Mobile Price dataset

Principal Components Analysis

Submitted by

Chirat Suwannachote 665020010-1

Present to

Dr.Prem Junsawang

This report is a part of the Multivariate Analysis Subject

Semester 1 2023 Khon Kean University

Data background

ข้อมูลที่นำมาใช้นำในงานครั้งนี้มีชื่อว่า ‘Mobile Price Classification’ จากเว็บ Kaggle เป็นข้อมูลเกี่ยวกับข้อมูลพื้นฐานของโทรศัพท์เครื่องที่ มีจำนวนทั้งสิ้น 2000 แถว และแบ่งการเก็บข้อมูลเป็นตัวแปร 21 ตัว ได้แก่

1. ID หมายถึง รหัสระบุลำดับของตัวอย่าง

2. battery_power หมายถึง พลังงานสูงสุดที่เก็บไว้ได้ในหนึ่งครั้ง

3. blue หมายถึง มีฟังก์ชัน Bluetooth หรือไม่

4. clock_speed หมายถึง ความเร็วในการประมูลผลของชิป

5. dual_sim หมายถึง รองรับสองซิม หรือไม่

6. fc หมายถึง คุณภาพของกล้องหน้า(หน่วย:ล้านพิกเซล)

7. four_g หมายถึง รองรับ 4G หรือไม่

8. int_memory หมายถึง หน่วยความจำ(หน่วย:กิกะไบท์)

9. m_dep หมายถึง ความหนา(หน่วย:เซนติเมตร)

10. mobile_wt หมายถึง น้ำหนัก(หน่วย:กรัม)

11. n_cores หมายถึง จำนวนของ Core

12. pc หมายถึง คุณภาพของกล้องหลัก(หน่วย:ล้านพิกเซล)

13. px_height หมายถึง ความสูงของการประมวลผลพิกเซล

14. px_width หมายถึง ความกว้างของการประมวลผลพิกเซล

15. ram หมายถึง ขนาดของ RAM

16. sc_h หมายถึง ความสูงของหน้าจอ(หน่วย:เซนติเมตร)

17. sc_w หมายถึง ความกว้างของหน้าจอ(หน่วย:เซนติเมตร)

18. talk_time หมายถึง เวลาที่นานที่สุดในการทำงานด้วยการชาร์จพลังงานหนึ่งครั้ง

19. three_g หมายถึง รองรับ 3G หรือไม่

20. touch_screen หมายถึง มี touch screen หรือไม่

21. wifi หมายถึง มี wifi หรือไม่

Principal Component Analysis (PCA)

PCA เป็นวิธีการลด Dimension ของ Dataset ที่มีขนาดใหญ่ ด้วยการแปลง Variables ที่มีจำนวนมาก ให้มีจำนวนน้อยลงแต่ยัง Contains ข้อมูลส่วนใหญ่ของชุดข้อมูลไว้ได้

การลดจำนวน Variables ของชุดข้อมูลย่อมแลกมาด้วยการสูญเสียความแม่นยำเล็กน้อย อย่างไรก็ตามการลด Dimension ของข้อมูลจะช่วยให้การวิเคราะห์ง่ายและสะดวกมากขึ้น เนื่องจากชุดข้อมูลที่มีขนาดเล็กกว่านั้นง่ายต่อการ Explore และ Visualize การวิเคราะห์ข้อมูลจึงรวดเร็วมากขึ้นสำหรับ Machine Learning Algorithms โดยไม่ต้องประมวลผล Variables จำนวนมาก

Clustering

การแบ่งกลุ่มข้อมูล(clustering) เป็นเทคนิคการแบ่งข้อมูลออกด้วยการใช้ โมเดลต่าง ๆหาความสัมพันธ์หรือรูปแบบของข้อมูลในแต่ละตัวแปร เพื่อจำแนกออกเป็นกลุ่มต่าง ๆ โดยใช้ลักษณะของข้อมูลที่ค้นพบ

PCA in R

เรียกใช้ library ที่จำเป็นสำหรับการ clustering

```
library(DataExplorer)
library(ClusterR)
library(cluster)
library(ggfortify)
library(stats)
library(fpc)
library(factoextra)
library(corrplot)
```

Data set

battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g
842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0
1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1
563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1
615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1
1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1
1859	0	0.5	1	3	0	22	0.7	164	1	7	1004	1654	1067	17	1	10	1
1821	0	1.7	0	4	1	10	0.8	139	8	10	381	1018	3220	13	8	18	1
1954	0	0.5	1	0	0	24	0.8	187	4	0	512	1149	700	16	3	5	1
1445	1	0.5	0	0	0	53	0.7	174	7	14	386	836	1099	17	1	20	1
509	1	0.6	1	2	1	9	0.1	93	5	15	1137	1224	513	19	10	12	1
769	1	2.9	1	0	0	9	0.1	182	5	1	248	874	3946	5	2	7	0
1520	1	2.2	0	5	1	33	0.5	177	8	18	151	1005	3826	14	9	13	1
1815	0	2.8	0	2	0	33	0.6	159	4	17	607	748	1482	18	0	2	1

