T.C
DOKUZ EYLÜL ÜNİVERSİTESİ
FEN FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

DÜNYA MUTLULUK VERİLERİ ANALİZ RAPORU

Ahmet ÖZTÜRK

Danışman: Prof. Dr. Neslihan DEMİREL

Ocak / 2025

İZMİR

İÇİNDEKİLER

Özet

Bu çalışma, Dünya Mutluluk Verisi kullanılarak ülkelerin mutluluk seviyelerini etkileyen temel faktörlerin incelenmesi ve bu faktörlerin ülkeler arası farklılıklarına ışık tutulması amacıyla gerçekleştirilmiştir. Veri seti, 150'den fazla ülkenin mutluluk skorlarını, kişi başına düşen gelir, sosyal destek, sağlıklı yaşam beklentisi, özgürlük, yardımseverlik ve yolsuzluk algısı gibi göstergelerle ilişkilendirmektedir.

Çalışmada, denetimsiz makine öğrenmesi yöntemleri uygulanmıştır. K-Means, K-Medoids ve Hiyerarşik Kümeleme algoritmaları kullanılarak ülkeler gruplandırılmış, Temel Bileşenler Analizi (PCA) ile veri boyutu azaltılarak analiz sonuçlarının daha net bir şekilde yorumlanması sağlanmıştır.

Analiz sonuçları, kişi başına düşen gelir ve sosyal destek gibi ekonomik ve sosyal göstergelerin mutluluk üzerindeki önemli etkisini ortaya koymuştur. Ayrıca, düşük yolsuzluk algısı ve bireylerin özgürlük düzeyi, mutluluk skorlarını artıran diğer önemli faktörler arasında yer almıştır.

Bu bulgular, ülkelerin sosyal ve ekonomik politikalarını geliştirmelerine katkı sağlamakta ve araştırmacılar ile politika yapıcılar için rehber niteliği taşımaktadır. Çalışma, mutluluk seviyelerinin çok boyutlu bir perspektiften ele alınması gerektiğini vurgulamakta ve sosyal refahı artırmaya yönelik stratejik kararlar için bilimsel bir temel sunmaktadır.

1.Bölüm

KÜMELEME YÖNTEMLERİ

1.1 Kümeleme Analizi

Kümeleme analizi, bir veri setinde farklı grupların varlığını belirlemek ve eğer varsa bu grupları tanımlamak amacıyla kullanılan çok değişkenli istatistiksel bir yöntemdir. Çok boyutlu uzayda verilerin özetlenmesi ve tanımlanması için rehberlik eden bir araştırma yöntemi olan Kümeleme analizi, heterojen gruplardaki farklı gözlem yapılarını veya homojen gruplardaki benzer gözlemleri uygun yöntemlerle gruplamayı sağlayan bir tekniktir. Kümeleme analizi, normallik, doğrusalık ve homojenlik gibi diğer çok değişkenli istatistiksel yöntemlerde önemli olan varsayımları prensipte korur, ancak uzaklık değerlerinin normalliğine daha fazla vurgu yapar. Bu yöntemde, gözlemler arasındaki benzerlikleri ölçen uzaklık metrikleri kullanılarak gruplar oluşturulur. Kümeleme analizinin genel amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya veri setindeki desenleri ve ilişkileri anlamada yardımcı olacak özetleyici bilgiler sağlamaktır.

1.2K-Means

K-means kümelemesi, belirli bir veri kümesini bir dizi k grubuna (yani k kümeleri) bölmek için en yaygın kullanılan denetimsiz makine öğrenimi algoritmasıdır; burada k, analist tarafından önceden belirtilen grup sayısını temsil eder. Nesneleri birden fazla grup halinde (yani kümeler) sınıflandırır, öyle ki aynı küme içindeki nesneler mümkün olduğunca benzerdir (yani, yüksek sınıf içi benzerlik), oysa nesneler daha az benzerlik gösterir. K-means kümelemesinde, her küme, kümeye atanan noktaların ortalamasına karşılık gelen merkezi (yani, ağırlık merkezi) ile temsil edilir.

1.2.1 K-Means Algoritması

K Means algoritması, verileri belirlenen k sayısına göre gruplandırır ve her küme için bir merkez (centroid) belirler. Sonuç olarak, her küme içindeki benzerlik yüksek, kümeler arasındaki benzerlik düşüktür. Bu sayede, müşteri segmentasyonu veya veri analizi gibi

alanlarda etkili bir kümeleme sağlanır. Başarılı bir sonuç, doğru k değerinin seçimine bağlıdır.

1.3 K-Medoids

K-medoids algoritması, bir veri kümesini k gruba veya kümeye ayırmak için k-means kümelemesi ile ilgili bir kümeleme yaklaşımıdır. K- medoids kümelemesinde her bir küme, kümedeki veri noktalarından biri ile temsil edilir. Bu noktalar **küme medoidleri(ortalama)** olarak adlandırılır. Medoid terimi, bir küme içindeki, kendisiyle kümenin diğer tüm üyeleri arasındaki ortalama farklılığın minimum olduğu bir nesneyi ifade eder. Küme içinde en merkezi konumda bulunan noktaya karşılık gelir. K-medoid, k-means kümelemesine sağlam bir alternatiftir. En yaygın k-medoids kümeleme yöntemi PAM algoritmasıdır.

1.3.1 PAM Kavramı

Ortalamaların kullanılması, k-means kümelemesinin aykırı değerlere karşı oldukça hassas olduğu anlamına gelir. Bu durum gözlemlerin kümelere atanmasını ciddi şekilde etkileyebilir. Daha sağlam bir algoritma PAM algoritması tarafından sağlanmaktadır.

1.3.2 PAM Algoritması

PAM algoritması, veri setindeki gözlemler arasında k adet temsili nesne veya medoid aramaya dayanır.

PAM algoritması şu şekilde ilerler :

1. Medoid olacak k nesneyi seçin veya bu nesnelerin sağlanmış olması durumunda bunları medoid olarak kullanın.

2. Sağlanmamışsa benzerlik matrisini hesaplayın.

3. Her nesneyi kendisine en yakın medoide atayın.

4. Her küme için, kümedeki herhangi bir nesnenin ortalama benzemezlik katsayısını azaltıp azaltmadığını araştırın; eğer azaltıyorsa, bu katsayıyı en çok azaltan varlığı bu küme için medoid olarak seçin.

5. En az bir medoid değişmişse (3)'e gidin, aksi takdirde algoritmayı sonlandırın.

1.4 Hiyerarşik Kümeleme

Verileri bir hiyerarşi içinde, iç içe geçmiş kümeler halinde organize eden kümeleme yöntemidir. Bu hiyerarşi genellikle bir dendrogram (ağaç diyagramı) ile görselleştirilir. Bu yöntem, verilerin yapısını anlamak ve istenilen detayda kümeleme sonuçları elde etmek için oldukça faydalıdır.

Hiyerarşik kümelemenin iki temel yaklaşımı vardır:

1.Birleştirici (Agglomerative) Kümeleme: Her veri noktasını başlangıçta ayrı bir küme olarak ele alır ve ardından en benzer kümeleri ardışık olarak birleştirerek yukarıdan aşağıya bir hiyerarşi oluşturur. (Aşağıdan yukarıya yaklaşımı)

2.Bölücü (Divisive) Kümeleme: Tüm veri noktalarını başlangıçta tek bir küme olarak ele alır ve ardından kümeyi ardışık olarak daha küçük kümelere bölerek yukarıdan aşağıya bir hiyerarşi oluşturur. (Yukarıdan aşağıya yaklaşımı)

Kümeler arasındaki mesafeyi ölçmek için farklı yöntemler kullanılır. Bunlara bağlantı (linkage) yöntemleri denir:

1. Tek Bağlantı (Single Linkage): Kümelerdeki en yakın iki nokta arasındaki mesafe.
2. Tam Bağlantı (Complete Linkage): Kümelerdeki en uzak iki nokta arasındaki mesafe.
3. Ortalama Bağlantı (Average Linkage): Kümelerdeki tüm nokta çiftleri arasındaki mesafelerin ortalaması.
4. Ward Yöntemi: Küme içi varyansı en aza indirmeye çalışan yöntem.

Hiyerarşik kümelemenin sonuçları genellikle bir dendrogram ile görselleştirilir. Dendrogram, kümelerin nasıl birleştirildiğini veya bölündüğünü gösteren bir ağaç diyagramıdır.

1.5 Temel Bileşenler Analizi(PCA)

PCA, özellikle büyük boyutlu veri setlerinde (çok sayıda değişken içeren) veriyi özetlemek için kullanılır. Amaç, veri setinin özündeki yapıyı koruyarak boyutunu azaltmaktır. PCA, veriyi farklı bir koordinat sistemine dönüştürür ve bu sistemdeki yeni eksenler (bileşenler) veri setindeki en büyük varyansı sırasıyla temsil eder. Yeni eksenler orijinal değişkenlerin doğrusal kombinasyonlarıdır.

Büyük veri setlerinde, birçok değişken birbiriyle yüksek korelasyon gösterebilir. Bu durum analizlerde bilgi fazlalığına yol açar. PCA, bu fazlalığı azaltır. Örneğin, görselleştirme için genellikle 2 veya 3 boyut yeterlidir. PCA, bu tür uygulamalarda kritik bir araçtır.

PCA'nın 4 temel amacı:

- **Boyut İndirgeme:** Veri setindeki değişken sayısını azaltmak
- **Özellik Çıkarımı:** Veriyi daha anlamlı özelliklere dönüştürmek
- **Korelasyonu Azaltma:** Yüksek korelasyonlu değişkenleri bağımsız bileşenlere dönüştürmek.
- **Veri Görselleştirme:** Çok boyutlu veriyi 2D veya 3D formatta anlamayı kolaylaştırmak

2.Bölüm

VERİ SETİ TANITIMI

2.1 Analizin amacı

Çalışmamızın temel amacı ; World Happiness veri seti üzerinde denetimsiz makine öğrenmesi uygulamaktır.

2.2 Veri Seti Hakkında Bilgi

Dünya Mutluluk Raporu Verisi, dünya genelinde ülkelerin mutluluk seviyelerini ve bu seviyeleri etkileyen faktörleri anlamak için kapsamlı bir kaynak sunar. 150'den fazla ülkeye ait kayıtlarla bu veri seti, ülkelerin refah düzeylerini karşılaştırmak, mutluluğu etkileyen faktörleri belirlemek ve sosyal politikalar geliştirmek isteyen araştırmacılar, veri bilimcileri ve politika yapıcılar için ideal bir araçtır.

DEĞİŞKENLER:

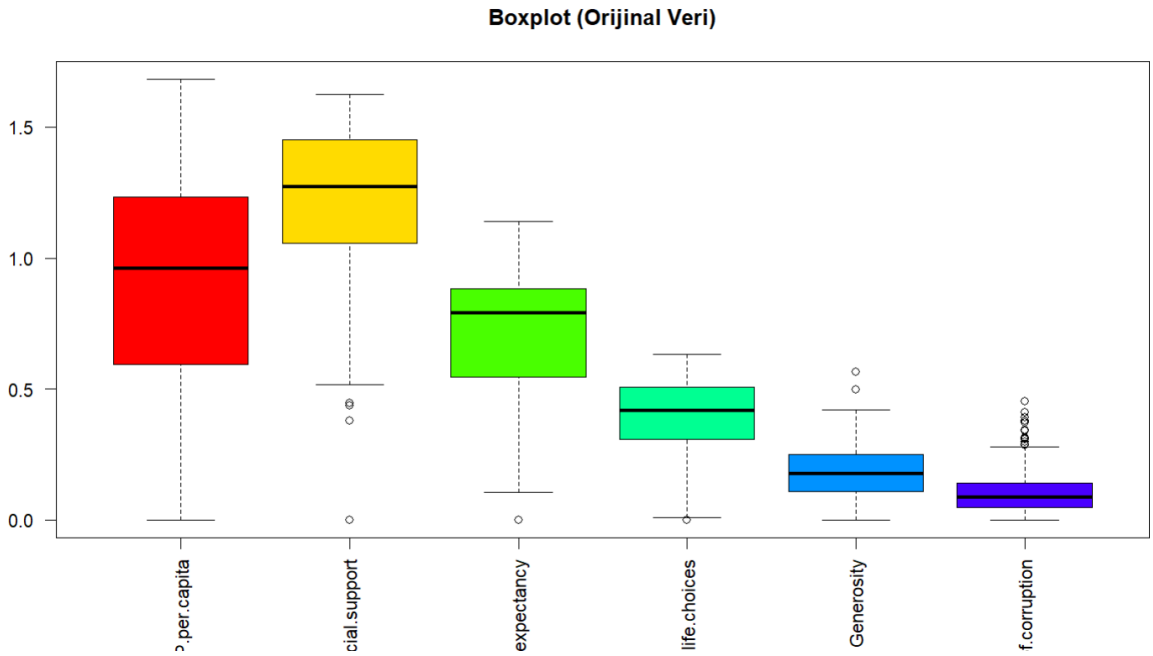
- **Overall Rank (Genel sıralama)**: Ülkelerin mutluluk skorlarına göre sıralandığı ve her bir ülkenin mutluluk düzeyine göre aldığı genel sıralamayı gösteren bir değişkendir.
- **Country (Ülke)**: Veri setindeki her gözlem bir ülkeyi temsil eder. Bu değişken, her bir ülkenin adını içerir ve ülkeler arasındaki mutluluk farklarını analiz etmek için temel bir kategorik değişkendir.
- **Happiness Score (Mutluluk Skoru)**: Her ülkenin mutluluk düzeyini 0 ile 10 arasında bir ölçekle gösterir. Bu skor, bir ülkenin genel yaşam kalitesi, sosyal refah ve ekonomik koşullarıyla doğrudan ilişkilidir.

