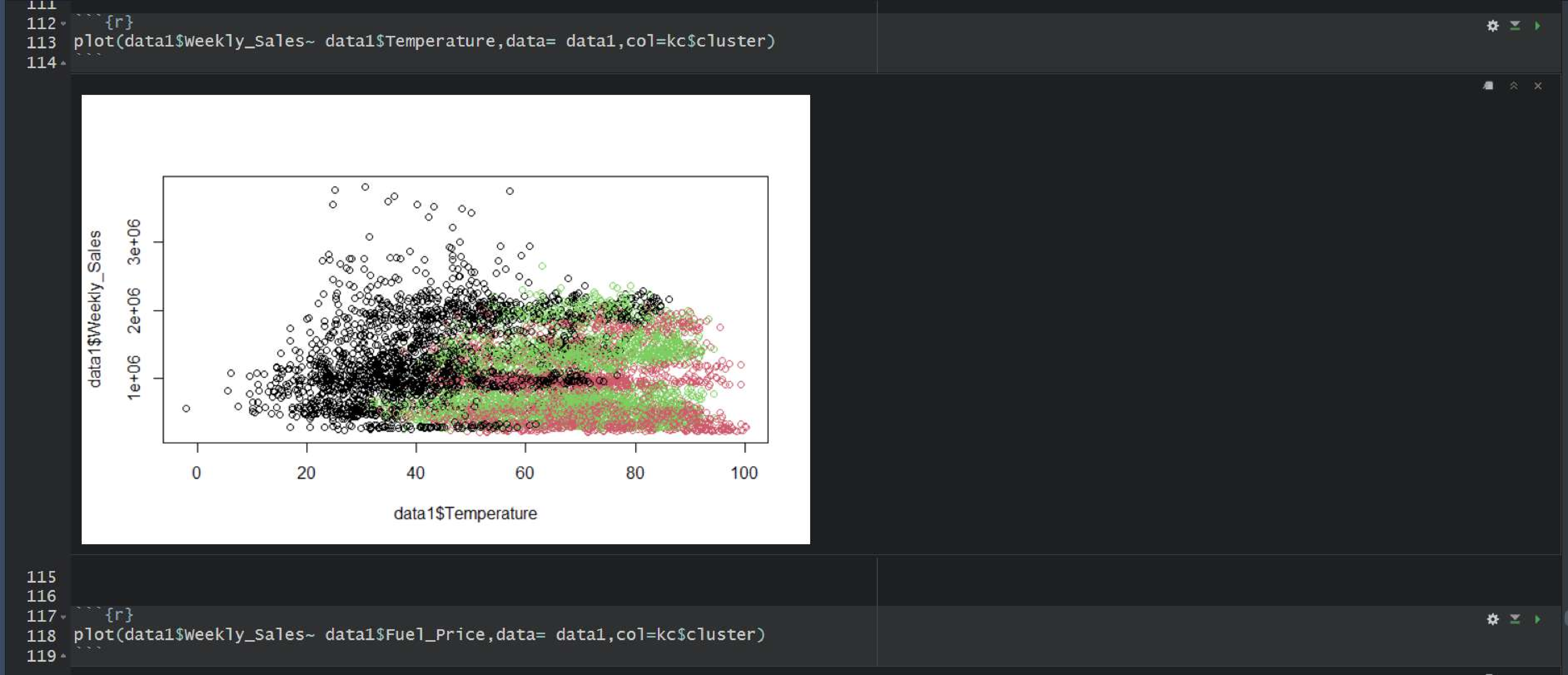


So in this data ideal number of clusters should be 3, 4, or 5.

~~~
```{r}
set.seed(123)
kc=kmeans(norm,3,nstart=25)
kc

```
~~~

