

R FONKSİYONUYLA KUTU GRAFIĞİNİN ÇİZİLMESİ ÜZERİNE YENİ BİR UYGULAMA

Yazar Alper Kilic

Gönderim Tarihi: 14-Eyl-2021 11:29AM (UTC+0300)

Gönderim Numarası: 1648143763

Dosya adı: Tam_Metin_R4.docx (74.78K)

Kelime sayısı: 2303

Karakter sayısı: 15338

R FONKSİYONUyla KUTU GRAFİĞİNİN ÇİZİLMESİ ÜZERİNE YENİ BİR UYGULAMA

Verilerin özetlenmesinde tablolar ve hesaplamaların yanında grafikler de sıklıkla kullanılmaktadır. Grafiklerin kullanımı sayesinde, veri kümelerinde fark edilmeyen desenler rahatlıkla belirlenebilmekte ve yorumlanabilmektedir. Özellikle açık kaynak kodlu programlama dillerinin gelişmesiyle birlikte veri görselleştirme aracı olan grafiklerin özelliklerinin de geliştiği görülmektedir. Bu gelişim, verilerin özetlenmesinde sıklıkla kullanılan bazı grafik türlerindeki sakıncaların ortadan kaldırılmasına olanak sağlamaktadır. Görünüşte oldukça sade ve basit olan kutu grafiği, ilk etapta bir verinin en düşük, ilk çeyrek, medyan, üçüncü çeyrek ve en büyük değerlerinden oluşan özetine dayalı olarak dağılımını gösteren çok kullanışlı bir grafik türüdür. Bununla birlikte kutu grafiği verinin simetrik veya çarpık olup olmadığı, verinin nerede yoğunlaştığı ve varsa aykırı değerler hakkında da bilgi vermektedir. Basit görüntüsüne rağmen kutu grafiğinin çok işlevsel bir gösterim olduğu söylenebilir. Kutu grafiği, değişkenlere ilişkin bazı betimsel istatistiklerin izlenmesini sağlamakla birlikte, bu grafikte aykırı değerler olarak belirlenen araştırma birimlerinin isimleri görünmemekte, bu birimler yıldız ve/veya nokta işaretleri ile temsil edilmektedir. Ayrıca, birden çok sürekli değişken için kutu grafikleri çizilmek istendiğinde çizilen kutu grafiklerini tek bir grafik çerçevesinde görmek mümkün olmamaktadır. Bu çalışmada, kutu grafiğindeki bu sakıncaları gidermek amacıyla, kutu grafiği açık kaynak kodlu bir programlama dili olan R'de fonksiyon yazılarak yeniden oluşturulmuştur. Böylece, yazılan yeni R kodu ile veri kümesindeki aykırı değerler doğrudan görülebilmektedir. R kodlarıyla yazılmış olduğumuz kutu grafiği fonksiyonu girdi olarak tüm veriyi kapsayan bir .csv dosyasını, verinin ait olduğu etiketi, sayısal ve faktör verilerinin bulunduğu sütun bilgisini ve ayrıca faktör seviyelerini kullanmaktadır. Bu bilgilere göre kod her bir faktör ve sayısal değer için çoklu kutu grafiği çizmekte, her bir kutu grafiği içinde varsa aykırı değerleri işaretleyerek isimlerini grafik üzerine yazmaktadır. Çalışmada, yazılan R kodu birkaç veri kümesi üzerinde denenmiş, elde edilen bulguların SPSS 21.0'den elde edilen bulgular ile örtüştüğü görülmüştür. Böylece çalışmada, araştırmacılar için daha kullanışlı olduğu düşünülen bir kutu grafiği kod sistemi tanıtılmıştır.

Anahtar Kelimeler: Kutu grafiği, R programlama, R fonksiyonu.

