Buse Nur BALTACIOĞLU
2019900540

**DSM 5008**

**DENETİMSİZ İSTATİSTİKSEL ÖĞRENME**

**TAKE HOME II**

1. **Hücre çekirdeğine ait özelliklerin tanımlayıcı istatistiklerini elde ederek, yorumlayınız.**

```
   yaricap           doku           cevre           alan          puruzsuzluk
Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5   Min.   :0.05263
1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637
Median :13.370   Median :18.84   Median : 86.24   Median : 551.1   Median :0.09587
Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9   Mean   :0.09636
3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530
Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0   Max.   :0.16340
   yogunluk         icbukeylik      icbukeynoktalar      simetri        fraktalboyut
Min.   :0.01938   Min.   :0.00000   Min.   :0.00000   Min.   :0.1060   Min.   :0.04996
1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619   1st Qu.:0.05770
Median :0.09263   Median :0.06154   Median :0.03350   Median :0.1792   Median :0.06154
Mean   :0.10434   Mean   :0.08880   Mean   :0.04892   Mean   :0.1812   Mean   :0.06280
3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957   3rd Qu.:0.06612
Max.   :0.34540   Max.   :0.42680   Max.   :0.20120   Max.   :0.3040   Max.   :0.09744
```

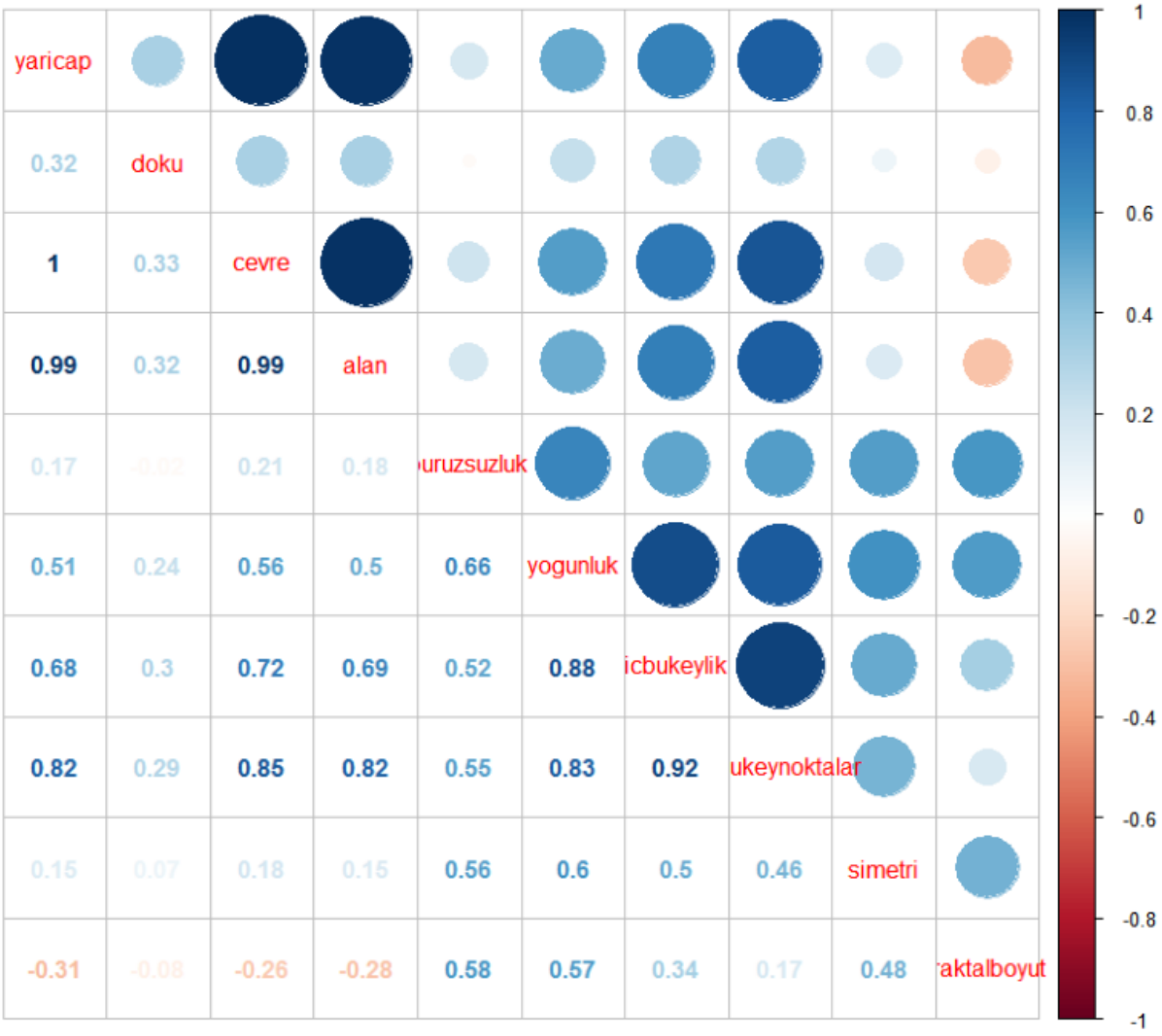
- Yarıçap değişkeni 14 ortalama ile yaklaşık 7 ve 28 arasında değişmektedir.
- Doku değişkeni 19 ortalama ile yaklaşık 10 ve 39 arasında değişmektedir.
- Çevre değişkeni 92 ortalama ile yaklaşık 44 ve 189 arasında değişmektedir.
- Alan değişkeni 655 ortalama ile yaklaşık 144 ve 2501 arasında değişmektedir.
- Pürüzsüzlük, yoğunluk, içbükeylik, içbükey noktalar, simetri, fraktal boyut değişkenlerinin ortalaması ve değişimi benzerdir.

```
     yaricap           doku           cevre           alan      puruzsuzluk
3.524049e+00   4.301036e+00   2.429898e+01   3.519141e+02   1.406413e-02
    yogunluk     icbukeylik icbukeynoktalar        simetri     fraktalboyut
5.281276e-02   7.971981e-02   3.880284e-02   2.741428e-02   7.060363e-03
```

- En yüksek değişime sahip değişken alan değişkenidir.
- En az değişime sahip değişken fraktal boyut değişenidir.