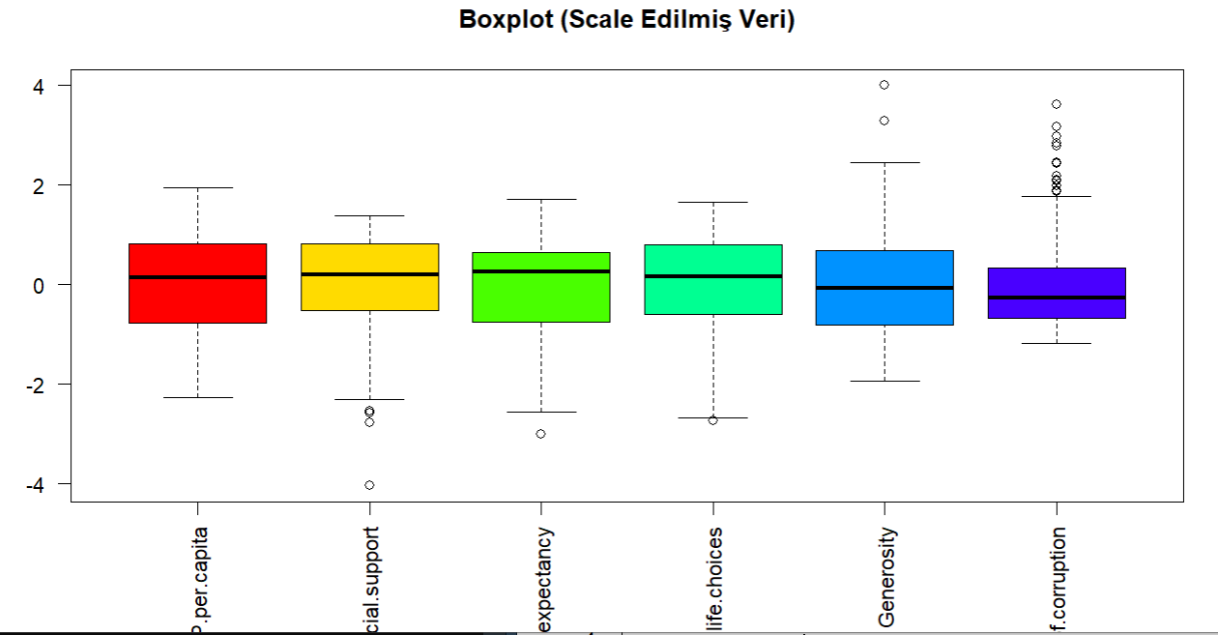
- **GDP per Capita (Kişi Başına Düşen Gelir)**: Bir ülkenin kişi başına düşen Gayri Safi Yurtiçi Hasıla (GSYİH) değeridir. Bu ekonomik gösterge, bir ülkenin zenginliğini ve ekonomik gelişmişliğini yansıtır.
- **Social Support (Sosyal Destek)**: Bireylerin, zor zamanlarında yardım alabileceği sosyal ağların gücünü gösterir. Bu değişken, sosyal bağlantıların mutluluk üzerindeki etkisini anlamak için kullanılır.
- **Healthy Life Expectancy (Sağlıklı Yaşam Beklentisi)**: Bir kişinin sağlıklı bir yaşam sürme süresinin beklenen değerini gösterir. Sağlık durumu, mutluluk düzeyini etkileyen önemli bir faktördür.
- **Freedom to Make Choices (Tercih Yapma Özgürlüğü)**: Bireylerin kendi yaşamlarını yönlendirme ve kararlar alma özgürlüğünü ifade eder. Yüksek özgürlük seviyesi, genellikle daha yüksek mutluluk seviyeleri ile ilişkilidir.
- **Generosity (Yardımseverlik)**: İnsanların başkalarına yardım etme ve toplumsal sorumluluk taşıma düzeyini ölçen bir göstergedir. Yardımseverlik, toplumsal bağları ve bireysel mutluluğu artırabilir.
- **Perceptions of Corruption (Yolsuzluk Algısı)**: Ülkelerdeki yolsuzluk seviyelerini algılayan bireylerin görüşlerini yansıtan bir değişkendir. Yüksek yolsuzluk algısı, genellikle düşük mutluluk skorlarıyla ilişkilidir.

2.2.1Veri Seti Çıkarılan Değişkenler

Veri setinde ilk 3 değişken kategorik olduğu için veri setimizden çıkardık ve 4. değişkenimiz de etiket değişkeni olduğu için veri setimizden çıkarılmasını uygun gördük.

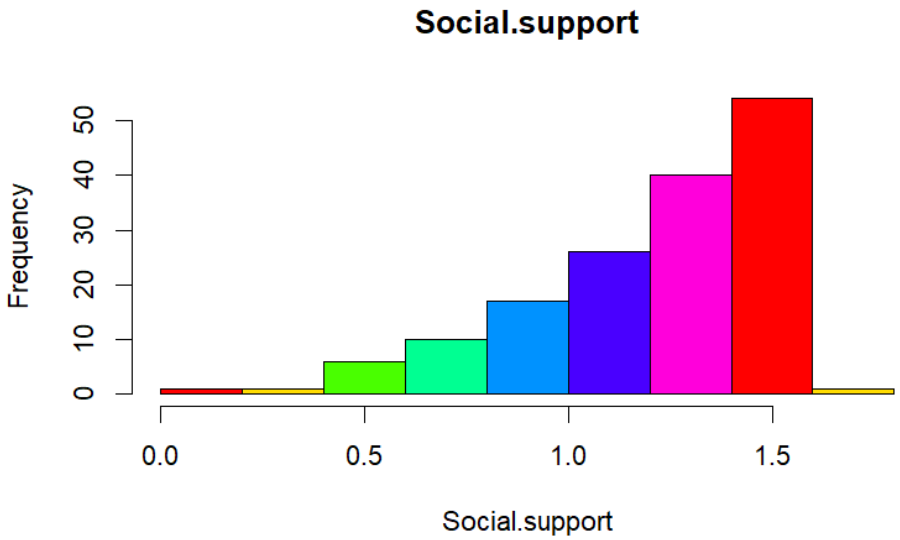
2.2.2 Veri Setinin Görüntüsü

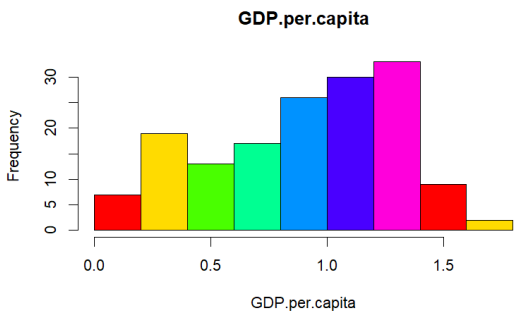


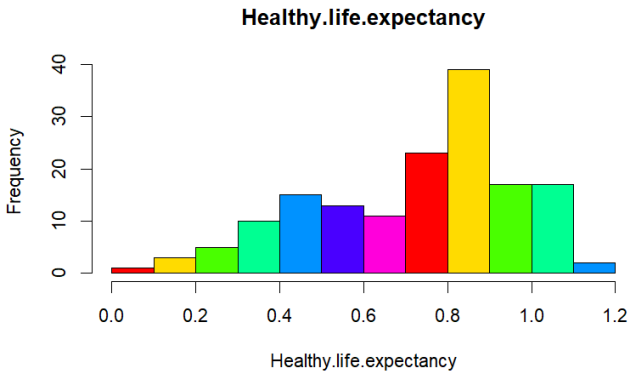


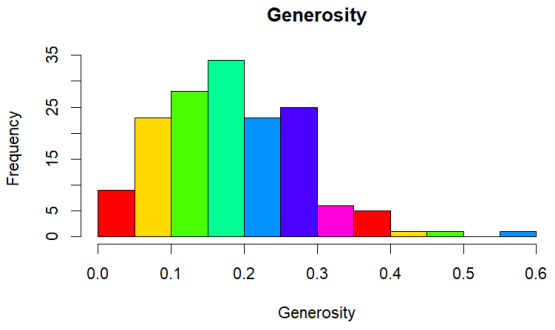
Bu bölümde veri setindeki değişkenlerin görselleştirilmesi için boxplot kullanılmıştır. Veri setinin orijinal halinin boxplotını ve standartlaştırılmış(scale) boxplotı görebilirsiniz.

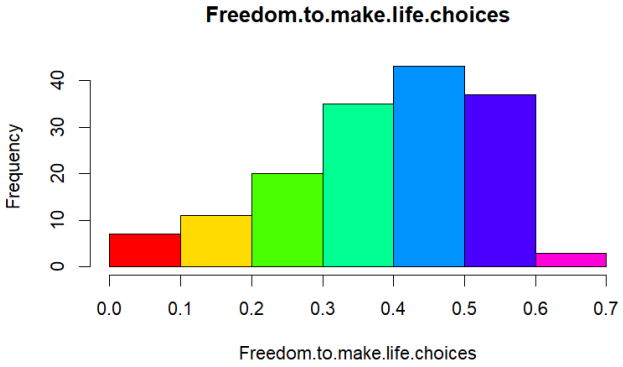
HİSTOGRAMLARI

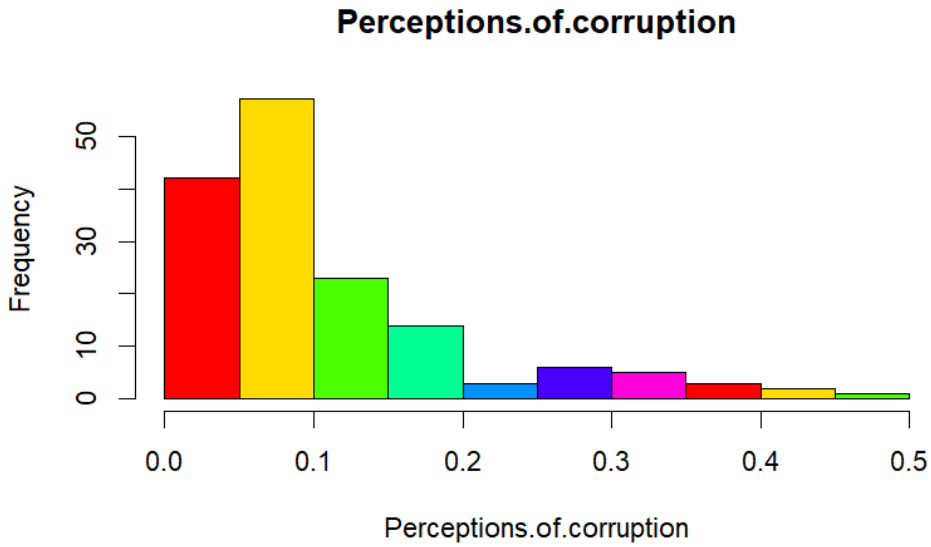












Bu kısımda veri setindeki değişkenlerin dağılımını görmek için çizilmiş histogramları görebilirsiniz.

2.2.3Özet İstatistikler

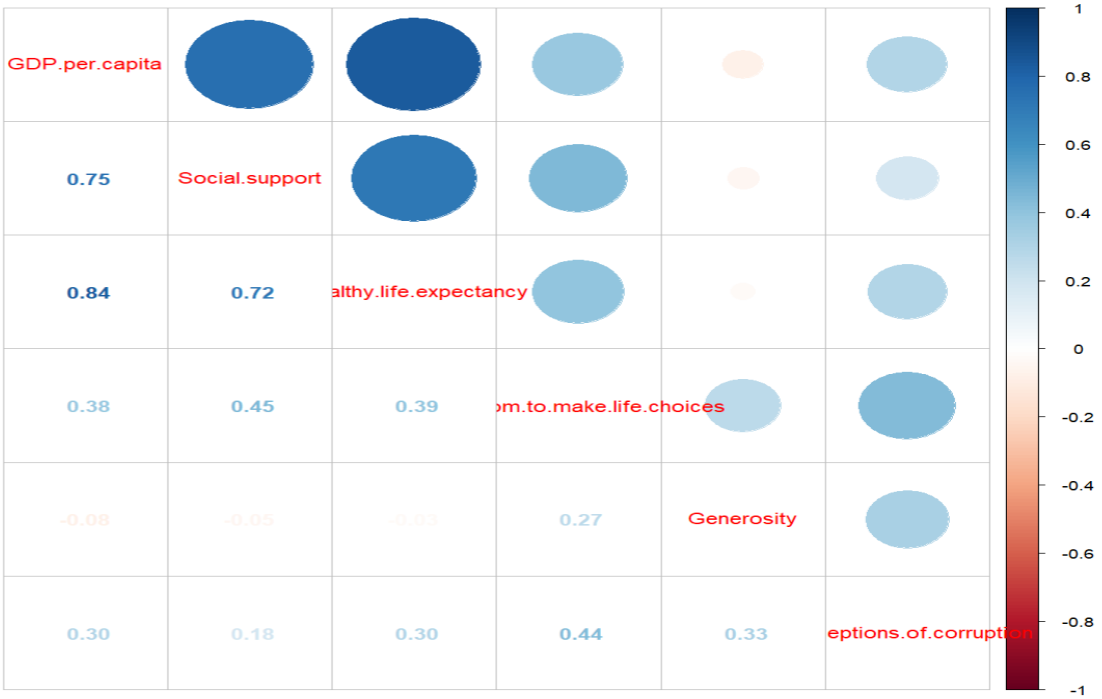
```
> summary(x)
 GDP.per.capita   Social.support  Healthy.life.expectancy
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477
 Median :0.9600   Median :1.272   Median :0.7890
 Mean   :0.9051   Mean   :1.209   Mean   :0.7252
 3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818
 Max.   :1.6840   Max.   :1.624   Max.   :1.1410
 Freedom.to.make.life.choices   Generosity     Perceptions.of.corruption
 Min.   :0.0000               Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.3080               1st Qu.:0.1087   1st Qu.:0.0470
 Median :0.4170               Median :0.1775   Median :0.0855
 Mean   :0.3926               Mean   :0.1848   Mean   :0.1106
 3rd Qu.:0.5072               3rd Qu.:0.2482   3rd Qu.:0.1412
 Max.   :0.6310               Max.   :0.5660   Max.   :0.4530
>
```

Veri setinin temel özelliklerini anlamak amacıyla özet istatistikler çıkarılmıştır. Yukarıda yer alan tablo, değişkenlerin merkezi eğilim ve dağılım ölçülerini içermektedir.

- **Ortalamalar**: Veri setindeki değişkenlerin genel eğilimlerini belirlemek için kullanılmıştır. Örneğin, mutluluk skoru değişkeninin ortalaması, ülkelerin genel mutluluk seviyelerini karşılaştırmak için önemli bir referans noktası sunmaktadır.
- **Standart Sapmalar**: Değişkenlerin dağılımlarını ve veri setindeki varyasyonu ölçmek için hesaplanmıştır. Örneğin, kişi başına düşen gelirdeki yüksek standart sapma, ülkeler arasındaki ekonomik farklılıkların belirgin olduğunu göstermektedir.
- **Minimum ve Maksimum Değerler**: Her değişkenin aldığı en düşük ve en yüksek değerler, veri setindeki ekstrem noktaları tespit etmek için incelenmiştir. Bu analiz, özellikle outlier'ların (aykırı değerler) belirlenmesi açısından önemlidir.
- **Çeyrek Değerler**: İlk (%25), medyan (%50), ve üçüncü çeyrek (%75) değerleri, veri setinin dağılımını daha ayrıntılı bir şekilde anlamamızı sağlamaktadır.

Bu özet istatistikler, veri setinin genel yapısını anlamada kritik bir rol oynamış ve analiz için temel oluşturmuştur.

