Veri Setleri ve Grafikler

Beste Ünal

17.11.2022

ÖZET

3 farklı veri seti verilmiştir. Bu veri setlerinin her birinde 3 durum ele alınmış ve uygun grafikler seçilerek grafiklendirilmiştir. Okunurluğu ve anlaşılırlığını arttırmak için estetikleri ile oynanmıştır. Grafiklerin her biri yorumlanmış ve kodları paylaşılmıştır.

Grafikleri çizdirebilmek için gerekli paketler ve kütüphaneler yüklendi.

```
library(readr)
library(readxl)
install.packages("ggplot2")
library(ggplot2)
install.packages("tidyverse")
library(tidyverse)
install.packages("readxl")
library(readxl)
install.packages("ggforce")
library(ggforce)
install.packages("dplyr")
library(dplyr)
install.packages("gridExtra")
library(gridExtra)
install.packages("latexpdf")
library(latexpdf)
options(repos = list(CRAN = "http://cran.rstudio.com/"))
```

```
library(readr)
library(readxl)
```

2.KISIM

Veri seti "Data2" olarak yeniden adlandırıldı. Data2 veri setindeki "Year" değişkeninin sınıfına bakıldı.

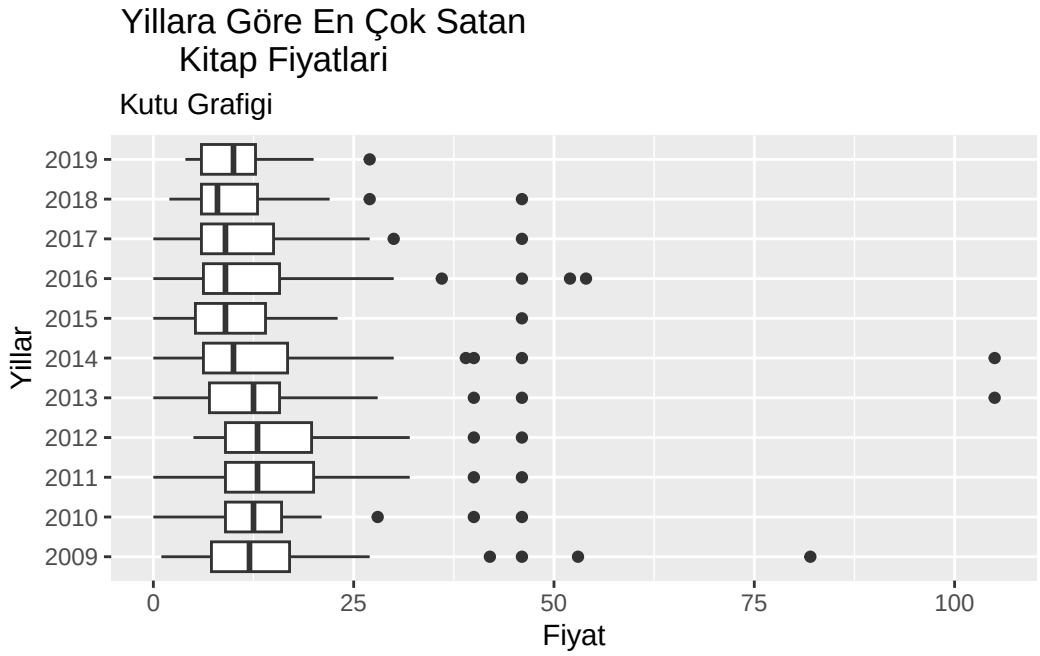
```
library(readr)
library(readxl)
Data2 <- read_excel("books.xlsx")
class(Data2$Year)
```

```
[1] "numeric"
```

2.1

Verilen veri setinden alınan Yıllar(Year) ve Fiyat(Price) değişkenlerine göre "Yıllara Göre En Çok Satan Kitap Fiyatları" kutu grafiği çizdirildi. Estetikleri ile oynanıp okunurluğu ve anlaşırlığı düzeltildi.

```
library(readr)
library(readxl)
ggplot(Data2, aes(y = as.factor(Year), x = Price)) +
  geom_boxplot() +
  labs( y  =  " Yillar " ,
        x  =  " Fiyat " ,
        title  =  " Yillara Göre En Çok Satan
        Kitap Fiyatlari " ,
        subtitle =  " Kutu Grafigi " )
```

