T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
FEN FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

Kümeleme Analizi ile Ülkelerin Refah Düzeylerinin Belirlenmesi

---

Bitirme Projesi Raporu

Canberk Yanık
Çağatay Gülmez
Yunus Ergün

Mayıs, 2022

Rapor Değerlendirme

Kümeleme Analizi ile Ülkelerin Refah Düzeylerinin Belirlenmesi başlıklı bitirme projesi raporu tarafımdan okunmuş, kapsamı ve niteliği açısından bir Bitirme Projesi raporu olarak kabul edilmiştir.

Doç. Dr. Neslihan Demirel

Teşekkür

Tüm çalışma süresince, yönlendiriciliği, katkıları ve yardımları ile yanımızda olan danışmanımız Doç. Dr. Neslihan Demirel'e, böyle bir çalışmayı yapmamız için bize fırsat tanıyan Dokuz Eylül Üniversitesi Fen Fakültesi İstatistik Bölümü'ne teşekkür ederiz.

Canberk Yanık
Çağatay Gülmez
Yunus Ergün

Özet

Bu projede, Legatum Enstitü Vakfı'nın (Legatum Insitute Foundation) sağladığı, 167 ülkeye ait 12 değişkene sahip refah düzeylerinin incelendiği veri seti üzerinde önce temel bileşenler analizi yapılmıştır. k-means, k-medoids ve hiyerarşik kümeleme teknikleri uygulanıp küme doğrulama ölçülerinden connectivity, dunn index, silhoutte yöntemleri kullanılarak karşılaştırılmıştır. Hiyerarşik kümeleme yöntemi seçilip tüm ülkeler refah düzeylerine göre 3 farklı kümeye atanmıştır.

Anahtar kelimeler: kümeleme analizi , k-ortalamalar, k-medoids, hiyerarşik, bağlantılılık, dunn indeksi, silhoutte katsayısı

Abstract

In this project, principal components analysis was first conducted on the data set provided by the Legatum Institute Foundation, in which the welfare levels of 167 countries with 12 variables were examined.
k-means, k-medoids and hierarchical clustering techniques were applied and cluster verification measures were compared using connectivity, dunn index and silhoutte methods. Hierarchical clustring technique was selected and all countries were assigned to 3 different clusters according to their welfare levels.

Keywords: clustering analysis, k-means, k-medoids, hierarchical, connectivity, dunn index, silhoutte

İçindekiler

Şekil Listesi

Tablo Listesi

Giriş

Refah zenginlikten çok daha fazlasıdır; tüm insanların gelişme fırsatına ve özgürlüğüne sahip olduğu zamandır. Refah, her bireyin temel özgürlüklerini ve güvenliğini koruyan güçlü bir sözleşmeyle kapsayıcı bir toplum tarafından desteklenir. Yoksulluktan sürdürülebilir yollar yaratmak için fikirlerden ve yeteneklerden yararlanan açık bir ekonomi tarafından yönlendirilir. Refahı teşvik eden bir toplum yaratmaya katkıda bulunan ve rol oynayan yetkilendirilmiş insanlar tarafından inşa edilmiştir (2021LegatumProsperityIndex™ (2021)).

Refahı ölçülmesi kolay değildir. Ölçülebilmesi için birçok farklı etkene değinilmesi gerekir. Örneğin yaşam koşullarını ölçmek için beslenme, barınma, korunma ve temel kaynaklara ulaşım gibi unsurlar göz önünde bulundurulmalıdır. Aynı zamanda temel kaynaklara ulaşım etkeni içerisinde elektrik ve içilebilir su gibi göstergelerin sayısal verilerle ölçülmesi gerekir.

Legatum Enstitü Vakfı'nın (Legatum Institute Foundation) oluşturduğu Legatum refah endeksi bir dönüşüm aracı olarak tasarlanmıştır ve dünya çapındaki liderlerin büyüme ve gelişme gündemlerini belirlemeye yardımcı olmak için kullanılması hedeflenmiştir. İndeks oluşturulurken 167 ülkenin çok çeşitli kamuya açık veri kaynakları kullanılarak 300'e yakın gösterge ile ölçülen 66 farklı unsurdan oluşan 12 değişken kullanılmıştır. Legatum refah indeksinin etkili bir şekilde analiz edilmesi için doğru makine öğrenmesi algoritması kullanılmalıdır.

Makine öğrenimi algoritmaları insan müdahalesi olmadan verilerden öğrenebilen ve deneyimler ile geliştirebilen programlardır. Makine öğrenimi algoritmaları denetimli ve denetimsiz öğrenme olarak 2 ayrı kategoriye ayrılır. Denetimli makine öğrenmesi belirsizlik dahilinde kanıta dayalı tahminler yapan bir model oluşturur. Ardından yeni verilere yanıt için makul tahminler oluşturmak üzere bir modeli eğitir. Denetimli makine öğrenmesi algoritmaları sınıflandırma teknikleri ve regresyon teknikleri olarak

ikiye ayrılır.

Denetimsiz öğrenme Denetimsiz Makine Öğrenmesi (UML: Unsupervised Machine Learning) modelinde sistem eğitilirken etiketsiz veri kullanılarak öğrenmesi sağlanır. Denetimsiz öğrenmede amaç veri setindeki örneklerin çıkışları bilinmediği için tanıma veya sınıflandırma değildir. Genellikle kümeleme, olasılık yoğunluk tahmini, öznitelikler arasındaki ilişkilerin bulunması ve boyut indirgeme gibi amaçlarla kullanılmaktadır. Ayrıca denetimsiz öğrenme algoritması ile elde edilen sonuçlar denetimli öğrenme için de kullanılabilmektedir (Chao (2011)).

Veri setindeki veriler elde etmek istediğimiz çıktının nasıl olduğu hakkında çok az ya da hiç fikir vermediği için denetimsiz makine öğrenmesi kullanılarak kümele yapılmıştır.

Projenin 1. Bölümünde uzaklık ölçümleri, küme sayısı belirleme yöntemleri, kümeleme algoritmaları ve küme doğrulama istatistiklerinden bahsedilmiştir. 2. Bölümde veri seti analize hazırlanıp öklid uzaklık temel alınarak kümeleme algoritmaları karşılaştırılıp 167 ülke kümelenmiştir.

Bölüm 1

Kümeleme Analizi

1.1 Kümeleme Uzaklık Ölçüleri

Kümeleme analizi, birimleri değişkenler arası benzerlik ya da uzaklıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplar oluşturmaya çalışır (Özdamar, 2004). Kümeleme analizinde birey ya da nesneler arasındaki uzaklıkları hesaplamak için en yaygın kullanılan uzaklık ölçüsü Öklid uzaklığıdır (Ünlükaplan, 2008).

Değişkenler arası benzerlik ya da farklılıklara dayanır. Analizde bir birime ait değişkenlerin birbiriyle olan uzaklıkları hesaplanılır. Kısaca, uzaklık ya da benzerlik matrisinden yararlanılır. Benzerlikler nesne çiftleri arasındaki uzaklığın ölçüsüdür. Benzerlik ya da uzaklık ölçüleri olarak Nicel kümeleme yapmak istendiği durumda öklid, manhattan, minkovski vb. ölçüler kullanılır. Nitel kümeleme yapıldığı durumda korelasyon uzaklık ölçüsü, pearson gibi yöntemler kullanılır.

Uzaklık ölçüsünün seçimi kümelemede kritik bir adımdır. İki objenin (x, y) birbirine benzerliğini hesaplar ve kümelerin şeklini etkiler. Aşağıda sık kullanılan uzaklık ölçüleri olan Öklit ve Manhattan uzaklıkları tanımlanmıştır (Kassambara (2017)).

1.1.1 Öklid Uzaklığı

$$d_{\text{euc}}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1.1}$$

1.1.2 Manhattan Uzaklığı

$$d_{\text{man}}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{1.2}$$

1.1.3 Pearson Korelasyon Uzaklığı

Çoğunlukla gen ekspresyonu veri analizinde kullanılan korelasyon temelli uzaklıklar da bulunmaktadır. Korelasyon temelli uzaklıklar eğer iki veri güçlü bir şekilde bağıntılı ise, bu verileri birbirine benzer kabul eder. Eğer birbirlerine tam olarak bağıntılı ise aralarındaki uzaklık 0 kabul edilir.
Pearson korelasyon uzaklığı sapan değerlere karşı hassastır.

$$d_{\text{cor}}(x,y) = 1 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{1.3}$$

1.1.4 Eisen Cosine Korelasyon Uzaklığı

Pearson's korelasyon uzaklığının x-bar ve y-bar'ın sıfır ile değiştirilmiş halidir.

$$d_{\text{eisen}}(x,y) = 1 - \frac{|\sum_{i=1}^{n} x_i y_i|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{1.4}$$

1.1.5 Spearman Korelasyon Uzaklığı

Spearman korelasyon uzaklığı yöntemi x ve y değişkenlerinin rankları arasındaki korelasyonu hesaplar.

$$d_{\text{spear}}(x,y) = 1 - \frac{\sum_{i=1}^{n} (x_i' - \bar{x}')(y_i' - \bar{y}')}{\sqrt{\sum_{i=1}^{n} (x_i' - \bar{x}')^2 \sum_{i=1}^{n} \left(y_i' - \overline{y'}\right)^2}} \tag{1.5}$$

$x_i =$ rank(x_i) ve $y_i =$ rank(y) olduğu durumda.

1.1.6 Kendall Korelasyon Uzaklığı

Kendall korelasyon yöntemi x ve y değişkenler arasındaki uygunluğu ölçer. x ve y gözlemlerindeki toplam olağan eşleşmeler $n(n-1)/2$ dir. Başlangıçta x değerlerine göre eşleşmeler sıralanır. Eğer x ve y bağıntılı ise, aynı bağıl rank sıralamasına sahiptirler. Ardından her y_i için y_i>y_i'ler
(concordant pairs (c)) ve y_i<y_i'ler (discordant pairs (d)) sayılır (Kassambara (2017)).

$$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{1.6}$$

- n_c: toplam uyumlu ikili
- n_d: toplam uyumsuz ikili
- $n : x$ ve y nin boyutu

1.2 En Uygun Küme Sayısının Belirlenmesi

En uygun küme sayısının belirlenmesi bölmeli kümelemenin temel konularından biridir. Örneğin k-means gibi bölmeli kümele algoritmaları analizcinin küme sayısına karar vermesini gerektirir. Küme sayısına kesin bir cevap yoktur. En uygun küme sayısı benzerliği ölçmede kullanılan yöntem ve bölme yapılırken kullanılan parametrelere göre değişiklik gösterir. En uygun küme sayısının belirlenmesi için Doğrudan ve İstatistiksel yöntemler kullanılabilir (Kassambara (2017)).

1.2.1 Doğrudan Yöntemler

Küme içindeki kareler toplamı veya ortalama silhouette gibi bir ölçütün optimize edilmesinden oluşur. Karşılık gelen yöntemler sırasıyla elbow ve silhouette yöntemleri olarak adlandırılır.

1.2.1.1 Elbow (Dirsek) Yöntemi

K- means gibi bölme yöntemlerindeki temel prensip, küme içi toplam kareler toplamı (total within-cluster sum of square (WSS)) minimize olacak şekilde küme sayısının seçilmesidir. Küme içi toplam kareler toplamı kümenin yoğunluğunu ölçer (Kassambara (2017)).

$$W(C_k) = \sum_{x_i \varepsilon C_k} (x_i - \mu_k)^2 \tag{1.7}$$

$$ToplamKümeİçiKarelerToplam = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{1.8}$$

k : küme sayısı

x : C_k kümesinin sahip olduğu veri noktası

μ : C_k kümesine atanan noktaların ortalama değeri

Elbow yönteminin uygulama adımları şunlardır;

1. Farklı k değerleri için kümeleme algoritması çalıştırılır.
2. Her k değeri için WSS hesaplanır
3. Küme sayısına göre (k) WSS eğrisi çizilir.

4. Grafik üzerinde toplamlar arasındaki farkın azalmaya başladığı dirsek noktası genellikle en uygun küme sayısını belirtir.

1.2.1.2 Average Silhouette Yöntemi

Average silhouette yöntemi yapılan kümelemenin kalitesini ölçer. Objelerin kümelerde ne kadar iyi konumlandığını belirler. Yüksek average silhouette genişliği iyi bir kümeleme yapıldığını gösterir. Average Silhouette yönteminin uygulama adımları şunlardır (Kassambara (2017));

1. Farklı k değerleri için kümeleme algoritması çalıştırılır.

2. Her k değeri için gözlemlerin average silhoutte değeri hesaplanır.

3. k değerine göre average silhoutte grafiği çizilir.

4. Grafikteki maksimum değerin konumu uygun küme sayısını belirtir.

1.2.2 İstatistiksel Yöntemler

1.2.2.1 Gap İstatistiği Yöntemi

Gap istatistiği yaklaşımı her kümeleme yöntemine uygulanabilir. Gap istatistiği farklı k değerlerindeki WSS'ler ile verilerin boş referans dağılımı altındaki beklenen değerlerini karşılaştırır. Gap istatistiğini maksimize eden değer en uygun küme sayısını verir. Gap İstatistiği yöntemi uygulama adımları şunlardır(Kassambara (2017));

1. $k = 1, 2, \ldots, k_max$ değerleri için gözlem verisi kümelenir ve kümelere karşılık gelen toplam küme içi kareler toplamı W_k hesaplanır.

2. Rastgele uniform dağılımı ile referans veri setleri oluşturulur. $k = 1, 2, \ldots, k_max$ değerleri için referans veri setleri kümelenir ve kümelere karşılık gelen toplam küme içi kareler toplamı W_kb hesaplanır.

3. Tahmini Gap istatistiği sıfır hipotezi $\text{Gap}(k) = \frac{1}{D}\sum_{b=1}^{B}\log\left(W_{kb}^{*}\right) - \log\left(W_k\right)$ altında, beklenen değerinden gözlemlenen sapması olarak hesaplanır.

4. Gap istatistiği $(k+1 : \text{Gap}(k) \geq \text{Gap}(k+1) - S_{k+1})$'in 1 standart sapma içerisinde olacak şekilde en küçük k değeri küme sayısı olarak belirlenir.

1.3 Kümeleme Algoritmaları

Çok boyutlu uzayda verilerin özetlenmesi ve tanımlanmasında yol gösterici bir araştırma yöntemi olan kümeleme analizi; nispeten heterojen olan farklı gruplardaki gözlem yapılarını ya da nispeten homojen olan benzer gruplardaki gözlemleri uygun yöntemlerle gruplamaya olanak sağlayan bir yöntem olarak bilinmektedir (Wierzchoń & Kłopotek, 2018).