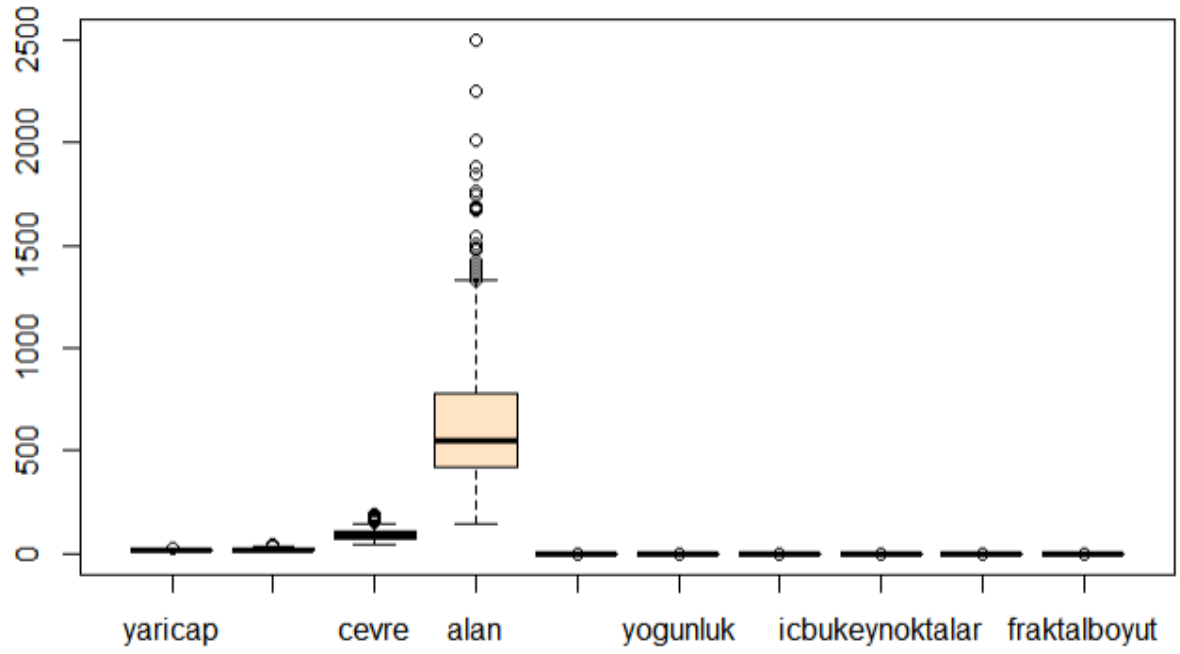
Bu değişkenlerin ölçekleri ve değişimleri birbirinden farklı olduğu için standartlaştırmalıyız.

**2. Korelasyon matrisi elde ederek, yorumlayınız.**

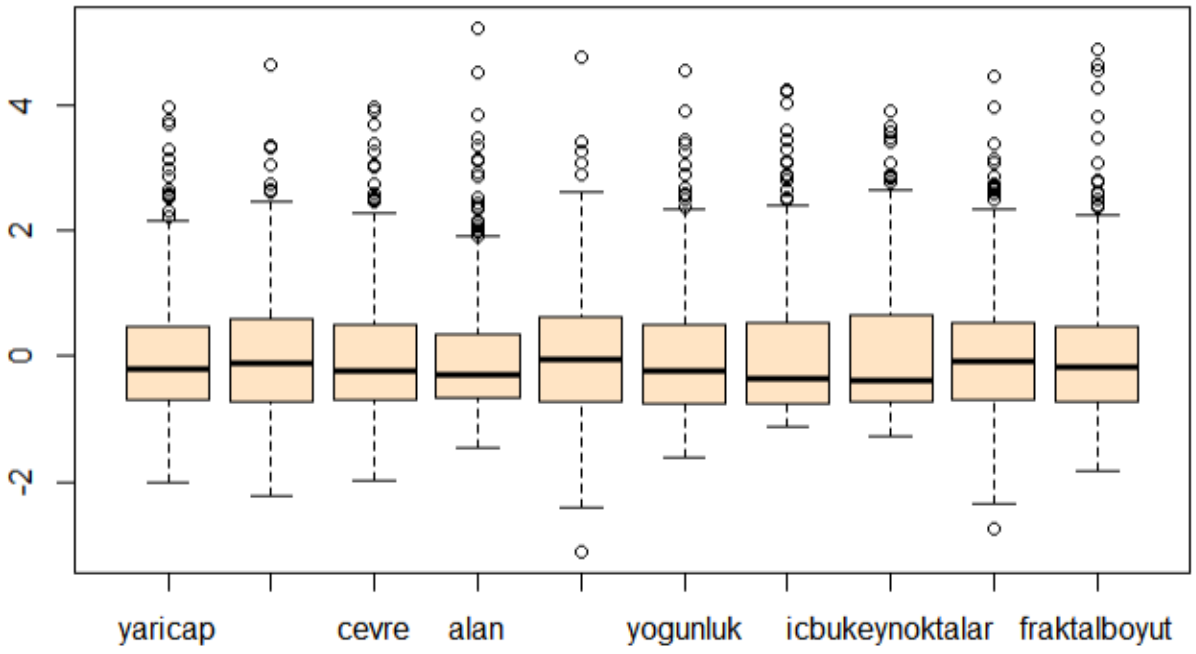



- Yarıçap ve çevre değişkenleri arasında doğrusal anlamda tam uyum bulunmaktadır.
- Yarıçap-alan, çevre-alan değişkenleri arasında da pozitif yönlü çok güçlü doğrusal bir ilişki bulunmaktadır.
- Doku ile pürüzsüzlük değişkenleri arasında doğrusal bir ilişki bulunmamaktadır.
- Yarıçap ile fraktal boyut değişkenleri arasında negatif yönlü güçsüz bir ilişki bulunmaktadır.

**3. Değişkenler için kutu grafiği (box-plot) çizdirerek, yorumlayınız.**

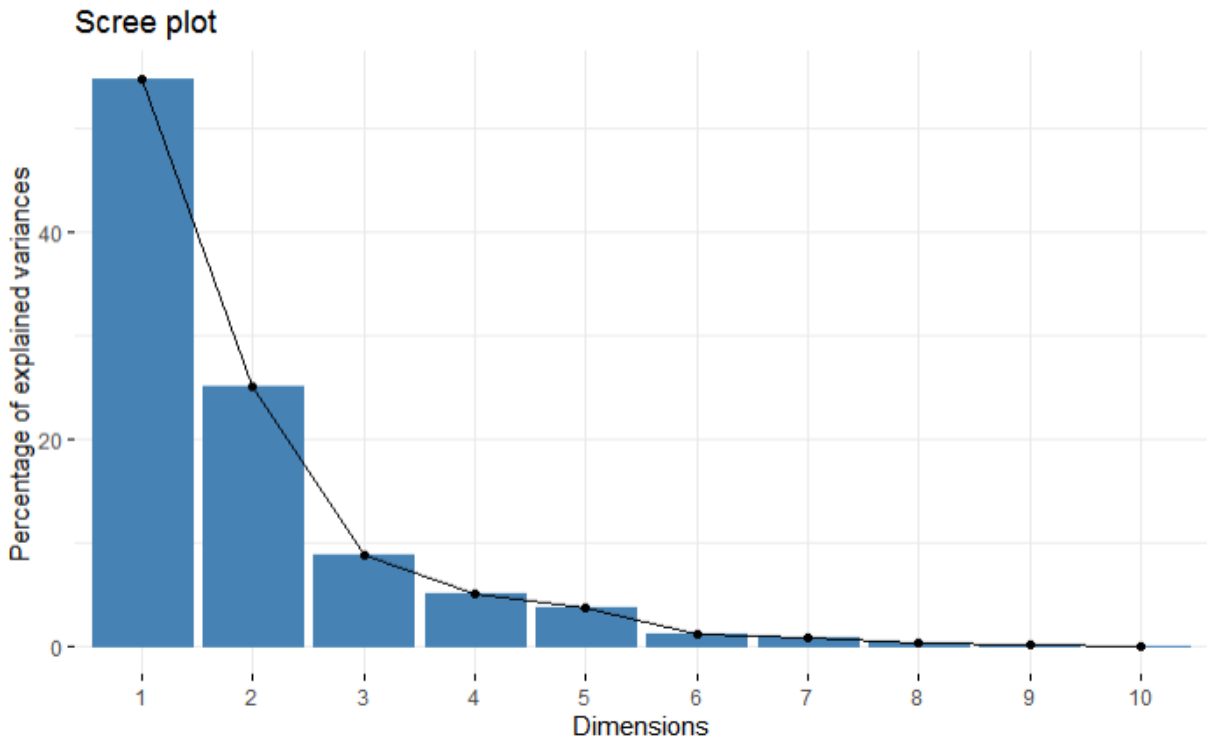



- Orijinal veri setimizin box-plotına baktığımızda alan değişkeninin diğer değişkenlere göre çok daha fazla değişime sahip olduğunu görmekteyiz.
- Değişkenlerin ölçek ve değişimleri birbirinden farklı olduğu için standartlaştırmalıyız.