ข้อมูลที่ผ่านการ standardize เพื่อให้ข้อตัวแปรอยู่ในมาตรฐานเดียวกัน

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc
[1,]	-0.902371578	-0.989802	0.83057170	-1.0189292	-0.76230402	-1.0437046	-1.38029832	0.340654310	1.348911455	-1.1016958	-1.30542363
[2,]	-0.495014765	1.009798	-1.25275089	0.9809318	-0.99264214	0.9576465	1.154735432	0.687376254	-0.120029419	-0.6646016	-0.64582728
[3,]	-1.537302029	1.009798	-1.25275089	0.9809318	-0.53196589	0.9576465	0.493422277	1.380820141	0.134210348	0.2095866	-0.64582728
[4,]	-1.418963737	1.009798	1.19821686	-1.0189292	-0.99264214	-1.0437046	-1.214970039	1.034098198	-0.261273734	0.6466808	-0.15113001
[5,]	1.325574343	1.009798	-0.39401218	-1.0189292	2.00175345	0.9576465	0.658750566	0.340654310	0.021214896	-1.1016958	0.67336542
[6,]	1.412052326	-0.989802	-1.25275089	0.9809318	-0.30162777	-1.0437046	-0.553656884	0.687376254	0.670938744	-1.5387899	-0.48092819
[7,]	1.325574343	-0.989802	0.21782976	-1.0189292	-0.07128965	0.9576465	-1.214970039	1.034098198	-0.035282830	1.5208690	0.01376907
[8,]	1.628247283	-0.989802	-1.25275089	0.9809318	-0.99264214	-1.0437046	-0.443438025	1.034098198	1.320662592	-0.2275075	-1.63522180
[9,]	0.469897462	1.009798	-1.25275089	-1.0189292	-0.99264214	-1.0437046	1.154735432	0.687376254	0.953427374	1.0837749	0.67336542
[10,]	-1.660191794	1.009798	-1.13020250	0.9809318	-0.53196589	0.9576465	-1.270079468	-1.392955409	-1.334730526	0.2095866	0.83826451
[11,]	-1.068500334	1.009798	1.68841041	0.9809318	-0.99264214	-1.0437046	-1.270079468	-1.392955409	1.179418277	0.2095866	-1.47032271
[12,]	0.640577691	1.009798	0.83057170	-1.0189292	0.15904847	0.9576465	0.052546841	-0.006067634	1.038173963	1.5208690	1.33296177
[13,]	1.311919925	-0.989802	1.56586202	-1.0189292	-0.53196589	-1.0437046	0.052546841	0.340654310	0.529694429	-0.2275075	1.16806269
[14,]	-0.991125297	1.009798	0.70802331	-1.0189292	0.61972472	-1.0437046	-0.829204032	1.727542085	1.631400085	-0.2275075	0.17866816

เนื่องจากข้อมูลของเรานั้นไม่มี missing data หรือความผิดพลาดอื่น ๆ จึงข้ามส่วนของการทำความสะอาดข้อมูล

PCA dataset โดยเราจะเลือกเก็บองค์ประกอบที่ครอบคลุมข้อมูลส่วนใหญ่ ซึ่งก็คือ PC14 ซึ่งมีค่า cumulative ที่ 0.83005 ซึ่งหมายถึง ข้อมูลจนถึงจุดนี้ เป็นข้อมูล 83% ของข้อมูลทั้งหมด

```
> res.pca = prcomp(standD_data, scale = TRUE)
> print(summary(res.pca))
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15
Standard deviation     1.41966 1.29492 1.26968 1.24303 1.19463 1.05789 1.03236 1.02514 1.00958 1.00738 0.99430 0.98805 0.9839 0.97559 0.9558