1.3.1 K-Means Kümeleme Algoritması

K-means algoritması eldeki veri setini k sayıda kümeye ayırabilmek için kullanılan en yaygın denetimsiz makine öğrenmesi algoritmasıdır. K-means kümelemede küme içi toplam değişkenliğin en aza indirilmesi için kümeleri belirlemesi esas alınır (Kassambara (2017)).

K-means algoritması aşağıdaki gibi özetlenebilir;

1. Analist tarafından oluşturulacak küme sayısı (k) belirlenir.
2. Başlangıç küme merkezini belirlemek üzere rastgele k rassal veri noktası seçilir.
3. Tüm gözlemleri her bir merkez noktaya olan uzaklıkları hesaplanarak en yakın merkeze atanır.
4. Her bir k küme için yeni ortalamalar hesaplanarak küme merkezleri güncelleştirilir.
5. Yinelemeli olarak toplam küme içi kareler toplamı en aza indirilir. Küme atamaları değişimi durana ya da azami yineleme sayısına ulaşılana kadar 3. ve 4. adım tekrar edilir. R yazılımı, varsayılan azami yineleme sayısını 10 kabul eder.

K-Means Kümelemesi Avantajları

- K-means kümelemesi basit ve hızlı bir algoritmadır.
- Büyük veri setleri üzerinde verimli bir şekilde sonuç alınabilir.

K-Means Kümelemesi Dezavantajları

- Veri üzerinde öncelikli bilgi sahibi olunduğunu varsayar ve analistin öncelikli olarak küme sayısını (k) belirlemesini gerektirir.

- Elde edilen nihai sonuçlar, küme merkezlerinin ilk rastgele seçimine karşı hassastır.
- Aykırı değerlere karşı hassastır.
- Veri yeniden düzenlenirse, verinin sıralamasının her değişiminde farklı bir sonuç alınması çok olasıdır.

1.3.2 K-Medoids Kümeleme Algoritması

K-Means kümelemesinde ortalamanın kullanılması bu yöntemi aykırı değerlere karşı oldukça hassas yapar. Bu gözlemlerin kümeleme atanmasını oldukça etkiler. PAM (partition around medoids) algoritması daha dayanıklı bir algoritma sunar (Kassambara (2017)).

K-Medoids (PAM) algoritması aşağıdaki gibi özetlenebilir;

1. Medoid olacak k objeleri seçilir.
2. Farklılık matrisi hesaplanır.
3. Her gözlem en yakın medoidse yerleştirilir.
4. Her küme için kümenin bir gözleminin ortalama farklılık katsayısını azaltıp azaltmadığı incelenir. Eğer azaltıyorsa ortalama farklılık katsayısını en çok azaltan girdi kümenin medoidi olarak seçilir.
5. Eğer en az bir medoid değiştiyse 3. Adıma gidilir. Diğer durumlarda algoritma sona erdirilir.

K-Medoids Kümelemesi Avantajları

- Verilerin işleniş sırası ve ilk atamada ki merkez seçiminin kümeleme üzerinde etkisi yoktur.
- Merkezi elemanların kümeyi temsil etmesinden dolayı gürültülü veriye karşı duyarlı değildir.

K-Medoids Kümelemesi Dezavantajları

- Uygun küme sayısının belirlenmesi için birden fazla deneme yapmak gerekir.

1.3.3 Hiyerarşik Kümeleme Algoritması

Hiyerarşik kümeleme, bölmeli kümelemeye alternatif bir yaklaşımdır. Bölmeli kümelemelerden farklı olarak hiyerarşik kümeleme küme sayısının önceden belirlenmesine ihtiyaç duymaz. Hiyerarşik kümeleme 2 alt gruba ayrılır (Kassambara (2017));

- Birleştirici Kümelemede her gözlem başlangıçta ayrı bir küme olarak varsayılır. Sonrasında birbirine en benzer kümeler büyük bir küme oluşturana kadar birleştirilir.
- Ayırıcı Kümelemede ise bütün gözlemler tek bir küme içerisine dahil edilir. Sonrasında birbiri arasında heterojenliği en yüksek olan kümeler her gözlemin kendine ait kümesi olana kadar bölünür.

Hiyerarşik kümelemenin sonucu bir ağacı temsil eden dendogramlar ile temsil edilir. Birden çok küme bağlantısı yöntemi vardır. En yaygın yöntemler aşağıda listelenmiştir.

- Complete linkage
- Single Linkage
- Average Linkage
- Centroid Linkage
- Ward's Minimum Variance Method

Hiyerarşik Kümeleme Avantajları

- Hiyerarşik kümelemenin en büyük avantjı algoritma için küme sayısına önceden karar vemeyi gerektirmemesidir.

Hiyerarşik Kümeleme Dezavantajları

- Büyük veri setlerinde kullanılması önerilmez.

1.4 Küme Doğrulama İstatistikleri

Küme doğrulaması terimi kümeleme algoritması sonuçlarının değerlendirme prosedürünü belirtir. Bu prosedürler birden fazla kümeleme algoritmasını karşılaştırmada kullanılır. Genellikle küme doğrulama istatistikleri 3 farklı sınıfta kategorize edilir (Kassambara (2017)).

1. Internal Cluster Validation (İçsel Küme Doğrulaması)
2. External Cluster Validation (Dışsal Küme Doğrulaması)
3. Relative Cluster Validation (Bağıl Küme Doğrulaması)

1.4.1 Küme Doğrulaması İçin İçsel Ölçümler

İçsel doğrulama ölçümleri küme bölmesinin, bölünmesini, yoğunluğunu ve bağlantılılığını yansıtır.

- **Yoğunluk**: Küme içerisindeki objelerin birbirlerine ne kadar yakın olduğunu ölçer. Düşük küme içi toplam kareler toplamı (WSS) iyi yoğunluğun göstergesidir.
- **Bölünme**: Bir kümenin diğer kümelerden ne kadar iyi bölündüğünü ölçer.

  Bölünme ölçümlerinin içerdiği indeksler şunlardır;

  - Küme merkezleri arasındaki uzaklık
  - Farklı kümelerdeki objelerin minimum ikili uzaklıkları
- **Bağlantılılık**: Öğelerin en yakın komşuları gibi ayını kümeye ne ölçüde yerleştirildiğine karşılık gelir. Bağlantılılık minimize edilmesi gerekir ve 0 ile sonsuz arasında bir değer alır.

1.4.1.1 Silhouette Katsayısı

Silhouette analizi bir gözlemin ne kadar iyi kümelendiğini ve kümeler arası ortalama uzaklığı hesaplar. Silhoutte grafiği bir kümedeki her bir noktanın komşu kümelerdeki noktalara ne kadar yakın olduğunun ölçüsünü gösterir. Her i gözlemi için, silhouette genişliği S_i aşağıdaki gibi hesaplanır.

1. Her i gözlemi için i ile i'nin dahil olduğu kümenin tüm noktaları arasındaki ortalama dissimilarity (benzeşmezlik) a_i hesaplanır.

2. i'nin dahil olmadığı diğer kümeler için C'nin tüm gözlemleri için ortalama dissimilarity $(d(i, C))$ hesaplanır. En küçük $d(i, C)$, $b_i = \min Cd(i, C)$ olarak tanımlanır. b_i değeri i ve i'nin komşusu olan küme arasındaki dissimilarity olarak gözlemlenebilir.

3. Son olarak i gözleminin silhouette genişliği şu formülle tanımlanır;
   $S_i = (b_i - a_i) / \max(a_i , b_i)$

Silhouette genişliği aşağıdaki gibi yorumlanabilir.

- Büyük S_i ye sahip gözlemler çok iyi kümelenmiştir.
- Küçük S_i gözlemin iki küme arasında olduğuna işaret eder.
- Negatif S_i değerine sahip gözlemler muhtemelen yanlış kümeye dahil edilmiştir.

1.4.1.2 Dunn İndeks

Dunn indeks aşağıdaki gibi hesaplanan bir başka içsel küme doğrulaması ölçüsüdür.

1. Her küme içerisindeki tüm objelerin aralarındaki ve diğer kümelerdeki objeler arasındaki uzaklıklar hesaplanır.
2. Bu eşli uzaklığın minimumu küme içi bölünme olarak kullanılır.
3. Her küme için aynı küme içerisindeki objelerin aralarındaki uzaklıklar hesaplanır.
4. Maksimum küme içi uzaklık küme içi yoğunluk olarak kullanılır.
5. Dunn indeks (D) aşağıdaki gibi hesaplanır.

$$D = \text{ min. Küme içi Bölünme / max. Küme içi uzaklık} \tag{1.9}$$

Eğer veri seti yoğun ve heterojen kümeler içeriyorsa kümelerin maksimum küme içi bölünmesinin küçük olması ve kümeler arası uzaklığın büyük olması beklenir. Bu sebeple Dunn indeksi maksimize edilmelidir.

Bölüm 2

Uygulama

2.1 Veri ve Yöntem

Bu bitirme projesi uygulama çalışmasında ülkelerin refah seviyelerini ölçme amacıyla oluşturulan Legatum refah indeksi üzerinde 1. Bölümde ele alınan yöntemler kullanılarak kümeleme analizi yapılmıştır.

2.1.1 Veri Setinin Tanıtılması

- **Emniyet ve Güvenlik**: Emniyet ve Güvenlik değişkeni, savaş, çatışma, terör ve suçun bireylerin güvenliğini hem anında hem de daha uzun süreli etkiler yoluyla istikrarsızlaştırma derecesini ölçer.
- **Kişisel Özgürlük**: Kişisel Özgürlük değişkeni, temel yasal haklara, bireysel özgürlüklere ve sosyal hoşgörüye yönelik ilerlemeyi ölçer.
- **Yönetim**: Yönetim değişkeni, güç üzerinde ne kadar kontrol ve kısıtlama olduğunu ve hükümetlerin yolsuzluk olmadan ve etkili bir şekilde çalışıp çalışmadığını ölçer.
- **Sosyal Sermaye**: Sosyal Sermaye değişkeni, bir ülkedeki kişisel ve sosyal ilişkilerin, kurumsal güvenin, sosyal normların ve sivil katılımın gücünü ölçer.
- **Yatırım Ortamı**: Yatırım ortamı değişkeni, yatırımların ne ölçüde yeterince korunduğunu ve kolayca erişilebilir olduğunu ölçer.
- **Kurumsal Koşullar**: Kurumsal koşullar değişkeni, işletmelerin kurulmasına, rekabet etmesine ve genişlemesine, düzenlemelerin ne derece olanak tanıdığını ölçer.

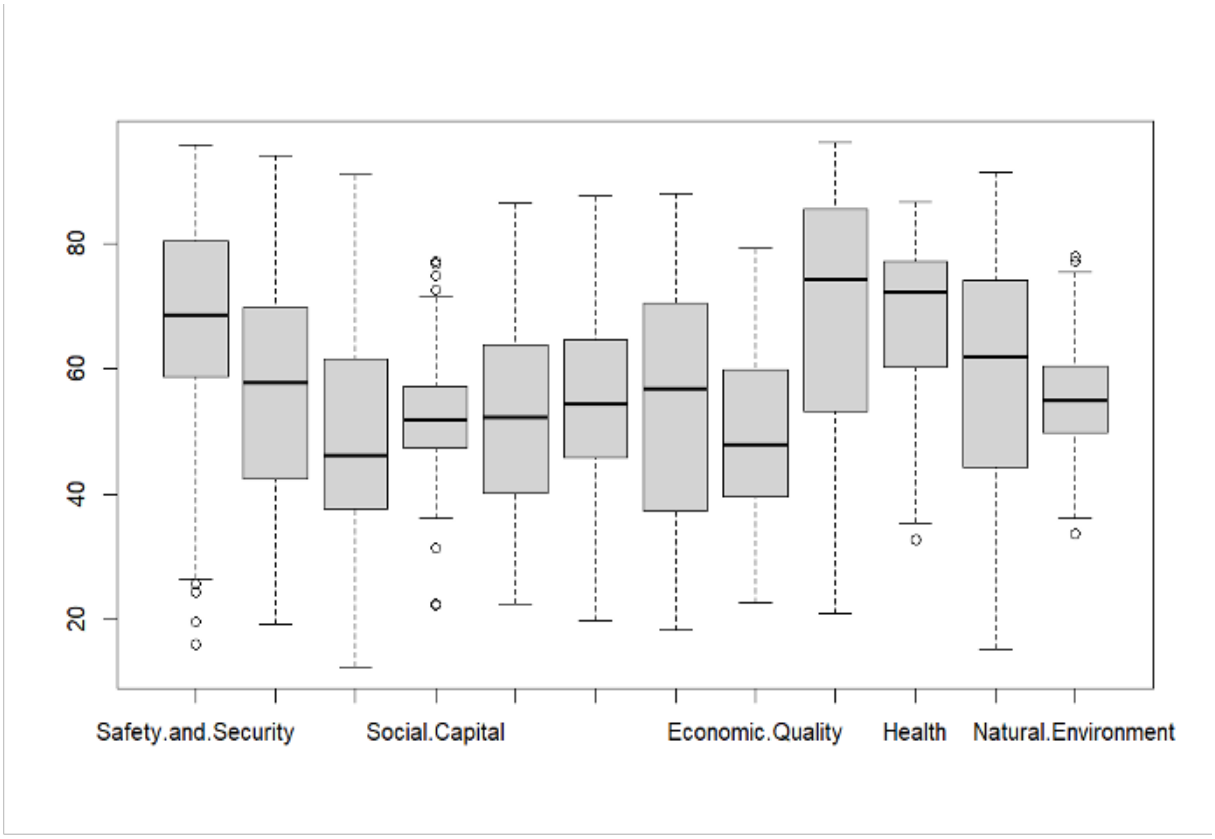
- **Pazar Erişimi ve Altyapı**: Pazar Erişimi ve Altyapı değişkeni, ticareti mümkün kılan altyapının kalitesini, mal ve hizmet pazarındaki bozulmaları ölçer.
- **Ekonomik Kalite**: Ekonomik Kalite değişkeni, bir ekonominin sürdürülebilir şekilde ve iş gücünün tam katılımıyla varlık yaratmak için ne kadar donanımlı olduğunu ölçer.
- **Yaşam Koşulları**: Yaşam koşulları değişkeni, materyal kaynakları, beslenme, temel hizmetler, barınma, sigorta hizmetleri dahil olmak üzere herkes tarafından makul bir yaşam kalitesinin ne kadar deneyimlediğine ölçer.
- **Sağlık**: İnsanların ne ölçüde sağlıklı olduğunu, sağlık sistemleri, sonuçları, hastalık ve risk faktörleri, ölüm oranları dahil olmak üzere sağlıklılığı sürdürmek için gerekli hizmetlere ne ölçüde erişilebildiğini ölçer.
- **Eğitim**: Eğitim, yetişkin nüfustaki becerilerin yanı sıra eğitimin 4 aşaması (ilkokul öncesi, ilköğretim, ortaöğretim, yükseköğretim) genelinde kayıtları, sonuçları ve kaliteyi ölçer.
- **Doğal Çevre**: Doğal çevrenin, insanlar üzerinde günlük yaşamlarında doğrudan etkisi olan yönlerini ve gelecek nesillerin refahını etkileyebilecek değişiklikleri ölçer.