- Veri setimizi standartlaştırdığımızda yaklaşık sıfır ortalamalı hale dönüştürdük ve box-ploatına baktığımızda değişkenlerimizin sağa çarpık olduğunu görmekteyiz ve dolayısıyla uç değerlerimiz bulunmaktadır.

4. **Temel bileşenler analizi uygulayınız.**
    a. **Bileşen sayısına gerekçelerinizi belirterek karar veriniz.**




- Scree plottaki dirsek noktası değeri bize 3 temel bileşen olarak görünmektedir.

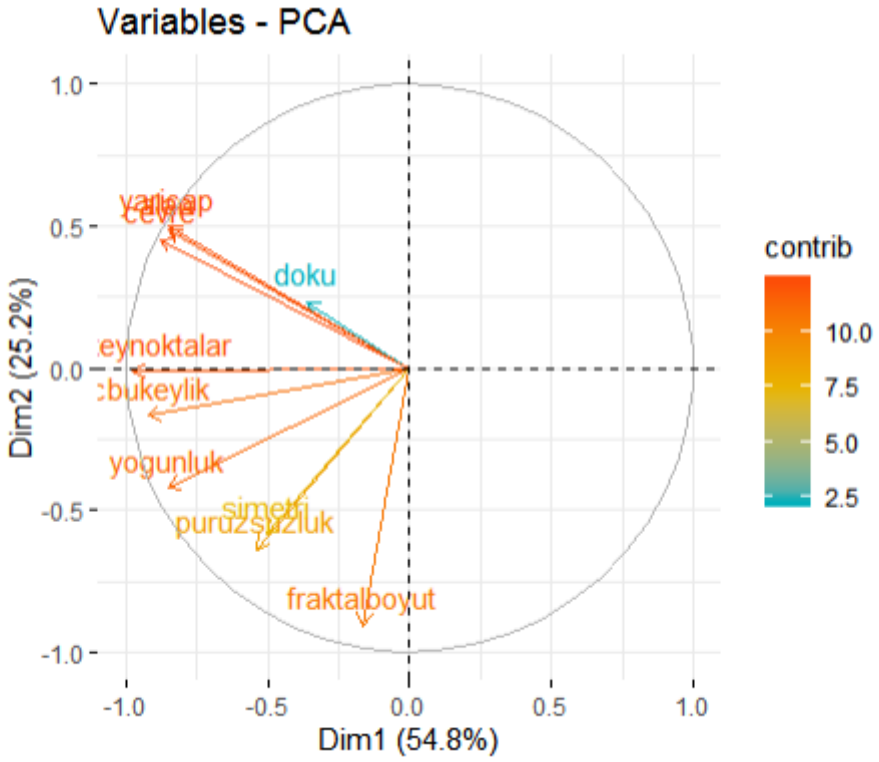
```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     2.3406 1.5870 0.93841 0.7064 0.61036 0.35234 0.28299 0.18679 0.10552 0.01680
Proportion of Variance 0.5479 0.2519 0.08806 0.0499 0.03725 0.01241 0.00801 0.00349 0.00111 0.00003
Cumulative Proportion  0.5479 0.7997 0.88779 0.9377 0.97495 0.98736 0.99537 0.99886 0.99997 1.00000
```

- Kümülatif varyans açıklama oranının %75'den büyük olması yeterlidir bu değerlere göre 2 temel bileşen (%80 açıklama oranı bulunmaktadır) uygun görülmektedir.

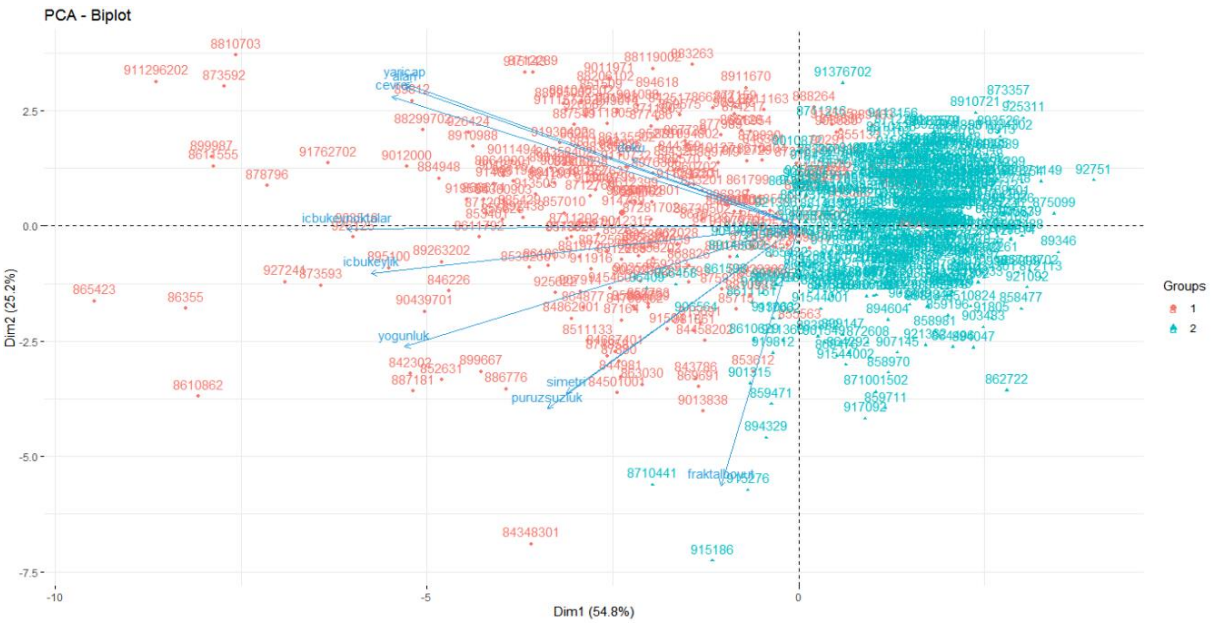
```
[1] 5.4785879917 2.5187135854 0.8806151792 0.4990094357 0.3725391897
[6] 0.1241417485 0.0800853104 0.0348897928 0.0111354606 0.0002823059
```

- Korelasyon matrisini kullandığımız için yukarıdaki özdeğerlerin 1'den büyük olanlarının sayısı yeterli bulunmaktadır bu özdeğerlere göre 2 temel bileşen yeterli bulunmaktadır.
- Bu üç kurala göre değerlendirdiğimizde 2 temel bileşen üzerinde çalışmaya karar verdik.