Proportion of Variance 0.09597 0.07985 0.07677 0.07358 0.06796 0.05329 0.05075 0.05004 0.04854 0.04832 0.04708 0.04649 0.0461 0.04532 0.0435
Cumulative Proportion  0.09597 0.17582 0.25259 0.32616 0.39412 0.44742 0.49817 0.54821 0.59675 0.64507 0.69215 0.73864 0.7847 0.83005 0.8736
                          PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.93615 0.70791 0.68724 0.64253 0.59237 0.20432
Proportion of Variance 0.04173 0.02386 0.02249 0.01966 0.01671 0.00199
Cumulative Proportion  0.91529 0.93915 0.96164 0.98130 0.99801 1.00000
```

Data set ที่ผ่านการทำ PCA

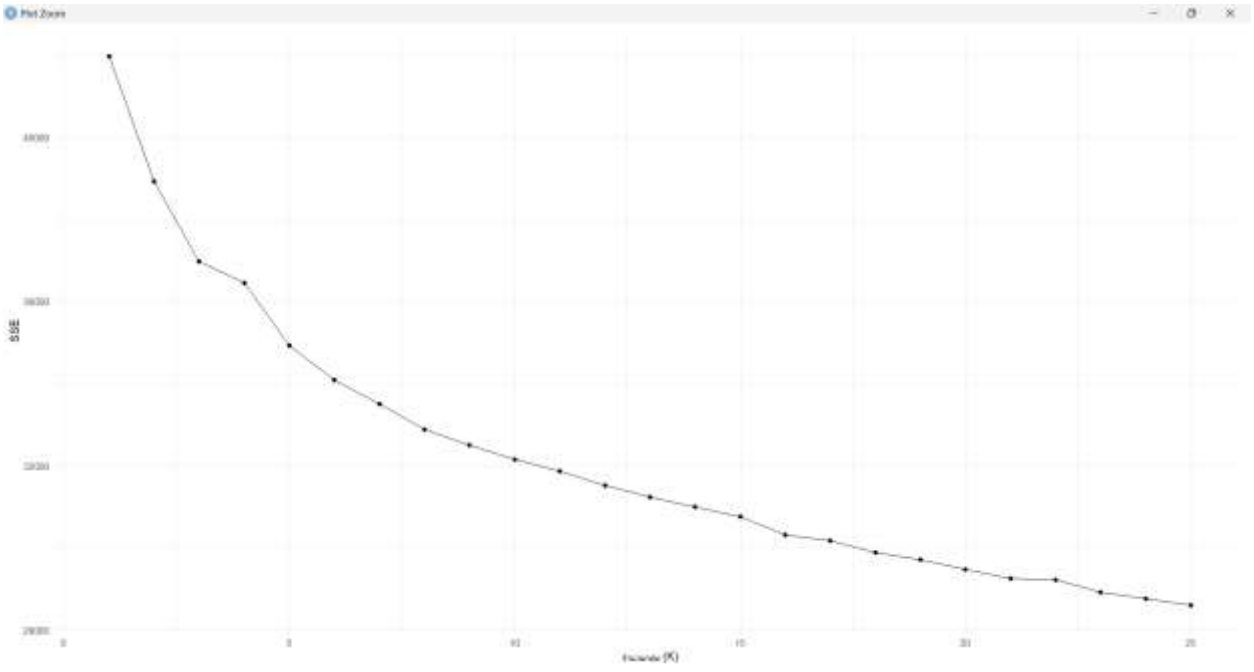
```
pca_components # with PCA
             PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9
 [1,] -1.139659021 -0.1674101265  2.0958494301  1.2214879456  1.3285800360  0.292795899 -1.6901994329 -2.0495159431 -0.7722414890
 [2,]  1.043429009  1.9278975951 -0.3214044853 -0.2011832757 -1.0761962311 -0.328050146  0.2712275266  1.5327633051  0.2034655711
 [3,]  0.751088922  1.2624842511 -0.0697118902  0.1666420517 -1.9309790462 -0.467085356  1.0763521509  0.9820245129  0.2563395808
 [4,]  0.978216957  1.0854703818  0.9563155743 -1.1953905692 -0.2605969153  0.605492417  0.8973921619  0.1546694163 -1.1312644075
 [5,] -0.250401379 -1.6739219266 -1.2774778226 -0.1554392444 -1.8800496619  1.859937624  0.0330199073  1.3274041276  0.0454474347
 [6,] -0.476742845  0.4085599621  0.4706533624 -0.7353076572 -1.2401702669  0.338941780 -2.5057552686 -0.1771164991 -0.1149679097
 [7,]  1.617975700 -0.0068539358 -0.8874106357  0.7668602366  0.4592178895  1.729358851 -0.1771005716 -1.8245954411  0.4838679343
 [8,] -1.866992744  1.4005811934  0.5056100867  0.0063550618 -0.3473121457 -0.044938052 -1.4180310552 -1.2936528139 -0.1178227047
 [9,] -1.740293177 -0.2726080031  0.1791277023  0.3056976379  0.6841749586  2.334401501 -0.2543915774  0.7684504288  1.0653327755
[10,] -1.441233018  1.2260436819 -1.4372541789 -1.9691575855  0.4572550349 -1.094663322  1.0994808629  0.4392347552  0.7802450317
[11,]  0.698561302 -0.2064636473  2.8195481746  2.9459766331  0.7705252617 -1.773786299  0.6794153091 -0.3705521392 -2.2456735694
[12,]  1.909290903 -0.7157770887 -1.4105475968  0.6639898263  1.2912927479  1.010775871  1.5538461073 -0.5958293463 -0.7889655437
[13,] -0.868745239 -0.8449786408 -0.1639798083  0.1512485073  0.4843039370  0.694002831 -1.5101812485  0.1699543789 -1.4875303038
[14,]  0.203201511 -1.4202129785  0.4119771480  0.8767164363 -0.9093527170 -0.635694165 -0.6088824228 -0.1268706122 -1.2891671057
```