2.2.4Korelasyon Matrisi



```
> cor(x)
                                GDP.per.capita Social.support Healthy.life.expectancy
GDP.per.capita                      1.00000000     0.75490573              0.83546212
Social.support                      0.75490573     1.00000000              0.71900946
Healthy.life.expectancy             0.83546212     0.71900946              1.00000000
Freedom.to.make.life.choices        0.37907907     0.44733316              0.39039478
Generosity                         -0.07966231    -0.04812645             -0.02951086
Perceptions.of.corruption           0.29891985     0.18189946              0.29528281
                                Freedom.to.make.life.choices  Generosity
GDP.per.capita                                     0.3790791 -0.07966231
Social.support                                     0.4473332 -0.04812645
Healthy.life.expectancy                            0.3903948 -0.02951086
Freedom.to.make.life.choices                       1.0000000  0.26974181
Generosity                                         0.2697418  1.00000000
Perceptions.of.corruption                          0.4388433  0.32653754
                                Perceptions.of.corruption
GDP.per.capita                                  0.2989198
Social.support                                  0.1818995
Healthy.life.expectancy                         0.2952828
Freedom.to.make.life.choices                    0.4388433
Generosity                                      0.3265375
Perceptions.of.corruption                       1.0000000
```

Veri setindeki değişkenler arasındaki doğrusal ilişkileri incelemek için korelasyon matrisi oluşturulmuştur. Korelasyon matrisi, değişken çiftleri arasındaki ilişkiyi Pearson korelasyon katsayısı ile ölçer ve -1 ile 1 arasında bir değer alır:

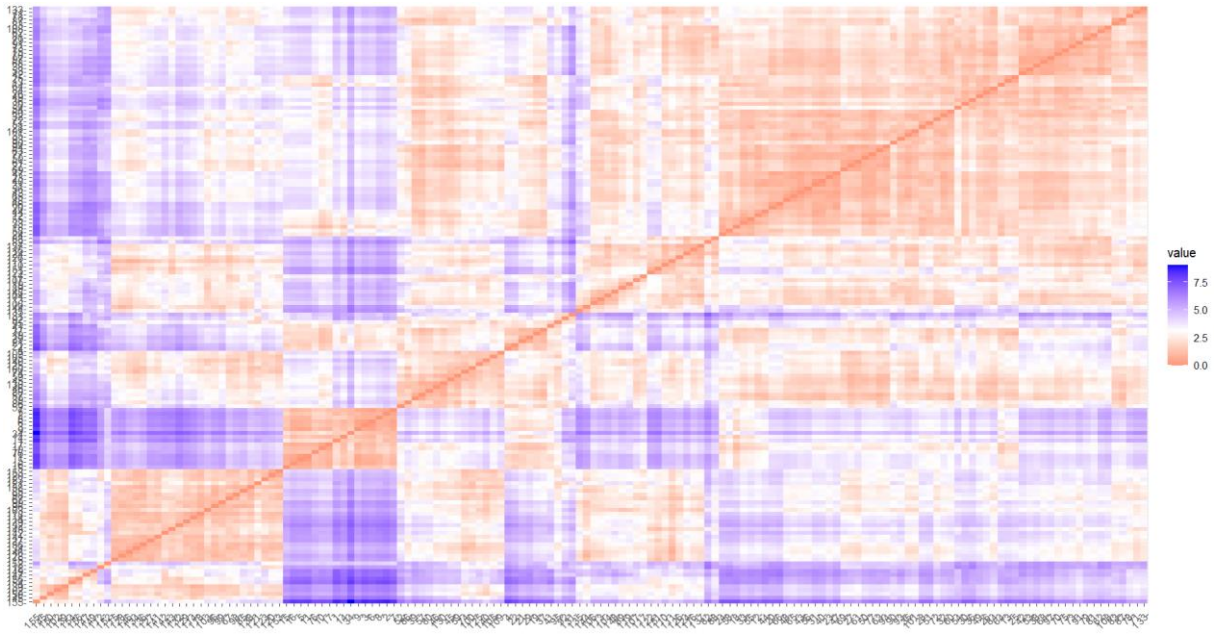
- **1**: Güçlü pozitif ilişkiyi ifade eder; bir değişken artarken diğeri de artar.
- **0**: İki değişken arasında hiçbir doğrusal ilişki olmadığını gösterir.
- **-1**: Güçlü negatif ilişkiyi ifade eder; bir değişken artarken diğeri azalır.

Korelasyon matrisi, veri setindeki önemli değişkenler arasındaki ilişkiyi şu şekilde özetlemiştir:

- **Kişi Başına Düşen Gelir (GDP per Capita)** ile **Sosyal Destek (Social Support)** arasında güçlü bir pozitif ilişki gözlenmiştir. Bu, ekonomik refahı yüksek ülkelerde sosyal ağların da güçlü olduğunu göstermektedir.
- **Sağlıklı Yaşam Beklentisi (Healthy Life Expectancy)** ile **Sosyal Destek (Social Support)** arasında anlamlı bir pozitif korelasyon bulunmuştur. Sağlıklı yaşam koşullarına sahip ülkelerde bireylerin sosyal destek alma oranlarının yüksek olduğu görülmüştür.

Korelasyon matrisi, değişkenler arasındaki temel ilişkileri belirlemek ve analizde kullanılacak değişkenlerin seçiminde rehberlik etmek için önemli bir araç olmuştur.

2.2.5Uzaklık Matrisi



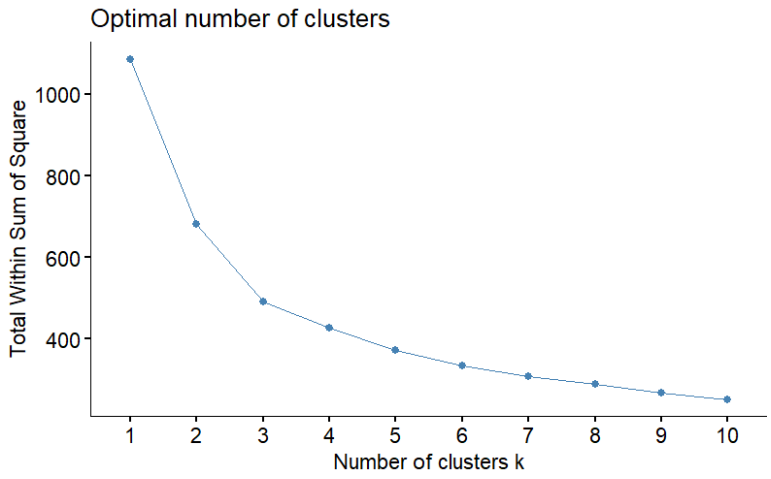
Gözlemlerin gruplar halinde sınıflandırılması bazı hesaplama yöntemleri gerektirir. Her bir gözlem çifti arasındaki mesafe veya benzemezlik. Sonucu bu hesaplama, farklılık veya uzaklık matrisi olarak bilinir. Yukarıdaki görselde uzaklık matrisini görebilirsiniz.

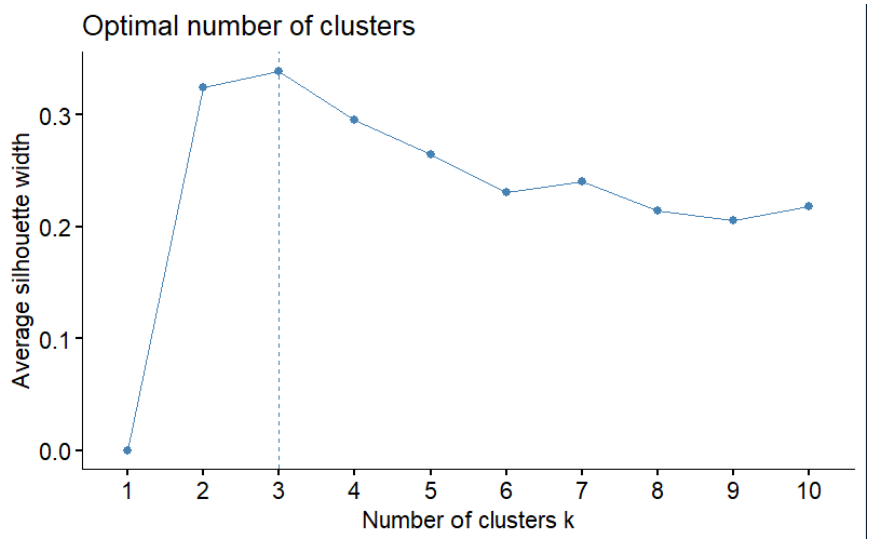
3.Bölüm

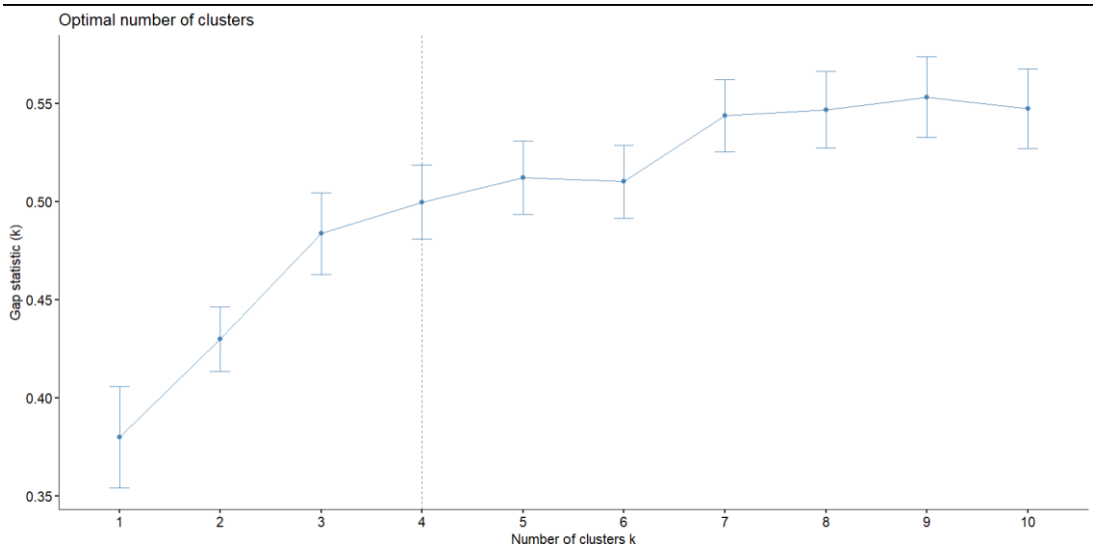
UYGULAMA

3.1K-Means Uygulaması

3.1.1Küme Sayısı Belirleme

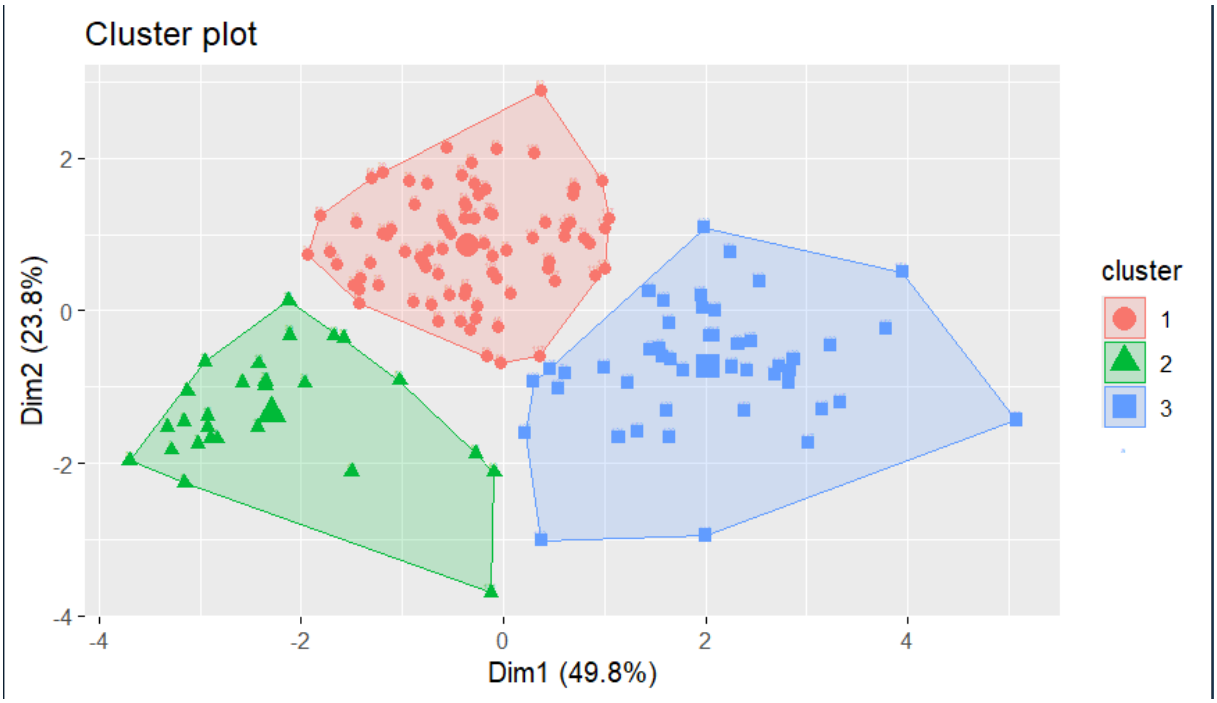




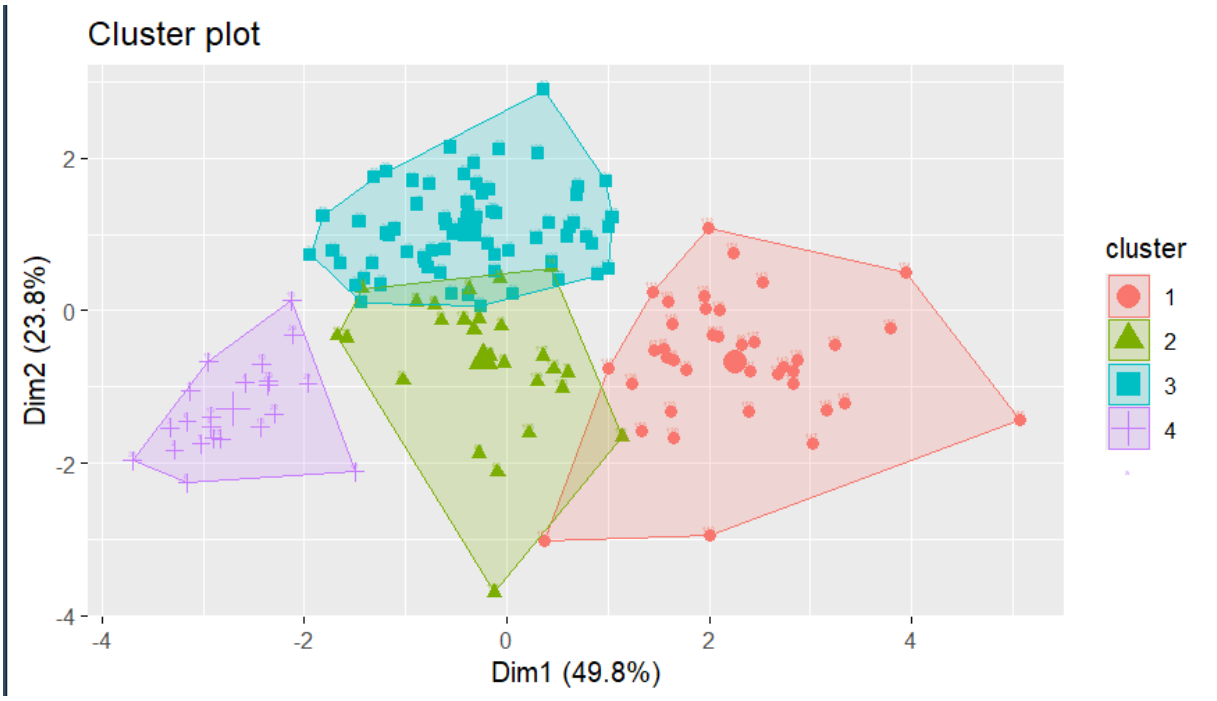


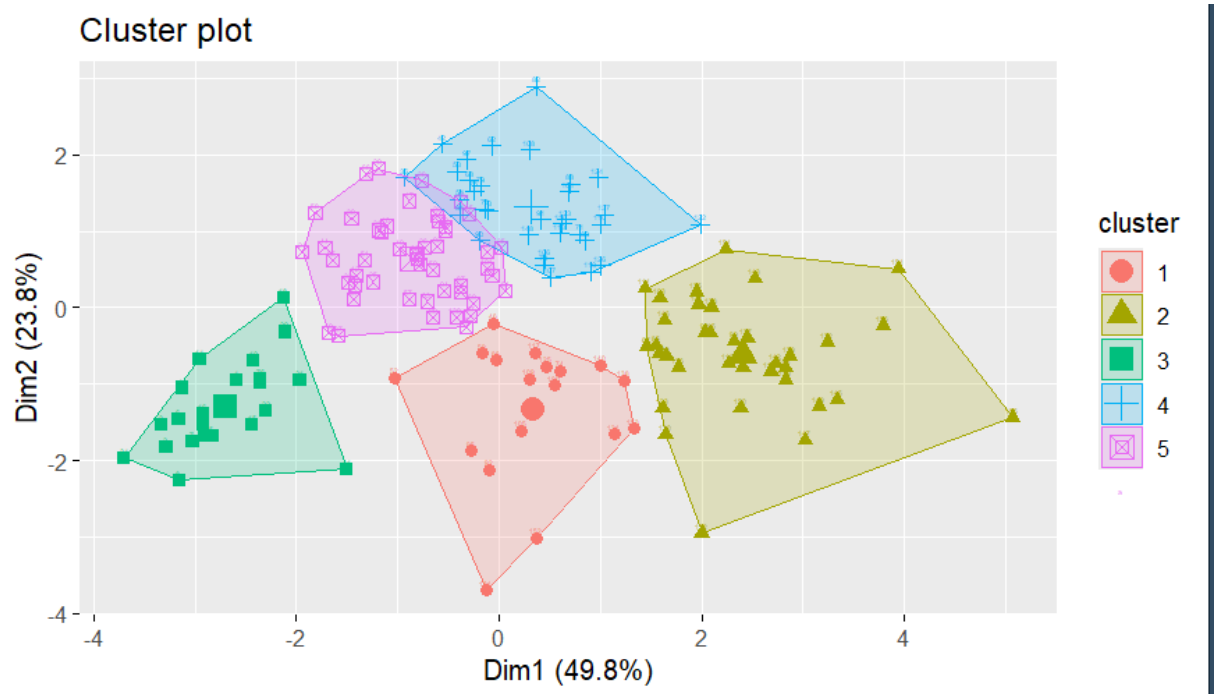
Yukarıdaki şekillerde optimal küme sayısını belirlemek için K-Means algoritmaları kullanılmış olup çıktıları verilmiştir. Görsellere baktığımızda optimal küme sayısını ilk iki görselimizde 3, son görselimizde 4 olarak görüyoruz. Bu algoritma bize küme sayısını belirlemek için yalnızca fikir verir. Görsellerdeki küme sayısını seçmek zorunda değiliz.