**b. Görseller üzerinden, değişkenler ve gözlemler için yorumlama yapınız.**



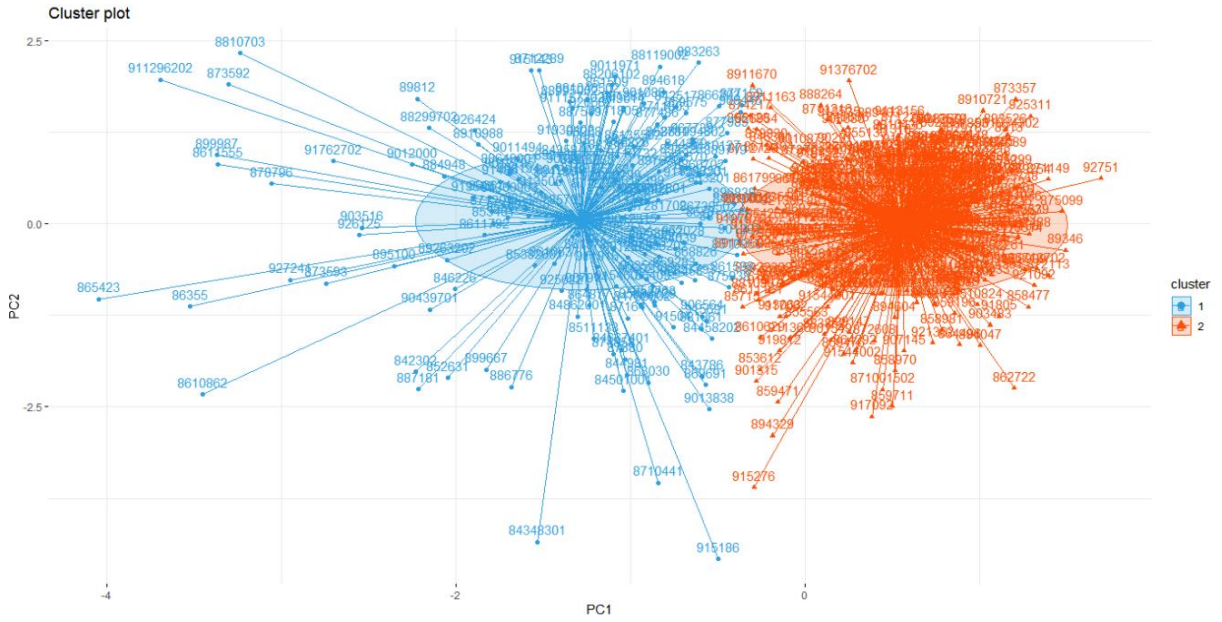

- Yarıçap-Çevre-Doku değişkenleri aynı yönde güçlü bir ilişkileri bulunmaktadır.
- Simetri ve pürüzlük değişkenleri arasında da aynı yönlü ilişki bulunmaktadır.
- Doku ve simetri-pürüzlük arasında ilişki bulunmamaktadır.
- Korelasyon matrisi ile bu grafik yorumları birbirini desteklemektedir.




- Grafiğe baktığımızda iki küme olabilir gibi durmaktadır hedef değişkenimizin de iki kategoride olması bu düşüncemizi desteklemektedir.

- Uç değerlerin olması kümeler içi homojen olma kuralını zorlamaktadır.
- 8810703 numaralı kişinin hücre çekirdeğine ait yarıçapı, çevresi ve alanı en yüksek değerlere sahiptir.
- 865423 numaralı kişinin hücre çekirdeğine ait iç bükeyliği ve iç bükey noktaları en yüksek değere sahiptir.
- 915186 numaralı kişinin hücre çekirdeğine ait en yüksek pürüzlük ve yüksek bir fraktal boyut değerine sahiptir.
- 925311 numaralı kişinin hücre çekirdeğine ait iç bükeylik, iç bükey noktalar, simetri ve fraktal boyutta en düşük değerlere sahiptir.
- 862722 numaralı kişinin hücre çekirdeğine ait yarıçap, çevre, alan, iç bükeylik ve iç bükey noktalar en düşük değerlere sahiptir.

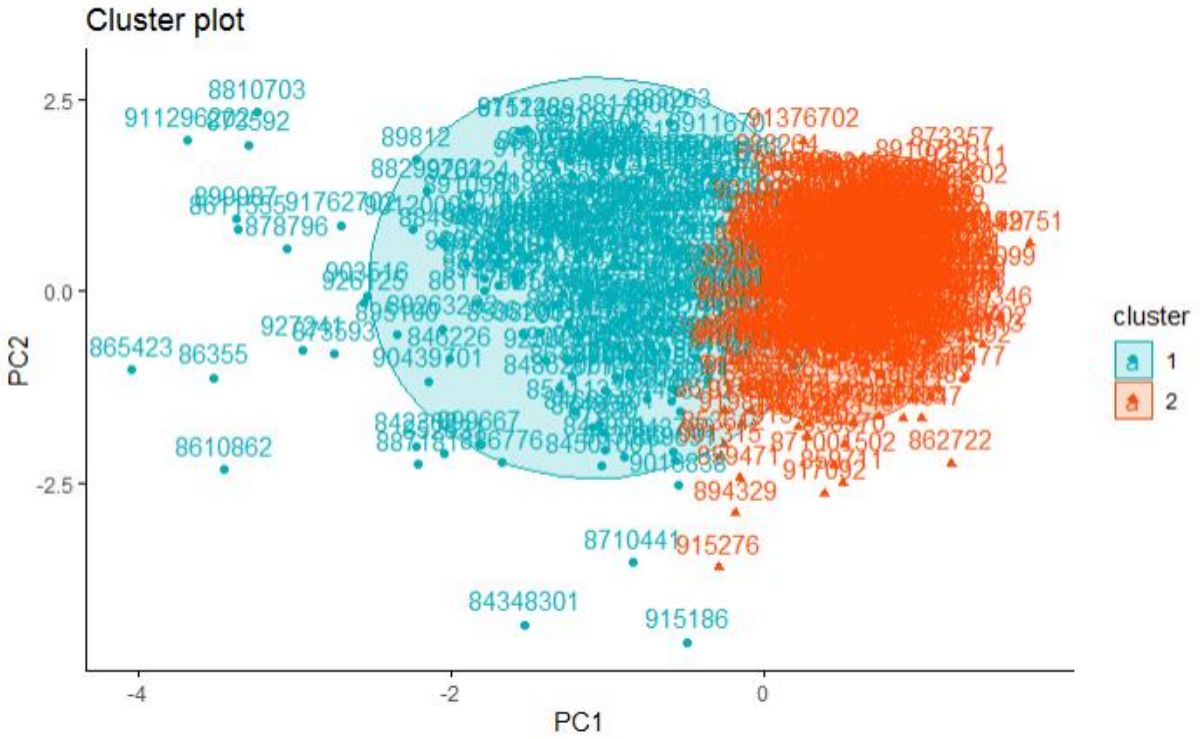
## 5. K-ortalamalar ile kümeleme analizi uygulayarak, yorumlayınız.




- Varyans açıklama oranı %48.5'tir bu oran yetersiz bulunmaktadır ve grafiğe baktığımızda 1. küme heterojen durmaktadır bu sebepten k-means için küme sayısı 2 olduğunda yetersiz durmaktadır.
- Hedef değişken ile k-meansin oluşturduğu kümelerin frekanslarını karşılaştırdığımızda 55 gözlem yanlış kümeye atanmıştır.
- Rand indeksi, 0.83 değerini verdi burada 1'e yakın çıkması iki kümeleme sonucunun özdeş olduğu anlamına gelir.