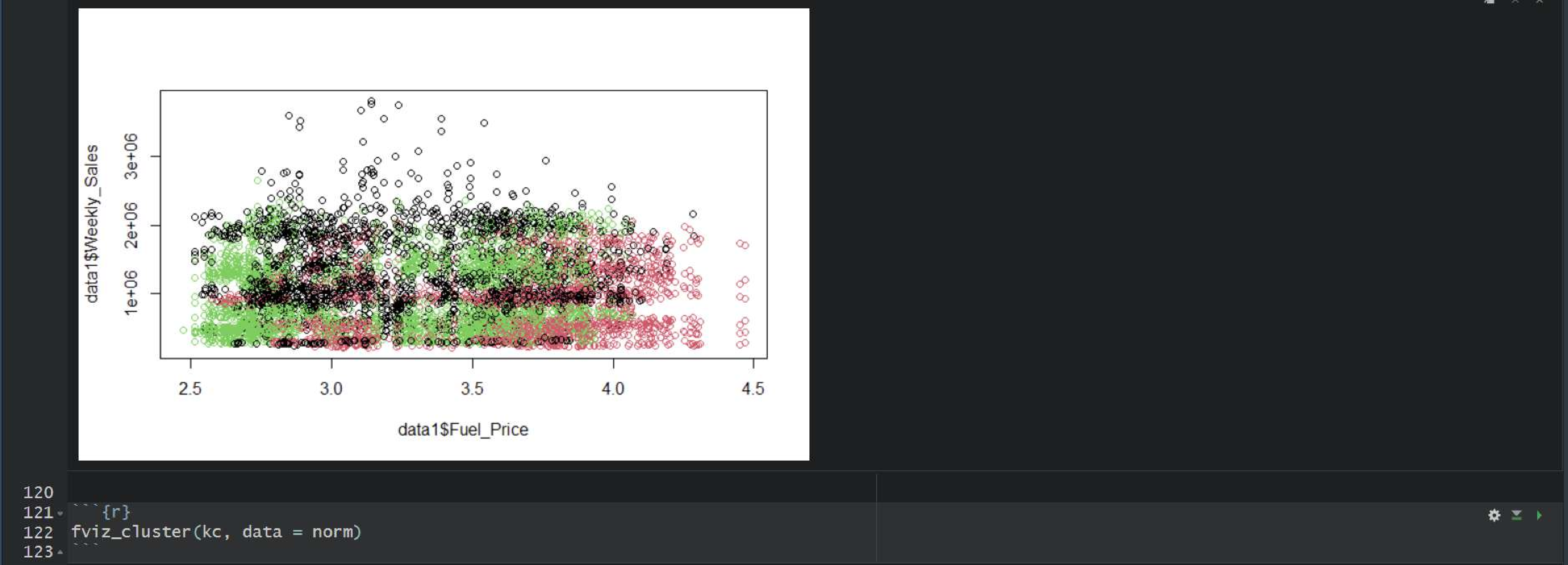


```
K-means clustering with 3 clusters of sizes 2101, 1608, 2726

Cluster means:
  Weekly_Sales Temperature Fuel_Price        CPI Unemployment
1    0.4621694  -0.8891838 -0.2008419 -0.5764466   -0.3536277
2   -0.3078190   0.4842206  0.5851376 -0.9746556    0.9417987
3   -0.1746313   0.3996876 -0.1903641  1.0192078   -0.2829936

Clustering vector:
   [1] 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 1 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3
  [69] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [137] 3 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3
 [205] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [273] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [341] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [409] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [477] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [545] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [613] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [681] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [749] 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 1 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3
 [817] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [885] 3 3 3 3 3 3 3 3 3 3 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1
 [953] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [ reached getOption("max.print") -- omitted 5435 entries ]

Within cluster sum of squares by cluster:
[1] 7994.993 5904.855 7413.419
 (between_SS / total_SS =  33.7 %)

Available components:

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

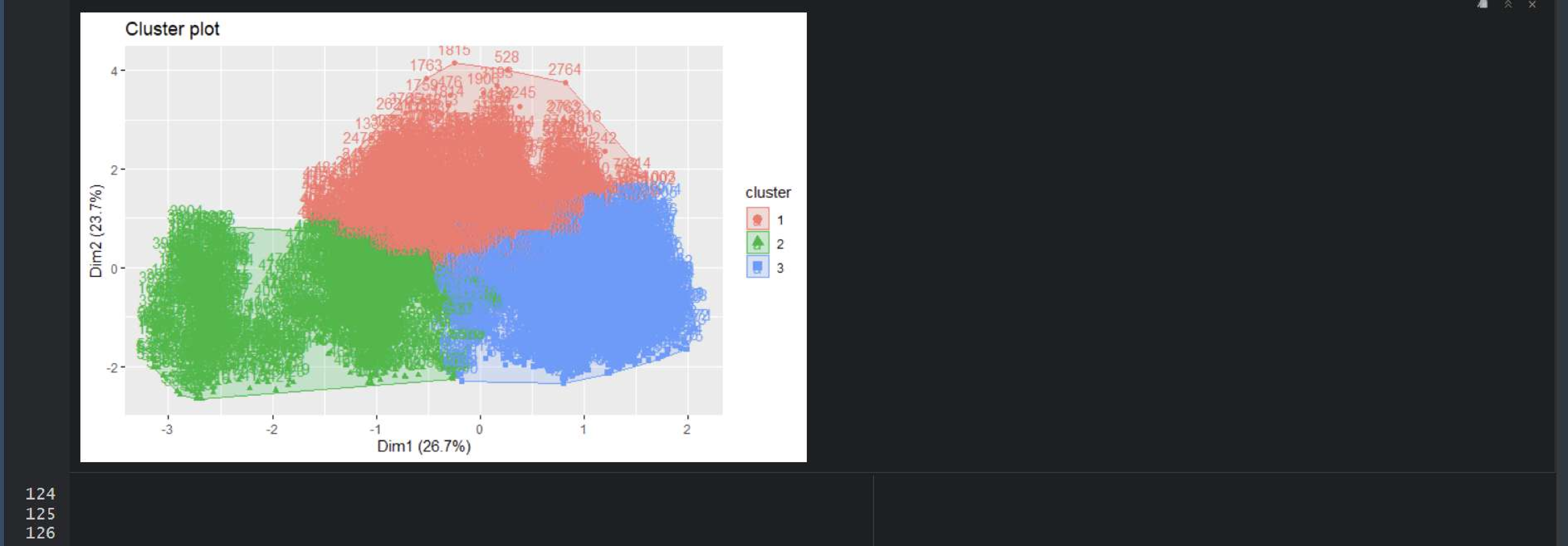


```{r}
plot(data1$Weekly_Sales~ data1$Temperature,data= data1,col=kc$cluster)
```

```{r}
plot(data1$Weekly_Sales~ data1$Fuel_Price,data= data1,col=kc$cluster)
```



~~~
```{r}
fviz_cluster(kc, data = norm)
```
~~~


Cluster plot
1815
1763
528
2764
1906
2245
Dim2 (23.7%)
4
2
0
-2
-3
-2
-1
0
1
2
Dim1 (26.7%)
cluster
1
2
3