BÜŞRA ÇALIŞKAN

20457164115

Dünya Genelindeki Kadınların Pist Rekorları Üzerine SPPSS'de Temel Bileşen Analizi

UMHB515-Uygulamalı Çok Değişkenli Analiz

Danışman: Dr.Öğr.Üyesi Beste Hamiye BEYAZTAŞ

İÇİNDEKİLER

1.GİRİŞ

Projenin Amacı:

Hedef; veri setinde sütunların sayısı indirgenerek mevcut sütundaki değişkenlerin kombinasyonundan oluşan yeni değişkenler yaratılarak sütun sayısı azaltılmaktır. Böylece veri setindeki tüm özellikler bir şekilde hala mevcut ancak değişken sayısı azaltılmış olacaktır.

Veri seti[1] Tanıtımı:

- 7 farklı etkinlikte 55 ülkeyi temsil eden Kadınların Uluslararası kendi Pist Rekor Kayıtlarını içeren bir CSV dosyasıdır.
- İlk satır, ülke isimleri temsil eder.Diğer satırlar(X1'den….. X7'ye kadarki sutunlar) kadınların aynı yarışmada aynı koşu mesafesindeki kendi rekorlarını kaç saniye veya dakikada kırdıklarını simgeler.

➢ Verisetin ilk beş satırı:

	COUNTRY	X1	X2	X3	X4	X5	X6	X7
0	Argentina	11.61	22.94	54.50	2.15	4.43	9.79	178.52
1	Australia	11.20	22.35	51.80	1.98	4.13	9.08	152.37
2	Austria	11.43	23.09	50.62	1.99	4.22	9.34	159.37
3	Belgium	11.41	23.04	52.00	2.00	4.14	8.88	157.85
4	Bermuda	11.46	23.05	53.30	2.16	4.58	9.81	169.98

➢ X1'den….. X7'ye kadarki sütunların koşu mesafesini ne kadar zamanda(s=saniye, min=dakika) aldığını temsil ettiği değerler:

X1: 100m (s)
X2: 200m (s)
X3: 400m (s)
X4: 800m (min)
X5: 1500m (min)
X6: 3000m (min)
X7: Marathon (min)

2.METOT

A-Veri seti üzerinden korelasyon kontrolü ve standartlaştırılma gerekli mi kontrolü yapma.

B-Eğer gerekiyorsa veriler 0-1 aralığına çekilerek standartlaştırılacaktır. Böylece programımızın performansı ve verimi artacaktır.

Verileri Standartlaştırma formülü:

İ.değeri, bulunduğu satırdaki değerlerin ortalamasından çıkarılarak varyansının karaköküne bölünür.

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, \; i = 1,2,\ldots, p$$

C-Standartlaştırılmış Veri seti üzerinden Faktör Analizi yapılıp yapılamayacağı kontolü:

1. Korelasyon matrisine bakmak
2. Bartlett Küresellik Testi ile Korelasyon matrisinin birim matrisine eşit olup olmadığını test etmek
3. Kaiser-Meyer-Olkin(KMO) ile de örneklem ölçümümüzün yeterli olup olmadığına bakmak

D-Temel bileşen analizi yapılacaktır. Boyut indirgenerek(yüksek korelasyona sahip değişkenler birleştirilecek ve en yüksek varyanslara sahip değişkenler oluşturulacak) birbirinden bağımsız **temel bileşenler** oluşturulacaktır.

Temel bileşenler,varyansları mümkün olduğu kadar büyük olan birbirinden bağımsız *Y1,Y2,…,Yp* lerin doğrusal kombinasyonlarıdır.