PC10	PC11	PC12	PC13	PC14
0.0240069037	-0.1553718173	-0.4336274376	-0.7294918939	-0.5035577054
0.9450093484	-0.6185808222	-0.2587212649	0.9120423513	-1.2831878317
1.8635528115	-0.4877941565	-0.8192636662	0.9719610678	-1.4605794818
1.5334785863	1.4216856548	0.0034645936	0.2245387814	0.1819368605
0.7746418933	-0.6155914802	0.1463140966	-0.0492475333	-0.6888668139
1.0096968889	-0.5619812718	1.5097055865	0.7067605840	0.0754727591
0.3498980317	0.0437756039	-0.0997322924	-0.2232144027	1.6532454574
0.1113961871	-1.9635541531	-0.7096933727	0.9707192875	1.1386515023
1.7065562014	-0.1895285023	-0.3369505282	-1.5911528949	-0.1815824606
1.4099705924	0.2659061743	2.4022402802	-0.3641434629	-0.4762294506
0.8868000858	0.8782218517	0.7415544593	-0.9197210715	0.1564156176
0.4125529971	-0.6088829052	-1.5308100829	-1.0160083759	0.6914041770
0.3992661885	1.8160087030	-0.1704696589	0.4680909139	1.4724326026

K-mean clustering

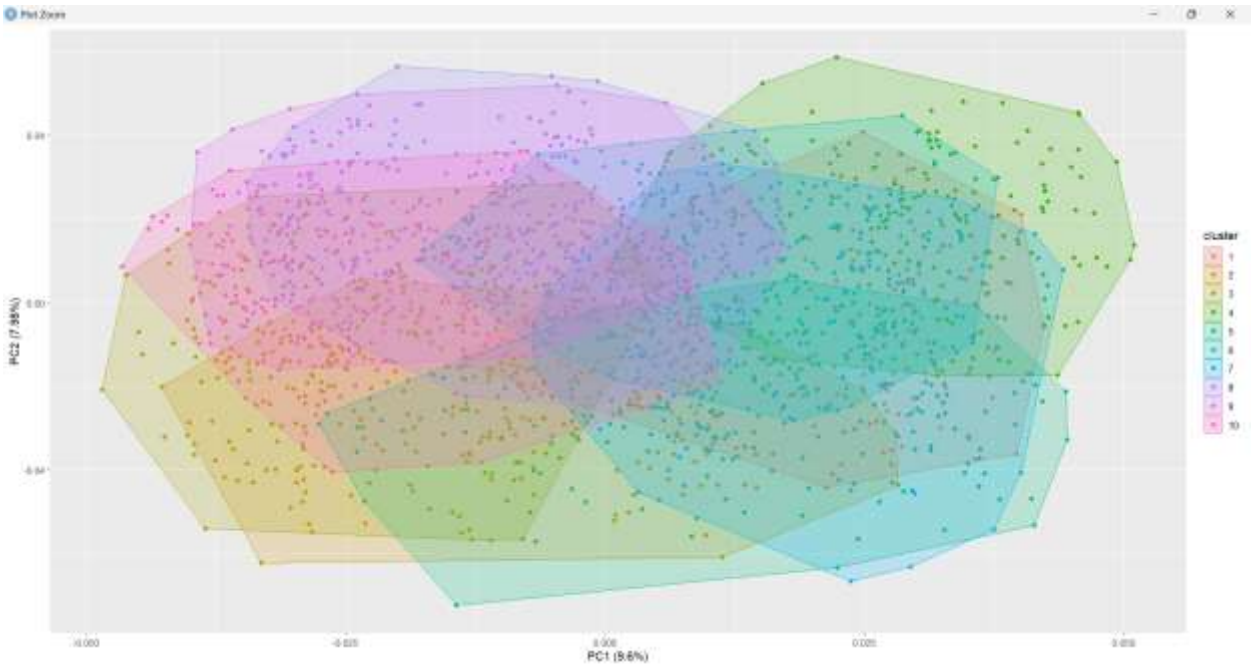
k-mean clustering of standardize data

```r
#K-mean no PCA
SR = 25
sse <-numeric(SR)
for (k in 1:SR) {
  kmeans_model <- kmeans(standD_data, centers = k)
  sse[k] <- kmeans_model$tot.withinss
}

ggplot(data.frame(K = 1:SR, SSE = sse), aes(x = K, y = SSE)) +
  geom_line() +
  geom_point() +
  labs(x = "จำนวนกลุ่ม (K)", y = "SSE") +
  theme_minimal()
```