2.1.2 Tanımlayıcı İstatistikler

Değişkenlerin tanımlayıcı istatistiklerine ve kutu grafiklerine bakıldığında sapan değerler ve çarpık dağılımlar gözlemlenmiştir.

Tablo 2.1: Tanımlayıcı İstatistikler

	n	mean	sd	median	min	max	skew	kurtosis	se
Safety and Security	167	67.21	17.52	68.6	16.00	95.70	-0.63	-0.08	1.36
Personel Freedom	167	57.49	18.23	57.9	19.30	94.10	0.05	-0.85	1.41
Governance	167	50.58	19.02	46.2	12.40	91.00	0.47	-0.58	1.47
Social Capital	167	52.72	9.02	51.9	22.30	77.20	0.08	1.14	0.70
Investment Enviroment	167	53.25	15.50	52.3	33.50	86.40	0.09	-0.96	1.20
Enterprise Conditions	167	55.02	14.08	54.4	19.90	87.50	0.17	-0.56	1.09
Market Access and Infrastructure	167	54.26	18.38	56.8	18.30	88.00	-0.07	-1.17	1.42
Economic Quality	167	49.73	13.28	48.00	22.70	79.30	0.35	-0.82	1.03
Living Conditions	167	68.93	19.88	74.4	21.10	96.20	-0.53	-0.90	1.54
Health	167	68.85	11.16	72.3	32.80	86.60	-0.86	0.13	0.86
Education	167	58.93	19.40	61.9	15.30	91.30	-0.25	-1.03	1.50
Natural Environment	167	55.57	8.82	55.00	33.70	78.00	0.27	-0.23	0.68

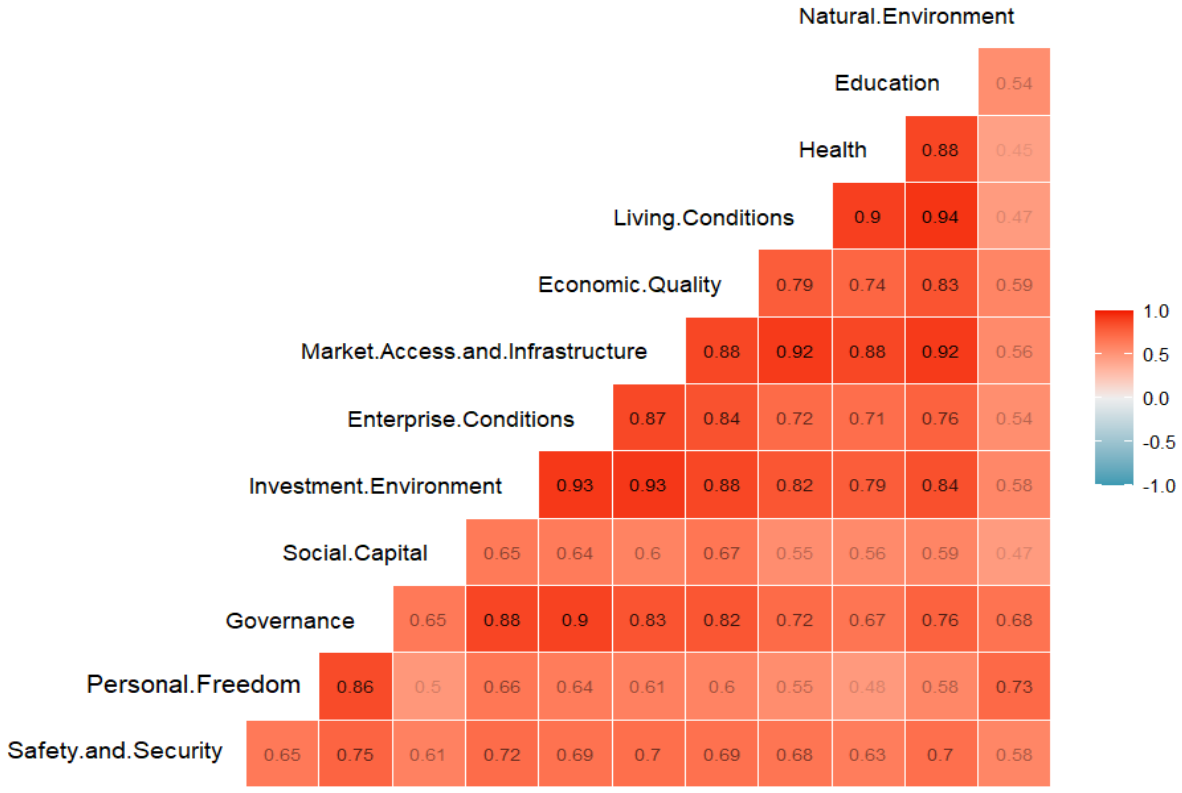


Şekil 2.1: Değişkenlerin Kutu Grafiği

2.1.3 Korelasyon Matrisi

Şekil 2.2 bakıldığı zaman; değişkenlerin her biri birbirleri ile pozitif yönlü doğrusal bir ilişki içerisinde oldukları gözlemlenmektedir. Değişkenler arasında negatif yönlü doğrusal hiçbir ilişki bulunmamaktadır. En yüksek pozitif yönlü doğrusal ilişki 0.942 korelasyon değeri ile eğitim değişkeni ile yaşam koşulları değişkeninin arasındadır. En düşük pozitif yönlü doğrusal ilişki 0.4457 korelasyon değeri ile doğal çevre değişkeni ile sağlık değişkeni arasındadır.

Verinin daha kolay ve doğru bir şekilde analiz edilip görselleştirilebilmesi için temel bileşenler analizi yapılmasına karar verilmiştir..



Şekil 2.2: Korelasyon Değerleri

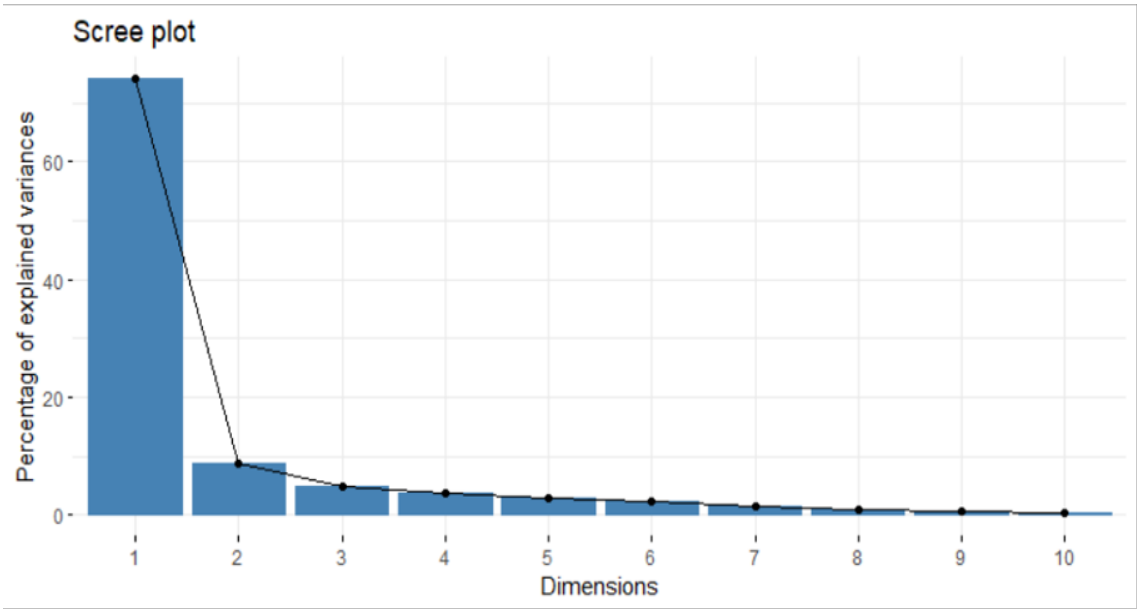
2.1.4 Temel Bileşenler Analizi

1. bileşen boyutunda açıklayıcılık %74.03'tür. Özdeğer ve toplam varyans yüzdesi değerlerine bakıldığında 2 temel bileşen ile %82.64 açıklayıcılığa ulaşıldığı görülmektedir.

Tablo 2.2: Temel Bileşenler Analizi Sonuçları

	eigenvalue	variance.percent	cumulative variance percent
Dim.1	8.88	74.02	74.0
Dim.2	1.03	8.62	82.64
Dim.3	0.56	4.72	87.37
Dim.4	0.44	3.69	91.06
Dim.5	0.35	2.93	93.99
Dim.6	0.27	2.27	96.26
Dim.7	0.16	1.33	97.60
Dim.8	0.10	0.88	98.49
Dim.9	0.06	0.56	99.05
Dim.10	0.05	0.42	99.48
Dim.11	0.03	0.29	99.77
Dim.12	0.02	0.22	100.0

Tablo 2.2'ye bakılığında, birinci özdeğer 8.88 ve ikinci özdeğer 1.03'tür. Üçüncü ise 0.56'dır, bu durumda ise ilk iki bileşen seçilmesi uygundur.

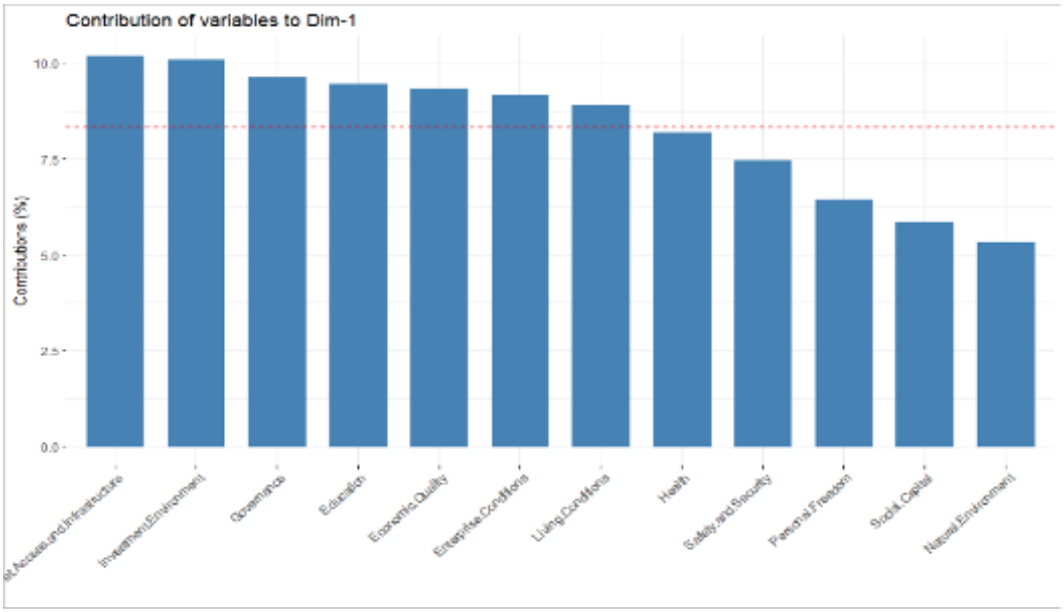


Şekil 2.3: Scree Plot Grafiği

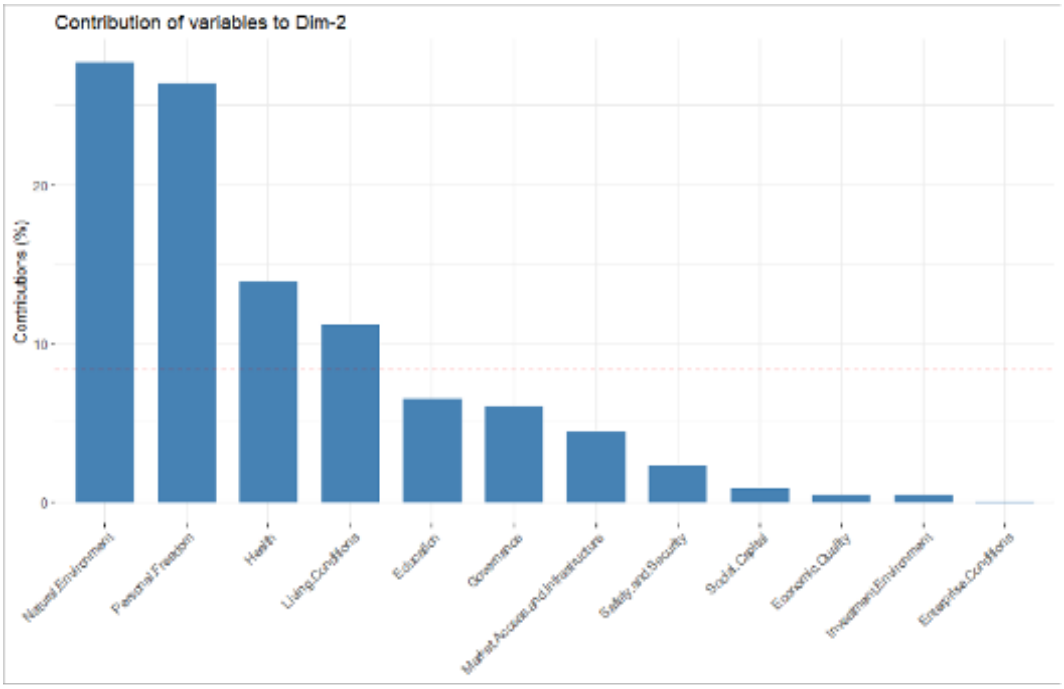
Şekil 2.3 incelendiğinde grafikte kırılmanın düzleştiği ilk yere kadar olan bileşen sayısı temel bileşen olarak belirlenebilir. Scree plotta dirsek noktası 2. Temel bileşendedir. %82,64 açıklayıcılığa sahip 2 temel bileşen oluşturulmasına karar verilmiştir.

2.1.5 Bileşenlerin Yorumlanması

1. Bileşenin açıklayıcılığı; 1.Bileşende, emniyet ve güvenlik, yönetim, sosyal sermaye, yatırım ortamı, kurumsal koşullar, pazar erişimi ve altyapı, ekonomik kalite ve eğitim değişkenlerinin açıklayıcılıkları daha yüksektir.
2.Bileşende, kişisel özgürlük, yaşam koşulları, sağlık, doğal çevre değişkenlerinin açıklayıcılıkları daha yüksektir.
1. bileşen Ekonomi ve Yönetim, 2. bileşen Yaşam Kalitesi olarak adlandırılmıştır.