3.1.2 Küme Sayısına Göre Görselleştirme



Küme sayısı 3 olduğunda Cluster Plot görselini görebiliyoruz. Dim1 ve Dim2 ye baktığımızda varyansın %73.6'sını açıkladığını görebiliyoruz bu da oldukça yüksek bir oran diyebiliriz.

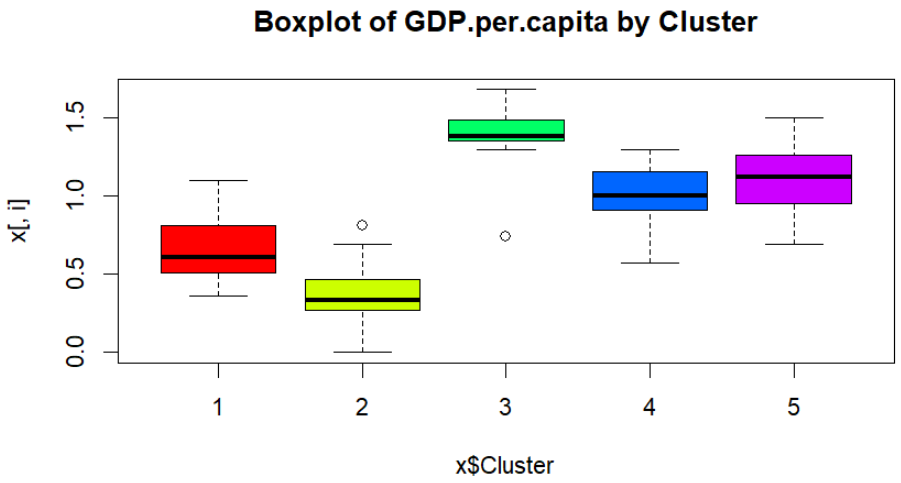


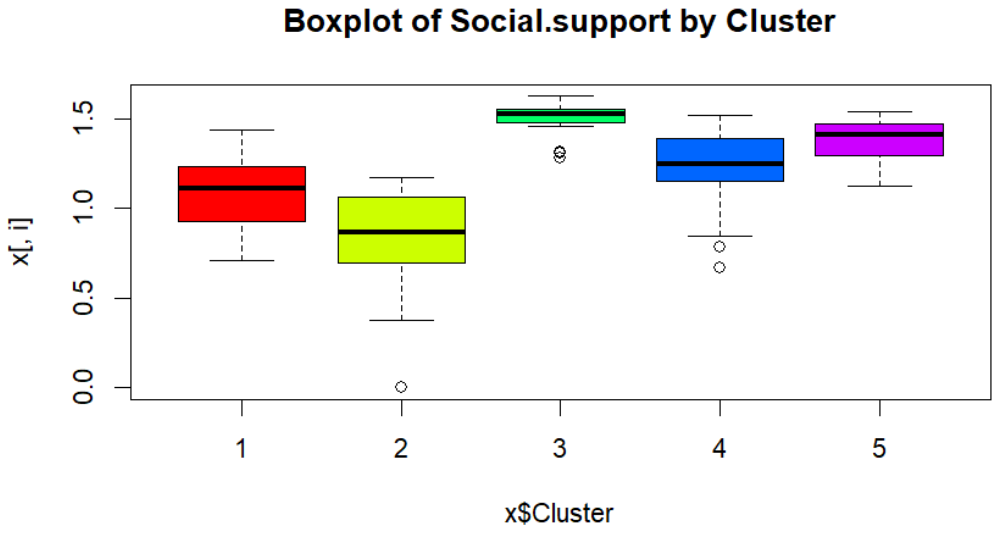


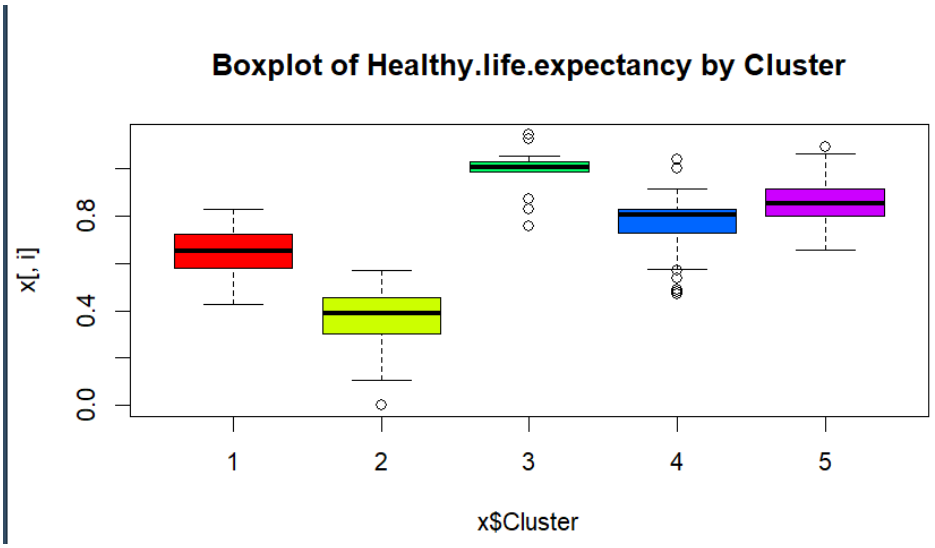
Yukarıdaki görsellerde küme sayısı 3,4 ve 5 olan görselleri görebiliyoruz. Küme sayısı 5 olan görselimizde kümeler arası farklılığı daha iyi açıklayabileceğimizi düşündüğümüz için küme sayısı 5 olanı alıyoruz.

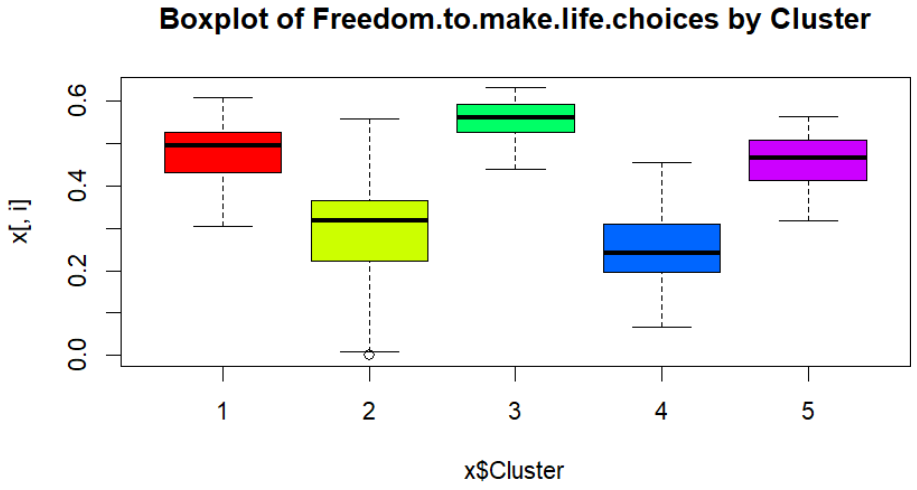
3.1.3Boxplot by Factor

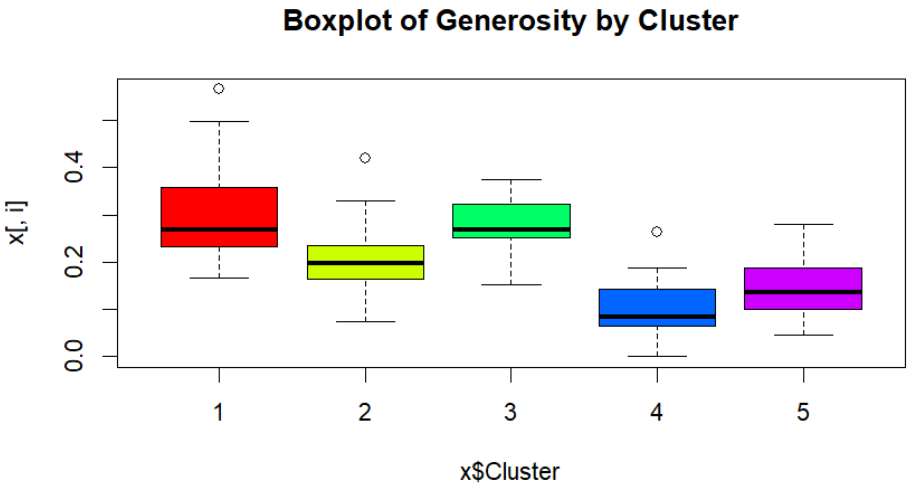
Küme sayısını 5 olarak belirlemiştik. Bu bölümde ise değişkenlerin kümeler arasındaki dağılımlarında fark olup olmadığına bakmak için Boxplot by Factor uyguladık. Aşağıdaki görsellere baktığımızda değişkenlerin dağılımının kümeler arasında heterojen olduğunu söyleyebiliriz.

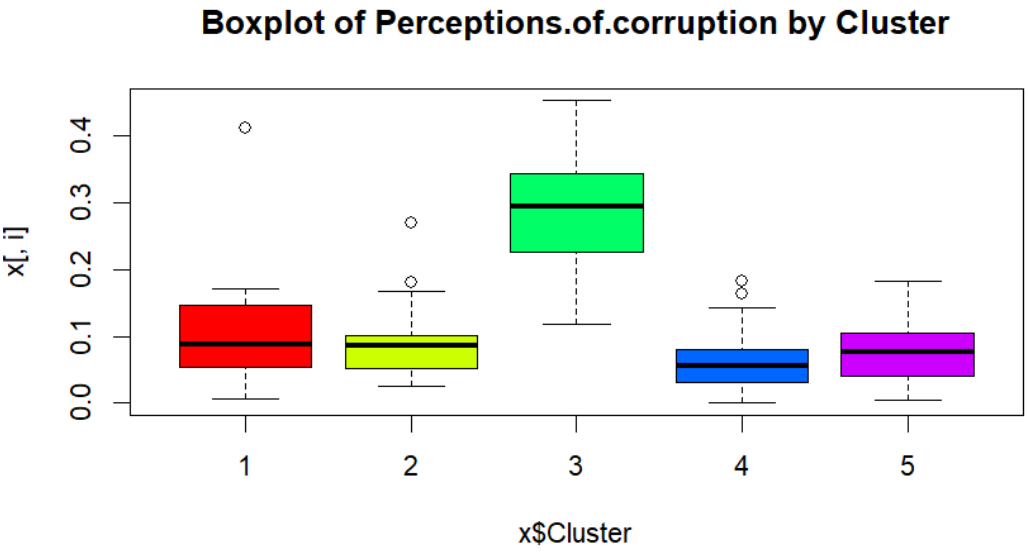





Boxplot of Healthy.life.expectancy by Cluster
x[, i]
0.0
0.4
0.8
1
2
3
4
5
x$Cluster

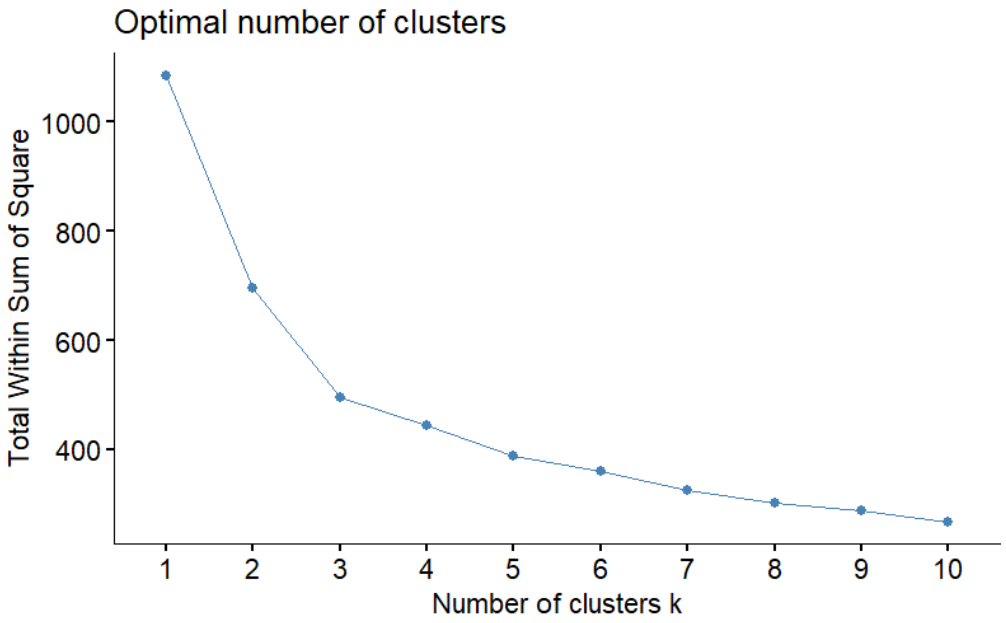

Boxplot of Freedom.to.make.life.choices by Cluster
x[, i]
0.0
0.2
0.4
0.6
1
2
3
4
5
x$Cluster