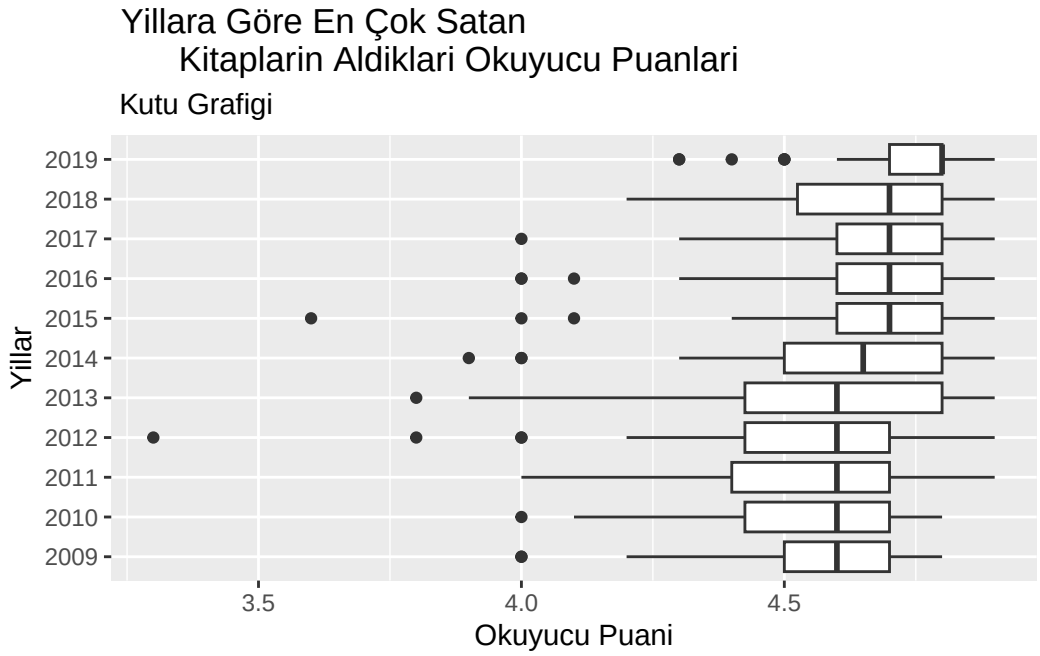


YORUM: Kitap fiyarlarının en ucuz olduğu yıl 2018 yılıdır. 2013 ve 2014 yıllarındaki kitap fiyatlarında 2 kitap fiyatı 100 TL'yi geçmiştir. Kitap fiyatlarında 0-25 TL fiyat aralığı dışında 25-50 TL fiyat aralığında da olan bir çok kitap mevcuttur.

2.2

Verilen veri setinden alınan Yıl(Year) ve Okuyucu Puanı(Reviews) değişkenleri kullanılarak "Yıllara Göre En Çok Satan Kitapların Aldıkları Okuyucu Puanları" kutu grafiği çizdirildi. Okunurluğunun ve anlaşırlığının artması için estetiklerele zenginleştirildi.["Year" değişkeninin sınıfı numaric olduğu için as.factor() kullanılarak factor yapıldı.]

```
library(readr)
library(readxl)
ggplot(Data2, aes(y = as.factor(Year), x = Data2$`User Rating`)) +
  geom_boxplot() +
  scale_x_continuous(labels = scales:: comma) +
  labs( y  =  " Yillar " ,
       x  =  " Okuyucu Puani " ,
       title  =  " Yillara Göre En Çok Satan
       Kitaplarin Aldiklari Okuyucu Puanlari " ,
       subtitle =  " Kutu Grafigi " )
```



Yorum: Okuyucuların Kitaplara verdiği en yüksek puanların dağılımı 2019 yılındadır. Değişim en yüksek 2013 yılındadır ve en düşük 20199 yılındadır.