Şekil 2.4: Değişkenlerin 1. Bileşeni Açıklayıcılık Grafiği

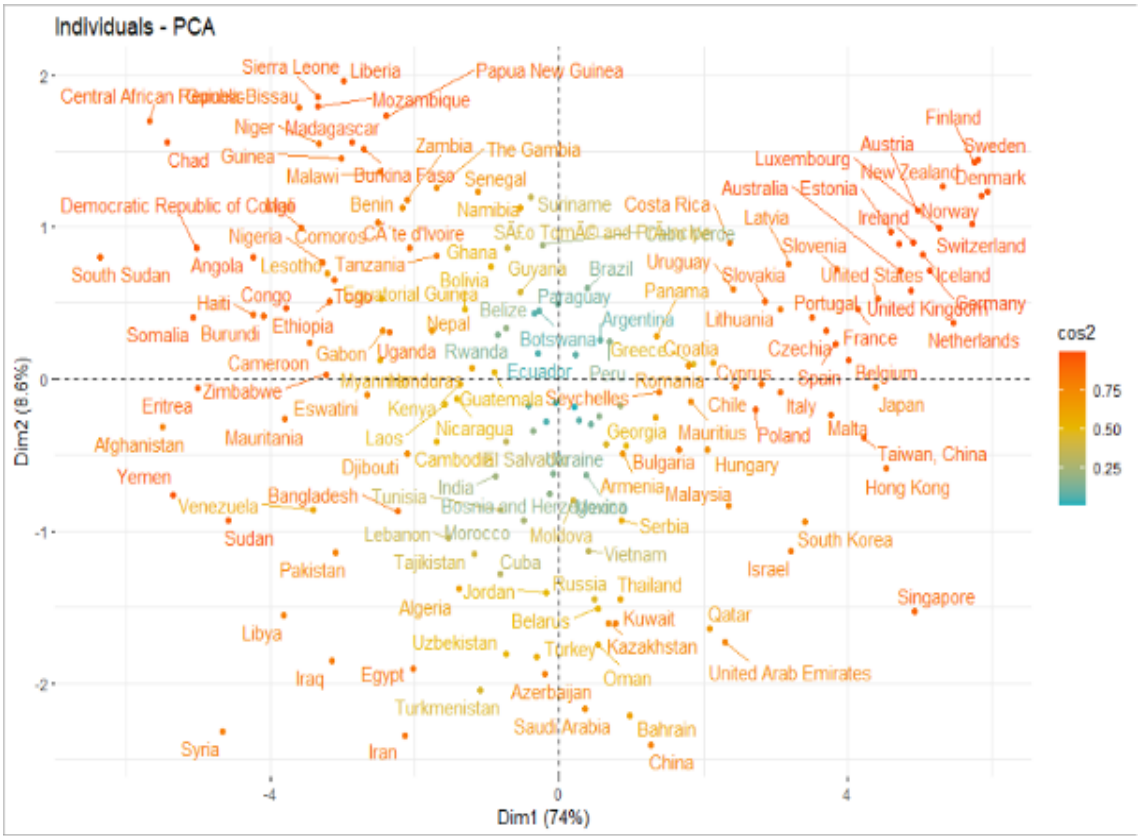


Şekil 2.5: Değişkenlerin 2. Bileşeni Açıklayıcılık Grafiği

2.1.6 Ülkelerin Konumları

```
'fviz\_pca\_ind(data.pca,
col.ind="cos2",
gradient.cols=c("#00AFBB", "E7B800", "FC4E07"))'
```

komutu ile elde edilen grafik aşağıdaki gibidir.



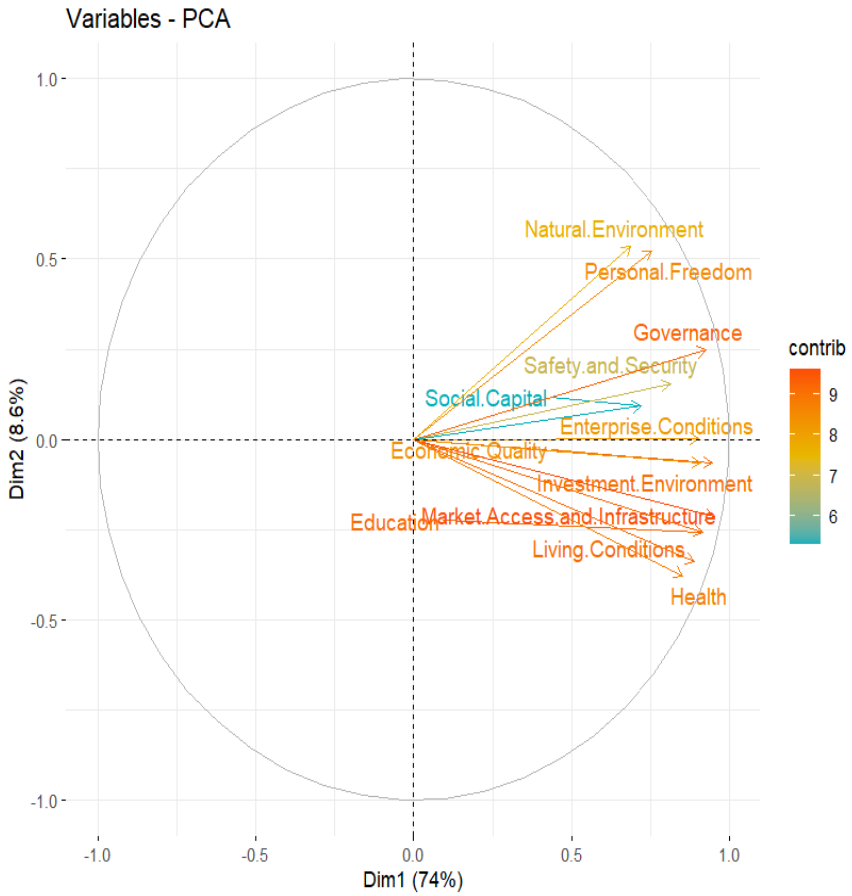
Şekil 2.6: Verilerin 1. ve 2. Bileşene göre Konumları-1

Şekil 2.6 daki grafikte bileşenler bazında değişkenlerin ve ülkelerin konumları gösterilmiştir. İsveç ve Finlandiya konumu ekonomi ve yonetim ile yaşam kalitesi açısından en yüksek konumda yer almaktadır. Suriye ve İran ekonomi ve yönetim ile yaşam kalitesi açısından en düşük konumda yer almaktadır. Yukarıdaki ülkelerin bileşenler üzerindeki yerleşimlerine bakıldığı zaman; Singapur ekonomi ve yönetim bakımında iyi olduğunun ama yaşam kalitesi açısından kötü olduğunu gözlemlemekteyiz. Orta Afrika, Çad ve Güney Sudan ülkelerinin ise ekonomi ve yönetim açısından kötü durumda olmasına rağmen yaşam kalitesi bakımından iyi koşulda olduğu gözlemlenmektedir. Türkiye'ye bakıldığı zaman ekonomi ve yönetim bakımından kötü olduğunu yaşam kalitesi bakımından ortalamada olduğu gözlemlenmektedir.

2.1.7 Değişkenlerin Açıklayıcılığa Katkıları

```
'fviz\_pca\_var(data.pca,
axes=c(1,2)
col.var="cos2",
gradient.cols=c("#00AFBB", "E7B800", "FC4E07"),
repel=TRUE)'
```

komutu ile elde edilen grafik aşağıdaki gibidir.



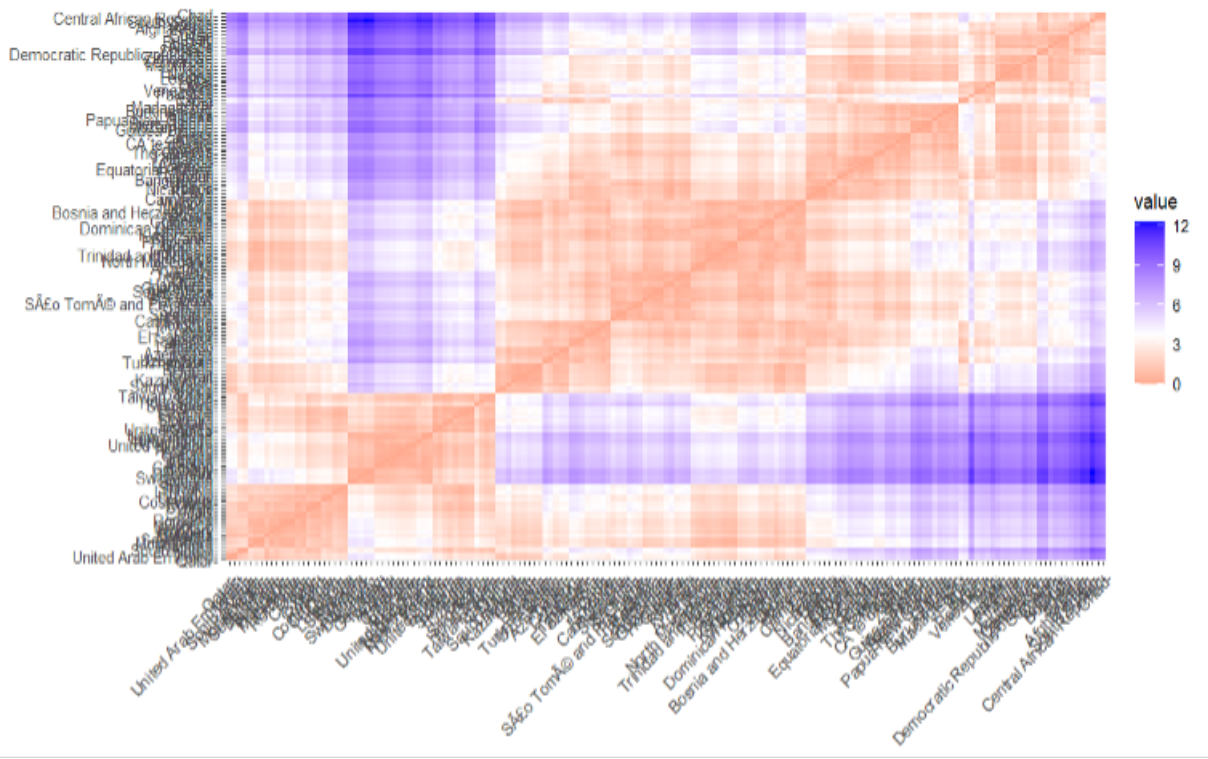
Şekil 2.7: Verilerin 1. ve 2. Bileşene göre Konumları-2

Hangi değişkenin hangi bileşende ne kadar açıklandığı yukarıdaki grafikte gözükmektedir. Vektörlerin uzunluklarına ve derecelerine bakıldığında, kişisel özgürlük ve doğal çevre değişkenlerinin Yaşam Kalitesi bileşenin açıklayıcılığına olan katkıları diğer değişkenlere göre daha yüksektir. Sosyal sermaye değişkeninin iki bileşene de olan katkısı diğer değişkenlere göre daha düşüktür.

2.2 Uzaklıklar

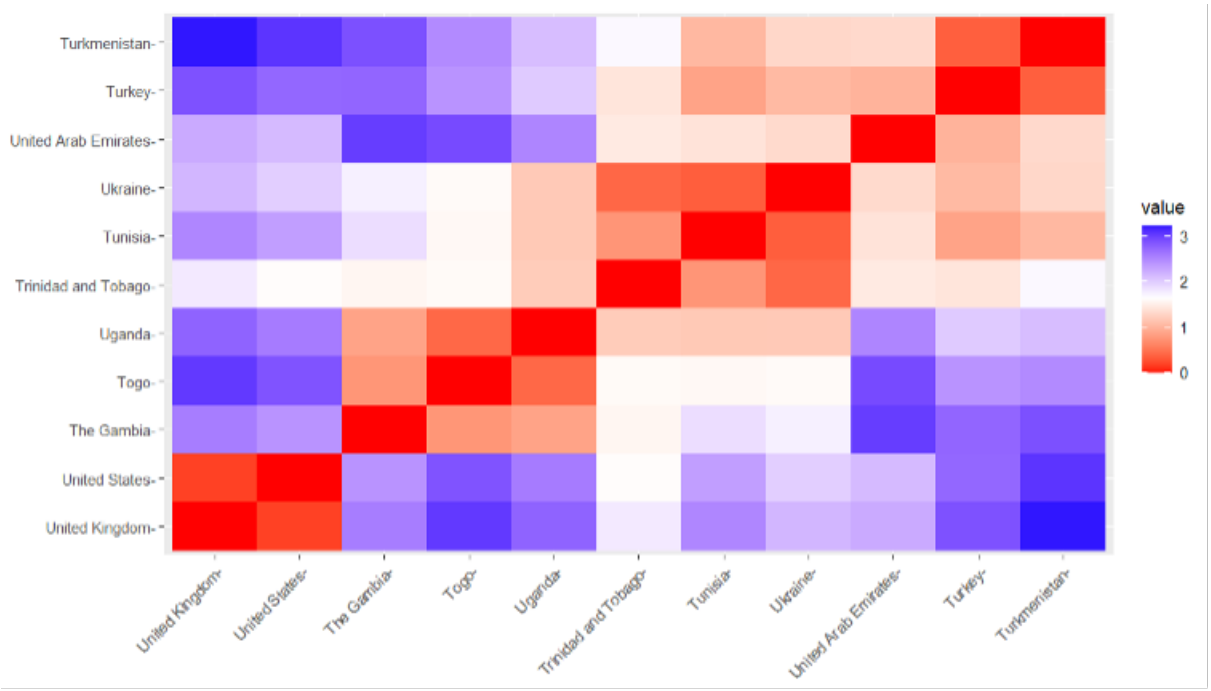
2.2.1 Veri Setinin Öklid Uzaklığı

Bütün ülkelerin öklid uzaklık grafiği aşağıda verilmiştir.



Şekil 2.8: Öklid Uzaklık Grafiği

Uzaklıkların Şekil 2.8 incelendiğinde kümeleme analizine uygun olduğu görülmektedir. Türkiye'nin uzaklık matrisinde bulunduğu konum aşağıdaki grafikte gösterilmiştir.