A NEW APPLICATION ON DRAWING BOX GRAPHS WITH THE R FUNCTION

In addition to tables and calculations, graphs are often used to summarize data. Thanks to the use of graphics, unnoticeable patterns in data sets can be easily identified and interpreted. Especially with the development of open source programming languages, it is seen that the features of graphics, which are data visualization tools, have also improved. This development allows to eliminate the drawbacks of some types of graphs that are frequently used in summarizing data. The seemingly plain and simple box chart is a very useful type of chart that initially shows the distribution of a data based on a summary of its lowest, first quartile, median, third quartile and highest values. In addition, the box plot also gives information about whether the data is symmetrical or skewed, where the data is concentrated, and outliers, if any. Despite its simple appearance, the boxplot can be said to be a very functional representation. Although the boxplot allows to monitor some descriptive statistics about the variables, the names of the research units determined as outliers are not visible in this graph, these units are represented by asterisks and/or dots. In addition, when it is desired to draw box plot for more than one continuous variable, it is not possible to see the drawn box plots in a single graphic frame. In this study, in order to eliminate these drawbacks in the boxplot, the boxplot was reconstructed by writing a function in R, an open source programming language, in order to make the names of these outliers appear in the boxplot and to eliminate this drawback. Thus, outliers in the dataset can be seen directly with the new R code written. The boxplot function that we wrote with R codes uses a .csv file that includes all the data, the label to which the data belongs, the column information containing the numerical and factor data, and also the factor levels. According to this information, the code draws a multi-box plot for each factor and numerical value, marks the outliers in each boxplot, and writes their names on the chart. In the study, the written R code was tested on several datasets, and it was seen that the findings were consistent with the findings obtained from SPSS 21.0. Thus, a boxplot code system, which is thought to be more useful for researchers, is introduced in the study.

Keywords: Boxplot, R programming, R function.

1. GİRİŞ

Verilerin özetlenmesinde tablolar, grafikler ve hesaplamalar olmak üzere kabaca üç yöntem kullanılır (Işığışık, 2018). Bu yöntemler arasında grafikler, verilerin görsel olarak özetlemesi bakımından önemli bir yer tutar. Özellikle büyük verinin etkisiyle verilerin görselleştirilmesi daha da önem kazanmıştır. Veri görselleştirme sayesinde büyük veri kümelerindeki ilişkileri belirlemek kolaylaşmıştır (Çelik & Akdamar, 2018). Diğer taraftan,

açık kaynak kodlu programlama dillerinin gelişmesiyle birlikte, veri görselleştirme aracı olan grafiklerin özelliklerinin de geliştiği görülmektedir. Python, R vb. programlama dilleri ile istatistiksel analizlerde kullanılan çok sayıda yeni grafik türleri ortaya çıkmıştır. Bunun yanında sıklıkla kullanılan grafiklerdeki eksikler, programlama dilleri kullanılarak giderilebilmektedir. Söz konusu eksikler, grafiklerin alışlagelmiş yapılarından kaynaklandığı gibi çalışmaya özgü sunum ihtiyaçlarından da (çoklu çizim, renklendirme, boyutlandırma, vb.) kaynaklanabilmektedir.

Verilerin özetlenmesinde sıklıkla kullanılan grafiklerden biri olan kutu grafiği, Excel, SPSS, MINITAB vb. çeşitli programlar yardımıyla oluşturulabilmektedir. Kutu grafiği ile verilerin birinci çeyrek, medyan, üçüncü çeyrek, en küçük ve en büyük değerleri izlenebilmektedir. Diğer taraftan veri setinde aykırı değerler olması durumunda da kutu grafiğinde görülebilmektedir (Gürsaka, 2013).

Kutu grafiği, düşey ekseninde yer alan bir sürekli değişken için yatay ekseninde yer alan farklı kategorilerde çizdirilebilmektedir. Ancak, birden çok sürekli değişken için kutu grafiklerine ihtiyaç duyulduğunda aynı grafik çerçevesinin içerisine kutu grafiklerini yerleştirmek mümkün olmamaktadır. Diğer taraftan, kutu grafiğinde görülen aykırı değerler yıldız ve/veya nokta ile temsil edilmekte, aykırı değerlerin hangi birime ait olduğu bazı programlarda görünmemektedir. Bu çalışmada, kutu grafiğinin kullanımındaki söz konusu eksiklerin giderilmesi amacıyla, R programlama dili kullanılarak kutu grafiği yeniden çizdirilmiştir. Bu çalışma ile önerilen R kodu ile birden fazla sürekli değişken için kutu grafiğinin aynı çerçeve içerisine çizilmesi mümkün olmaktadır. Ayrıca grafiğin üzerindeki aykırı değerlerin hangi birime ait olduğu yazdırılabilmektedir.