Korelasyon Formülü:

$\boldsymbol{Y_i}$ kovaryans matrisi kullanılarak elde edilen temel bileşen

$\boldsymbol{X_k}$ k. değişken

$\boldsymbol{e_{ik}}$, *k.* değişkenin *i.*temel bileşene katkısıdır yani **özvektördür** ve Y_i ve X_k arasındaki korelasyon katsayısına orantılıdır.
$\boldsymbol{\Lambda_i}$ = İ. özdeğer

i. temel bileşen Y_j ile k. değişken X_k arasındaki korelasyon katsayısı

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\ldots,p$$

Temel Değişkenlerin Hesaplanma Formülü:

Temel Değişkenler = k. değişken ile i.temel bileşen arasındaki korelasyon katsayısı($\rho_{Xi,XK}$) X i. standartize Edilmiş Veri(Zi)

3-UYGULAMA

A-)Pearson Correlation'u:

Correlations

		X1	X2	X3	X4	X5	X6	X7
X1	Pearson Correlation	1	,887**	,710**	,727**	,586**	,650**	,689**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000
	N	55	55	55	55	55	55	55
X2	Pearson Correlation	,887**	1	,710**	,717**	,559**	,633**	,670**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000
	N	55	55	55	55	55	55	55
X3	Pearson Correlation	,710**	,710**	1	,772**	,500**	,630**	,583**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000
	N	55	55	55	55	55	55	55
X4	Pearson Correlation	,727**	,717**	,772**	1	,665**	,785**	,780**
	Sig. (2-tailed)	,000	,000	,000		,000	,000	,000
	N	55	55	55	55	55	55	55
X5	Pearson Correlation	,586**	,559**	,500**	,665**	1	,587**	,636**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,000
	N	55	55	55	55	55	55	55
X6	Pearson Correlation	,650**	,633**	,630**	,785**	,587**	1	,708**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,000
	N	55	55	55	55	55	55	55
X7	Pearson Correlation	,689**	,670**	,583**	,780**	,636**	,708**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	
	N	55	55	55	55	55	55	55

**. Correlation is significant at the 0.01 level (2-tailed).

Değişkenler arasında %99 güven aralığında yüksek korelasyon vardır.

Değişkenlerin ortalamalarının diğer değişkenlerin ortalamalarına bakıp standardize etmelimiyiz kararı buradan veriliyor.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X1	55	10,79	12,90	11,6200	,44968
X2	55	21,52	27,10	23,5791	1,14717
X3	55	47,99	63,60	53,6105	2,99009
X4	55	1,89	2,33	2,0764	,10822
X5	55	3,87	5,81	4,3982	,39798
X6	55	3,92	13,04	9,3567	1,10986
X7	55	142,72	306,00	173,2531	30,46171
Valid N (listwise)	55				

Değişkenlerin ortlamalarına bakıldığında özellikle X7 değişkeni ile X4 değişkeni arasında ciddi ortalama farkı var, bu yüzden verileri standardize etmeliyiz

B-)Standartlaştırılmış Veriler:

ZX1	ZX2	ZX3	ZX4	ZX5	ZX6	ZX7
-,02224	-,55710	,29747	,68041	,07995	,39039	,17290
-,93399	-1,07141	-,60552	-,89041	-,67386	-,24934	-,68555
-,42252	-,42635	-1,00015	-,79801	-,44772	-,01507	-,45576
-,46700	-,46993	-,53863	-,70561	-,64874	-,42954	-,50565
-,35581	-,46121	-,10386	,77281	,45686	,40841	-,10745
-,68937	-,35661	-,27108	,21840	,23071	,37237	-,14783
1,15637	,77662	,46469	,95761	,13020	,13810	,58325
-1,37875	-1,15858	-1,00684	-,70561	-,84975	-,49261	-,78141
,84504	-1,79493	,43124	-,24360	-,42259	,01196	-,06149
,73385	,72431	,45465	,03360	-,17132	-,04210	-,15669
,04448	,36691	,11724	,31080	,12107	,09305	,25715

C.Faktör Analizi

C.1.Standartlaştırılmış verilerin korelasyon matrisi:

Correlation Matrix								
		Zscore(X1)	Zscore(X2)	Zscore(X3)	Zscore(X4)	Zscore(X5)	Zscore(X6)	Zscore(X7)
Correlation	Zscore(X1)	1,000	,887	,710	,727	,586	,650	,689
	Zscore(X2)	,887	1,000	,710	,717	,559	,633	,670
	Zscore(X3)	,710	,710	1,000	,772	,500	,630	,583
	Zscore(X4)	,727	,717	,772	1,000	,665	,785	,780
	Zscore(X5)	,586	,559	,500	,665	1,000	,587	,636
	Zscore(X6)	,650	,633	,630	,785	,587	1,000	,708
	Zscore(X7)	,689	,670	,583	,780	,636	,708	1,000

Eğer değişkenler arasında yüksek bir ilişki(korelasyon) varsa, aynı faktör altında toplanmalıdır.Korelasyon matrisinde görüldüğü üzere birim matris değildir ve değişkenler diğer değişkenler ile bağımlıdır.(ilk verilerimiz ile korelasyon matrisi aynıdır)

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,890
Bartlett's Test of Sphericity	Approx. Chi-Square	314,016
	df	21
	Sig.	,000

- ✓ Yeterli örneklem büyüklüğü Kaiser-Meyer-Olkin (KMO) ile test edilir. Eğer Kaiser-Meyer-Olkin (KMO) testi sonucu elde ettiğimiz değerimiz 0,5 ten büyük çıkarsa örneklem büyüklüğümüz faktör analizi için yeterlidir.**Görüldüğü üzere**

KMO 0.890 > 0.5 olduğundan dolayı örneklem ölçümü yeterlidir. Faktör analizi yapılabilir.

- ✓ Bununla birlikte Bartlett'in küresellik testinde de p-değeri(sig.) 0,000 < 0,05 olduğundan H0 hipotezi red edilir.Örneklemin alındığı kitleye ilişkin değişkenler arasındaki kitle korelasyon matrisi ρ birim matrisden farklıdır. Yani bazı değişkenler arasındaki korelasyonlar anlamlıdır. **Korelasyonları yüksek olan değişkenler bir çatı altında toplanmalı ve veri setine TBA'nın yapılması gerekliliğini gösterir.**

Communalities		
	Initial	**Extraction**
Zscore(X1)	**1,000**	**,784**
Zscore(X2)	**1,000**	**,763**
Zscore(X3)	**1,000**	**,682**
Zscore(X4)	**1,000**	**,841**
Zscore(X5)	**1,000**	**,568**
Zscore(X6)	**1,000**	**,705**
Zscore(X7)	**1,000**	**,726**
Extraction Method: Principal Component Analysis.		

Bu tablo, değişkenlerin temel bileşen tarafından açıklanma oranıdır. Örneğin elde edilen temel bileşen birinci değişkenin %78,4'ünü açıklamaktadır. Tabloda çıkarma değeri 0,5'ten küçük olan değerlere karşılık gelen değişkenlerin çıkarılarak analizin yapılması daha iyi sonuç verir. Bizim verilerimizde hiçbiri 0.5 altı değildir. Tüm sütunlar kullanılarak analiz yapılmalıdır.

D- **Temel Bileşen Sayısının Belirlenmesi**

Standartlaştırılmış verilere göre Kovaryans matris hesaplandıktan sonra özvektörler ve özdeğerler bu matristen elde edilebilir. Her boyut için özvektörleri ve karşılık gelen özdeğerleri hesaplanır.Özdeğerler yüksek değerden düşük değere doğru sıralanır. Amaç bileşenleri veriyi temsil etme oranına göre sıralamaktır. Böylelikle en önemli bileşenden en az önemli bileşene doğru bir sıralama yapılır.

- Öz değerlerin 1'den büyük olanlara bakarak faktör sayısına karar verilir.Bir faktör tarafından varyansın açıklanma yüzdesine bakarak da karar verilebilir.
- Her bir özdeğerin toplam özdeğerlere bölünmesi ile temel bileşenlerin toplam varyansı açıklama yüzdeleri elde edilir.