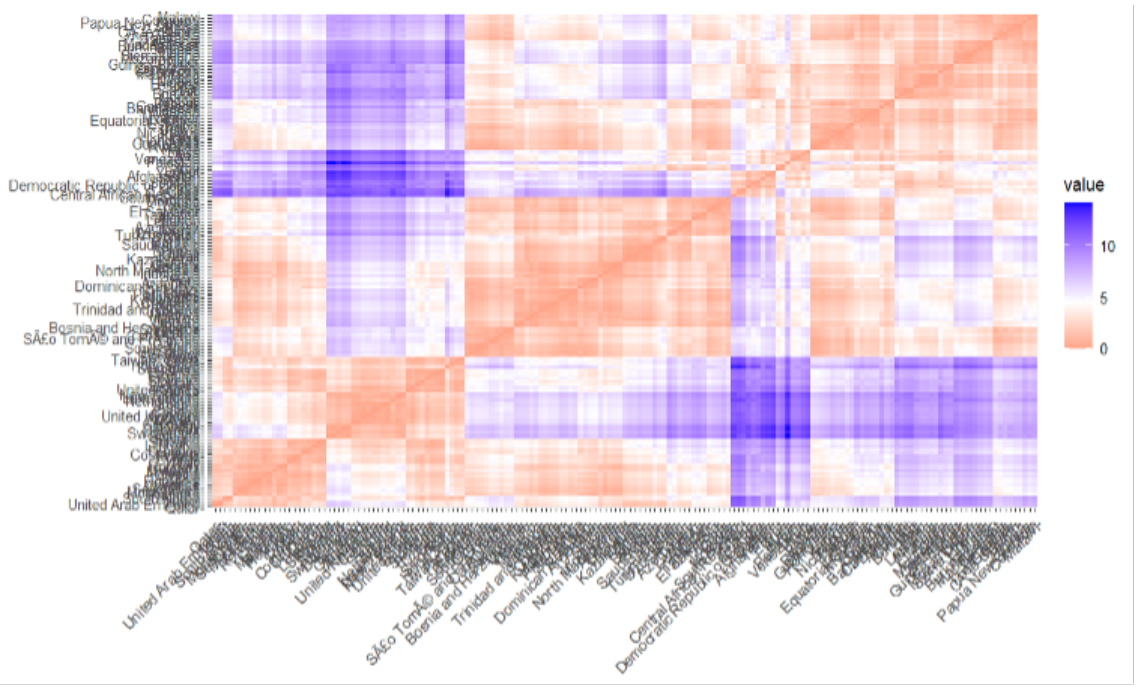


Şekil 2.9: Türkiye'nin Öklid Uzaklık Grafiği

Şekil 2.9 bakıldığında Türkiye'nin Birleşik Krallık ve Amerika ile olan öklid uzaklığı yüksektir. Tunus ve Ukrayna ile olan öklid uzaklığı ise diğer ülkelere kıyasla daha düşüktür.

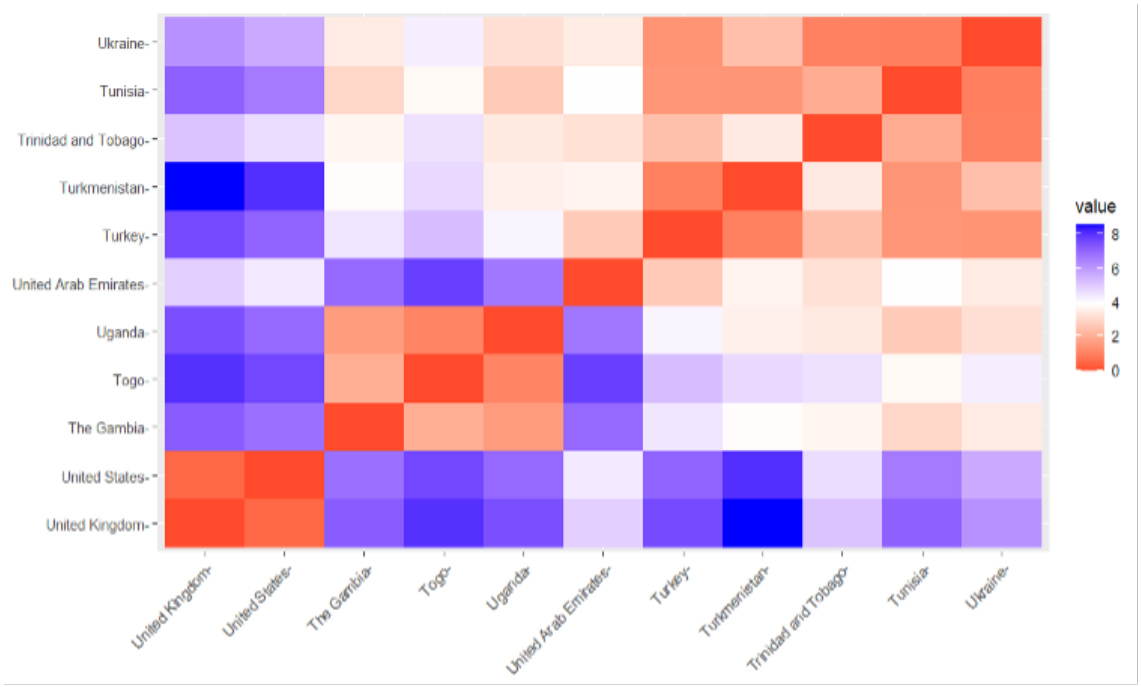
2.2.2 Veri Setinin Manhattan Uzaklığı

Bütün ülkelerin manhattan uzaklık grafiği aşağıda verilmiştir.



Şekil 2.10: Manhattan Uzaklık Grafiği

Uzaklıkların Şekil 2.10 incelendiğinde kümeleme analizine uygun olduğu görülmektedir. Türkiye'nin uzaklık matrisinde bulunduğu konum aşağıdaki grafikte gösterilmiştir.

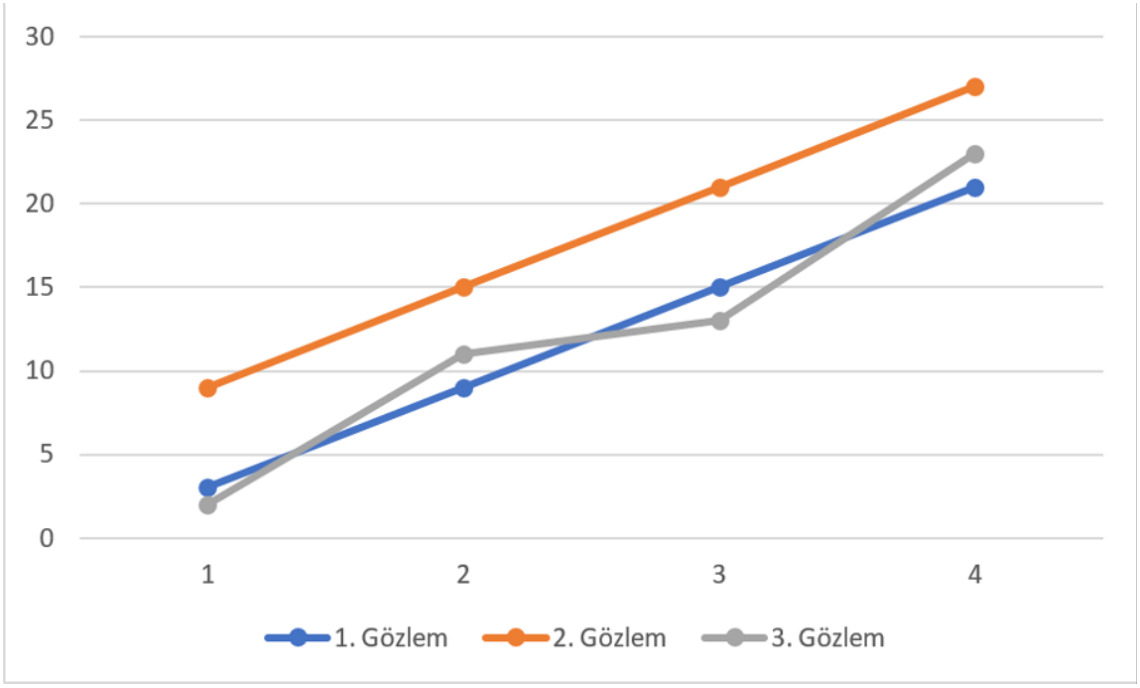


Şekil 2.11: Türkiye'nin Manhattan Uzaklık Grafiği

Şekil 2.11 bakıldığında Türkiye'nin Birleşik Krallık ve Amerika ile olan öklid uzaklığı yüksektir. Tunus, Türkmenistan ve Ukrayna ile olan öklid uzaklığı ise diğer ülkelere kıyasla daha düşüktür.

2.2.3 Pearson Korelasyon Uzaklığı

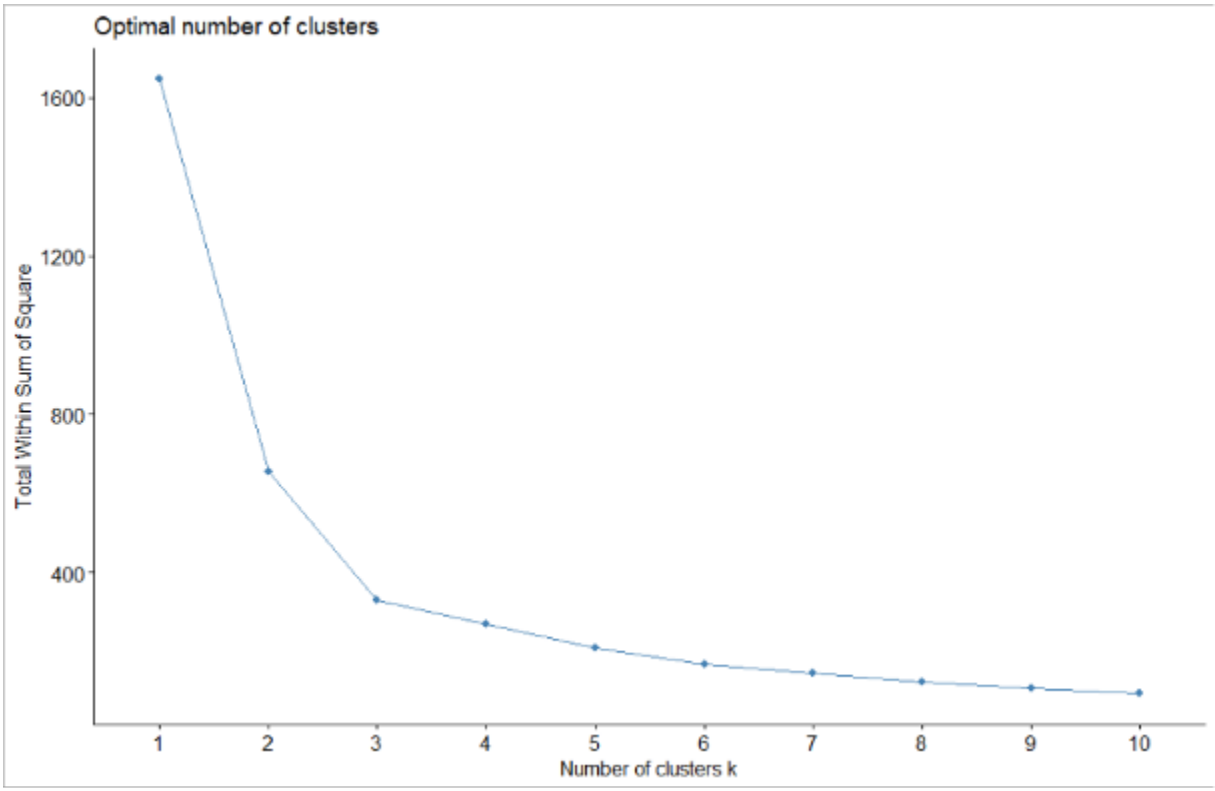
Korelasyon temelli uzaklık ölçüleri, iki ayrı veriyi eğer aralarında yüksek bir ilişki var ise aralarındaki değer farkına bakılmaksızın birbirine benzer kabul eder. Bu iki ayrı veri arasında nicel olarak çok fark olması korelasyon uzaklığında bir anlam ifade etmez. Aşağıdaki tabloda görüleceği gibi 1. Gözlem ile 2. Gözlem Pearson Korelasyon uzaklığına göre birbirine benzer sayılır ve aynı kümeye dahil edilir. Lakin 1. Gözlem ile 3. Gözlem in grafiğine bakıldığında değerlerin birbirine daha yakın olduğu gözlemlenmektedir. Veri setinde ülkelerin nicel olarak refah seviyeleri anlam ifade ettiği için bu veri seti için Pearson Korelasyon uzaklığı kullanılmamıştır.



Şekil 2.12: Uzaklık Yöntemlerinin Kıyaslaması(Gözlemler veri setinden alınmamıştır)

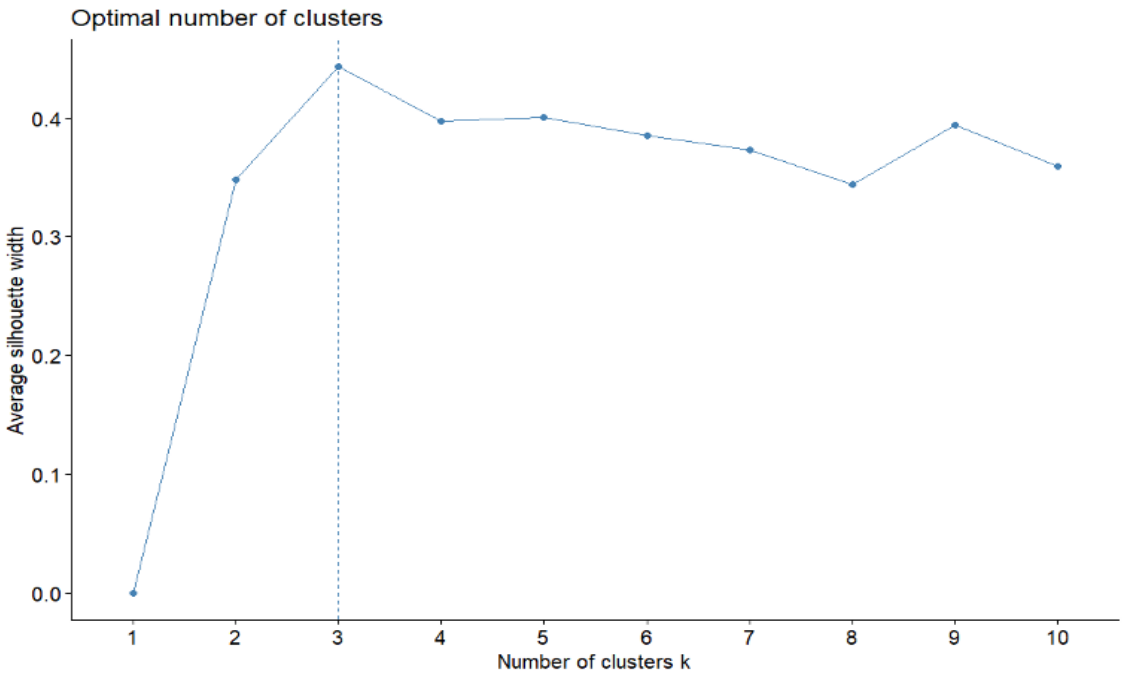
2.3 Küme Sayısının Belirlenmesi

Elbow Yöntemi kullanılarak oluşturulan Şekil 2.13 incelendiğinde dirsek noktasının $k = 3$ konumunda olduğu gözlemlenmektedir.