2. VERİ VE YÖNTEM

2.1. Çalışmada Kullanılan Veriler

Bu çalışmada kutu grafiği uygulaması ülkelerin lojistik performansları üzerinden gerçekleştirilmiştir. Ülkelerin lojistik performansının göstergesi olarak Dünya Bankası tarafından iki yılda bir yayınlanan Lojistik Performans Endeksi (Logistics Performance Index- LPI) kullanılmıştır. LPI, gümrük (Customs), altyapı (Infrastructure), uluslararası gönderiler (International Shipments), lojistiğin kalitesi (Logistics Competence), takip ve izleme (Tracking&Tracing) ve zamanındalık (Timeliness) olmak üzere altı alt göstergeden oluşmaktadır (World Bank, 2021).

Bu çalışmada kutu grafikleri SPSS ve Excel'de LPI'nın alt boyutlarından olan altyapı değişkeni için R'de ise çoklu çizimi göstermek amacıyla altyapı ve zamanındalık değişkenleri için çizdirilmiştir. Ülkeler, yataksız ekseninde ise insani gelişmişlik performanslarına göre dört sınıfa ayrılmıştır. Bu ayrım Birleşmiş Milletler tarafından her yıl yayınlanan İnsani Gelişmişlik Endeksi'ne (Human Development Index- HDI) göre gerçekleştirilmiştir (UNDP, 2021).

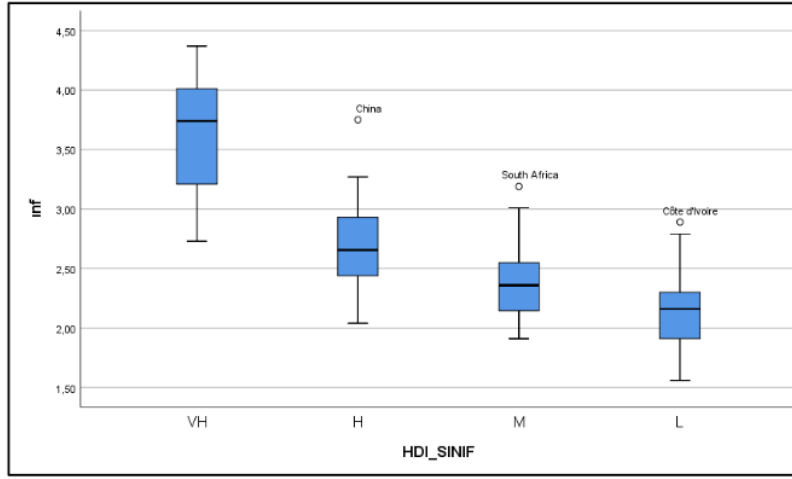
Kutu grafiği 154 ülkenin verileri üzerinden hem SPSS hem Excel'de çizilmiş, daha sonra ise R kodu yazılarak elde edilmiştir.

2.2. SPSS ile Kutu Grafiği

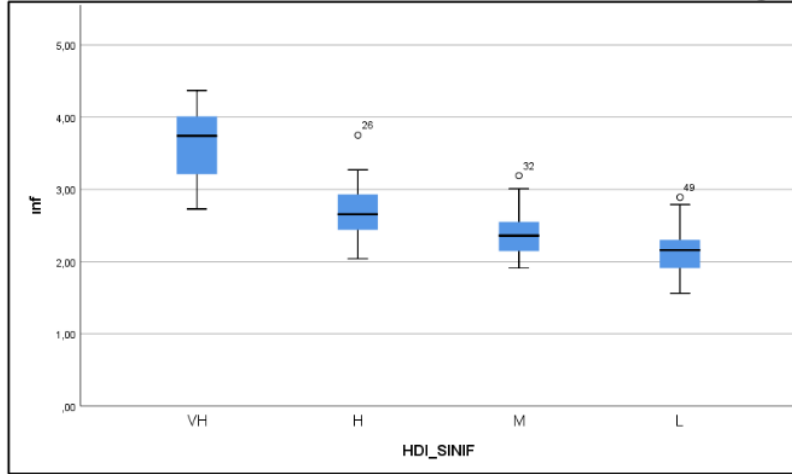
SPSS'de kutu grafiği çizdirmek için Graphs menüsünün alt sekmeleri olan Legacy Dialogs ve Chart Builder olmak üzere iki farklı seçenek kullanılabilir. Şekil 1.'de Legacy Dialogs menüsü kullanılarak çizilen kutu grafiği, Şekil 2.'de ise Chart Builder menüsü kullanarak çizilen kutu grafiği görülmektedir. Grafiklerde düşey ekseninde LPI'nın alt boyutlarından biri olan altyapı (infrastructure- inf) bulunurken, yatay ekseninde ise ülkelerin