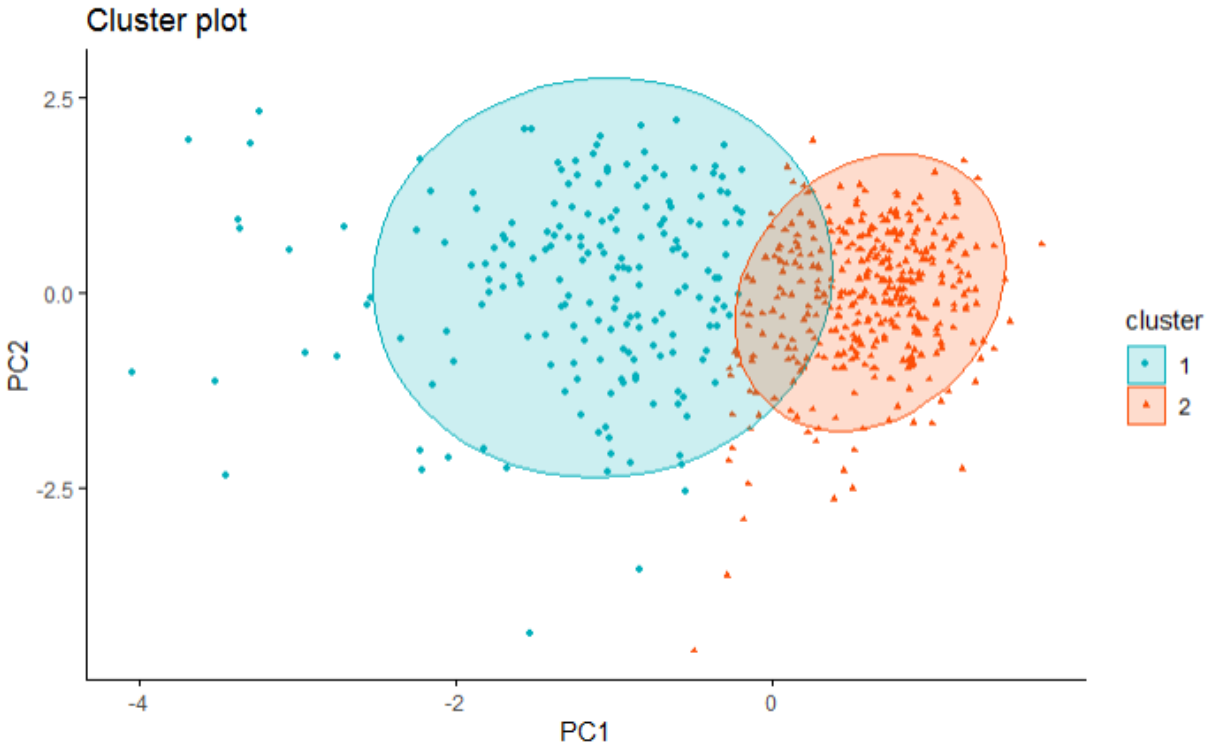
**6. K-medoids ile kümeleme analizi uygulayarak, yorumlayınız.**

## I. Pam Algoritması




- Hücre çekirdeklerini pam algoritmasını kullanarak kümelediğimizde build 1.807 ve swap 1.700 değerini vermektedir.
- Hedef değişken ile pam algoritmasının oluşturduğu kümelerin frekanslarını karşılaştırdığımızda 42 gözlem yanlış kümeye atanmıştır.
- Rand indeksi, 0.86 değerini verdi burada 1'e yakın çıkması iki kümeleme sonucunun özdeş olduğu anlamına gelir.

## II. CLARA



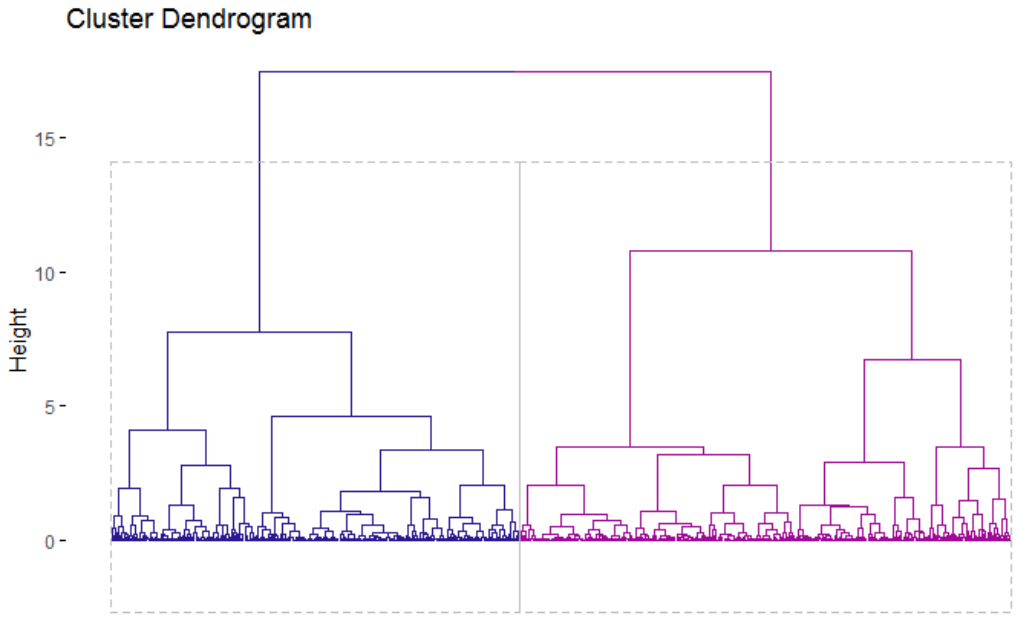

- Hücre çekirdeklerini Clara kullanarak kümelediğimizde amaç fonksiyonu 1.704 değerini vermektedir.
- Hedef değişken ile Clara algoritmasının oluşturduğu kümelerin frekanslarını karşılaştırdığımızda 40 gözlem yanlış kümeye atanmıştır.
- Rand indeksi, 0.87 değerini verdi burada 1'e yakın çıkması iki kümeleme sonucunun özdeş olduğu anlamına gelir.