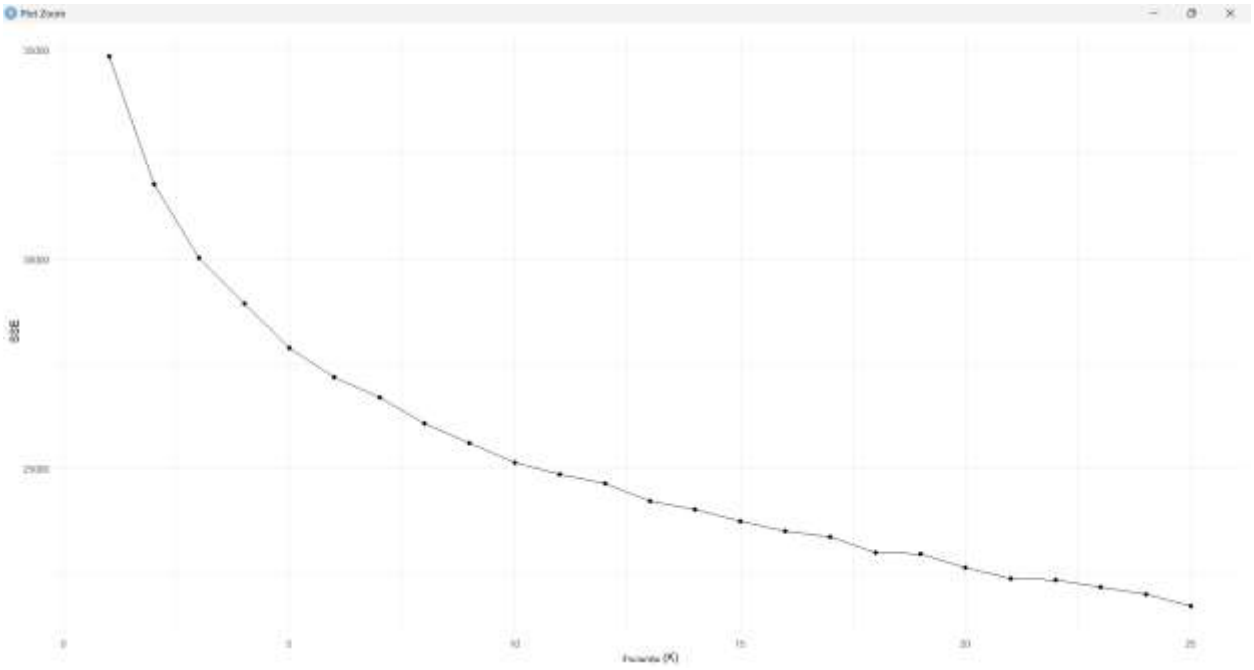
จากวิธีการ Elbow คำนวณหาค่า K ที่เหมาะสมในการแบ่งกลุ่ม เลือก K = 10

ภาพการแบ่งกลุ่มด้วย k-mean



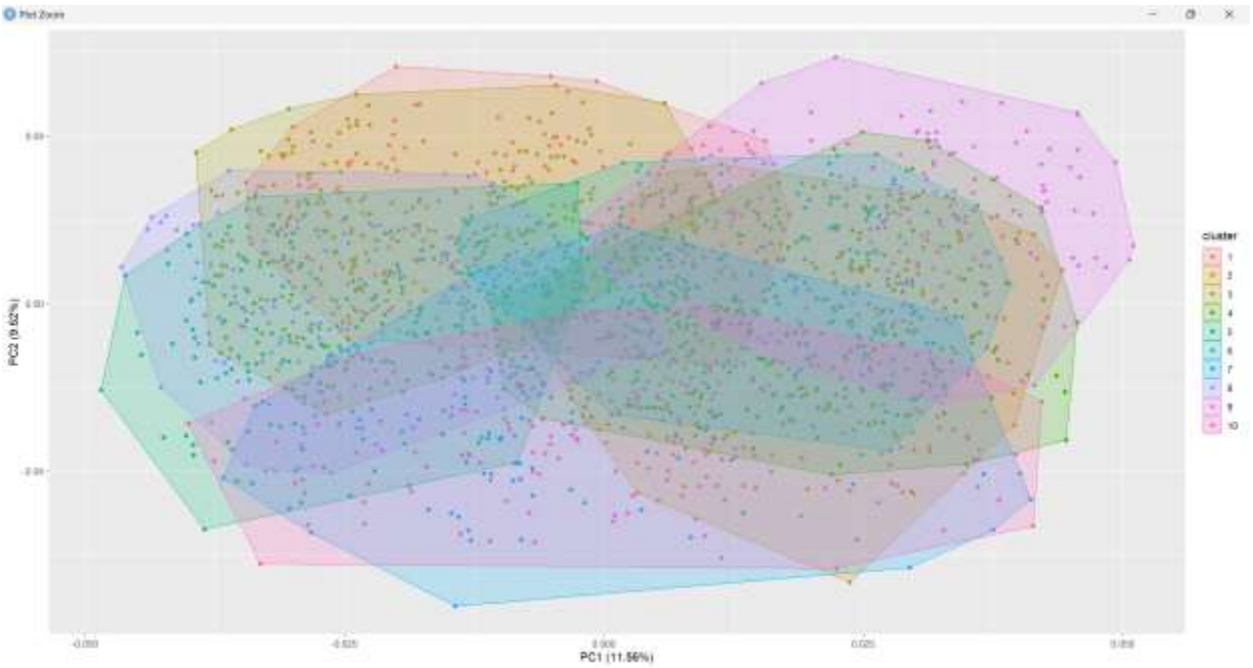
k-mean clustering of PCA data

```r
#K-mean PCA
SR = 25
sse <-numeric(SR)
for (k in 1:SR) {
  kmeans_model <- kmeans(pca_components, centers = k)
  sse[k] <- kmeans_model$tot.withinss
}

ggplot(data.frame(K = 1:SR, SSE = sse), aes(x = K, y = SSE)) +
  geom_line() +
  geom_point() +
  labs(x = "จำนวนกลุ่ม (K)", y = "SSE") +
  theme_minimal()
```