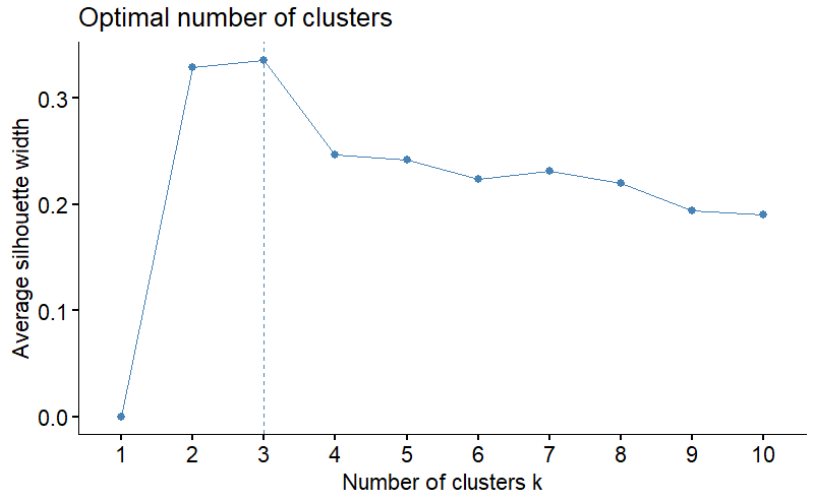

Boxplot of Generosity by Cluster
x[, i]
0.0
0.2
0.4
1
2
3
4
5
x$Cluster

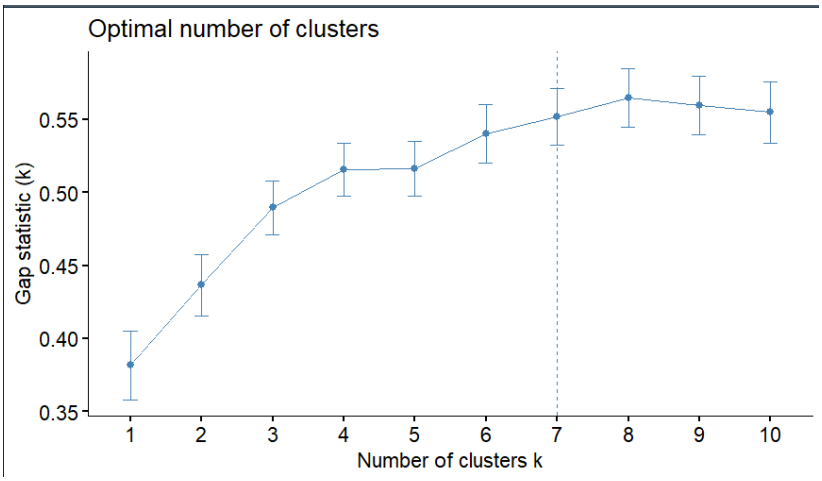


3.2K-Medoids Uygulaması

3.2.1 Küme Sayısı Belirleme

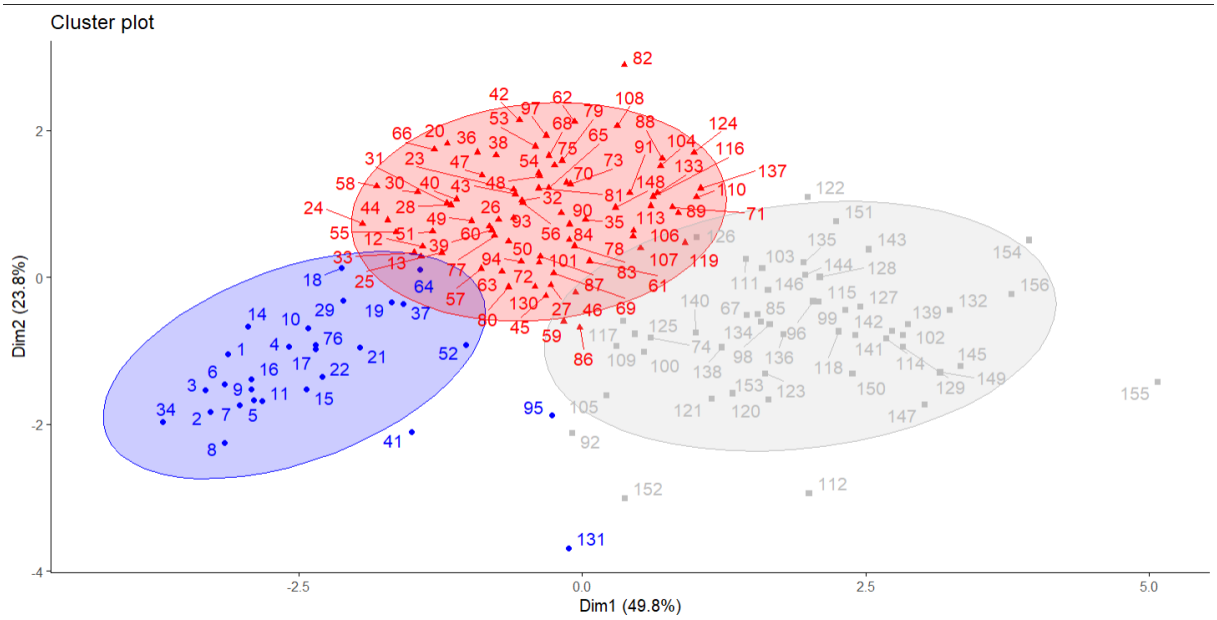


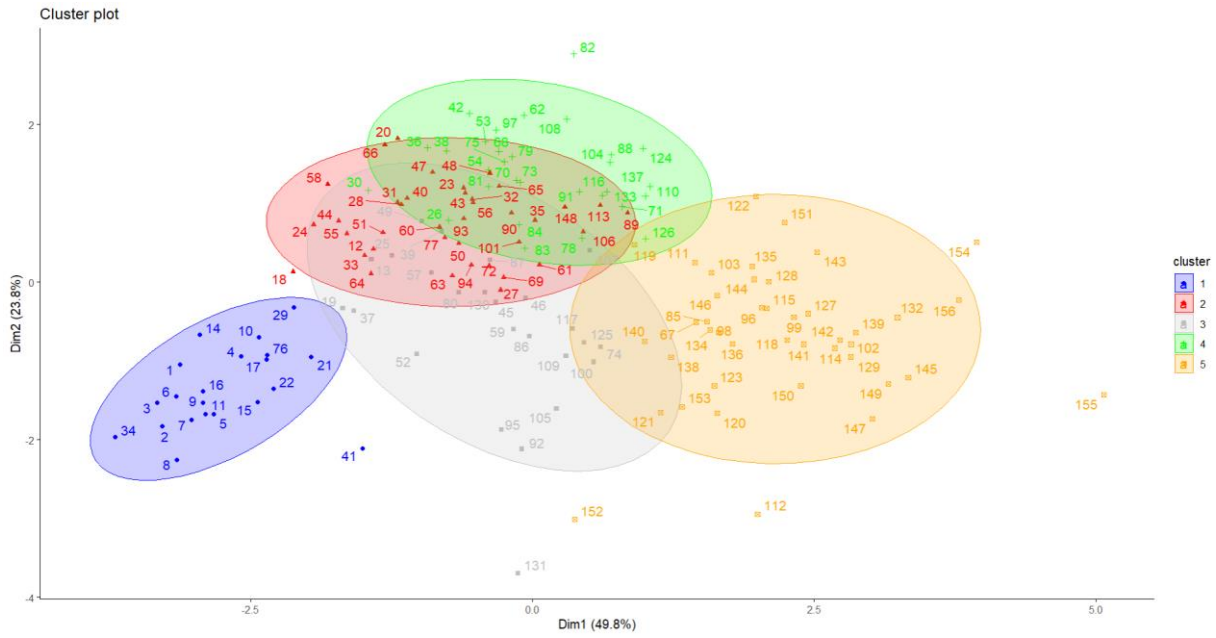


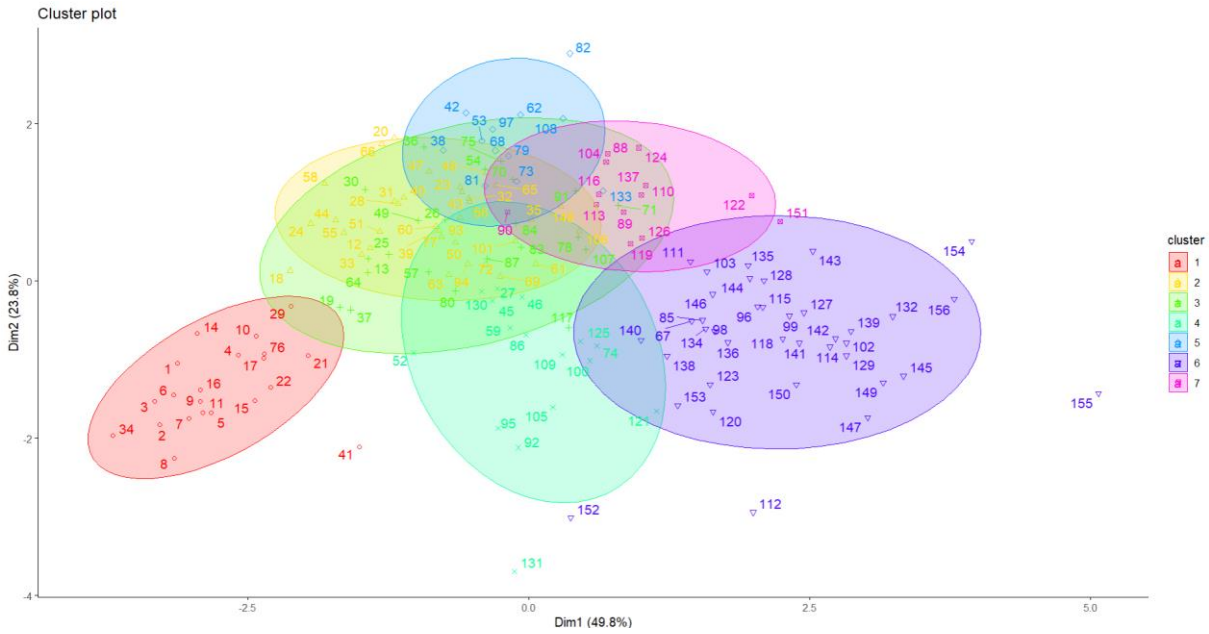


Optimal küme sayısının belirlenmesi için kullanılan yöntemler ve görsellerden anlaşıldığı üzere, 3 ve 7 küme sayısını verdiğini görebiliriz. Bu algoritma bize küme sayısını belirlemek için yalnızca fikir verir. Görsellerdeki küme sayısını seçmek zorunda değiliz. Görselleştirme yaparak en uygun küme sayısını seçeceğiz.

3.2.2 Küme Sayısına Göre Görselleştirme

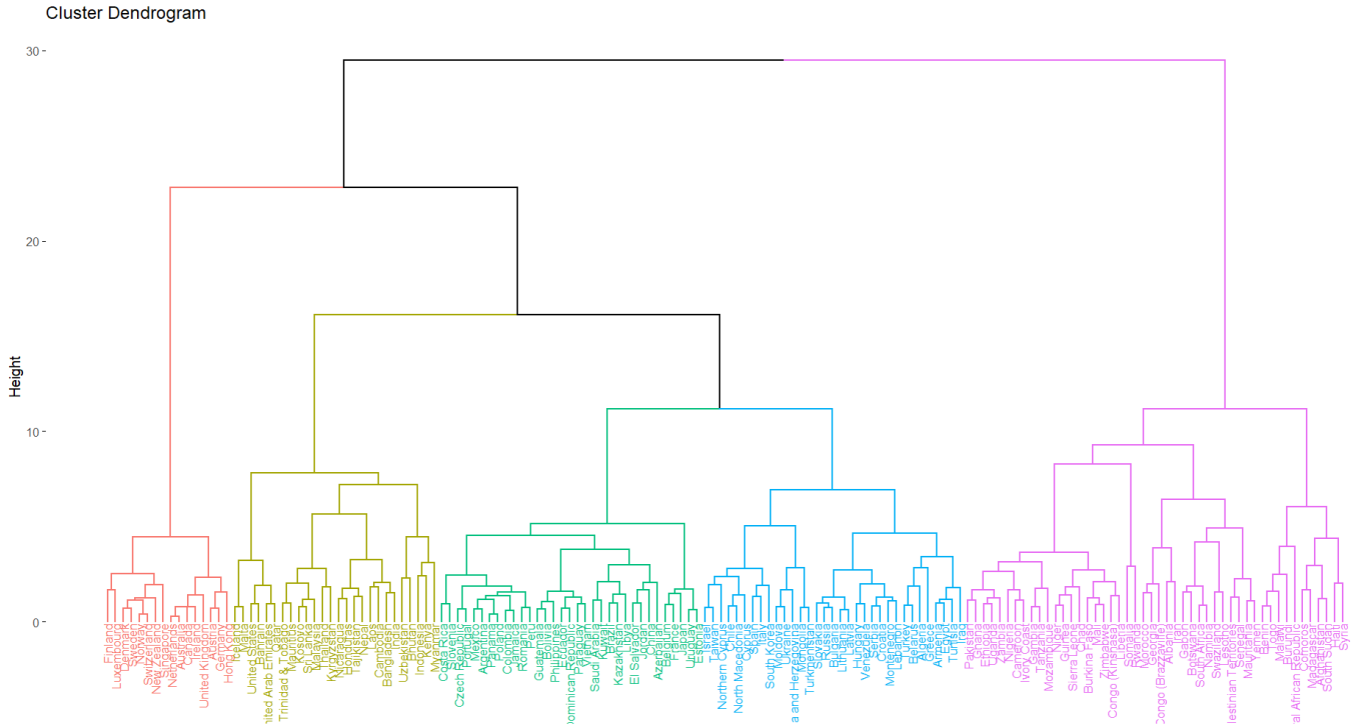






Çıktılardaki görseller, her kümenin net bir şekilde ayrıldığını ve veri noktalarının homojen bir şekilde kümelendiğini göstermektedir. Küme sayısını 3 olarak belirliyoruz.

3.3Hiyerarşik Kümeleme Uygulaması

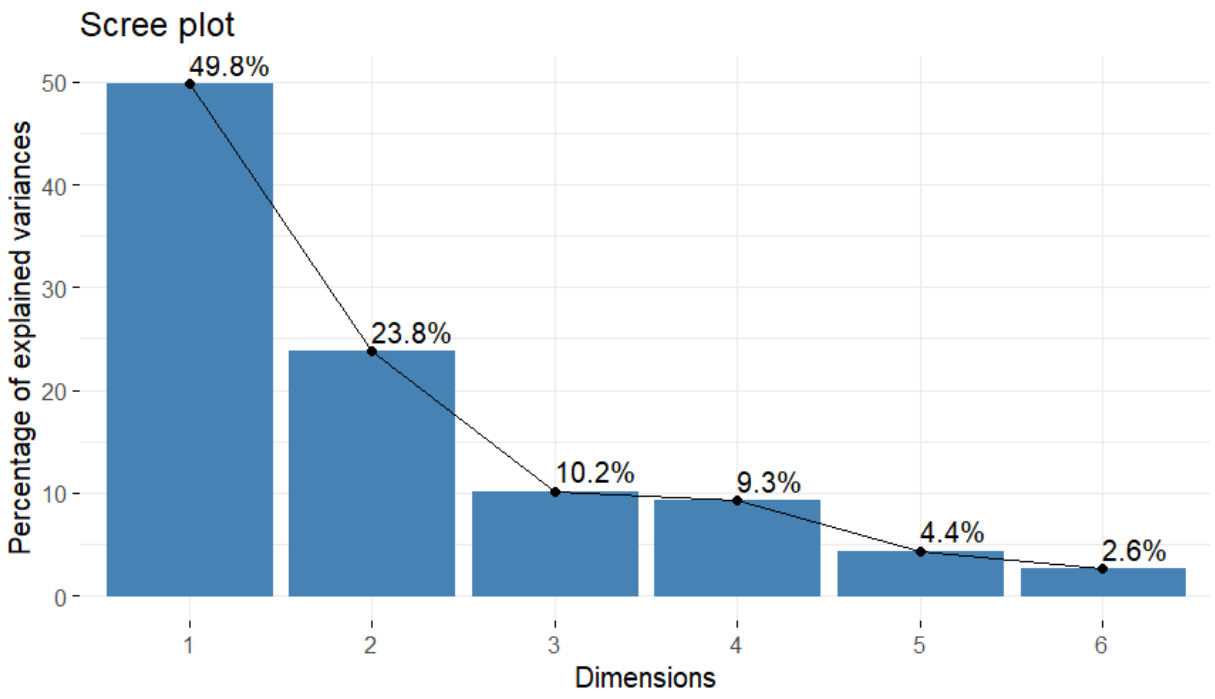


Hiyerarşik kümeleme ile ülkeler arasındaki benzerlikler dendrogram kullanılarak görselleştirilmiştir. Kesme seviyesinin uygun bir noktadan belirlenmesiyle 5 küme elde edilmiştir. Bu kümeler, ekonomik ve sosyal faktörlere dayalı olarak grupların homojenliğini sağlamış ve veri setindeki farklılıkları net bir şekilde ortaya koymuştur.