Özdeğerlere göre toplam Varyansların Açıklanma tablosu:

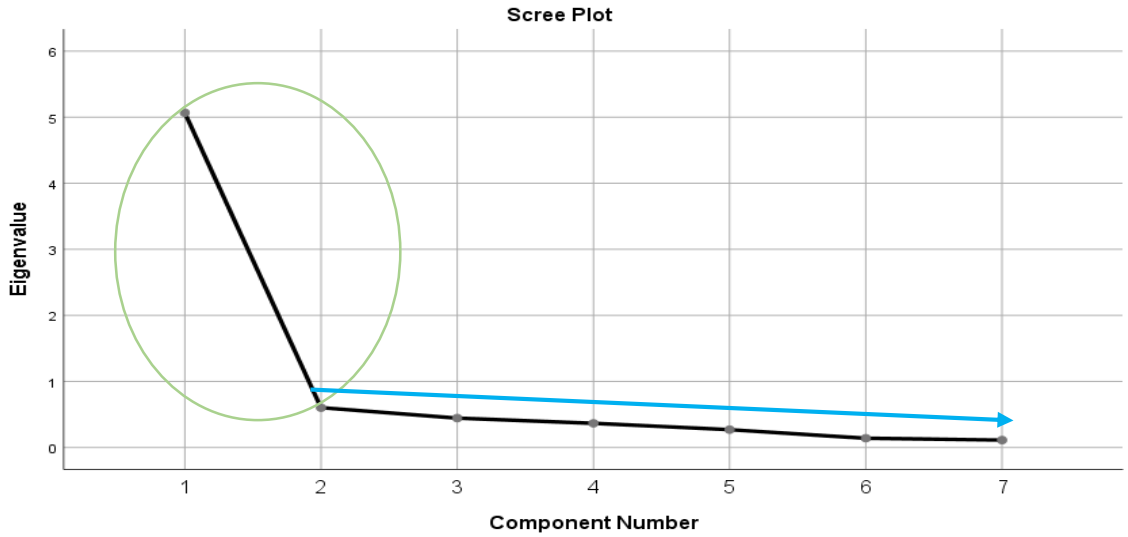
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,068	72,394	72,394	5,068	72,394	72,394
2	,602	8,600	80,995			
3	,444	6,347	87,342			
4	,366	5,227	92,569			
5	,269	3,847	96,416			
6	,139	1,990	98,406			
7	,112	1,594	100,000			

Extraction Method: Principal Component Analysis.

- Öz değerlerden 1 tanesi 1' den büyük olduğu için 1 faktörlü çalışabilirim yorumu yapılabilir.
- Birinci Temel bileşen toplam varyansın %72,394'ünü açıklamaktadır.
- **Özdeğerler:** 5,068 0,602 0,444 0,366 0,269 0,139 0,112
- **Açıklanan varyans oranı: %**72,394 %8,600 %6,347 %5,227 %3,847 %1,990 %1,594
- **Birikimli açıklanan varyans oranı:**
  72,394 %80,995 %87,342 %92,569 %96,416 %98,406 %100,000

- Yedi tane temel bileşen olduğundan p=7'dir. Ancak uygulamada bu temel bileşen çıkmıştır. Ancak işlemin kontrolü için temel bileşen sayısının belirlenmesine ilişkin bazı kriterler vardır. Bunlardan biri bir temel bileşen, toplam değişimin en az 2/3'nü açıklaması gerekir. Buradan:
- $\lambda_l / \lambda_{toplam}$ =5,068/ (5,068+ 0,602+0,444+0,366+0,269+0,139+0,112)=**0.724>0.67**
- Toplam varyansın **%72,394**'ü ilk temel bileşen yardımıyla açıklamaktadır. Bu oran 2/3 den büyük olma ölçütünü sağladığından önemli temel bileşen sayısı bir olarak alınır.



Yukarıdaki Scree Plot'a göre de uygun temel bileşen sayısına karar verilebilinir.

- Dirseğin **ÜSTÜNDE** kalan öz değer adeti kadar faktörümüz vardır.
- Grafiğe bakıldığında, belli bir yere kadar bir düşüş var ve daha sonra değişkenlik oldukça azalarak sanki sabit noktaya ulaşıyormuş gibi olduğunu görüyoruz. Bu değişkenliğin yok denilecek kadar azaldığı noktaya biz dirsek diyoruz ve dirseği kapatıp, dirseğin üstünde kalan öz değer adeti kadar da faktör sayımız vardır diyoruz. Burada 1 adet faktör sonucuna varılır.