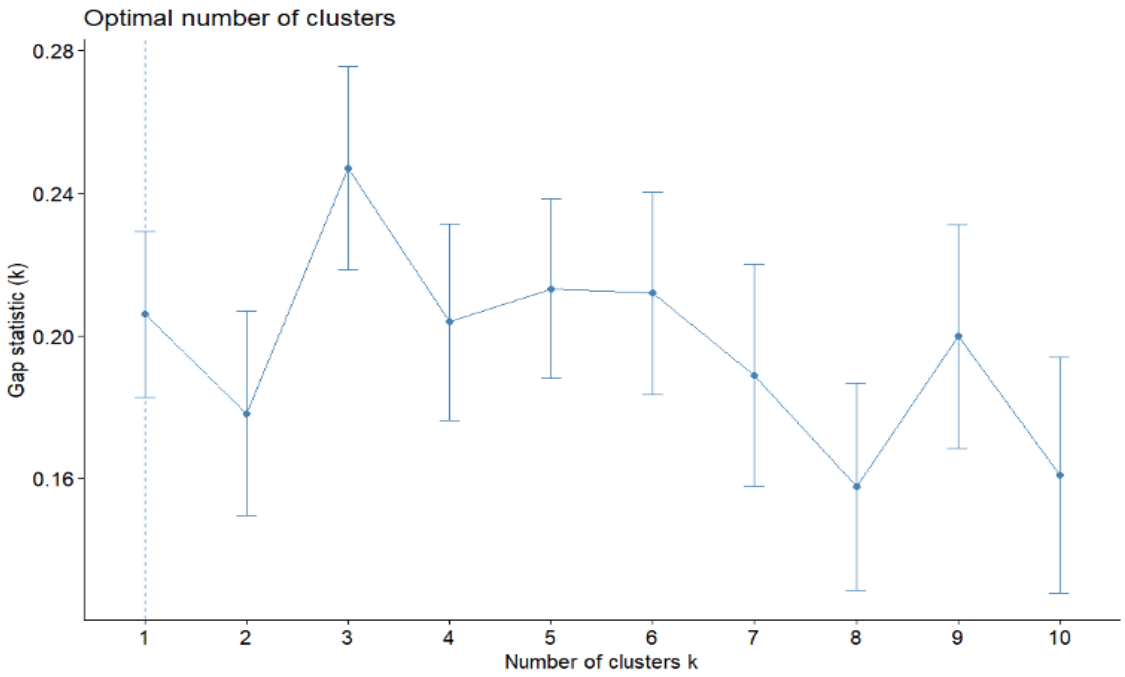
Şekil 2.13: Elbow Methodu Grafiği

Average Silhouette yöntemi grafiği Şekil 2.14 incelendiğinde maksimum değer olan 3 en uygun küme sayısı olarak belirtilmiştir.



Şekil 2.14: Average Silhouette Yöntemi Grafiği

Şekil 2.15 görüldüğü üzere GAP istatistiğinin 1 standart sapma içerisinde olacak şekilde en küçük değer 3'tür. Bu en uygun küme sayısını belirtir.



Şekil 2.15: GAP Yöntemi Grafiği

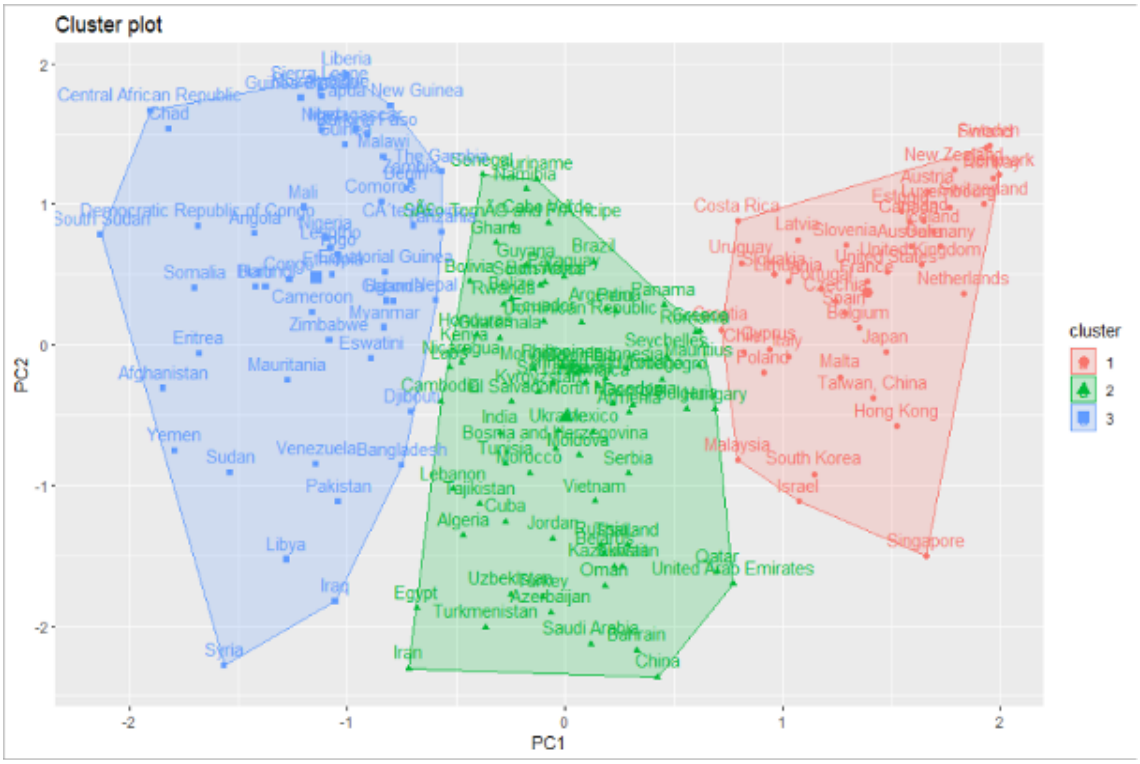
Üç yöntemin de sonucu olan 3 değeri oluşturulacak küme sayısı olarak seçilmiştir.

2.4 Kümeleme

2.4.1 K-Means Kümeleme

```
fviz\_cluster(list(data=(df\_numpca), cluster=km.clusters ))
```

komutu ile elde edilen grafik aşağıda verilmiştir.



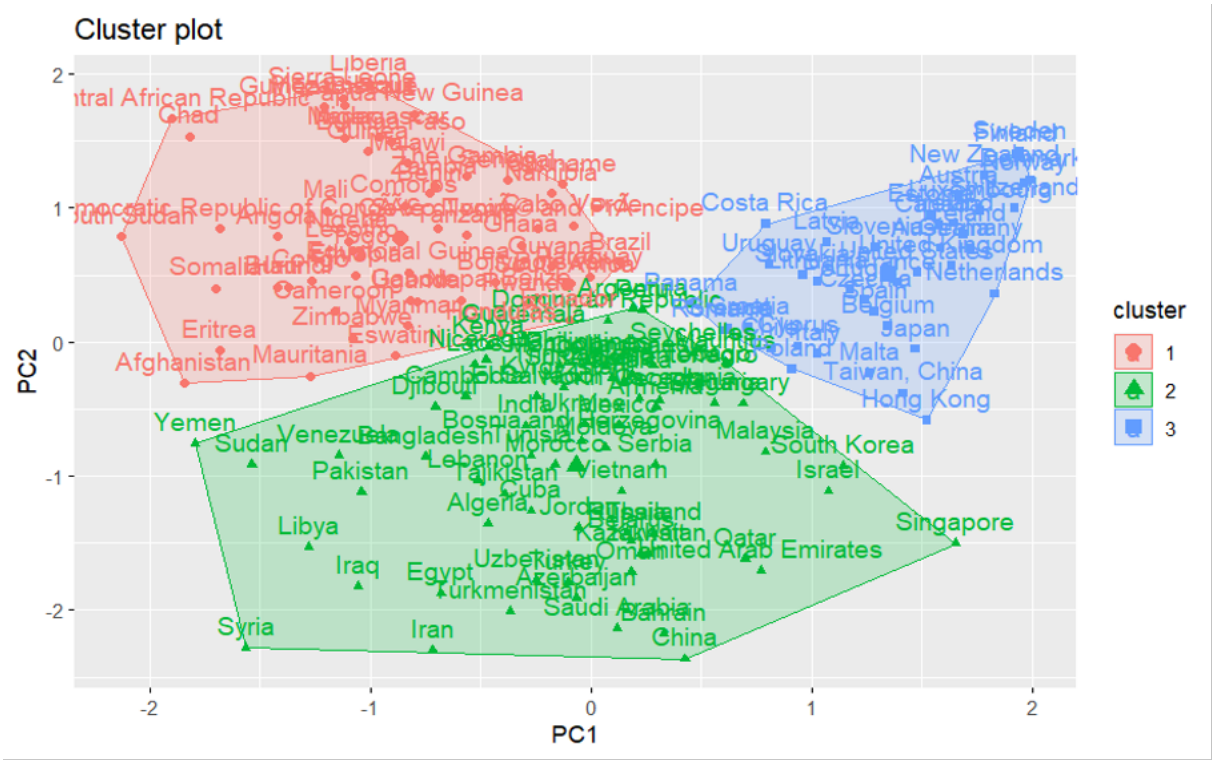
Şekil 2.16: K-means Yöntemi Kümeleme Grafiği

K-Means kümeleme algoritması kullanılarak yukarıdaki Şekil 2.16 elde edilmiştir. Şekil 2.16 görüldüğü üzere 1.bileşen ekseni üzerinde kümeler aralarında açıklayıcı bir farklılığa sahiptirler. Fakat 2.bileşen ekseninde bu farklılığa ulaşılamamıştır.

2.4.2 K-Medoids Kümeleme

fviz_cluster(list(data=scale(df_numpca), cluster=kmed.clusters))

komutu ile elde edilen grafik aşağıda verilmiştir.



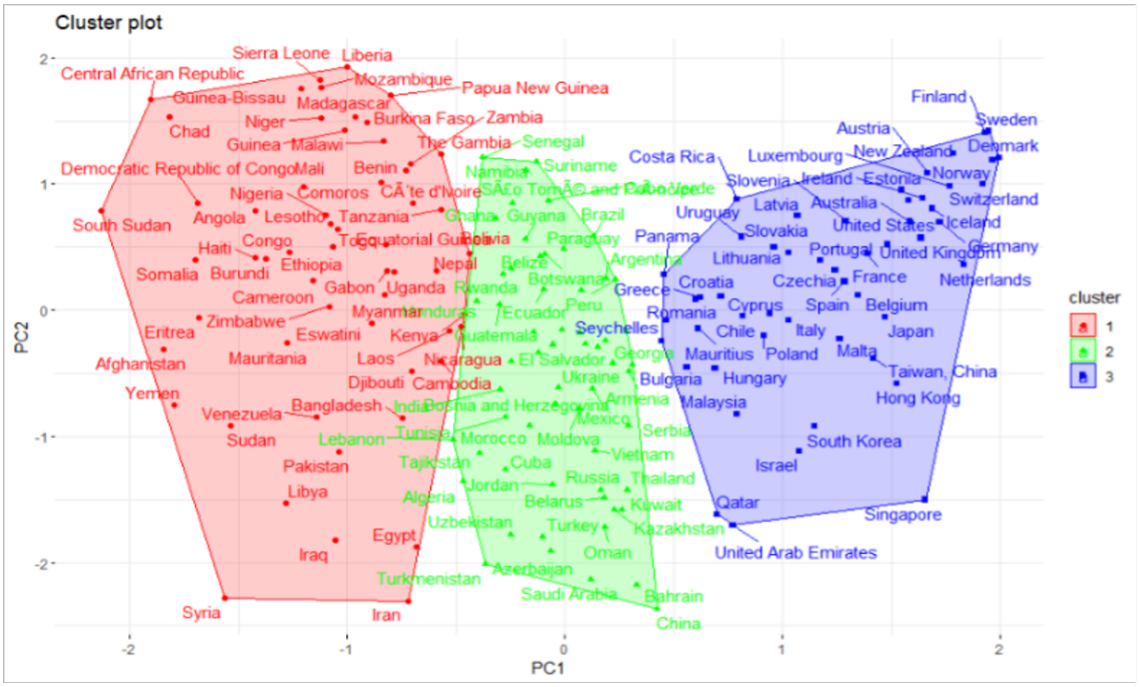
Şekil 2.17: K-medoids Yöntemi Kümeleme Grafiği

K-Medoids kümeleme algoritması kullanılarak yukarıdaki şekil 2.17 elde edilmiştir. Şekil 2.17 görüldüğü üzere ikinci kümede küme içi homojenlik yeterince sağlanamamıştır.

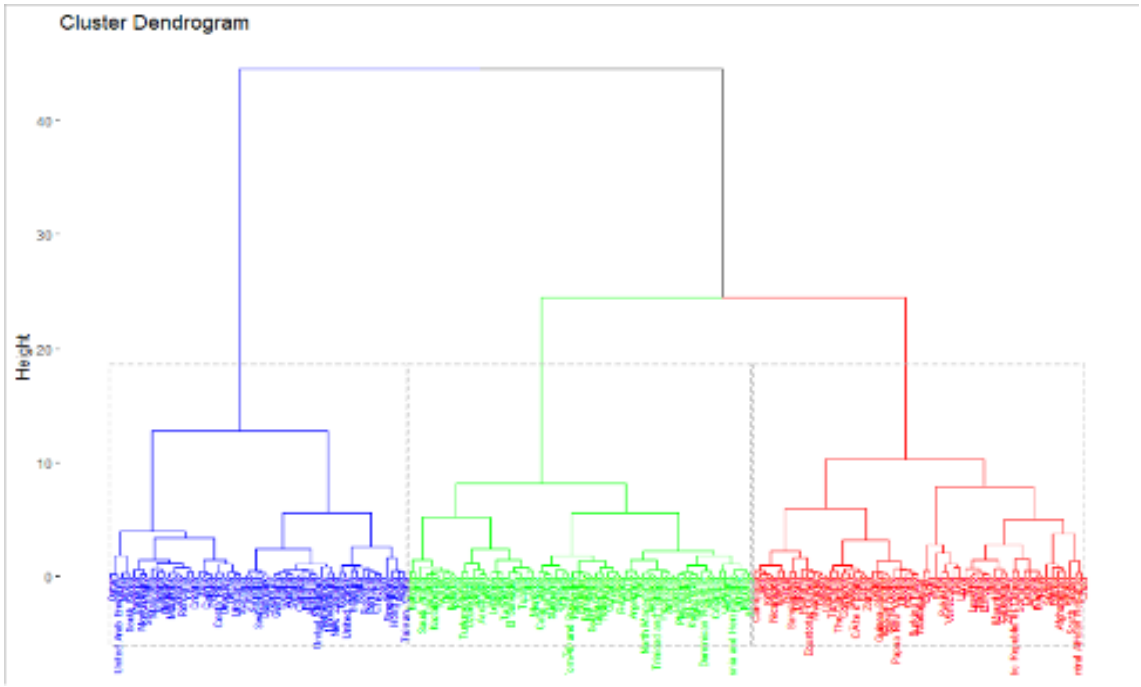
2.4.3 Hiyerarşik Kümeleme

```
fviz\_dend(hc.res, show\_labels=FALSE, palette="jcp", as.ggplot=TRUE)
fviz_cluster(list(data = df_numpca, cluster = grp),
palette = c("red", "green", "blue", "black"),
ellipse.type = "convex",
repel = TRUE,
show.clust.cent = FALSE, ggtheme = theme_minimal())
```

komutu ile elde edilen grafik ve dendogram aşağıda verilmiştir.