3.4 Temel Bileşenler Analizi (PCA)

3.4.1Bileşen Varyansları ve Bileşen Seçimi

```
> # PCA uygulama
> pca_result <- prcomp(df, center = TRUE, scale. = TRUE)
> # PCA sonuçlarını inceleyelim
> summary(pca_result)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.7290 1.1940 0.7809 0.74584 0.51120 0.39668
Proportion of Variance 0.4983 0.2376 0.1016 0.09271 0.04355 0.02623
Cumulative Proportion  0.4983 0.7359 0.8375 0.93022 0.97377 1.00000
>
```

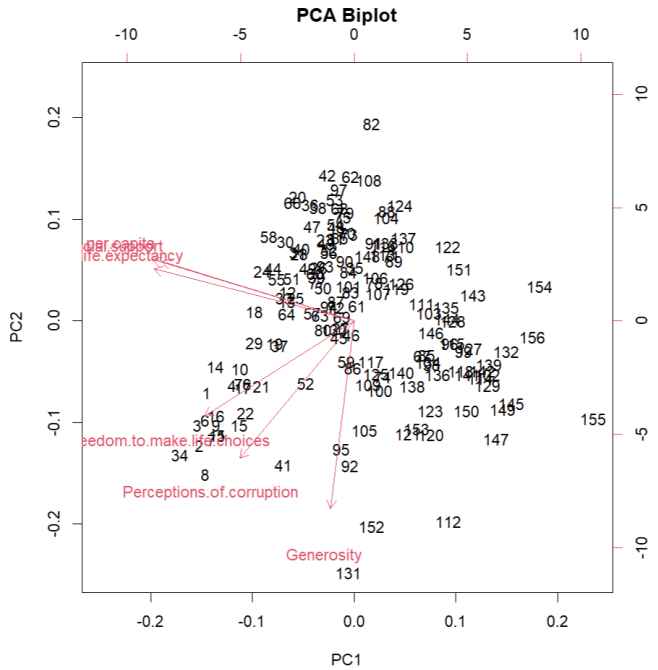


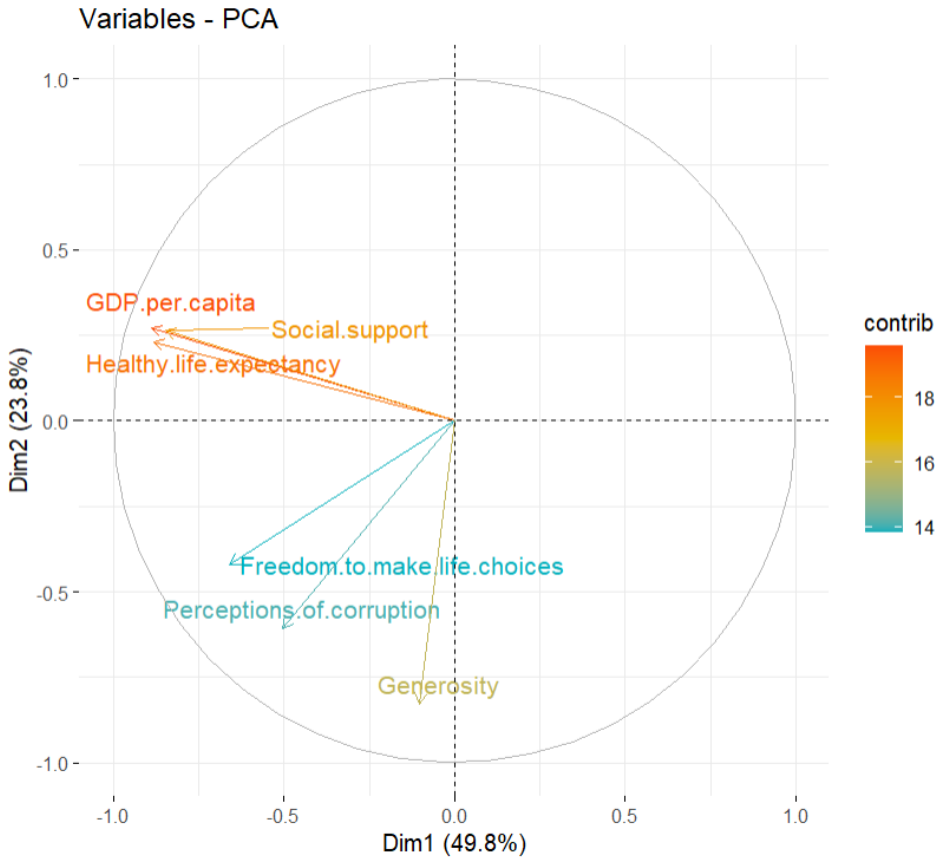
Temel bileşenler analizi uyguladığımız verimize baktığımızda, ilk iki bileşenin varyansın %73.6'sını açıkladığını görebiliyoruz. Bu durum, verinin büyük bir kısmının ilk iki bileşen ile temsil edilebileceğini ve analizde daha az sayıda bileşenle çalışmanın uygun olduğunu göstermektedir. Bu bileşenler, verideki temel desenleri ve değişkenler arasındaki ilişkileri özetleyerek, kümelerin ayrışmasını daha net bir şekilde ortaya koymuştur.

```
> transformed_data <- pca_result$x[, 1:2]
> summary(transformed_data)
      PC1                 PC2
 Min.   :-3.6998   Min.   :-3.6974
 1st Qu.:-1.1220   1st Qu.:-0.8261
 Median :-0.1516   Median : 0.1177
 Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 1.0691   3rd Qu.: 0.9613
 Max.   : 5.0733   Max.   : 2.8934
>
```

PCA analiziyle veri setindeki değişkenlerin boyutları azaltılmış ve varyansı en iyi açıklayan bileşenler seçilmiştir. İlk iki bileşenin toplam varyansın büyük bir kısmını (73.6) açıklaması, bu bileşenlerin analize dahil edilmesi için yeterli bir kriter olduğunu göstermiştir. Daha az varyans açıklayan bileşenler elenerek analizi sadeleştirdim.

3.4.2 Görselleştirme

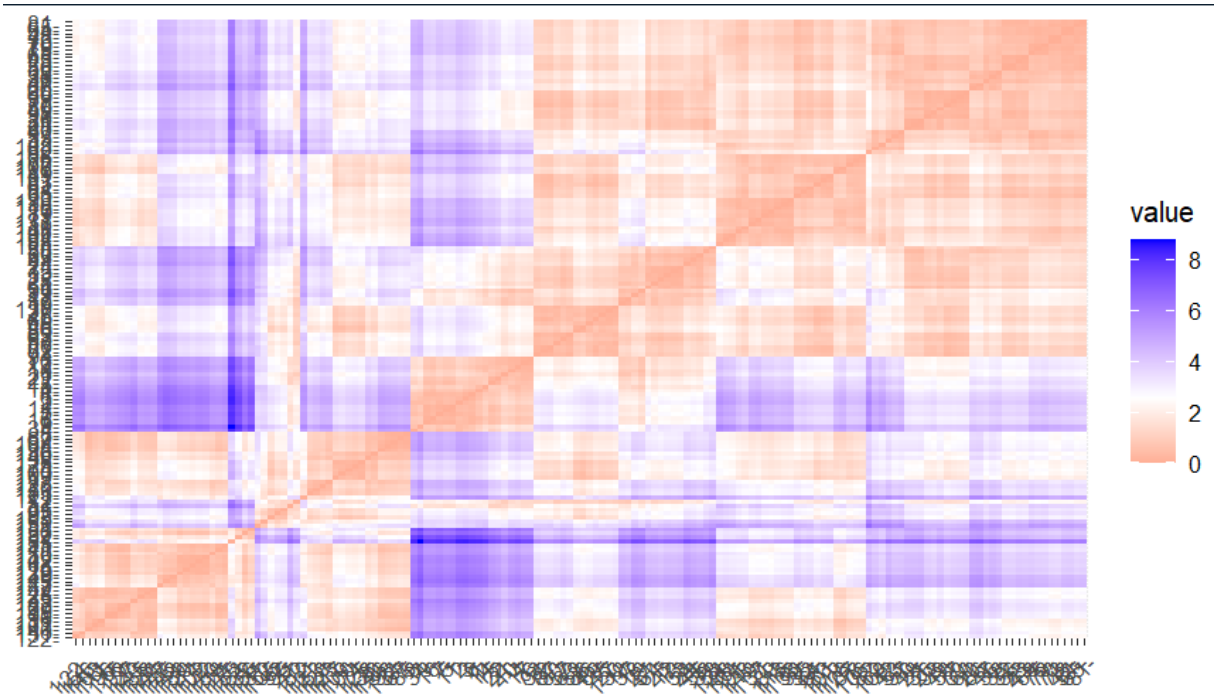




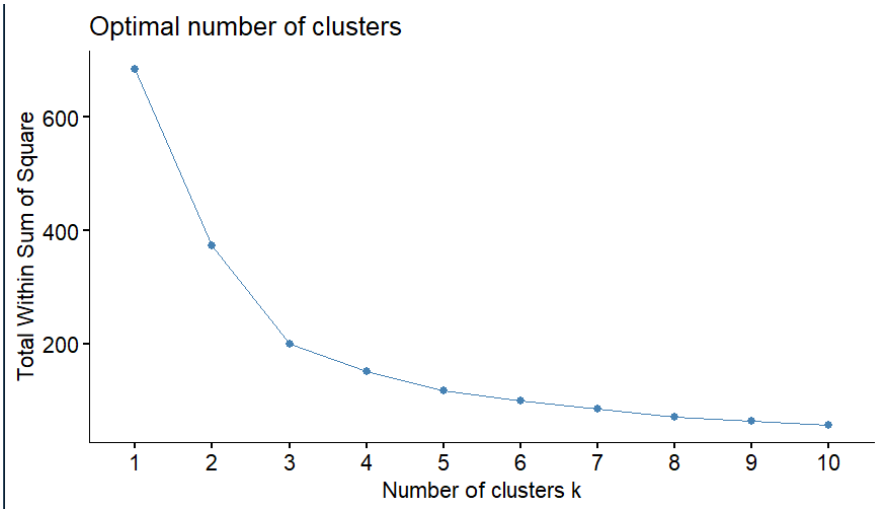
PCA sonuçları iki boyutlu bir grafik üzerinde görselleştirilmiştir. Burada değişkenlerin yüklerinin (vektörlerinin) farklı yönlerde veya açılarda yer alması, değişkenler arasındaki ilişkiler hakkında önemli bilgiler sunmaktadır: 90 derece açıda (dik) yer alan değişkenler aralarında anlamlı bir ilişki bulunmadığını, yani birbirlerinden bağımsız olduklarını ifade ediyor.

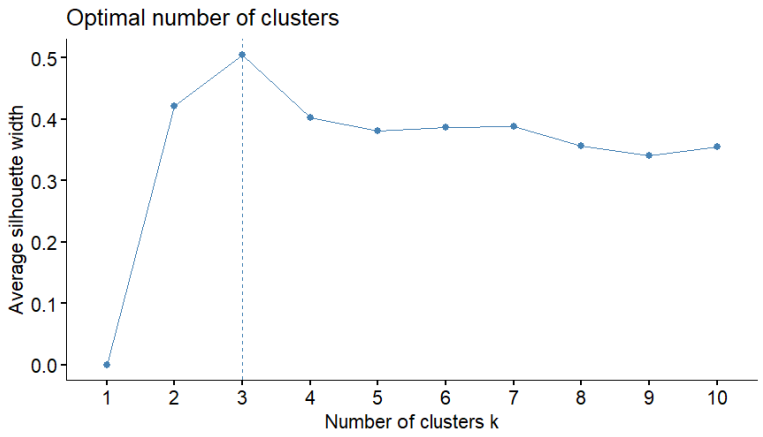
3.5Temel Bileşenler Analizi (PCA) Sonrası Kümeleme Analizi

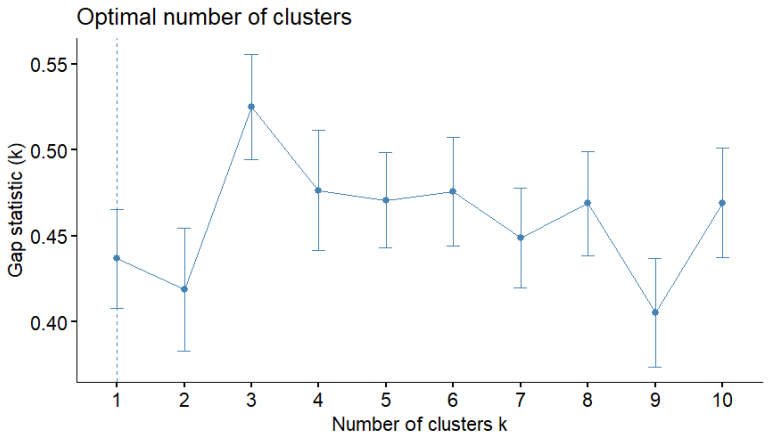
3.5.1 K-Means



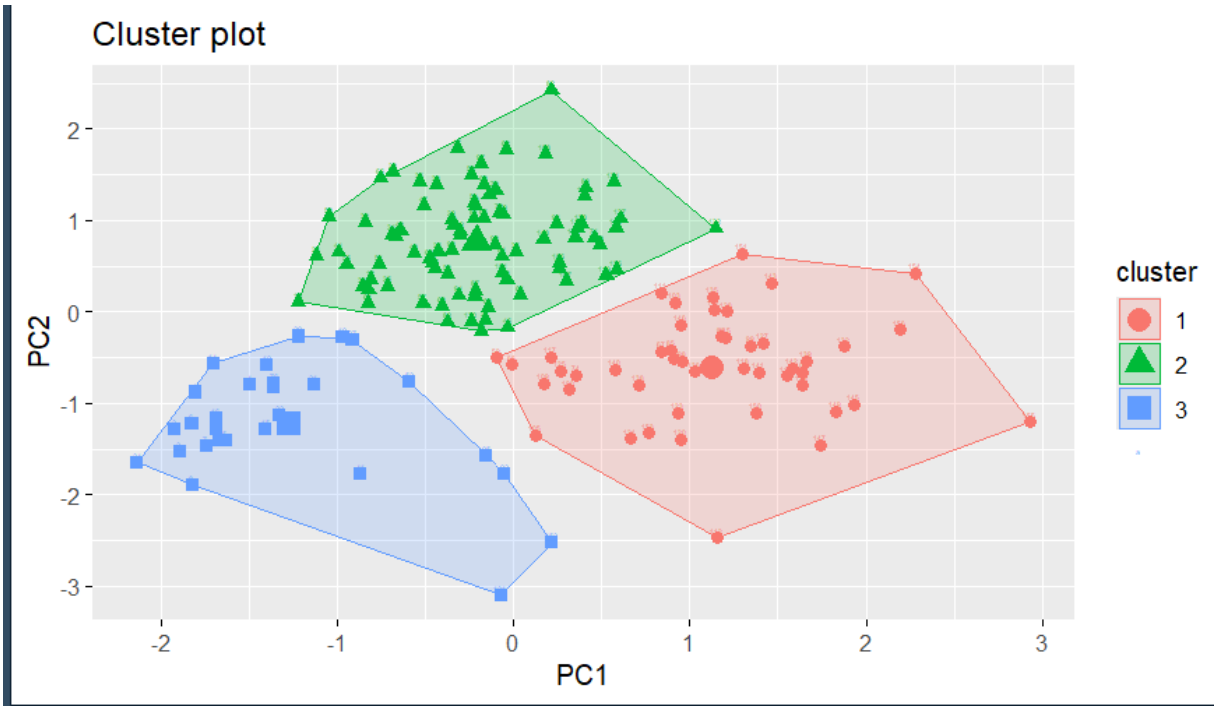
Temel bileşenler analizi uyguladıktan sonra verimizin uzaklık matrisini yukarıda görebiliyoruz. Kümelenebilir olduğunu söyleyebiliriz.