Yeni bağımsız değişkenlerimiz ile standartlaştırılmış verilerimizin Korelasyon matrisi(benzerlik kontolü)

Component Matrix[a]	
	Component
	1
Zscore(X1)	,886
Zscore(X2)	,873
Zscore(X3)	,826
Zscore(X4)	,917
Zscore(X5)	,753
Zscore(X6)	,839
Zscore(X7)	,852
Extraction Method: Principal Component Analysis.	
a. 1 components extracted.	

Bu matriste değişkenlerle bileşenlerin ilişkileri yer alır. Hangi değişkenin hangi faktöre yüklendiğine karar veririz.Faktör olarak 1 temel bileşen vardır.

1. değişkenin temel bileşendeki yükü: 0.086

2. değişkenin temel bileşendeki yükü: 0.873

3. değişkenin temel bileşendeki yükü: 0.826

4. değişkenin temel bileşendeki yükü: 0.917

5. değişkenin temel bileşendeki yükü: 0.753

6. değişkenin temel bileşendeki yükü: 0.839

7. değişkenin temel bileşendeki yükü: 0.852

(Tek temel bileşen olduğu için rotasyon matrisi oluşmadı)

Temel bileşenin katsayı skorları

Component Score Coefficient Matrix	
	Component
	1
Zscore(X1)	,175
Zscore(X2)	,172
Zscore(X3)	,163
Zscore(X4)	,181
Zscore(X5)	,149
Zscore(X6)	,166
Zscore(X7)	,168
Extraction Method: Principal Component Analysis. Component Scores.	

Bu tablo, tüm verilerin küçük bir kayıp ile temsil eden Temel bileşenin hesaplanmasında kullanılan değişkenler ile.temel bileşen arasındaki korelasyon katsayısılarıdır.

Temel bileşenlerin, değişkenlerle olan korelasyonları genellikle bileşenleri yorumlamaya yardımcı olsada, korelasyonlar sadece tek başına bir değişkeninin temel bileşenine olan tek değişkenli katkısını ölçer. Bu, diğer standartlaştırılmış verilerin varlığı durumunda, Standartlaştırılmış verilerin temel bileşenindeki önemini göstermez. Bu nedenle, bazı istatistikçiler korelasyona değilde $\mathbf{e_{ik}}$ özvektör ile orantılı olan korelasyon katsayılarına bakarak bileşen yorumlanması gerektiğini belirtirler.

Temel bileşenin hesaplanmasında bütün standartlaştırılmış verilerin korelasyon katsayıları hemen hemen yakındır. İlk üç önem düzeyinde olan sırasıyla X4, X2 ve X1'dir.

Temel Bileşen = Y_1 = $0.175*Z_1+0.172*Z_2+0.163*Z_3+0.181*Z_4+0.149*Z_5+0.166*Z_6+0.168*Z_7$

Sonuç olarak bir temel bileşenin uygun olduğunu kabul edersek, bu temel bileşenlere ilişkin 55 bayan atletin temel bileşen değerleri (skorları)

FAC1_1
,17733
-,86439
-,60036
-,63065
,09880
-,11543
,72523
-1,07162
-,20669
,27444
,04681
2,52643
1,01921
-2,05291
,08344
,91968
-,91722
-,87893
-1,44640
-1,22575
-1,16880
,31053
1,28941
-,61586
,41934
,85590
-,45753

FAC1_1
-,08150
-,88000
,34189
,21109
,50594
,24253
,53513
,92536
1,70995
-,04339
-,38998
-,24875
-,58188
1,61268
,63740
-1,12164
-,05729
-,81082
,81583
-,12052
-,77088
-,54762
-,26103
,74043
,69360
-1,40663
-1,43305
3,28912