Şekil 2.18: Hiyerarşik Kümeleme Yöntemi Grafiği



Şekil 2.19: Hiyerarşik Kümeleme Yöntemi Dendogram

Hiyerarşik kümeleme algoritması kullanılarak yukarıdaki Şekil 2.18 ve Şekil 2.19 elde edildi. Dendogram incelendiğinde boyutları birbirine yakın 3 adet kümeye ayrılması uygun görülmüştür.

2.5 Kümeleme Algoritmalarının Karşılaştırılması

```
clmethods= c("hierarchical", "kmeans", "pam")
intern=clValid(df\_numpca, nClust=3, clMethods=clmethods,
validation="internal")
summary(intern)
```

Aşağıda verilen çıktıda, veri setine uygulanan kümeleme algoritmalarının doğrulama ölçüleri incelenerek en uygun kümse sayısı ve algoritmaya karar verilmiştir. En düşük bağlantılılık katsayısına, en yüksek dunn indeks katsayısına sahip olan hiyerarşik kümeleme yöntemi silhoutte katsayısında da ortalama bir değer aldığı için sonuçlar doğrultusunda hiyerarşik yöntem kullanılarak 3 kümeye oluşturulmasına karar verilmiştir.
Yukarıdaki komutlar ile çıktı elde edilmiştir.

Clustering	Methods	:			
hierarchical	kmeans	pam			
Cluster	sizes	:			
3 4 5 6					
Validation	Measures	:			
		3	4	5	6
hierarchical	Connectivity	14.6171	19.7163	25.8984	32.3032
	Dunn	0.0836	0.0926	0.0952	0.1026
	Silhouette	0.4639	0.4325	0.3910	0.3992
kmeans	Connectivity	20.5163	33.7730	37.1310	51.3524
	Dunn	0.0602	0.0602	0.0727	0.0518
	Silhouette	0.4898	0.4471	0.3974	0.3778
pam	Connectivity	20.5163	25.5976	40.0286	53.5222
	Dunn	0.0602	0.0457	0.0457	0.0492
	Silhouette	0.4898	0.4285	0.3951	0.3768
Optimal Scores	:				
	Score	Method	Clusters		
Connectivity	14.6171	hierarchical	3		
Dunn	0.1026	hierarchical	6		
Silhouette	0.4898	kmeans	3		

Veri seti 3 kümeye ayrılmıştır ve kümeler aşağıdaki gibi adlandırılmıştır.

- 1.Küme: “Refah Seviyesi Düşük Olan Ülkeler” toplam 57 ülke
- 2.Küme: “Refah Seviyesi Orta Olan Ülkeler” toplam 59 ülke
- 3.Küme: “Refah Seviyesi Yüksek Olan Ülkeler” toplam 51 ülke

2.6 Kümelerin Tanımlayıcı İstatistikleri

Refah Seviyesi Düşük Olan Ülkeler Kümesi ;

Kümede bulunan ülkelerin birkaçı şunlardır. Afganistan, Irak, Suriye, Pakistan ve Zimbabve.

Tablo 2.3: Refah Seviyesi Düşük Olan Ülkeler Kümesi Tanımlayıcı İstatistikler

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Safety And Security	57	52.49	15.69	58.1	53.36	13.94	16	78.7	62.7	-0.50	-0.84	2.08
Personal Freedom	57	44.49	11.27	43	45.13	14.08	19.3	63.2	43.9	-0.39	-0.65	1.49
Govarnance	57	33.61	9.03	34.7	33.86	9.04	12.4	50	37.6	-0.28	-0.76	1.20
Social Capital	57	46.5	6.93	46.9	47.08	6.23	22.3	57.3	35	-1.32	2.78	0.92
Investment Enviroment	57	37.32	7.45	36.9	37.37	9.19	22.5	54.8	32.3	0.04	-0.71	0.99
Enterprise Conditions	57	42.09	8.14	42.9	42.39	8.90	19.9	56.3	36.4	-0.42	-0.29	1.08
Market Access and Infrastructure	57	34.07	8.18	34.3	33.86	8.60	18.3	56.3	38	0.30	-0.30	1.08
Economic Quality	57	37.66	6.29	37.9	37.74	6.38	22.7	51.8	29.1	-0.10	-0.42	0.83
Living Conditions	57	46.89	13.07	46.1	46.47	12.01	21.1	76.6	55.5	0.31	-0.57	1.73
Health	57	56.73	8.95	56.8	57.13	8.60	32.8	74.8	42	-0.36	0.13	1.19
Education	57	37.94	10.89	37.3	37.80	12.75	15.3	63.9	48.6	0.16	-0.55	1.44
Natural Environment	57	50.82	6.49	52.0	51.28	5.93	33.7	60.5	26.8	-0.69	-0.40	0.86

Refah Seviyesi Orta Olan Ülkeler Kümesi;

Kümede bulunan ülkelerin birkaçı şunlardır. Brezilya, Hindistan, Rusya ve Çin

Tablo 2.4: Refah Seviyesi Orta Olan Ülkeler Kümesi Tanımlayıcı İstatistikler

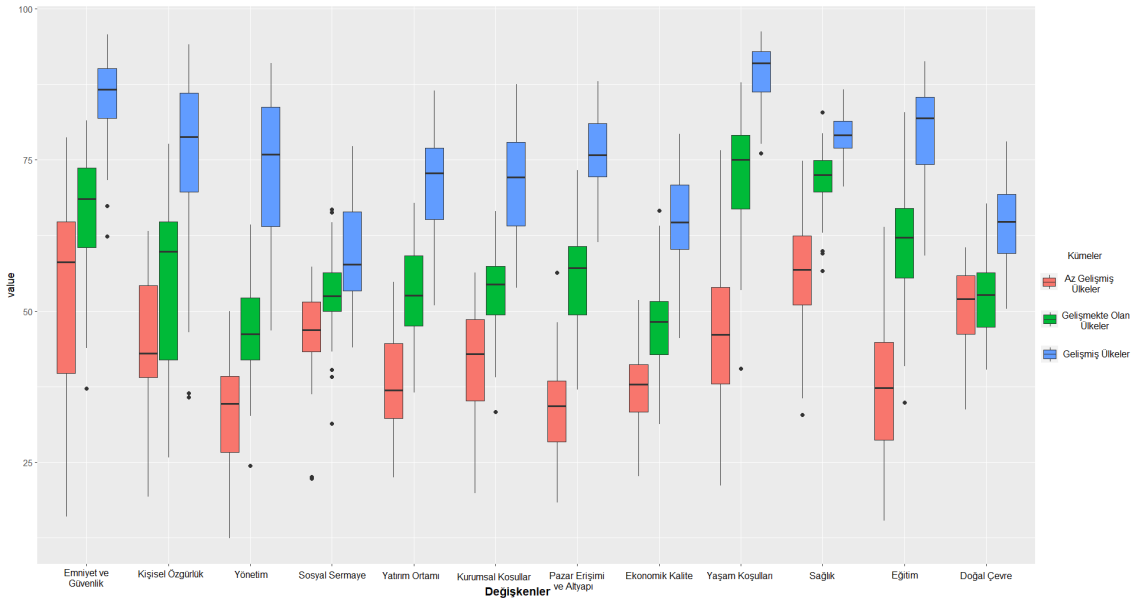
	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Safety And Security	59	66.03	10.17	68.5	66.84	9.64	37.2	81.5	44.3	-0.81	0.06	1.32
Personal Freedom	59	53.99	13.88	59.8	54.58	10.67	25.8	77.6	51.8	-0.47	-1.03	1.81
Govarnance	59	46.93	7.83	46.2	47.07	7.26	24.4	64.3	39.9	-0.22	-0.02	1.02
Social Capital	59	52.74	6.45	52.5	52.81	4.89	31.4	66.8	35.4	-0.34	1.13	0.84
Investment Enviroment	59	52.99	7.07	52.6	53.27	9.34	36.5	67.9	31.4	-0.29	-0.64	0.92
Enterprise Conditions	59	53.61	6.88	54.4	53.74	6.08	33.3	66.5	33.2	-0.31	0.20	0.90
Market Access and Infrastructure	59	55.40	8.26	57.1	55.55	8.75	37	73.3	36.3	-0.27	-0.64	1.08
Economic Quality	59	48.01	7.38	48.2	47.80	6.67	31.3	66.6	35.3	0.30	-0.17	0.96
Living Conditions	59	72.68	9.51	75	73.45	8.15	40.5	87.8	47.3	-0.95	0.78	1.24
Health	59	71.78	5.03	72.5	72.15	3.85	56.6	82.8	26.2	-0.83	0.85	0.65
Education	59	61.09	9.83	62.1	61.22	9.64	34.9	82.8	47.9	-0.26	-0.10	1.28
Natural Environment	59	52.59	6.22	52.7	52.34	6.38	40.3	67.8	27.5	0.26	-0.41	0.81

Refah Seviyesi Yüksek Olan Ülkeler Kümesi

Kümede bulunan ülkelerin birkaçı şunlardır. Danimarka, Finlandiya, Fransa, Almanya, Japonya, ABD ve Malezya.

Tablo 2.5: Refah Seviyesi Yüksek Olan Ülkeler Kümesi Tanımlayıcı İstatistikler

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Safety And Security	51	85.02	7.13	86.6	85.77	5.93	62.3	95.7	33.4	-1.01	0.84	1.00
Personal Freedom	51	76.08	13.34	78.8	77.81	11.86	35.7	94.1	58.4	-1.18	1.26	1.87
Govarnance	51	73.76	12.12	75.9	74.38	14.23	46.8	91.0	44.2	-0.36	-1.04	1.70
Social Capital	51	59.66	8.64	57.7	59.39	8.90	44.0	77.2	33.2	0.27	-0.84	1.21
Investment Enviroment	51	71.34	7.40	72.8	71.69	7.26	50.9	86.4	35.5	-0.44	-0.47	1.04
Enterprise Conditions	51	71.09	8.52	72.1	71.48	9.93	53.8	87.5	33.7	-0.35	-0.98	1.19
Market Access and Infrastructure	51	75.51	6.37	75.8	75.87	5.93	61.4	88	26.6	-0.40	-0.51	0.89
Economic Quality	51	65.20	8.02	64.7	65.67	8.90	45.5	79.3	33.8	-0.42	-0.50	1.12
Living Conditions	51	89.23	4.95	91	89.82	3.85	76.1	96.2	20.1	-0.95	-0.06	0.69
Health	51	79.01	3.40	79.1	79.03	3.41	70.6	86.6	16.0	-0.11	-0.28	0.48
Education	51	79.90	6.91	81.9	80.53	5.63	59.1	91.3	32.2	-0.80	0.04	0.97
Natural Environment	51	64.34	7.20	64.8	64.50	6.97	50.3	78	27.7	-0.17	-0.87	1.01



Şekil 2.20: Kümelerin Boxplot Karşılaştırması

Hiyerarşik kümeleme yöntemi ile elde edilen Refah Seviyesi Düşük Olan Ülkeler, Refah Seviyesi Orta Olan Ülkeler ve Refah Seviyesi Yüksek Olan Ülkeler kümelerinin kutu grafiği yukarıda verilmiştir.

- Emniyet ve Güvenlik açısından Refah Seviyesi Düşük Olan Ülkeler diğer ülkelere göre yüksek standart sapmaya ve sola çarpık bir dağılıma sahiptir. Refah Seviyesi Yüksek Olan Ülkeler Ülkeler ve Refah Seviyesi Orta Olan Ülkeler ile arasında büyük bir fark olduğu gözlemlenmiştir.

- Kişisel Özgürlük açısından Refah Seviyesi Düşük Olan Ülkeler ile Refah Seviyesi Orta Olan Ülkelerin ortalamaları arasında büyük bir fark olsa da dağılımlarının çarpıklığı sebebiyle birbirlerine yakın olan değerlere sahip olduğu gözlemlenmiştir.

- Yönetim açısından Refah Seviyesi Yüksek Olan Ülkelerin değişkenliğinin yüksek olduğu gözlemlenmektedir. Kümeler arasında büyük farklar olduğu gözlemlenmiştir.

- Sosyal Sermaye açısından Refah Seviyesi Orta Olan Ülkelerin değişkenliğinin az olduğu ve sapan değerlerin olduğu gözlemlenmiştir.

- Yatırım Ortamı açısından üç kümenin de dağılımları birbirine benzerdir. Refah Seviyesi Orta Olan Ülkeler Refah Seviyesi Düşük Olan Ülkelerden, Refah Seviyesi Yüksek Olan Ülkeler de Refah Seviyesi Orta Olan Ülkelerden daha yüksek değerler almıştır.

- Kurumsal Koşullar açısından bakıldığında Refah Seviyesi Yüksek Olan Ülkeler ile Refah Seviyesi Düşük Olan Ülkelerin değişkenlik açısından birbirine benzediği Refah Seviyesi Orta Olan Ülkelerin değişkenliğinin diğerlerine oranla daha düşük olduğu gözlemlenmiştir.

- Pazar Erişimi ve Altyapı açısından bakıldığında kümeler arasında bariz farkların olduğu ve Refah Seviyesi Düşük Olan Ülkeler kümesinde yüksek değer almış aykırı gözlem olduğu gözlemlenmiştir.

- Ekonomik Kalite açısından bakıldığında Refah Seviyesi Düşük Olan Ülkeler ile Refah Seviyesi Orta Olan Ülkeler arasındaki farkın az olduğu lakin Refah Seviyesi Yüksek Olan Ülkeler kümesinin diğer kümeler ile aralarında fark olduğu gözlemlenmiştir.