2.3

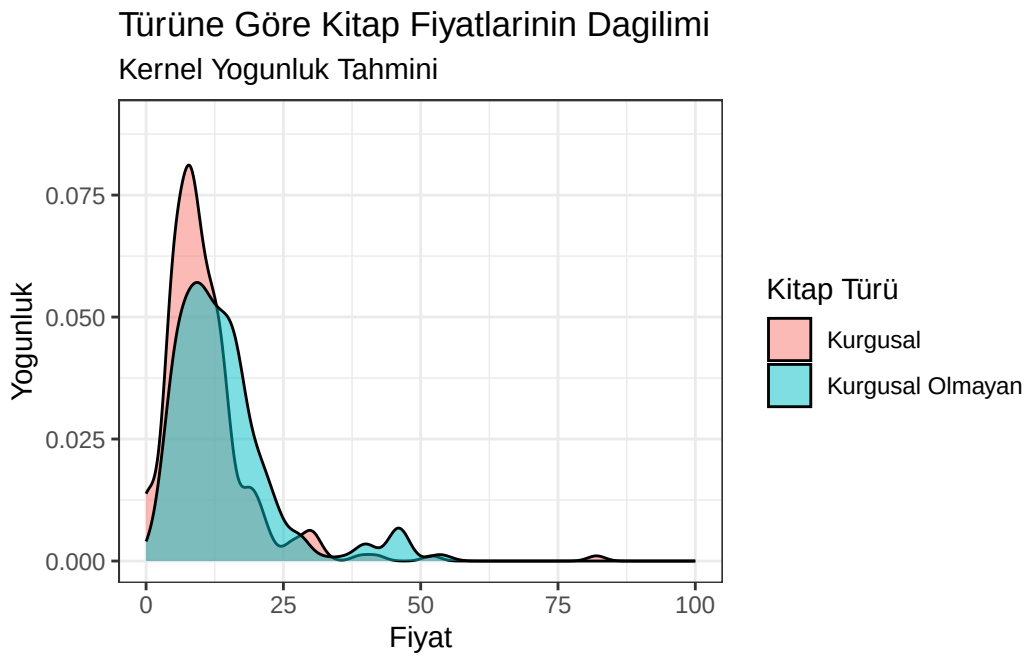
Data2 veri seti "Data2.1" ismine atanmıştır. Veri setindeki NA değerleri çıkartılmıştır. Sonra Genre değişkenindeki "Fiction" ve "Non Ficiton" çıktıları alınarak Yıl(Year) ile gruplandırılmıştır. Ve Fiyatlar(Price) özetlenmiştir. Bunların hepsi Data2.1 içerisine atılmıştır.

```
library(readr)
library(readxl)
Data2.1 <- Data2 %>%
  drop_na() %>%
  filter( Genre %in% c("Fiction", "Non Fiction")) %>%
  group_by(Year,Genre) %>%
  summarise(Price)
```

Dağılım garfiğindeki grafik renklerini değiştirebilmek için gerekli olan paketler ve kütüphaneler yüklenmiştir.

Oluşturulan yeni Data2 veri setinden Fiyat(Price) ve Kitap Türü(Genre)değişkenleri kullanılarak bir yoğunluk grafiği çizdirildi. Estekikleri düzeltildi ve böylece okunurluğu arttırılmıştır.

```
library(readr)
library(readxl)
ggplot(Data2.1, aes(x = Data2.1$Price, fill = Data2.1$Genre)) +
  geom_density(alpha = 0.5) +
  labs(x = "Fiyat",
       y = "Yogunluk",
       title = "Türüne Göre Kitap Fiyatlarinin Dagilimi",
       subtitle = "Kernel Yogunluk Tahmini",
       fill = "Kitap Türü") +
  lims(x = c(0, 100), y = c(0, 0.09)) +
  scale_fill_discrete(labels = c("Kurgusal","Kurgusal Olmayan")) +
  theme_bw()
```



Yorum: Kurgusal Kitap türleri 0-25 TL fiyat aralığında daha çok tercih edilmiştir. Kurgusal olmayan Kitap Türleri 25-50 TL fiyat aralığında Kurgusal Kitap Türlerinden daha çok tercih edilmiştir.