**7. Aşamalı kümeleme analizi uygulayarak, yorumlayınız.**

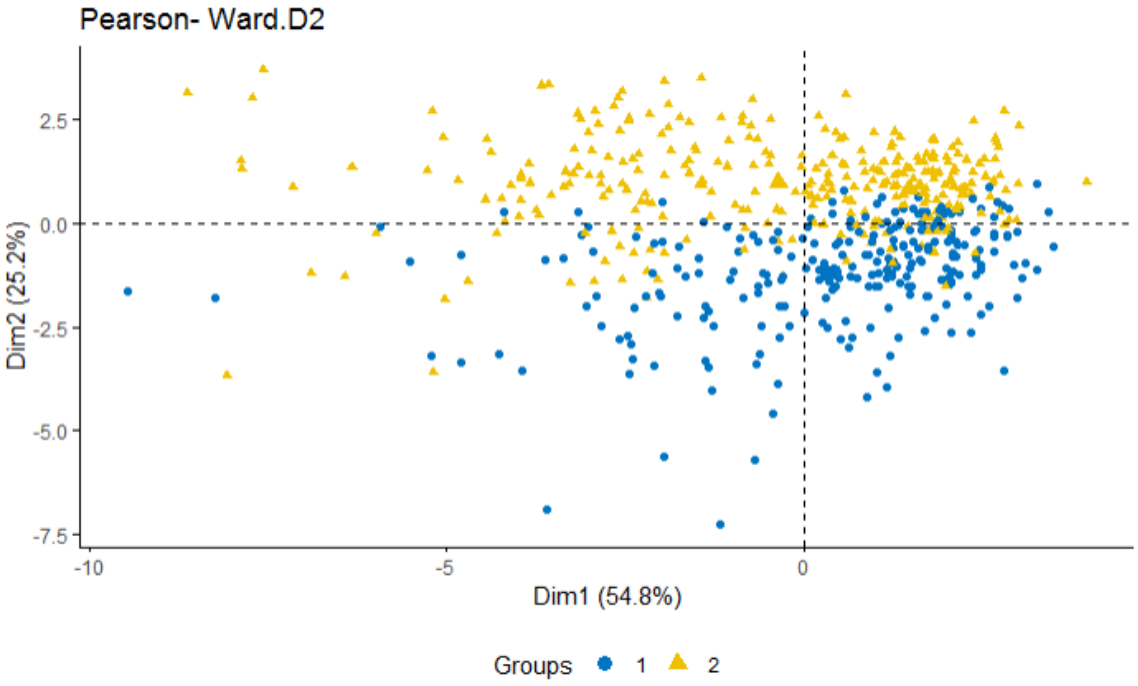
- Uzaklık ölçüleri olarak öklit, manhattan ve pearson ölçülerini bununla birlikte ward.d2, average ve median bağlantı yöntemleri kullanılarak uygulanmıştır.
- Bu tabloda bağlantı yöntemleri ve uzaklık metotları arasındaki kojenetik korelasyonu vermektedir.

| | Ward.D2 | Average | Median |
|---|---|---|---|
| Öklit | 0.67 | 0.80 | 0.57 |
| Manhattan | 0.60 | 0.76 | 0.65 |
| Pearson | 1 | 1 | 1 |

- 0.75 üzerinde ilişkisi olan bağlantı yöntemleri ve uzaklık metotları uygulanmıştır.
- Uygulanan bu yöntemlerden pearson-ward.D2 ve pearson-median kümelerinin rand indeks değeri 0.77'dir.
- Bu iki kümeleme yöntemi hedef değişkene göre 76 gözlemi yanlış kümeye atamıştır.

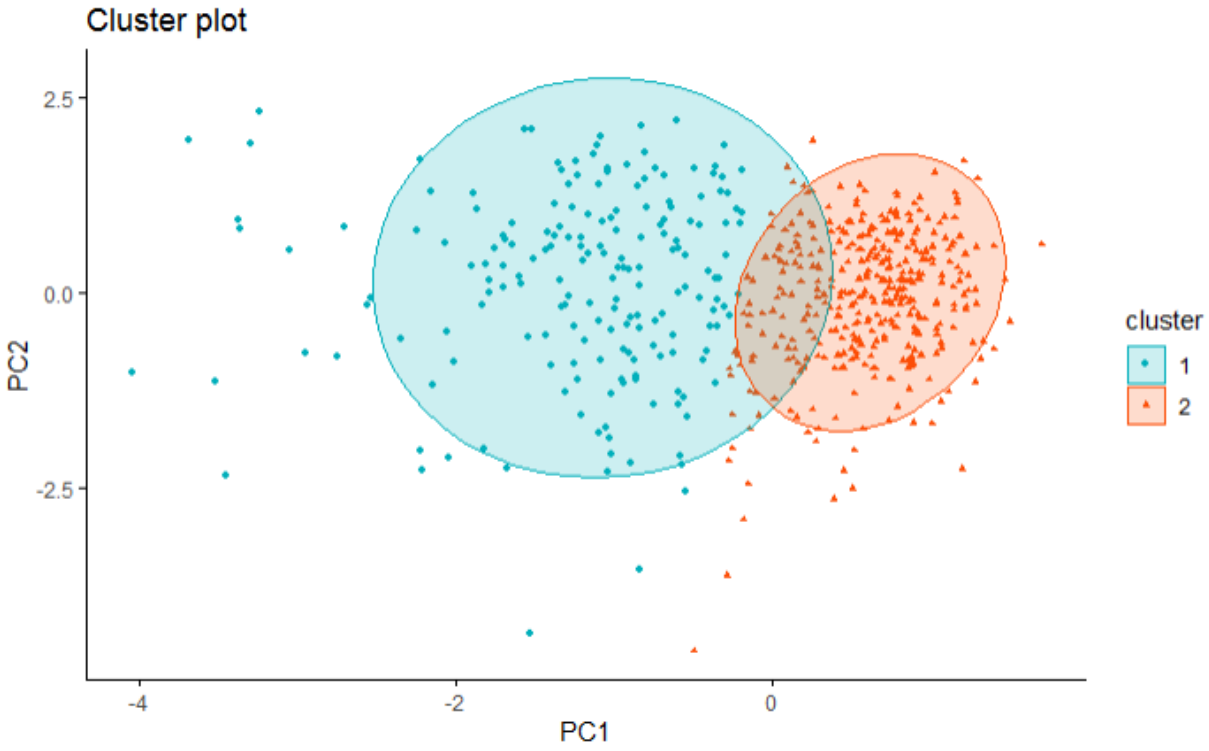



- Yukarıdaki grafik scale edilmiş veri setinden oluşturulmuştur.