จากวิธีการ Elbow คำนวณหาค่า K ที่เหมาะสมในการแบ่งกลุ่ม เลือก K = 10

ภาพการแบ่งกลุ่มด้วย k-mean

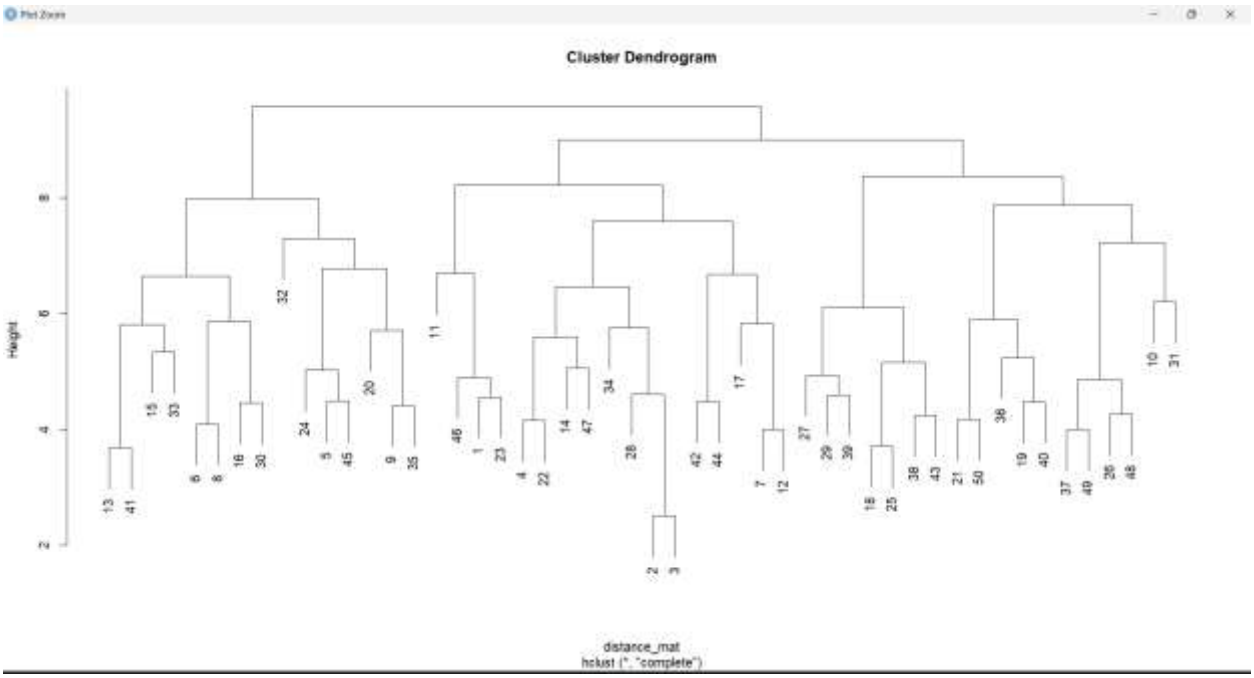


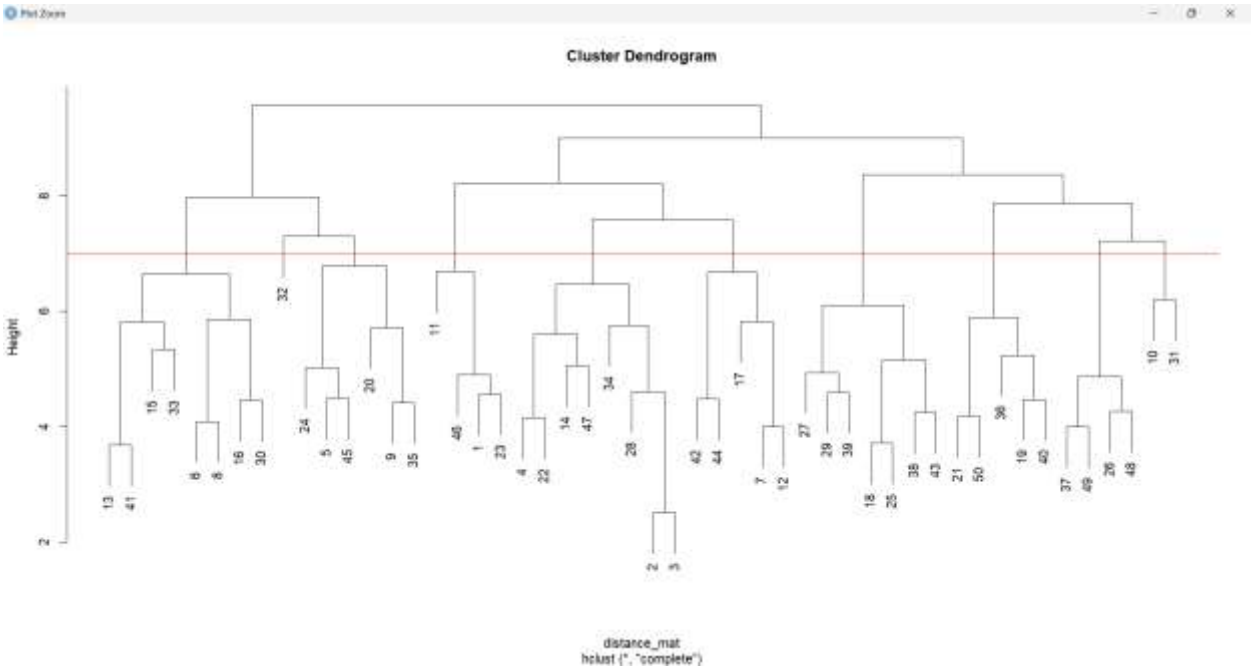
Hierarchical clustering

Hierarchical of standardize data

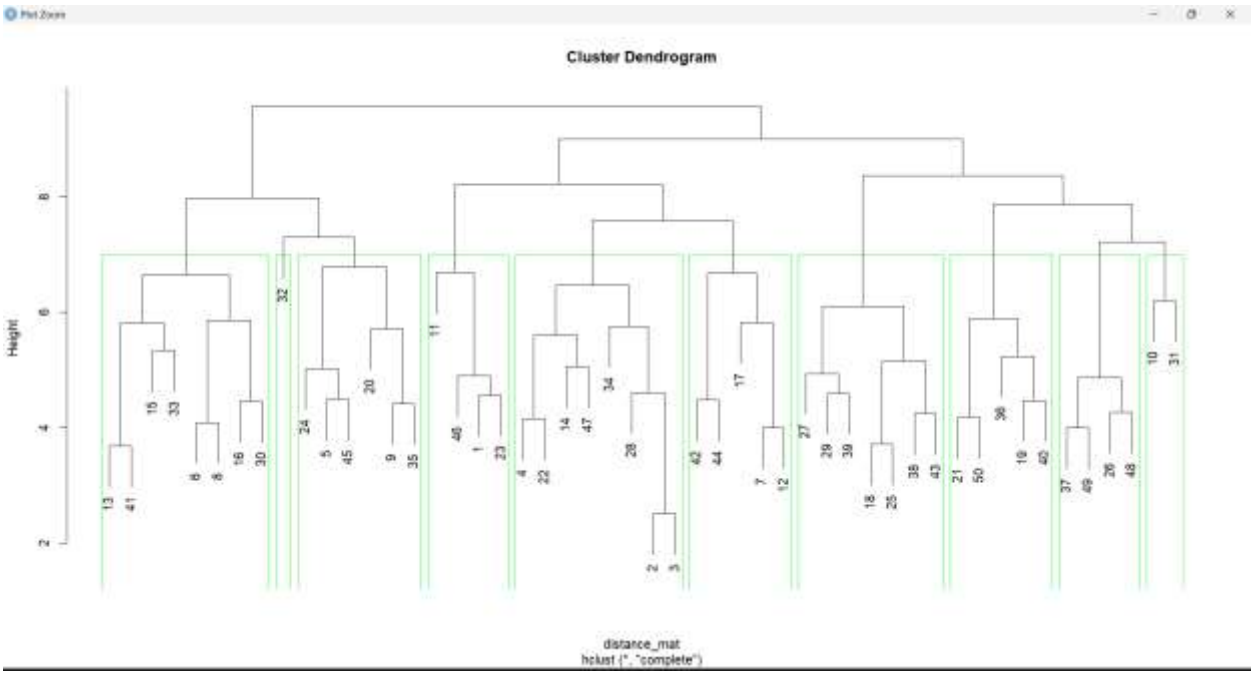
แบ่งข้อมูลออกมาส่วนหนึ่งเพื่อทำ Hierarchical ให้เห็นการแบ่งของข้อมูลอย่างชัดเจน

```
#Hierarchical no PCA
test = (standD_data[1:50,])
test

distance_mat <- dist(test,method = 'euclidean')
distance_mat
Hierar_cl <- hclust(distance_mat)
Hierar_cl

plot(Hierar_cl)
```