3.KISIM

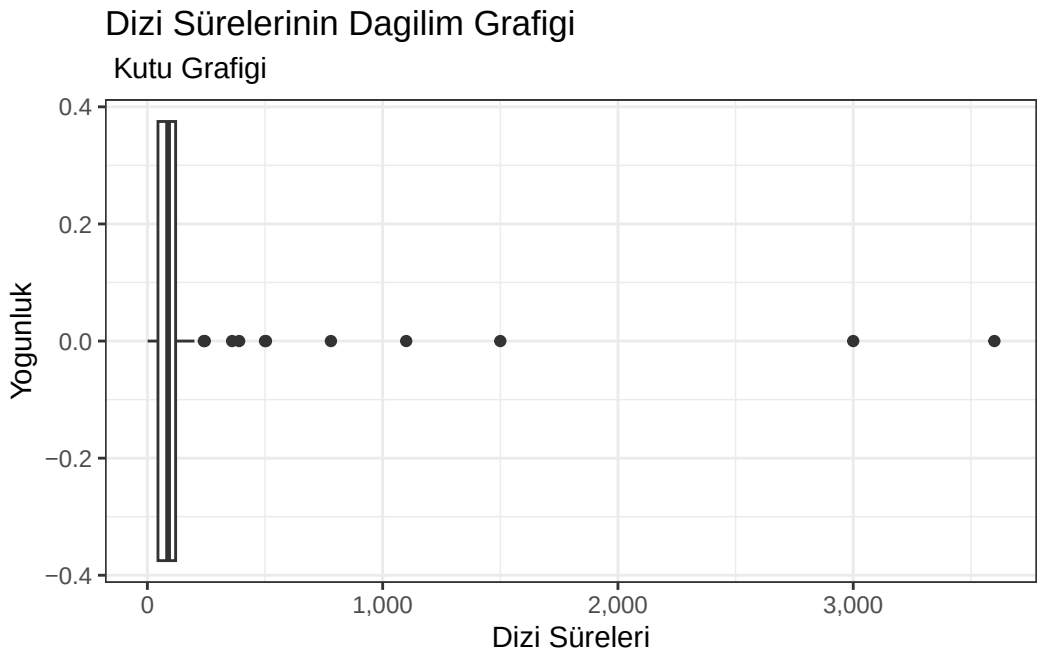
Veri seti “Data3” olarak yeniden adlandırılmıştır. Data2 veri setindeki “original title” değişkeni original_title olarak yeniden adlandırılmıştır. Data2 veri setindeki “number of seasons” değişkeni number_of_seasons olarak yeniden adlandırılmıştır.

```
library(readr)
library(readxl)
DataTv <- read_excel("DataTv.xlsx")
Data3 <- read_excel("turkish_tvseries.xlsx")
colnames(Data3)[1] <- c("original_title")
colnames(Data3)[13] <- c("number_of_seasons")
colnames(DataTv)[2] <- c("number_of_seasons")
```

3.1

NA değerleri çıkartılarak Dizilerin Sezon Sayıları(Number Of Season) ve Dizi Süreleri(runtimes) yeni bir excel dosyası(DataTv) haline getirilmiştir. R’a yüklenmiştir ve tanımlanmıştır. Daha sonra DataTv veri setinde bulunan Dizi Süreleri(runtimes) değişkeni kullanılarak Dizi Sürelerinin Dağılımı Grafiğini gösteren bir kutu grafiği çizilmiştir.

```
library(readr)
library(readxl)
  ggplot(DataTv, aes(x =DataTv$runtimes)) +
  geom_boxplot() +
   scale_x_continuous(labels = scales:: comma) +
  labs( y  =  " Yogunluk " ,
        x  =  " Dizi Süreleri " ,
        title  =  "Dizi Sürelerinin Dagilim Grafigi " ,
        subtitle =  " Kutu Grafigi " ) +
  theme_bw()
```