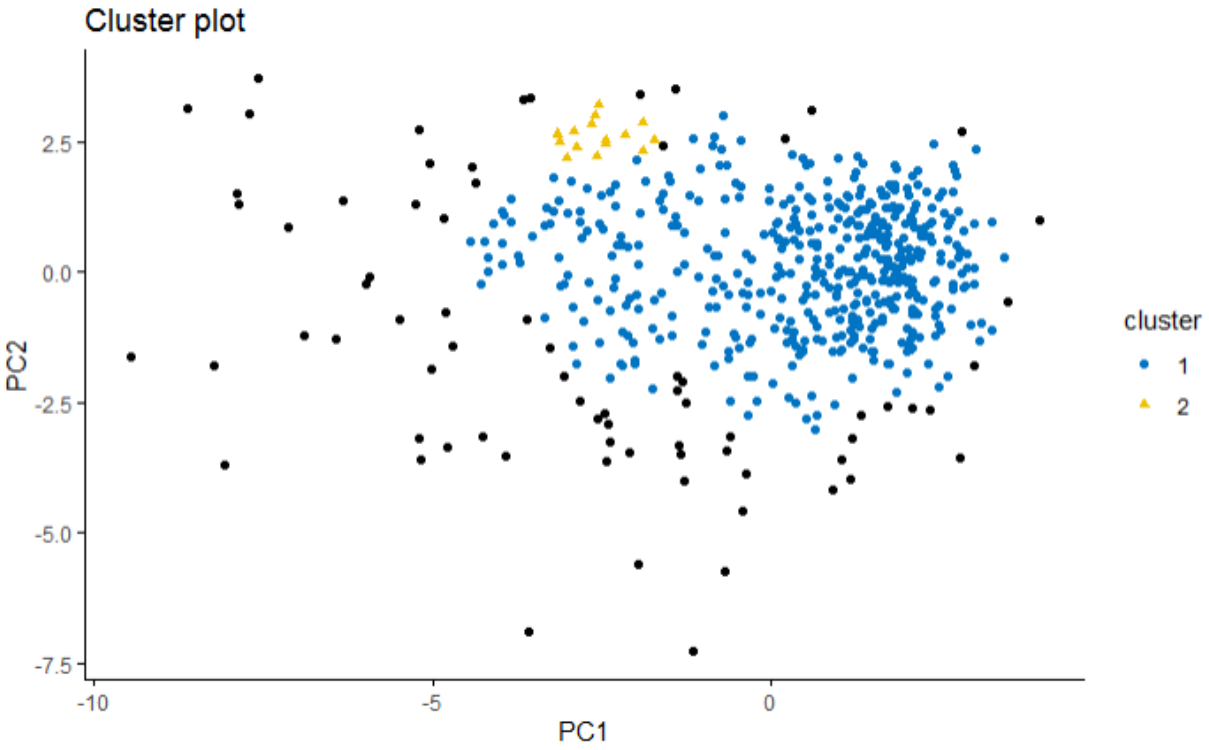

- Yukarıdaki grafik kümeler arası homojen küme içi heterojen durmaktadır.

**8. Model temelli kümeleme analizi uygulayarak, yorumlayınız.**




- Model temelli kümeleme, verilerin bir model tarafından oluşturulduğunu varsayar ve kümeleme, veriden orijinal modele erişmeye çalışır. Erişilen model ile kümeler tanımlanır.
- k-ortalamadan farklı olarak, model tabanlı kümeleme, her veri noktasının her bir kümeye ait olma olasılığına sahip olduğu bir atama kullanır.
- Model parametreleri, hiyerarşik model tabanlı kümelemeden yararlanılarak başlatılan Beklenti Maksimizasyonu (EM) algoritması kullanılarak tahmin edilebilir.
- Bu veri seti için VVI, kümelerin hacim ve şekli değişken olduğu ve benzer yönelime sahip oldukları anlamına gelir.
- En iyi model Bayesian Bilgi Ölçütü (BIC) kullanılarak seçilir. BIC puanı -4510.449, karşılık gelen model için güçlü kanıtlar olduğunu gösterir.
- Bu grafikte büyük semboller daha belirsiz olan gözlemleri gösterir.
- Hedef değişken ile model temelli kümelemenin oluşturduğu kümelerin frekanslarını karşılaştırdığımızda 77 gözlem yanlış kümeye atanmıştır.
- Rand indeksi 0.77 değerini vermiştir.

**9. Yoğunluk temelli kümeleme analizi uygulayarak, yorumlayınız.**




- Gürültü ve aykırı değerler içeren bir veri setinin herhangi bir şekildeki kümelerini tanımlamak için geliştirmişlerdir. Bu grafikte siyah noktalar herhangi bir kümeye atanamayan gözlemlerdir.
- Ana fikir, bir kümenin her noktası için, belirli bir yarıçapın komşusunun en az minimum sayıda nokta içermesi gerektiğidir.

```
DBSCAN clustering for 569 objects.
Parameters: eps = 0.55, minPts = 7
The clustering contains 2 cluster(s) and 74 noise points.

  0   1   2
 74 480  15

Available fields: cluster, eps, minPts
```

- Yoğunluk temelli kümeleme için epsilon değeri 0.55 ve minpoint değeri 7 olarak belirlenmiştir.
- Hedef değişken ile yoğunluk temelli kümelemenin oluşturduğu kümelerin frekanslarını karşılaştırdığımızda 180 gözlem yanlış kümeye atanmıştır.
- Rand indeksi 0.58 değerini vermiştir. Bu iki kümenin birbirine çok benzemediğini ifade etmektedir.

**10. Küme geçerliliği istatistiklerini de dikkate alarak seçtiğiniz en uygun kümeleme analizi yöntemini gerekçelerinizle belirtiniz. (Diğer şıklar içinde değerlendirildiyse burada özet bilgi şeklinde verilebilir.)**

Tüm kümeleme yöntemleri için farklı sayılarda küme sayıları belirlenmiştir fakat biz hedef değişkenin 2 kategoride olduğu için 2 küme oluşturmayı daha uygun bulduk. R kodlarında tüm kümeleme yöntemleri için uygulanmıştır.

| | Score <dbl> | Method <fctr> | Clusters <fctr> |
|---|---|---|---|
| Connectivity | 10.0647 | hierarchical | 2 |
| Dunn | 0.0637 | hierarchical | 2 |
| Silhouette | 0.5363 | hierarchical | 2 |

3 rows

- Yukarıdaki tabloya baktığımızda 2 küme için hiyerarşik yöntemi önermiştir.

| | Score <dbl> | Method <fctr> | Clusters <fctr> |
|---|---|---|---|
| APN | 0.0231723 | hierarchical | 2 |
| AD | 3.2302855 | kmeans | 2 |
| ADM | 0.2229903 | kmeans | 2 |
| FOM | 0.7963785 | clara | 2 |

4 rows

- Bu tabloda ise 2 küme için hiyerarşik, k-means ve Clara'yı önermiştir.

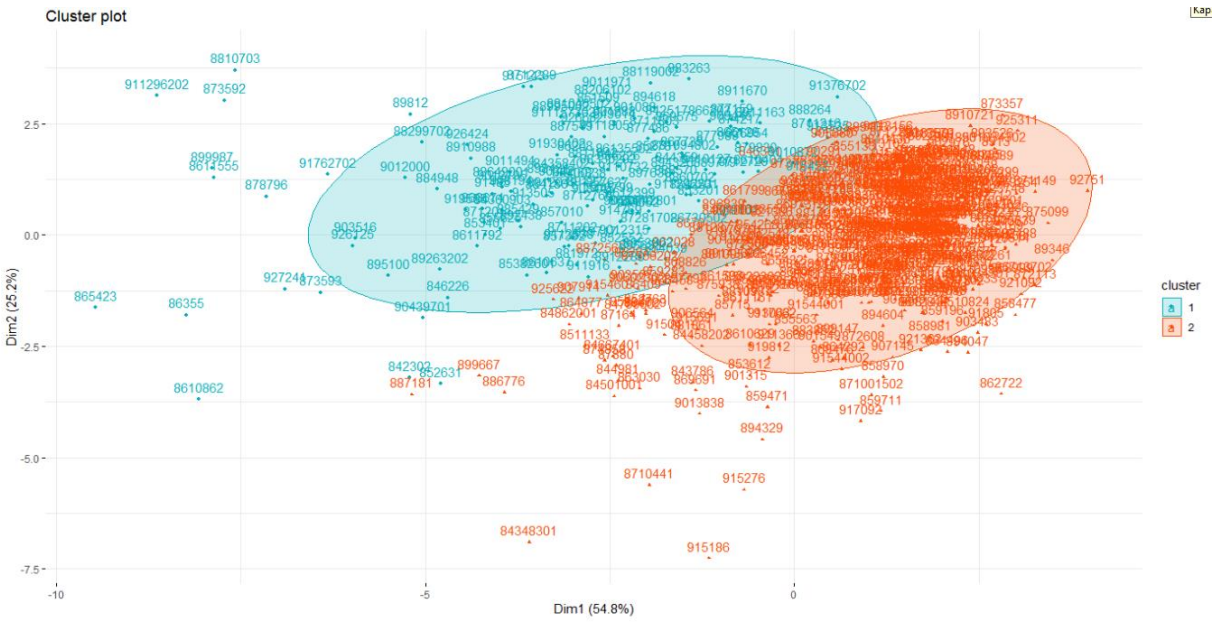
| | K-means | PAM | Clara | Hiyerarşik Pearson Ward.D2 | Hiyerarşik Euclidean Average | Hiyerarşik Manhattan Average | Hiyerarşik Pearson Average | Hiyerarşik Pearson Median | Model Tabanlı | Yoğunluk Tabanlı |
|---|---|---|---|---|---|---|---|---|---|---|
| Rand Endeks Değeri | 0.825 | 0.863 | 0.869 | 0.768 | 0.556 | 0.549 | 0.556 | 0.768 | 0.766 | 0.585 |