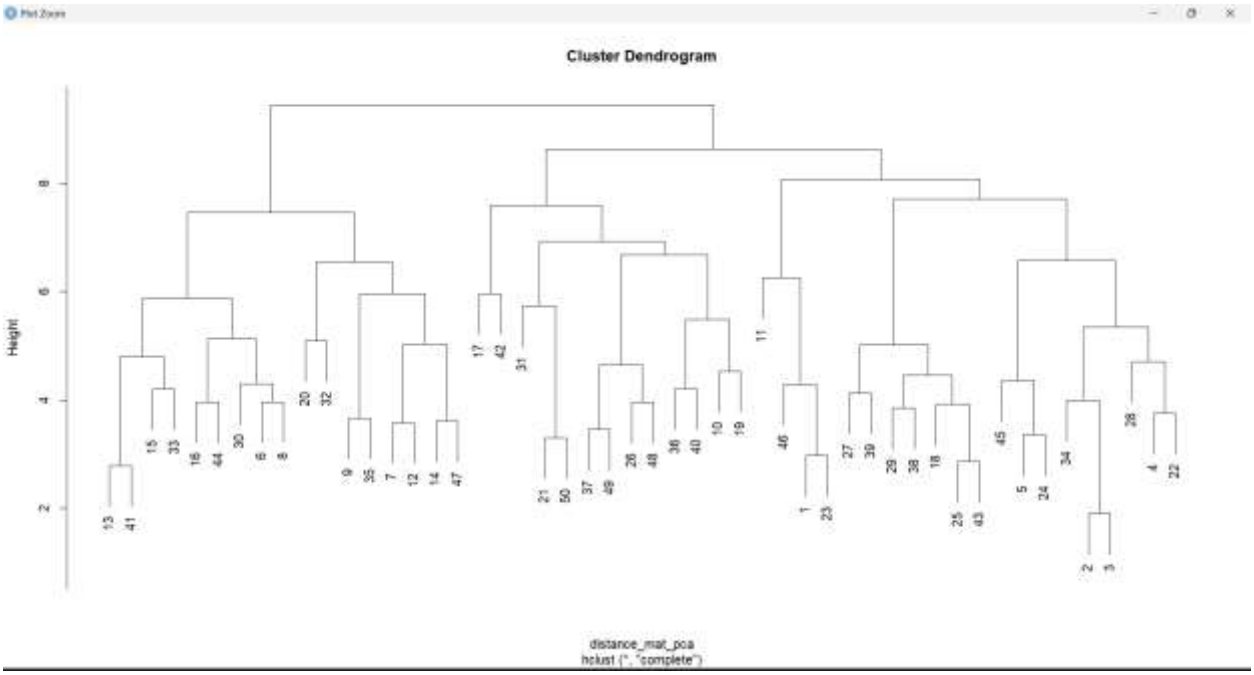
ทำการขีดเส้นการแบ่งกลุ่ม ที่ height = 7

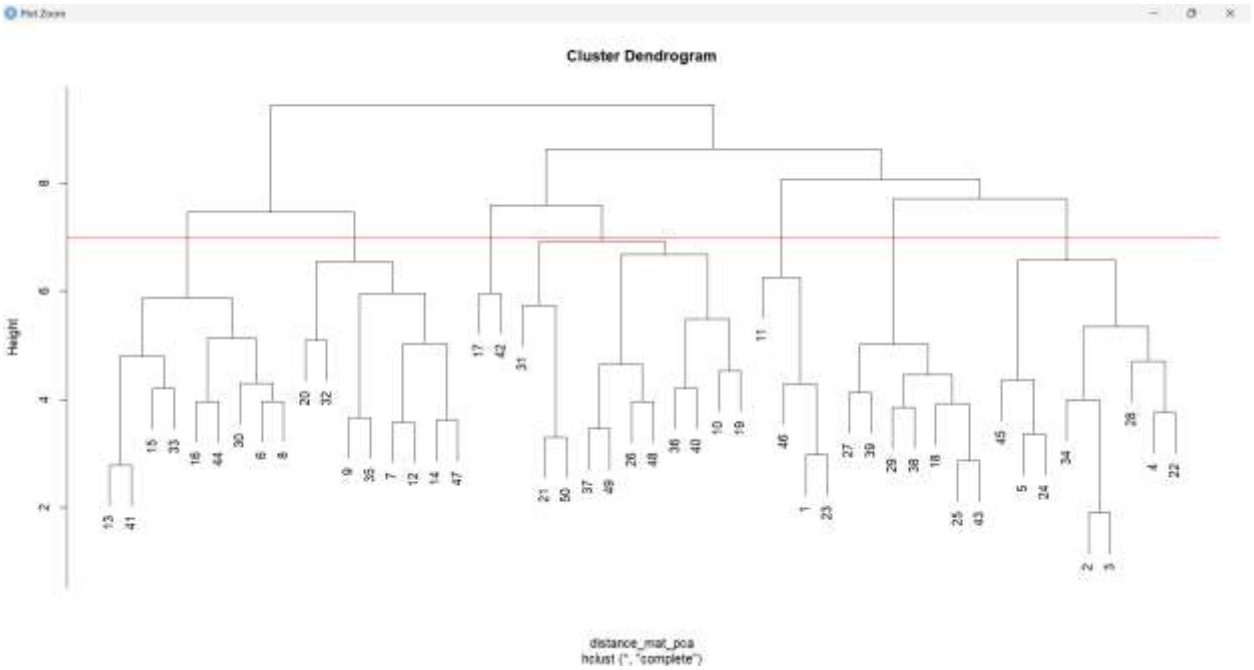


Hierarchical of PCA data

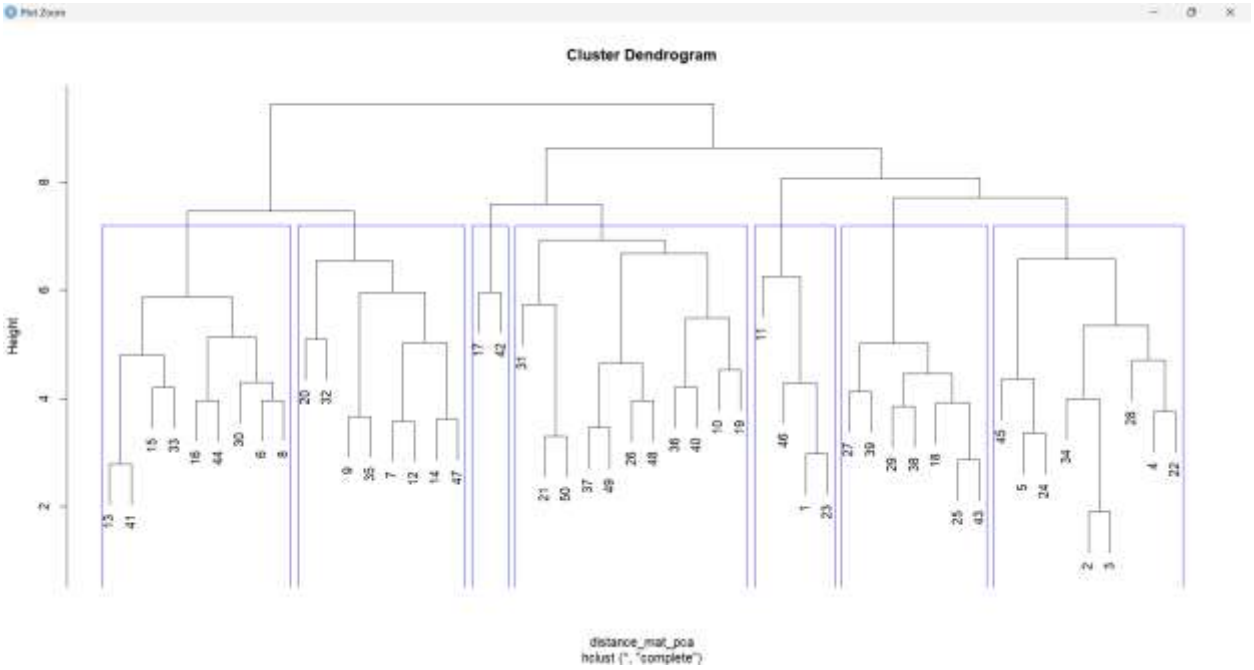
แบ่งข้อมูลออกมาส่วนหนึ่งเพื่อทำ Hierarchical ให้เห็นการแบ่งของข้อมูลอย่างชัดเจน

```
#Hierarchical PCA
test_pca = (pca_components[1:50,])
test_pca

distance_mat_pca <- dist(test_pca,method = 'euclidean')
distance_mat_pca
Hierar_cl_pca <- hclust(distance_mat_pca)
Hierar_cl_pca

plot(Hierar_cl_pca)
```