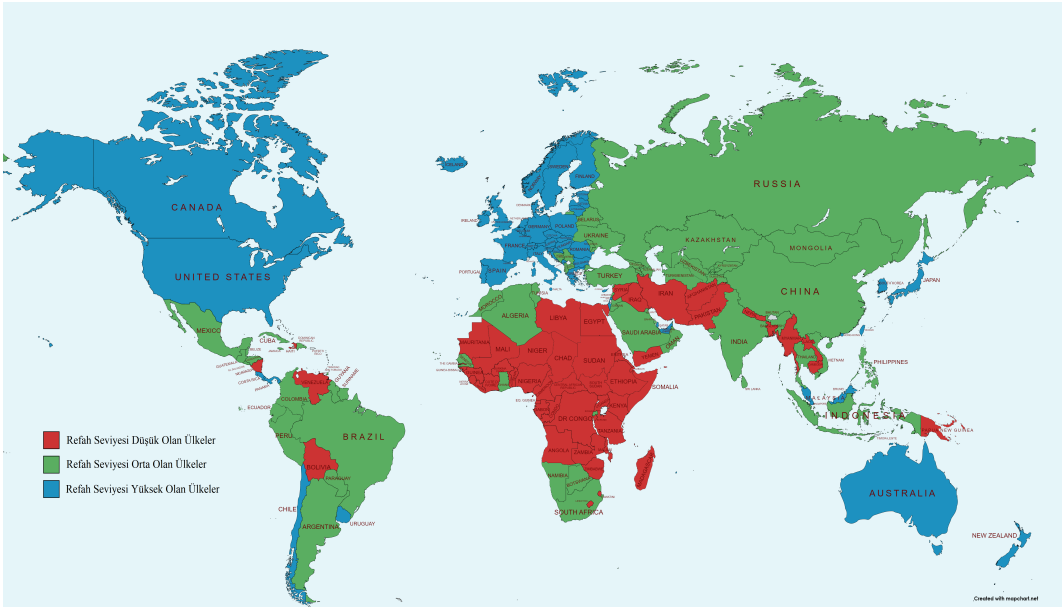
- Yaşam Koşulları açısından bakıldığında kümelerin kendi aralarında refah düzeyine göre sıralandığı ve kümelerin refah düzeyi arttıkça yaşam koşullarının da arttığı ve değişkenliğin azaldığı gözlemlenmiştir.

- Sağlık açısından bakıldığında kümelerin kendi aralarında refah düzeyine göre sıralandığı ve kümelerin refah düzeyi arttıkça sağlık düzeyinin de arttığı ve değişkenliği azaldığı gözlemlenmiştir.
- Eğitim açısından bakıldığında Refah Seviyesi Düşük Olan Ülkelerin değişkenliğinin yüksek olduğu gözlemlenmiştir.
- Doğal Çevre açısından bakıldığında Refah Seviyesi Orta Olan Ülkelerin Refah Seviyesi Düşük Olan Ülkeler ile benzer olduğu gözlemlenmiştir.

Bölüm 3

Sonuç

Bu projede, Legatum Enstitü Vakfı'nın sağladığı refah düzeylerinin incelendiği veri seti üzerinde tanımlayıcı istatistikler ve korelasyon matrisinden ulaşılan bilgiler aracılığıyla ilk önce temel bileşenler analizi yapılmıştır. Oluşturulan iki bileşen Ekonomi ve Yönetim ve Yaşam Kalitesi olarak adlandırılmıştır. Veri birimlerinin Öklid ve Manhattan uzaklıkları hesaplanmıştır. Sonrasında sırasıyla dirsek, Silhouette ve GAP istatistiği yöntemleri kullanılarak üç olan en uygun küme sayısı belirlenmiştir. Ardından k ortalamalar, k medoids ve hiyerarşik kümeleme algoritmaları veriye uygulanıp küme doğrulama istatistikleri yöntemlerinden Silhouette katsayısı, Dunn indeksi ve bağlantılılık hesaplanarak kümeleme algoritmaları karşılaştırılmıştır. Bunun sonucunda en uygun olan hiyerarşik kümeleme kullanılarak ülkeler Refah Seviyesi Düşük Olan Ülkeler, Refah Seviyesi Orta Olan Ülkeler ve Refah Seviyesi Yüksek Olan Ülkeler olarak üç ayrı kümeye ayrılmıştır. Şekil 3.1 kümelenen ülkelerin dünya haritası üzerinde hangi kümeye atandığı gösterilmiştir.



Şekil 3.1: Kümelerin Dünya Haritası Üzerinde Gösterimi

Yapılan kümeleme sonucunda, Refah Seviyesi Yüksek Olan Ülkelerin diğer ülkelere kıyasla daha iyi değerlere sahip olduğu gözlemlenmiştir. Kuzey Amerika, Batı Avrupanın büyük bir çoğunluğunun Refah Seviyesi Yüksek Olan Ülkeler kümesine atandığı gözlemlenmiştir. Afrika kıtasının büyük bi çoğunluğu Refah Seviyesi Düşük Olan Ülkeler kümesine dahil olsa da Gana, Fas ve Senegal gibi bazı ülkelerin Refah Seviyesi Orta Olan Ülkeler kümesine atandığı gözlemlenmiştir. Asya kıtasının genelinin Refah Seviyesi Düşük Olan Ülkeler ve Refah Seviyesi Orta Olan Ülkeler kümelerine dahil olsa da Japonya, Güney Kore, Malezya ve Katar gibi bazı ülkelerin Refah Seviyesi Yüksek Olan Ülkeler kümesine atandığı gözlemlenmiştir.

Türkiye Asya komşuları Suriye, Irak, İran gibi Refah Seviyesi Düşük Olan Ülkeler, Avrupa komşuları Yunanistan, Bulgaristan gibi Refah Seviyesi Yüksek Olan Ülkeler kümesine dahil olmayıp Azerbeycan gibi Refah Seviyesi Orta Olan Ülkeler Kümesine dahil olmuştur.

Kaynaklar

2021LegatumProsperityIndex™ (2021). The legatum prosperity index™. URL `http://www.prosperity.com`

Chao, W.-L. (2011). Machine learning tutorial. *Digital Image and Signal Processing*.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*, vol. 1. Sthda.

Özdamar, K. (2004). Paket programlar ile istatistiksel veri analizi (çok değişkenli analizler). *Kaan Kitabevi, Eskişehir*, *574*.

Ünlükaplan, Y. (2008). Çok değişkenli istatistiksel yöntemlerin peyzaj ekolojisi araştırmalarında kullanımı. *Yayınlanmamış Doktora Tezi, Adana: Çukurova Üniversitesi Fen Bilimleri Enstitüsü*.

Wierzchoń, S. T., & Kłopotek, M. A. (2018). *Modern algorithms of cluster analysis*, vol. 34. Springer.

Ek A
Kodlar

Gerekli Paketler

```
library(lattice)
library(reshape2)
library(dplyr)
library(fpc)
library(NbClust)
library(GGally)
library(ggplot2)
library(tidyverse)
library(psych)
library(pastecs)
library(philentropy)
library(usedist)
library(ggpubr)
library(factoextra)
library(cluster)
library(FactoMineR)
library(corrplot)
library(clValid)
```

Veri Girişi İçin Gerekli Kodlar

```
df <- read.csv("BP_Veri_Seti_2021.csv")
View(df)
df_num<-data.frame(df[3:14])
Country_Names=df[,1]
rownames(df_num)=Country_Names[as.integer(rownames(df_num))]
View(df_num)
```

Tanımlayıcı İstatistikler İçin Gerekli Kodlar

```
describe(df_num)
corrplot(cor(df_num))
ggcorr(df_num, method = c("everything", "pearson"))
ggcorr(df_num, label = TRUE,hjust = 1, label_size = 3,
label_round = 2, label_alpha = TRUE)
pairs(df_num)
ggpairs(df_num)
boxplot(df_num)
boxplot(scale(df_num))
```

Temel Bileşenler Analizi İçin Gerekli Kodlar

```
KMO(df_num)
bartlett.test(df_num)
data.pca <- prcomp(df_num, center = TRUE, scale. = TRUE)
summary(data.pca)
sqrt(data.eigen$values)
data.pca$sdev
data.pca$rotation
data.pca$center
data.pca$scale
data.pca$x
data.pca$x [1:2,]
predict(data.pca)[1:2,]
(scale(data)%*%data.pca$rotation)[1:2,]
fviz_eig(data.pca)
```

```
# Eigenvalues
eig.val <- get_eigenvalue(data.pca)
eig.val

# Results for Variables
res.var <- get_pca_var(data.pca)
res.var$coord # Coordinates
res.var$contrib # Contributions to the PCs
res.var$cos2 # Quality of representation
data.pca
sum(res.var$contrib[,2])

##Variable contributions to the principal axes:
# Contributions of variables to PC1
fviz_contrib(data.pca, choice = "var", axes = 1, top = 12)
# Contributions of variables to PC2
fviz_contrib(data.pca, choice = "var", axes = 2, top = 12)

# Results for individuals
res.ind <- get_pca_ind(data.pca)
res.ind$coord # Coordinates
res.ind$contrib # Contributions to the PCs
fviz_pca_var(data.pca, axes = c(1, 2), col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
fviz_pca_ind(data.pca,
col.ind = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
df_numpca<- predict(data.pca)[,1:2]
df_num<-
res.ind$cos2
describe(df_numpca)
```

Uzaklık Hesaplamaları İçin Gerekli Kodlar

```
df_num.matrix <- (as.matrix(df_num.scaled))

##EUCLIDEAN##
dist.eucl <- dist(df_num, method = "euclidean" , order("FALSE"))
round(as.matrix(dist.eucl))
t(dist.eucl)
view(round(as.matrix(dist.eucl)))
fviz_dist(dist.eucl)

##MANHATTAN##
dist.manh <- dist(df_num, method = "manhattan")
dist.manh.matrix <- round(as.matrix(dist.manh))
view(dist.manh.matrix)
t(dist.manh)
fviz_dist(dist.manh)

##EUCLIDEAN.tr#
df_num.t <- scale(df_num[150:160,])
dist.eucl <- dist(df_num.t, method = "euclidean")
rounded_dist<-round(as.matrix(dist.eucl))
distname<-dist_setNames(dist.eucl, df[150:160,1])
fviz_dist(distname,show_labels = TRUE)

##MANHATTAN.tr##
dist.manh <- dist(df_num[150:160,], method = "manhattan")
dist.manh.matrix <- round(as.matrix(dist.manh))
distname<-dist_setNames(dist.manh.matrix, df[150:160,1])
fviz_dist(distname,show_labels = TRUE)
fviz_dist(dist.manh)
```

K-ortalamalar Hesaplaması için Gerekli Olan Kodlar

```
fviz_nbclust((df_numpca),kmeans,method = "wss")
km.res<-kmeans((df_numpca),3,nstart = 25)
print(km.res)
km.clusters<-km.res$cluster
fviz_cluster(list(data=(df_numpca),cluster=km.clusters))


for (i in 1:3){
        a=data.frame()
        a=(df_numpca[which(km.clusters==i),])
        assign(paste("Clusterpca",i,sep="_"),a)

}
```

K-Medoids Hesaplaması için Gerekli Olan Kodlar

```
b <- pam(df_numpca,3,metric="euclidean", stand=TRUE)

fviz_cluster(list(data=scale(df_numpca),cluster=b$cluster))

for (i in 1:3){
        a=data.frame()
        a=(df_numpca[which(b$clustering==i),])
        assign(paste("K_MedoidsCluster",i,sep=""),a)
}
```

Alternatif Küme Sayısı Belirlemek için Gerekli Olan Kodlar

```
fviz_nbclust(scale(df_numpca),kmeans,method = "silhouette")

fviz_nbclust(scale(df_numpca),kmeans,method = "gap_stat")
```

Kümelerin Tanımlayıcı İstatistiklerini Hesaplamak İçin Gerekli Olan Kodlar

```
boxplot(scale(Clusterpca_1))
boxplot(scale(Clusterpca_2))
boxplot(scale(Clusterpca_3))
boxplot(PC1~df_numpca$km.clusters)

describe(Clusterpca_1)
describe(Clusterpca_2)
describe(Clusterpca_3)
```

Hiyerarşik Kümeleme İçin Gerekli Olan Kodlar

```
grp <- cutree(hc.res, k = 3)
fviz_dend(hc.res, k = 3,
cex = 0.5,
k_colors = c("blue", "green", "red", "black"),
color_labels_by_k = TRUE,
rect = TRUE
)
fviz_cluster(list(data = df_numpca, cluster = grp),
palette = c("red", "green", "blue", "black"),
ellipse.type = "convex",
repel = TRUE,
show.clust.cent = FALSE, ggtheme = theme_minimal())

fviz_cluster(hc.res, geom = "point",
palette = "jco", ggtheme = theme_minimal())
hc.res <- eclust((df_numpca), "hclust", k = 3,
hc_metric = "euclidean",hc_method = "ward.D2", graph = FALSE)
fviz_dend(hc.res, show_labels = FALSE,
palette = "jco", as.ggplot = TRUE)
for (i in 1:3){
        a=data.frame()
        a=(df_numpca[which(hc.res$cluster==i),])
        assign(paste("H_Clust",i,sep=""),a)        }
```

En Uygun Yöntemin Belirlenmesi İçin Gerekli Olan Kodlar

```
#silhouette
fviz_silhouette(hc.res, palette = "jco",
ggtheme = theme_classic())

silinfo <- hc.res$silinfo
silinfo
sil <- km.res$silinfo$widths[, 1:3]
neg_sil_index <- which(sil[, "sil_width"] < 0)
sil[neg_sil_index, , drop = FALSE]

#dunn index
km_stats <- cluster.stats(dist(df_numpca), hc.res$cluster)
km_stats$dunn # Dun index



clmethods <- c("hierarchical","kmeans","pam")
intern <- clValid(df_numpca, nClust = 3,
clMethods = clmethods, validation = "internal")
summary(intern) # En Uygun Yöntemin belirlenmesi



#30 indices for choosing the best number of clusters
nb <- NbClust(df_numpca, distance = "euclidean", min.nc = 2,
max.nc = 9, method = "kmeans")
fviz_nbclust(nb)

nb$All.index
nb$All.CriticalValues
nb$Best.nc
nb$Best.partition
```

Kümeleme Sonuçları İçin Gerekli Kodlar

```
df_num$Clusters <- hc.res$cluster

describe(H_Clust1)
boxplot(H_Clust1)
describe(H_Clust2)
boxplot(H_Clust2)
describe(H_Clust3)
boxplot(H_Clust3)



Cluster1 <- df_num[which(df_num[,13]==1),1:12]
Cluster2 <- df_num[which(df_num[,13]==2),1:12]
Cluster3 <- df_num[which(df_num[,13]==3),1:12]

describe(Cluster1)
describe(Cluster2)
describe(Cluster3)

boxplot(Cluster1)
boxplot(Cluster2)
boxplot(Cluster3)

df_nummelt <- melt(df_num, id = "Clusters")

ggplot(df_nummelt, aes(x = variable, y = value,
fill=factor(Clusters))) +  # ggplot function
geom_boxplot()

bwplot(value ~ variable, df_nummelt)
```