- Yukarıdaki tablolardan k-means, pam, clara, hiyerarşik ve model tabanlı kümeleme yöntemleri orijinal veri seti ile tekrar analiz yapılmıştır.
- Yapılan analizlerde yanlış kümeye atanan gözlem sayıları ve dolayısıyla rand indeks değerleri değişmiştir.
- Hücrenin çekirdek özellikleri üzerinden, o hücrenin iyi huylu ya da kötü huylu kanser hücresi olup olmadığını belirlemek üzere oluşturulan bu veri seti için en uygun küme sayısı 2 olarak belirlemiştik. Bununla birlikte yukarıdaki tabloda görüldüğü üzere hedef değişken ile bizim oluşturduğumuz kümeler arasındaki benzerliğe baktığımızda Clara yöntemine karar verdik.

**11. Finalde elde etmiş olduğunuz kümelerin tanımlayıcı istatistiklerini elde ederek yorumlayınız. (PCA skorlarını orijinal değerlerine çevirmeyi unutmayınız.)**

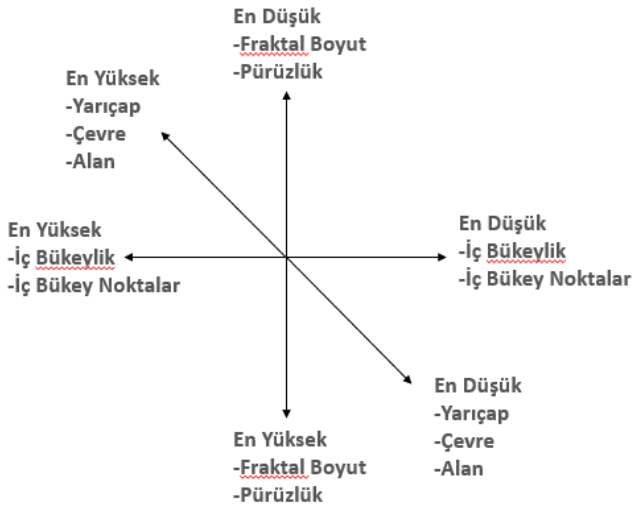
```
Call:     clara(x = data, k = 2, metric = "manhattan", samples = 50, pamLike = TRUE)
Medoids:
       yaricap  doku  cevre   alan puruzsuzluk yogunluk icbukeylik icbukeynoktalar simetri fraktalboyut
866203   19.00 18.91 123.40 1138.0     0.08217  0.08028    0.09271         0.05627  0.1946      0.05044
87930    12.47 18.60  81.09  481.9     0.09965  0.10580    0.08005         0.03821  0.1925      0.06373
Objective function:      155.7308
Clustering vector:       Named int [1:569] 1 1 1 2 1 2 1 2 2 2 2 2 1 2 2 2 2 2 ...
 - attr(*, "names")= chr [1:569] "842302" "842517" "84300903" "84348301" "84358402" "843786" "844359" ...
Cluster sizes:           136 433
Best sample:
 [1] 84300903 84667401 857374   861799   863270   86355    864729   866203   868682   86973702 871642   874662
[13] 87556202 87930    8811523  8812818  88147102 88330202 884437   88466802 885429   889403   8913     892438
[25] 892604   893783   898690   9010877  907145   908194   909445   9110732  9111843  9112085  9112366  91376702
[37] 914769   914862   91505    916799   921362   921644   924632   926424

Available components:
 [1] "sample"     "medoids"    "i.med"      "clustering" "objective"  "clusinfo"   "diss"       "call"
 [9] "silinfo"    "data"
```

- Clara yöntemini uygularken öklit, manhattan ölçüleri denenmiştir ve uç değerlerimizde fazla olduğu için manhattan uzaklık ölçüsü kullanılmıştır.
- Clara yöntemine göre 866203 ve 87930 gözlemleri küme merkezleri olarak belirlenmiştir.
- Amaç fonksiyon değeri 155.7308 çıkmıştır. Bu değer oldukça büyüktür fakat orijinal kümeyi en iyi açıklayan yöntem budur.




- Clara yöntemiyle elde ettiğimiz kümelerin frekansları ile hedef değişken frekanslarını karşılaştırdığımızda 84 gözlem yanlış kümeye atanmıştır.
- Örnek olarak 91376702 gözlemi hedef değişkende iyi huylu kanser hücresi iken bizim oluşturduğumuz kümelemede kötü huylu kanser hücresi kümesine atanmıştır.
- Örnek olarak 84667401 gözlemi hedef değişkende kötü huylu kanser hücresi iken bizim oluşturduğumuz kümelemede iyi huylu kanser hücresi kümesine atanmıştır.

- Rand indeksine baktığımızda %74 benzerlik bulunmuştur.
- 8810703 numaralı kişinin hücre çekirdeğine ait yarıçapı, çevresi ve alanı en yüksek değerlere sahiptir.
- 865423 numaralı kişinin hücre çekirdeğine ait iç bükeyliği ve iç bükey noktaları en yüksek değere sahiptir.
- 915186 numaralı kişinin hücre çekirdeğine ait en yüksek pürüzlük ve yüksek bir fraktal boyut değerine sahiptir.
- 925311 numaralı kişinin hücre çekirdeğine ait iç bükeylik, iç bükey noktalar, simetri ve fraktal boyutta en düşük değerlere sahiptir.
- 862722 numaralı kişinin hücre çekirdeğine ait yarıçap, çevre, alan, iç bükeylik ve iç bükey noktalar en düşük değerlere sahiptir.




- Küme yorumlarını kolaylaştırmak için yukarıdaki grafiği oluşturduk böylelikle belirgin olarak yorumlanabilmektedir.
- 1. küme yani kötü huylu kanser hücresi için yarıçap, çevre, alan, iç bükeylik ve iç bükey noktalar değişkenleri yüksek değerlere sahip çıkmıştır.
- 2. küme yani iyi huylu kanser hücreleri yukarıda belirtilen değişkenlerin düşük değerlerine sahip olduğu görülmektedir.

***Öneri:***

Kötü huylu kanser hücrelerini iyi huylu yapabilme imkânımız varsa bunun için yarıçap, çevre, alan, iç bükeylik, iç bükey noktalar değişkenlerinin değerlerini azaltacak yöntemler kullanılır ya da geliştirilirse kötü huylu kanser sorunlarına çözüm bulunulabilir.