ทำการขีดเส้นการแบ่งกลุ่ม ที่ height = 7

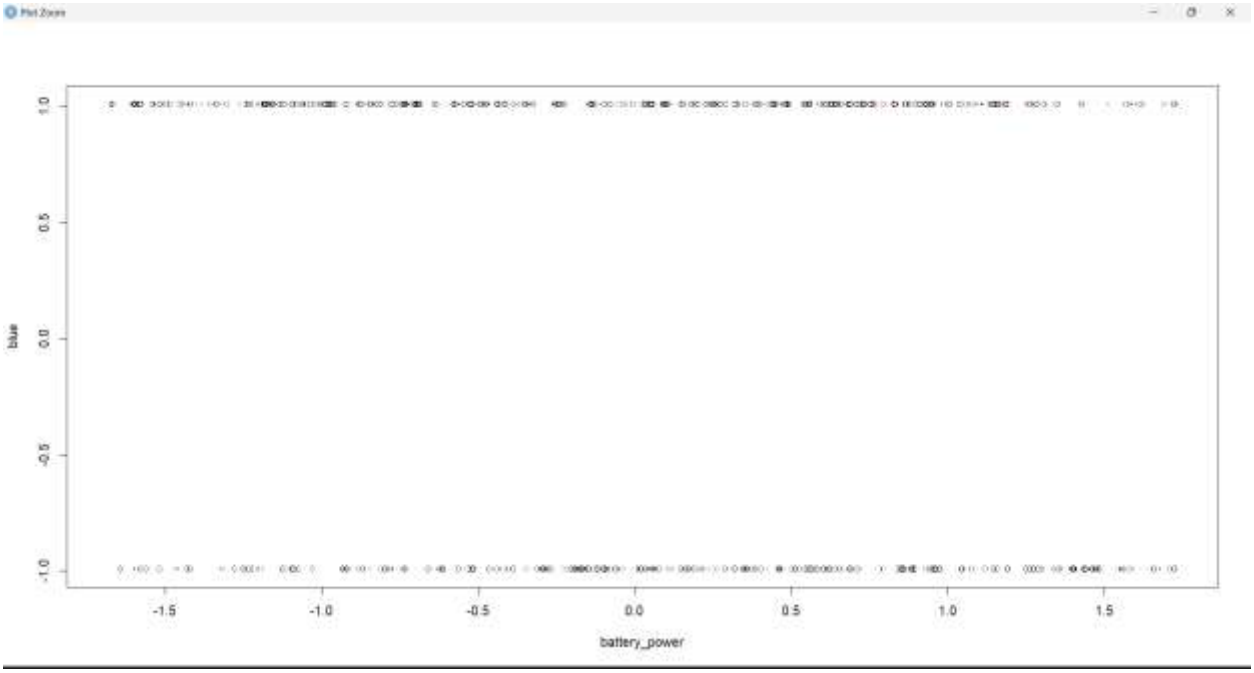


DBScan Clustering

DBScan of standardize data

```
dist_matrix <- proxy::dist(standD_data,method = "Euclidean")

Db_cl <- dbscan::dbscan(dist_matrix,eps = 4, minPts = 40)
Db_cl

Db_cl$cluster

plot(standD_data, col = Db_cl$cluster)
```

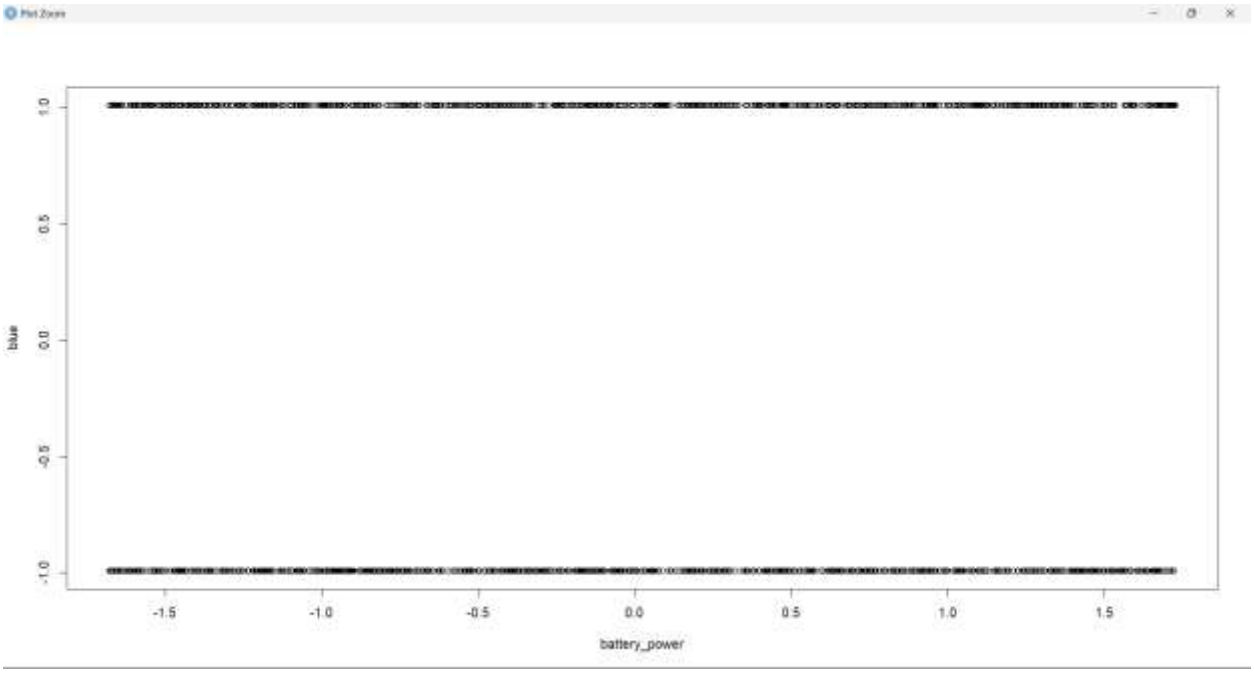
กำหนด parameter ของ DBScan ให้ ε = 4 และ จำนวนขั้นต่ำในการกลุ่ม = 40



DBScan of PCA data

```r
##DBScan euclidean with pca
dist_matrix_pca <- proxy::dist(pca_components,method = "Euclidean")

Db_cl_pca <- dbscan::dbscan(dist_matrix_pca,eps = 4, minPts = 40)
Db_cl_pca

Db_cl_pca$cluster

plot(standD_data, col = Db_cl_pca$cluster)
```

กำหนด parameter ของ DBScan ให้ ε = 4 และ จำนวนขั้นต่ำในการกลุ่ม = 40



สรุปผล

ในส่วนของ K-mean นั้น เราใช้จำนวน K ที่เท่ากัน กับข้อมูลทั้งสองชุด ซึ่งเห็นได้จากสีว่าตำแหน่งของ cluster ของข้อมูลนั้นสองนั้นมีความแตกต่างกัน

มาที่ Hierarchical เราจะเห็นได้ชัด ว่าแม้จะขีดเส้นตัดกลุ่มที่ตำแหน่งเดียวกัน จำนวนของกลุ่มที่แบ่งได้นั้นแตกต่างกัน โดย ข้อมูลที่ทั่วไป ที่ผ่าน standardize นั้นจะแบ่งได้ 7 กลุ่ม แต่ในส่วนของข้อมูลที่ผ่าน PCA มานั้น แบ่งได้ทั้งสิ้น 6 กลุ่ม

สุดท้าย DBScan จากกราฟทั้งสองจะเห็นได้ชัดถึงจำนวนที่ลักษณะที่แตกต่างกัน ข้อมูลที่ผ่านการ PCA จะมีการกระจุกตัวของข้อมูลมากกว่าข้อมูลที่ผ่านเพียงแค่ standardize