HDI'ya göre dört gelişmişlik seviyesi olan Çok yüksek (Very High- VH), Yüksek (High- H), Orta (Medium- M) ve Düşük (Low- L) seviyeleri görülmektedir.

Şekil 1. SPSS'de Legacy Dialogs menüsü kullanılarak elde edilen kutu grafiği.



Şekil 2. SPSS'de Chart Builder menüsü kullanılarak elde edilen kutu grafiği.

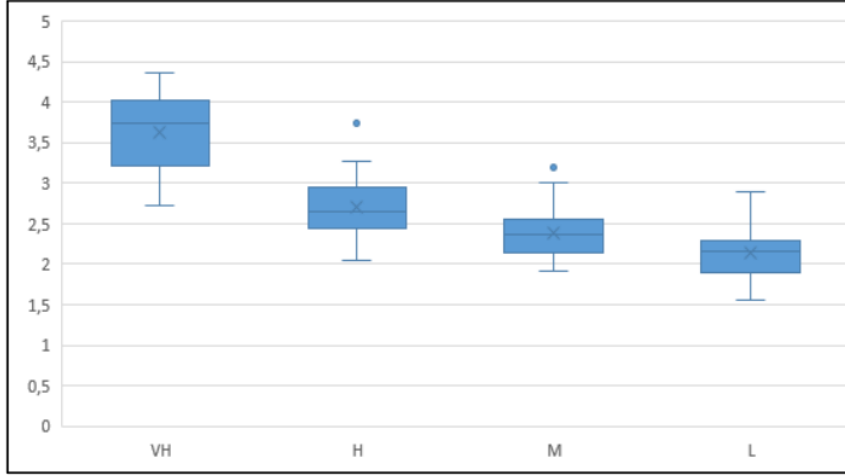


Şekil 1.'de çizilen kutu grafiğinde aykırı değerlerin hangi birimlere ait olduğu görülebiliyorken, Şekil 2.'deki grafikte ise aykırı değerlerin sadece numaraları görülmekte, ülke adları okunamamaktadır. Diğer taraftan SPSS'de, LPI'nın diğer alt boyutlarına ilişkin kutu grafikleri oluşturmak mümkün iken; Bu şekilde oluşturulan birden çok kutu grafiğini tek bir grafik çerçevesi içinde sunmak mümkün olmamaktadır.

2.3. Excel ile Kutu Grafiği

Sıklıkla kullanılan programlardan biri olan Excel'de çizilen kutu grafiği Şekil 3.'de sunulmuştur. Kutu grafiğinin çiziminde Şekil 1. ve Şekil 2.'de kullanılan değişkenler kullanılmıştır.

Şekil 3. Excel ile elde edilen kutu grafiği.



Excel’de çizilen kutu grafiğinde de birimlere ait aykırı değerler görülememektedir. Aynı çerçeve içinde birden çok kutu grafiği çizdirememe problemi burada da devam etmektedir.

2.4. R ile Kutu Grafiği

R fonksiyonuyla çizilen kutu grafiğine ait kodlar aşağıda sunulmuştur. Bu kodlar ile sürekli ve kategorik değişkenler içeren bir veri dosyasından verilerin okunarak, her bir kategorik değişken ile her bir sürekli değişkeni eşleyerek aynı grafik içinde birden fazla kutu grafiğini çizen bir fonksiyon oluşturulması hedeflenmiştir. Bu kodlar ile istenen işlemleri yapan bir fonksiyon bir defaya mahsus çalıştırıldıktan sonra, tanımlanan fonksiyon kullanılarak sadece fonksiyon ismi ile girdi değişkenleri değiştirilerek farklı veriler için kutu grafiği elde edilebilecektir.