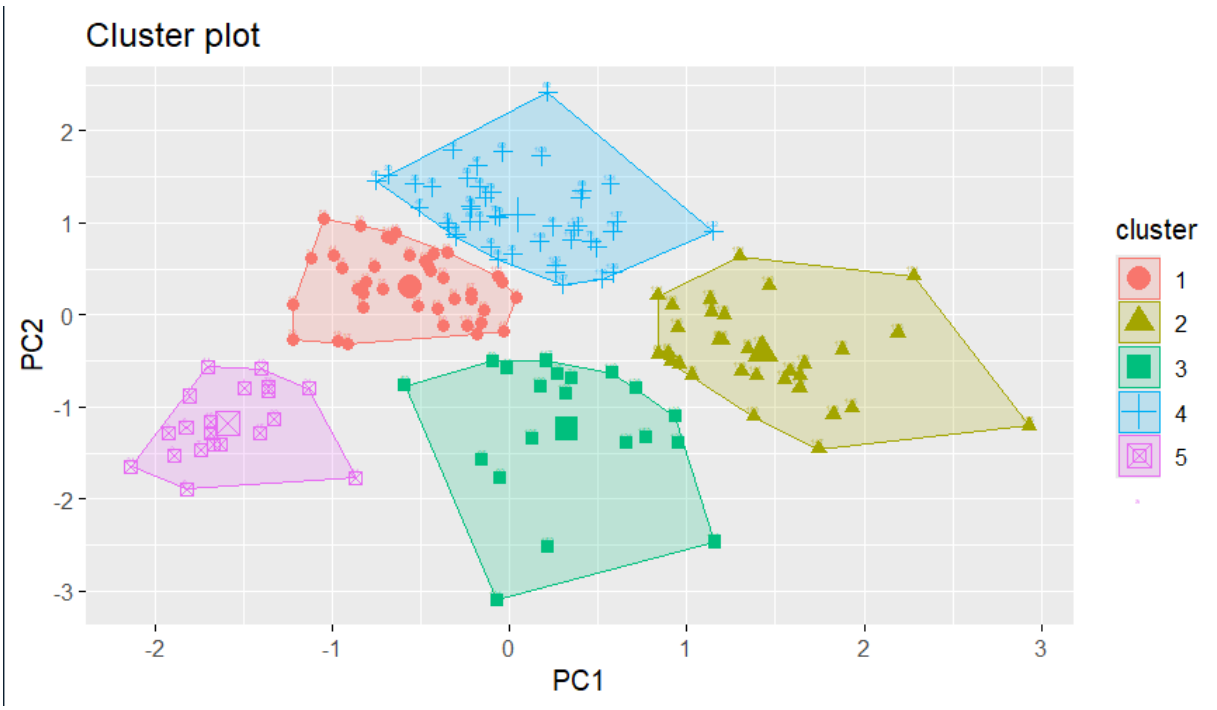






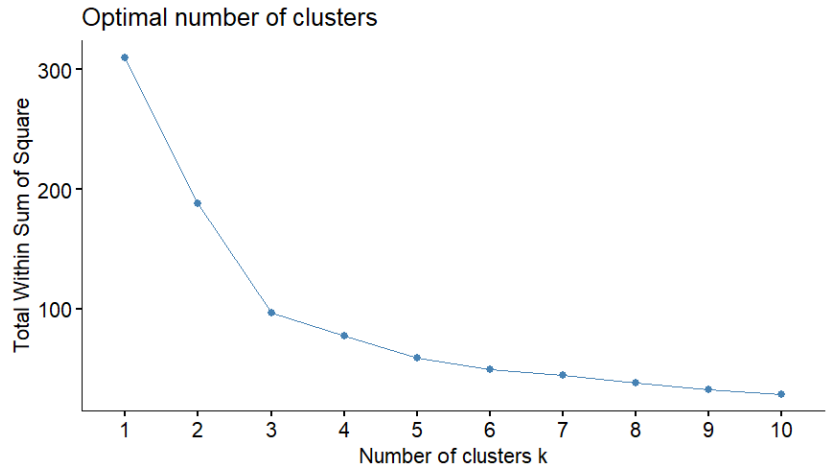
Yukarıdaki şekillerde optimal küme sayısını belirlemek için PCA sonrası K-Means algoritmaları kullanılmış olup çıktıları verilmiştir. Görsellere baktığımızda optimal küme sayısını 3 olarak görüyoruz. Bu algoritma bize küme sayısını belirlemek için yalnızca fikir verir. Görsellerdeki küme sayısını seçmek zorunda değiliz. Küme sayısının 3 ve 5 olduğu görseller aşağıdaki gibidir.

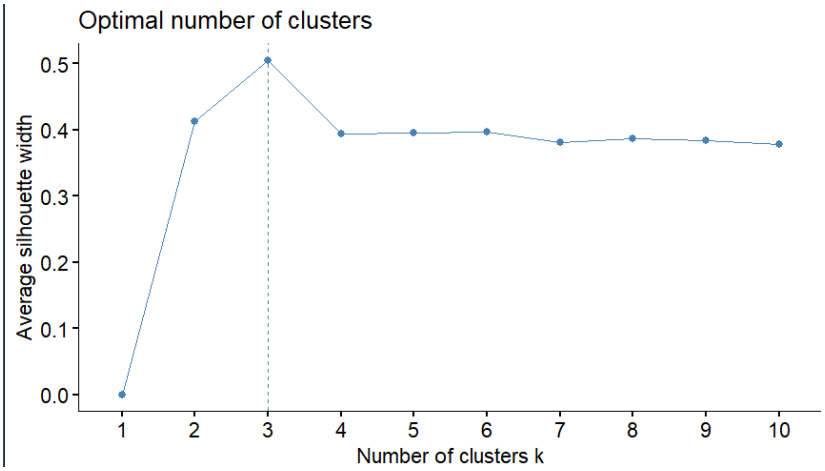


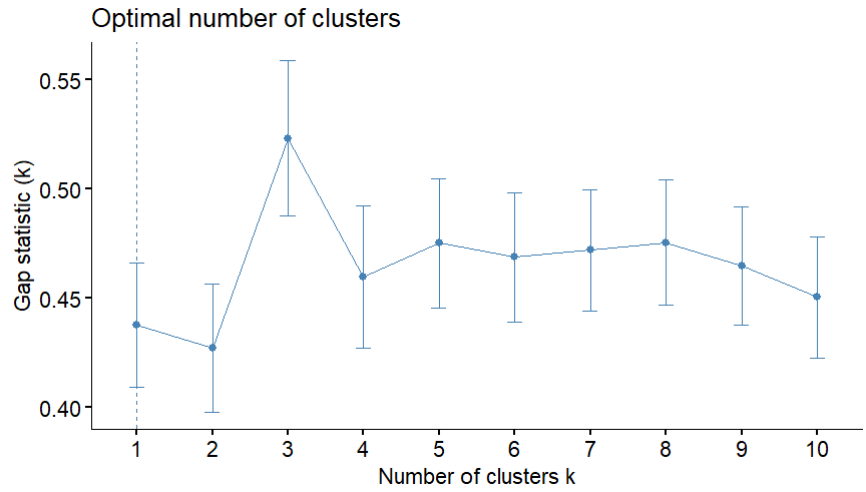


Yukarıdaki görsellerde küme sayısı 3 ve 5 olan görselleri görebiliyoruz. Küme sayısı 5 olan görselimizde kümeler arası farklılığı daha iyi açıklayabileceğimizi düşündüğümüz için küme sayısı 5 olanı alıyoruz.

3.5.2 K-Medoids

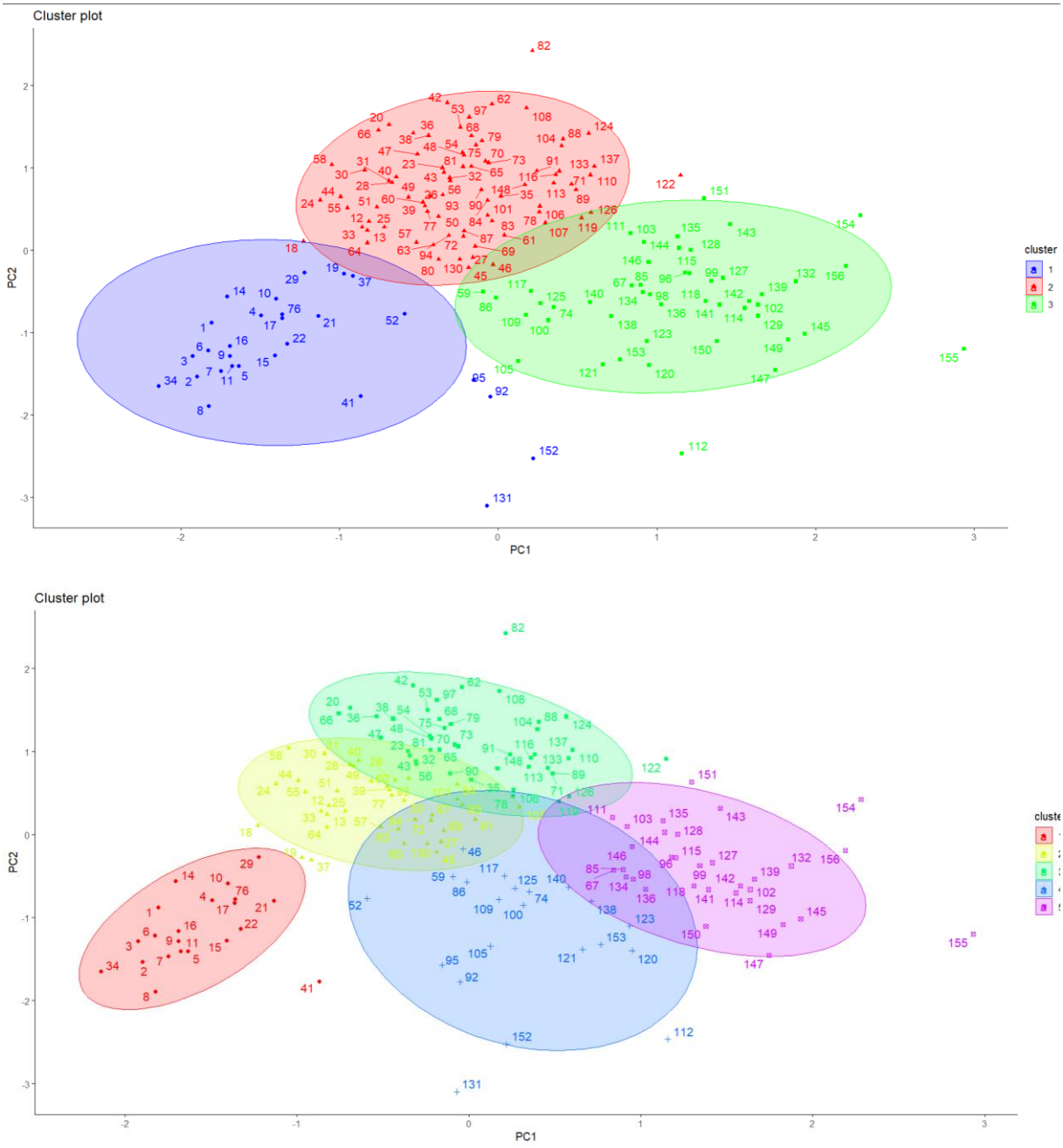






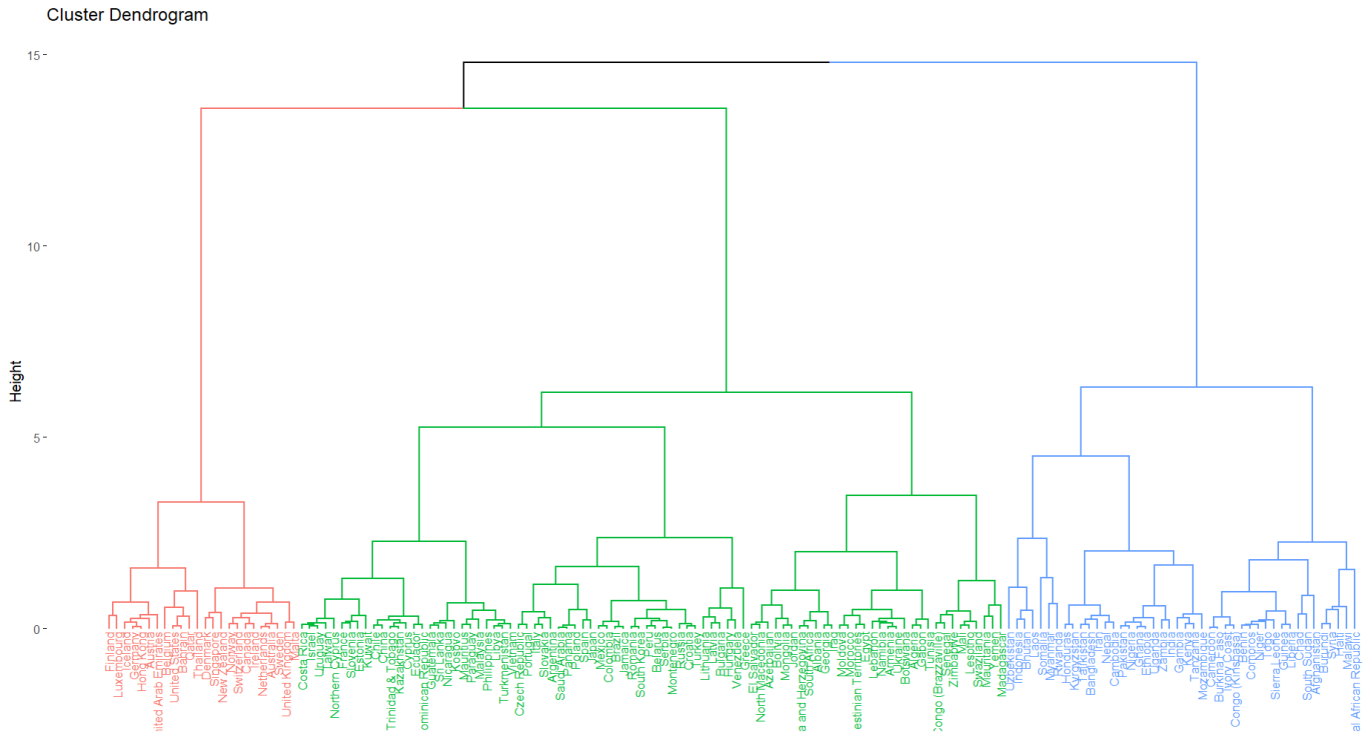
Optimal küme sayısının belirlenmesi için PCA sonrası K-Medoids algoritmasının görsellerinden anlaşıldığı üzere küme sayısının 3 olduğunu görebiliriz. Bu algoritma

bize küme sayısını belirlemek için yalnızca fikir verir. Görsellerdeki küme sayısını seçmek zorunda değiliz. Görselleştirme yaparak en uygun küme sayısını seçeceğiz



Çıktılardaki görseller, her kümenin net bir şekilde ayrıldığını ve veri noktalarının homojen bir şekilde kümelendiğini göstermektedir. Küme sayısını 3 olarak belirliyoruz.