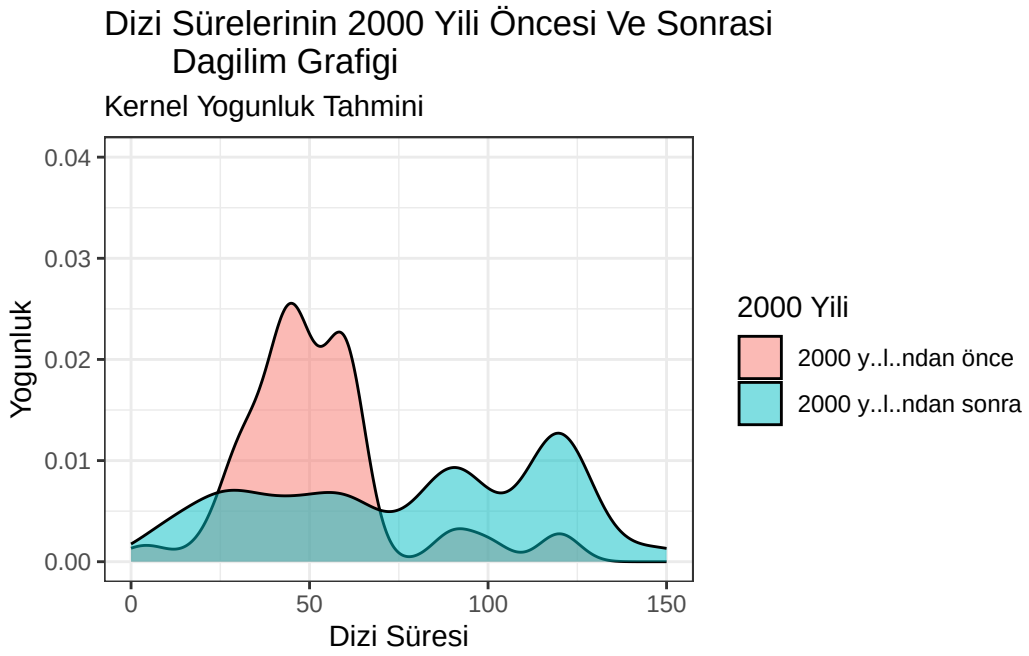


3.2

Data3 veri setinden yıl(Year) değişkeni alınmış ve dizi ye atanarak yeniden veri seti oluşturulmuştur. Data3 verisetinde bulunan yıl(Year) değişkeni 2000 yılından önce ve sonra şeklinde ayrılarak grup(group) değişkenine atanmıştır. Data3 verisetinde bulunan Dizilerin Süresi(runtimes) değişkeni ve dizi değişkeni alınıp yeni bir dizi2 veri seti oluşturulmuştur.

```
library(readr)
library(readxl)
dizi <- data.frame(Data3$year)

dizi$group <- as.factor(ifelse(dizi$Data3.year < 2000, "2000 yılından önce", "2000 yılında


dizi2 <- data.frame(Data3$runtimes, dizi)

dizi2 <- dizi2[-c(302, 327, 415, 585, 676, 735, 827, 1018, 1166, 1167, 1238, 1271, 1346, 1

dizi2 = dizi2[complete.cases(dizi2), ]
```

```r
library(readr)
library(readxl)
ggplot(dizi2, aes(group =dizi2$group , fill = dizi2$group, x =dizi2$Data3.runtimes )) +
  geom_density(alpha = 0.5) +
  labs(x = "Dizi Süresi",
       y = "Yogunluk",
       title = "Dizi Sürelerinin 2000 Yili Öncesi Ve Sonrasi
       Dagilim Grafigi",
       subtitle = "Kernel Yogunluk Tahmini",
       fill = "2000 Yili") +
  lims(x = c(0, 150), y = c(0, 0.04)) +
  theme_bw()
```



YORUM: 2000 yılından önceki yıllarda dizi süresi yoğunluğu 40 ile 60 dakika arasındadır. 2000 yılından sonraki yıllarda dizi süreleri 100 ie 150 dakika aralığında yoğunlaşmıştır.

3.3

NA değerleri çıkartılarak Dizilerin Sezon Sayıları(Number Of Season) ve Dizi Süreleri(runtimes) yeni bir excel dosyası(DataTv) haline getirilmiştir. R'a yüklenmiştir ve tanımlanmıştır. Daha sonra DataTv veri setinde bulunan Sezon sayıları(Number Of Season) değişkeni kullanılarak

Dizilerin Sezon Sayılarının Dağılım Grafiği , Kernel Yoğunluk Tahmini grafiği ile gösterilmiştir.

```
library(readr)
library(readxl)
ggplot(DataTv, aes(x = number_of_seasons)) +
  geom_density(alpha = 0.5, fill = "skyblue", color = "darkblue") +
  labs(x = " Dizilerin Sezon Sayilari",
       y = "Yogunluk",
       title = "Dizilerin Sezon Sayilarinin Dagilim Grafigi",
       subtitle = "Kernel Yogunluk Tahmini") +
  theme_bw()
```

YORUM: Verilen veri setindeki Türk dizilerinin büyük bir çoğunluğunun yoğunluğu 1 ile 4. sezon arasındadır. 4 sezondan daha fazla olan dizi sayısı yoğunluğu oldukça azdır. Grafikte görüldüğü üzere dizilerin sezon yoğunluğu 1 ve 2 sezon aralığında en yüksektedir.