Fonksiyonun yapısını açıklamak üzere her bir satır başına (satır numarası) şeklinde ardışık numara verilmiş olup, bu numaralar sadece açıklama amacıyla konmuştur. R ile kodlama sırasında bu numaralar kullanılmamalıdır. Fonksiyon kodlarında birden fazla satırda yazılması gereken kodlar için süslü parantez yapıları bulunmakta olup, bunların takibini yapabilmek üzere aynı renkler kullanılmıştır.

Öncelikle belirtilmesi gereken konu şudur ki; R programı fonksiyonel bir program olup, herhangi bir işlemi gerçekleştiren tüm komutlar aslında birer fonksiyondur ve R programında fonksiyonlar matematiksel işlemlerde olduğu üzere `fonsiyon_ismi` (girdi değişkenleri ve argüman değerleri) şeklinde yazılmaktadır.

İlk ve ikinci satırda fonksiyonun ismi “`kutu_diagram`” olarak belirtilmiş ve sonrasındaki yazılan tüm kodların bu isimli bir fonksiyon olarak tanımlanması amacıyla R’de mevcut bulunan `function` () isimli fonksiyon kullanılmıştır. Fonksiyonun girdi değişkenleri olarak kutu grafiklerini çizdirmek istediğimiz dosyanın adı, her bir değer isminin bulunduğu etiket isimli tek boyutlu vektörün bulunduğu verideki sütun numarası, sayısal değerlerin bulunduğu sütunların hangileri olduğu (burada 2. ve 4. sütunların arası), kategorik verilerin bulunduğu sütun numaraları (burada 5. ve 6. sütunlar) ile kategorik değişkenlerin ne şekilde isimlendirileceği (burada “VH” ile başlayan 4 farklı seviye mevcuttur) belirtilmektedir.

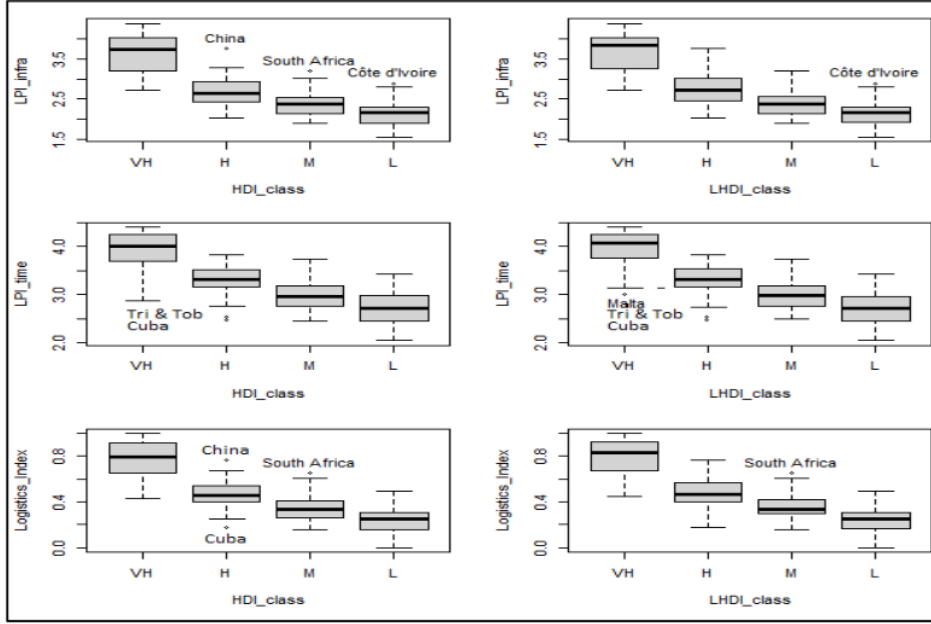
Üçüncü satırda verinin bulunduğu .csv dosyası okunarak veri isimli değişkene atanmış olup, bundan sonraki aşamalarda hesaplamaların hepsi bu değişken üzerinden yapılacaktır.

Dördüncü ve altıncı satırlar arasında veri içindeki kategorik değişkenin bulunduğu sütunlara faktör seviyeleri atanmıştır.