- ✓ Temel bileşenlerin katsayılarında eksi işaretinin olması, faktörlerin iki kutuplu olduğunu göstermektedir. Yapılan işlemler sonucunda, temel bileşenler analizinin amacına uygun olarak, bileşen sayısı yedi boyutlu ilişkili değişkenden, bir boyutlu ilişkili değişkene indirgenmiştir. Bunlardan en büyük varyansa sahip olan seçilmiştir.
- ✓ **Verilere rotasyon yapılmasına ve dakika cinsîndeki veriler saniyeye çevrilmesine rağmen sonuçlar değişmemiştir.**

4. TARTIŞMA

Birincisi: ''İlk birkaç temel bileşen için temel bileşenin toplam varyansını açıklama oranı %80 veya %90 elde edilirse , fazla bilgi kaybı olmadan temel bileşenler orijinal *p* değişkenin yerini alır.'' hipotezine dayanarak projede beklentilerden biri buydu. Ancak sonuçlarda 1 temel bileşen toplam varyansın %72,394'ünü açıkladı. %80 yakın bir oran ancak tek bileşen olduğu için 7 değişkenden 1 temel bileşene düşürülmesi verinin özelliklerini yansıtmıyor.Burda verilerin temsilinde yüksek bir kayıp var.

İkincisi: Verilerdeki ilk 3 değişkenin cinsi saniye , geriye kalan 4 değişkenin cinsi dakikadır. Sonuçlardan beklenen en az 2 tane temel bileşen çıkmasıydı.Ancak 1 tane temel bileşen bulundu. Muhtemelen '' Korelasyonları yüksek olan değişkenler bir çatı altında toplanmalı'' hipotezine dayanaraktan dakika cinsinden olan değişkenlerdendi içerinde yüksek korelasyon elde edemeyip bir arada toplanamadı(büyük ihtimalle dakika cinsinden veriler içlerinde birbirleriyle anlamlı ilişkileri yok) ve tüm saniye cinsinden olan değişkenlerle daha yüksek korelasyon hesaplandı.

Üçüncüsü: Dakikalar saniyelerle beraber açıklanan tek temel bileşenle ifade edildiği için toplam varyans açıklama oranı da bu yüzden %80 altında kalarak veriyi tanımlamada büyük kayıp verildi.

Özetle;kendi içerinde yüksek korelasyonlu verilerdi.Ancak cinsleri farklı olan değişkenlerin kendi içlerinde anlamlı olamamasından ve programın hepsini tek cins olarak algılayıp işlem yapmasından dolayı beklenen sonuçları elde edilemedi.

5.REFERANSLAR

1. ***https://www.datasciencearth.com/python-ile-temel-bilesenler-analizi-pca/***
2. ***https://www.veribilimiokulu.com/makine-ogrenmesine-cok-degiskenli-istatistiksel-yaklasimlar-temel-bilesenler-analizi/2/***
3. ***https://www.veribilimiokulu.com/makine-ogrenmesine-cok-degiskenli-istatistiksel-yaklasimlar-temel-bilesenler-analizi/***
4. ***http://www.zafercomert.com/Medya/2015_05_21_2_152_9274a3ea.pdf#viewer.action=download***
5. ***https://medium.com/@gulcanogundur/pca-principal-component-analysis-temel-bile%C5%9Fenler-analizi-bf9098751c62***
6. ***https://towardsdatascience.com/factor-analysis-on-women-track-records-data-with-r-and-python-6731a73cd2e0***
7. ***https://medium.com/swlh/machine-learning-guide-principal-component-analysis-pca-on-breast-cancer-dataset-efebec0531d9***
8. ***https://towardsdatascience.com/principal-component-analysis-algorithm-in-real-life-discovering-patterns-in-a-real-estate-dataset-18134c57ffe7***
9. ***https://www.veribilimiokulu.com/faktor-analizi-nedir-nasil-uygulanir/***

- **Veri seti:**
  [1]

  https://drive.google.com/file/d/1S3Ve0pshb42UIcOKaZaBHHJZ7Wbo4Kx/view