3.5.3 Hiyerarşik Kümeleme



PCA sonrası uygulanan hiyerarşik kümeleme, verinin boyutları azaltıldıktan sonra daha net bir ayrışma sağlamıştır. Kümeleme sonucunda, ülkeler arasındaki sosyal ve ekonomik farklılıklar daha belirgin bir şekilde ortaya çıkmış ve kümeler arasındaki ayrım daha açık olmuştur. Bu, PCA' nın veri setindeki gürüntüyü azaltarak daha anlamlı küme yapıları ortaya çıkarmasına olanak tanımıştır.

3.6PCA Öncesi ve Sonrası Kümeleme Yöntemleri Karşılaştırılması

3.6.1Uygun Yöntemin Seçilmesi

Uygun kümeleme yöntemini seçmek için silüet skoru kullanılmıştır. Silüet skoru, her bir veri noktasının kendi kümesindeki diğer noktalara ne kadar yakın olduğunu ve diğer kümelerle ne kadar uzak olduğunu ölçer. Silüet skoru +1 ile -1 arasında değer alır:

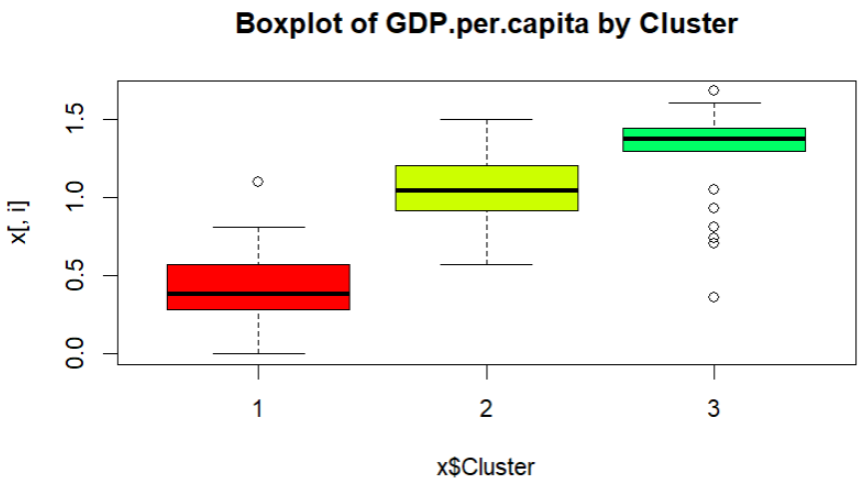
- +1: Veri noktası, kendi kümesine yakın ve diğer kümelerden uzak.
- 0: Veri noktası, iki küme arasında geçiş yapabilecek durumda.
- -1: Veri noktası, yanlış kümeye atanmış.

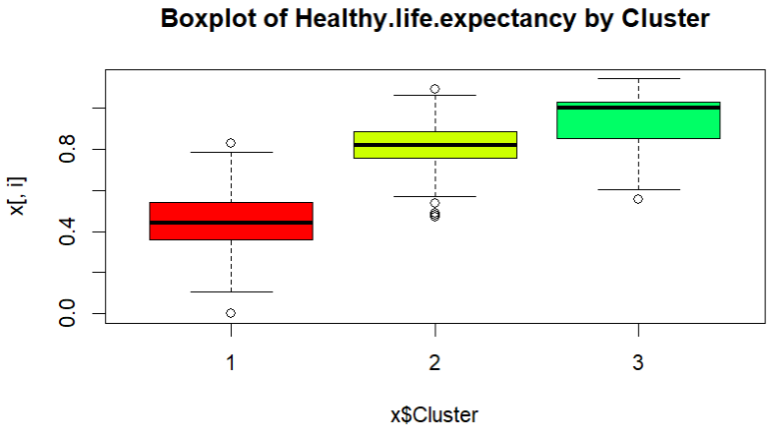
Kümeleme algoritmalarının doğruluğunu karşılaştırmak için farklı küme sayılarındaki silüet skorları hesaplanmış ve en yüksek skoru veren kümeleme yöntemi seçilmiştir. Bu yöntem, doğru küme sayısının ve algoritmanın belirlenmesinde rehberlik etmektedir. Eğer silüet skoru yüksekse, kümeler arasındaki ayrışma belirgindir ve seçilen kümeleme yöntemi uygun demektir.

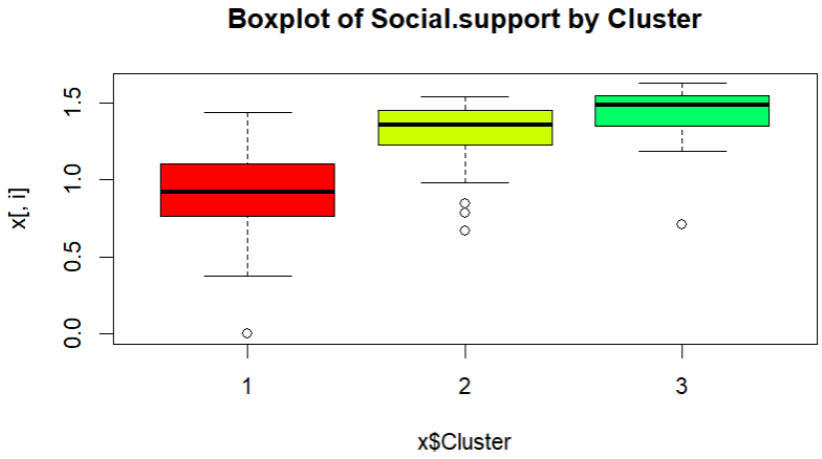
```
> print(silhouette_results)
          Model    PCA_Pre  PCA_Post
1       K-Means 0.2569508 0.5044241
2     K-Medoids 0.3359401 0.5044241
3 Hierarchical 0.2352986 0.4668286
>
> en_uygun_model <- silhouette_results[which.max(silhouette_results$PCA_Post), ]
> print(en_uygun_model)
    Model    PCA_Pre  PCA_Post
1 K-Means 0.2569508 0.5044241
```

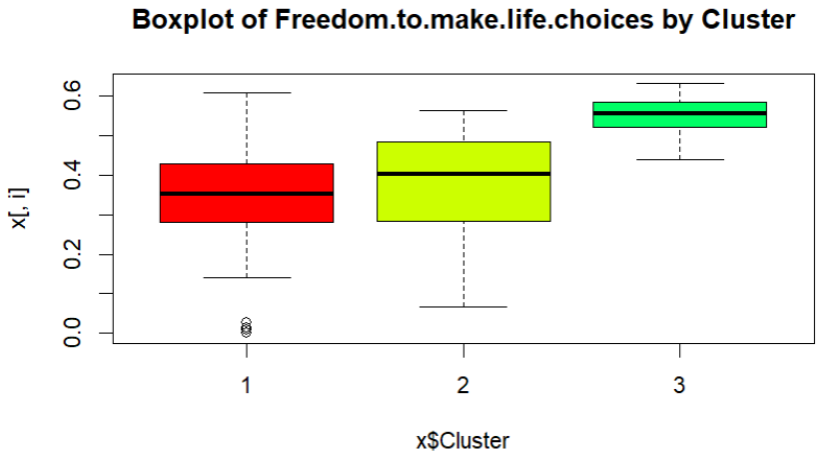
Yukarıdaki çıktıya baktığımızda en uygun yöntemin PCA sonrası K-Means olduğunu görebiliyoruz.

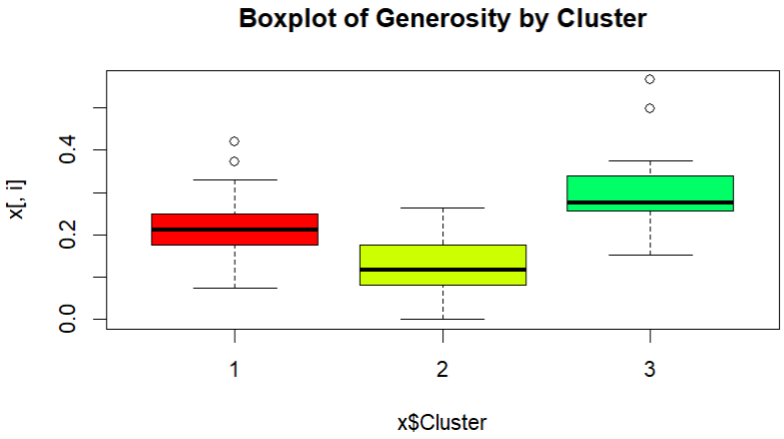
3.6.2Uygun Yöntem İçin Boxplot by Factor

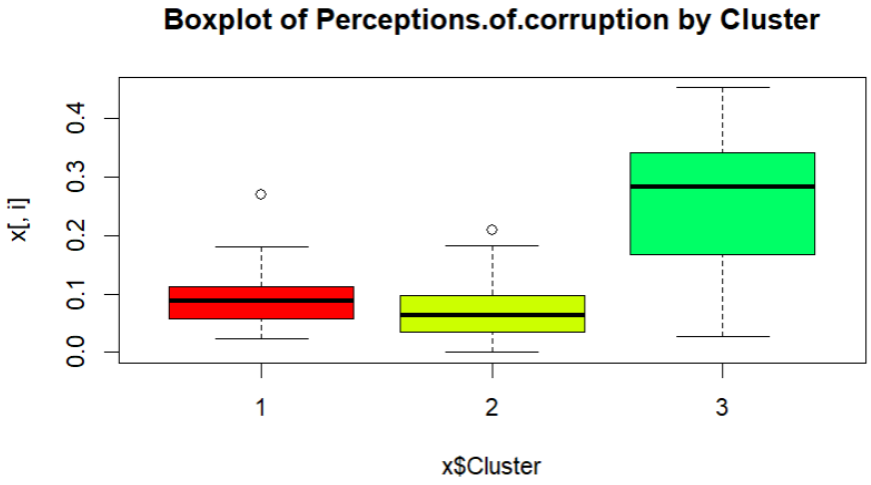








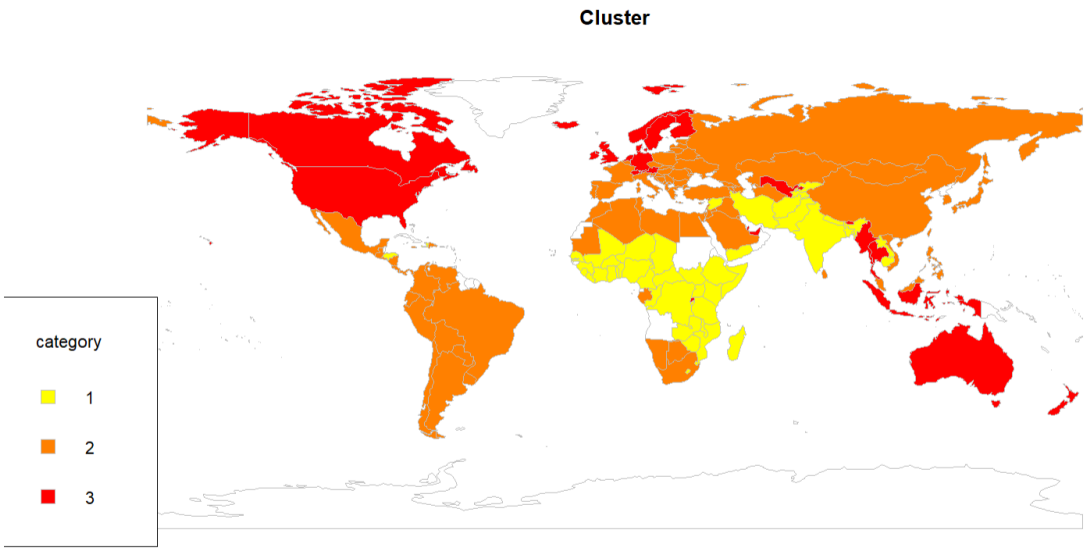




Boxplot'lar kümeler arasındaki heterojenliği belirgin bir şekilde ortaya koymaktadır.

4.Bölüm

SONUÇ



Bu çalışma, Dünya Mutluluk Verisi üzerinde denetimsiz makine öğrenmesi yöntemleri kullanarak ülkelerin mutluluk seviyelerini etkileyen faktörleri incelemiştir. K-Means, K-Medoids ve hiyerarşik kümeleme gibi kümeleme algoritmaları ve Temel Bileşenler Analizi (PCA) gibi boyut indirgeme yöntemleri ile ülkeler arasındaki benzerlikler ve farklılıklar detaylı bir şekilde analiz edilmiştir.

Kümeleme sonuçları, ülkelerin ekonomik ve sosyal göstergeler doğrultusunda anlamlı bir şekilde gruplandığını ortaya koymuştur. PCA sonrası uygulanan hiyerarşik kümeleme, veri setindeki gürültüyü azaltarak daha net ve anlamlı kümeler oluşturulmasına katkı sağlamıştır. Silüet skoru kullanılarak en uygun kümeleme yöntemi belirlenmiş ve bu yöntem ile ülkeler arasındaki ayrışma net bir şekilde gözlemlenmiştir.

Analizler, kişi başına düşen gelir, sosyal destek, sağlıklı yaşam beklentisi, özgürlük ve yolsuzluk algısı gibi faktörlerin mutluluk seviyeleri üzerinde önemli bir etkisi olduğunu göstermektedir. Özellikle, yüksek gelir ve güçlü sosyal destek sistemlerine sahip ülkelerin daha yüksek mutluluk skorlarına sahip olduğu bulunmuştur. Ayrıca, düşük yolsuzluk algısı ve yüksek bireysel özgürlük seviyelerinin mutluluğu artırıcı faktörler olduğu ortaya çıkmıştır.

Bu bulgular, sosyal ve ekonomik politikaların geliştirilmesinde önemli bir rehber olabilecek niteliktedir. Gelecekteki araştırmalarda, mutluluğun belirleyicilerinin daha derinlemesine incelenmesi ve farklı faktörlerin etkileşimlerinin daha iyi anlaşılması faydalı olacaktır.

KAYNAKÇA

- Kassambara,A. 2017. Practical Guide To Cluster Analysis in R Unsu pervised Machine Learning. Sthda
- https://www.kaggle.com/datasets/unsdsn/world-happiness
- https://www.geeksforgeeks.org/hierarchical-clustering-in-r-programming/
- https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/#google_vignette