Yedinci satırda her bir kategorik değişken ile her bir sürekli değişken değeri eşleştirilerek sürekli değişken sayısı kadar satır ve faktör sayısı kadar sütuna sahip bir grafik çerçevesi oluşturulmaktadır. Bu örnekte sürekli değişken sayısı 3 ve faktör sayısı 2 olmak üzere 3x2' lik bir grafik çerçevesi oluşturulmaktadır. Sekiz ve dokuzuncu satırlarda i olarak sürekli değişken ve j olarak faktör sayıları kadar iç içe döngü oluşturularak, sonraki kodlarla her bir alt grafik çizdirilmektedir. Onuncu satırdan itibaren plot () fonksiyonuyla her bir alt grafiğin içine kutu diyagramları çizilmeye başlanmaktadır. Kutu diyagramları çizilirken, fonksiyona girdi olarak tanımladığımız faktör seviyeleri ("VH", "H", "M", "L") ve sürekli değişken değerlerine göre kutu grafikleri çizdirilmektedir. Bunlar çizilirken verideki isimler kullanılarak yatay ve düşey eksen isimleri de yerleştirilmektedir. Onbirinci satırda her bir kutu diyagramı için aykırı değerler belirlendikten sonra aykırı değerlerin gerçekten sayısal bir değere karşılık gelip gelmediğini kontrol eden mantıksal sınamadan sonra uç değerlerin hangi birime karşılık geldiği on altıncı satırda belirlenerek on yedinci satır kullanılarak uç değerlerin olduğu noktanın civarında istenilen bir pozisyonda yazdırılmaktadır.

```
(1) kutu_diagram <- function(dosya_adi="Boxplot.csv", etiket=1, deger=2:4,
(2) faktor=5:6, faktor_seviye=c("VH", "H", "M", "L")) {
(3) veri <- read.csv(file=dosya_adi, header=TRUE, stringsAsFactors=FALSE)
(4) for (i in faktor) {
(5) veri[,i] <- factor(veri[,i], levels=faktor_seviye)
(6) }
(7) par (mfrow=c(length(deger),length(faktor)))
(8) for (i in deger) {
(9) for (j in faktor) {
(10) plot (veri[,i] ~ veri[,j], data = veri, xlab=names(veri)[j], ylab=names(veri)[i])
(11) for (k in 1:length (faktor_seviye)) {
(12) boxplot.stats (veri [veri[,j] == faktor_seviye[k], i ]) $out -> uc_deger
(13) if (identical(uc_deger, numeric(0))) {
(14) } else {
(15) for (m in 1: length (uc_deger) ) {
(16) veri [veri[,i]== uc_deger[m] & veri[,j]== faktor_seviye[k], etiket] -> u
(17) text(k, uc_deger , u, pos=3)
(18) }
(19) }
(20) }
(21) }
(22) }
(23) }
```

Şekil 4. Önerilen R kodu ile elde edilen kutu grafiği.



Şekil 4.'de, LPI_infra (LPI'nin alt yapı boyutu), LPI_time (LPI'nin zamanındalık boyutu) ve Logistics_Index (Lojistik endeksi) olmak üzere üç farklı sürekli değişken için, HDI_class (HDI Birleşmiş Milletler sınıfları) ve LHDI_class (HDI farklı bir sınıflandırma yöntemi) kategorilerinde kutu grafikleri çizilmiştir. Burada altı farklı kutu grafiğinin tek bir grafik çerçevesinde görünmesi ve aykırı değerlerin grafik üzerinde okunabilmesi önemli bir avantajdır. Böylece farklı sınıflandırma yöntemlerine göre sürekli değişkenlerin istatistikleri rahatlıkla takip edilebilmektedir. Örneğin ilk satırdaki grafikler incelendiğinde, HDI Birleşmiş milletler sınıflamasına göre (HDI_class) Çin aykırı değer olarak görülmekte ancak LHDI sınıflamasına göre (LHDI_class) Çin kendi grubunun aykırı değeri olmaktan çıkmaktadır. Bu matris formunda, hem sürekli değişkenlerin değişimi durumundaki istatistikler hem de kategorik değişkenin değişimi durumundaki istatistikler, grafiklerin ayrı ayrı çizilmesine göre daha kolay izlenebilmektedir. Benzer şekilde örneğin, dört farklı sürekli değişken ve üç farklı kategorik değişken için 4x3'lük bir grafik çerçevesinde kutu grafikleri elde edilebilir ve yorumlanabilir.

3. SONUÇ

Bu çalışmada, kutu grafiğinin SPSS ve Excel uygulamalarında ortaya çıkan bazı eksiklerini gidermek amacıyla bir R kodu tanıtılmıştır. R fonksiyonu ile çizilen kutu grafiği birden fazla değişken için aynı çerçeve içerisinde oluşturulabilmektedir. Bu yapısı sayesinde, sunum kolaylığı sağlamakta ve görsel gücü artmaktadır. Diğer taraftan SPSS'in bir menüsünde ve EXCEL'deki uygulamalarda kutu grafiğindeki aykırı değerlerin adları görünmemekte, nokta/yıldız işaretleriyle temsil edilmektedir. Bu çalışmadaki uygulamalarda kutu grafiğindeki aykırı değerlerin adları görünmemekte, nokta/yıldız işaretleriyle temsil edilmektedir. Bu çalışmada yazılan R fonksiyonuyla aykırı değerlerin hangi birimlere ait olduğu grafikte nokt/yıldız etrafına yazdırılmaktadır.

R'nin bir programlama dili olması dolayısıyla çizilen kutu grafiği istenilen şekilde geliştirilebilir. R'de renklendirme, boyutlandırma, konumlandırma vb. şekil özelliklerinin

yanında, kutu grafiđi üzerine ortalama gibi farklı istatistikler de eklenebilir. Yazılan kod fonksiyon yapısıyla yazıldıđı için bir defa kodlandıđında yalnızca girdiler deđiştirilerek hızlıca sonuç alınabilir.

Verilerin özetlenmesinde kullanılan diđer grafiklerin R'de yazılması, verilerin özetlenmesinde kullanılan bütün görsel araçların bir R paketine dönüştürülmesi ve analistlerin kullanımına açılması önerilmektedir.

KAYNAKÇA

Çelik, S., & Akdamar, E. (2018). Büyük Veri ve Veri Görselleştirme. *Akademik Bakış Dergisi*, 253-264.

Gürsakal, N. (2013). *Betimsel İstatistik*. Bursa: Dora Yayıncılık.

İşıđıçok, E. (2018). *Betimsel İstatistik*. Bursa: Alfa Aktüel Yayınları.

UNDP. (2021). *United Nations Development Programme Human Development Reports*. Human Development Index (HDI): <http://hdr.undp.org/en/content/human-development-index-hdi> adresinden alınmıştır

World Bank. (2021). *World Bank*. <https://lpi.worldbank.org/international/global> adresinden alınmıştır

R FONKSİYONUYLA KUTU GRAFİĞİNİN ÇİZİLMESİ ÜZERİNE YENİ BİR UYGULAMA

ORJİNALLİK RAPORU

%3

BENZERLİK ENDEKSİ

%2

İNTERNET KAYNAKLARI

%1

YAYINLAR

%1

ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1

www.scribd.com

İnternet Kaynağı

%1

2

9lib.net

İnternet Kaynağı

%1

3

Submitted to Tobb University of Economics & Technology

Öğrenci Ödevi

%1

4

AKAR, Gülşah SEZEN and ÇAKIR, Engin. "LOJİSTİK SEKTÖRÜNDE BÜTÜNLEŞTİRİLMİŞ BULANIK AHP - MOORA YAKLAŞIMI İLE PERSONEL SEÇİMİ", Balıkesir Üniversitesi, Bandırma İİBF, 2016.

Yayın

<%1

5

ILGAR, Rüştü. "Çanakkale'de kuraklık durumu ve eğilimlerinin standartlaştırılmış yağış indisi ile belirlenmesi", TUBITAK, 2010.

Yayın

<%1

6

openaccess.inonu.edu.tr:8080

İnternet Kaynağı

<%1

Alıntıları çıkart

Kapat

Eşleřmeleri çıkar

< 5 words

Bibliyografyayı Çıkart

üzerinde