Farklı Veri Setlerinin İncelenip Görselleştirilmesi

Rümeysa Kurt

16.11.2022

ÖZET

Bu raporda 3 farklı veri seti kaggle.com'dan alınmıştır. İlk veri seti en çok satan kitaplar ile ilgilidir ve 3 ayrı konuda görselleştirilmiştir. İkinci veri seti Game of Thrones dizisini içermektedir. Bu veri setide yine 3 ayrı konuda görselleştirilmiştir. Son veri seti ise türk televizyon dizilerini içermektedir, 3 ayrı konuda görselleştirme yapılmıştır. Görelleştirme araçlarını kullanmak için install.packages("ggplot2"), install.packages("ggridges"), install.packages("ggforce") paketleri indirilmiştir. Oluşan grafikleri renklendirmek için ise install.packages("MetBrewer") paketi indirilip kullanılmıştır. Son olarak ortaya çıkan görseller yorumlanmıştır.

Gerekli paketlerin yüklenmesi ve çağırılması

```
install.packages("ggplot2")
install.packages("ggridges")
install.packages("ggforce")
install.packages("dplyr")
install.packages("tidyr")
library(ggplot2)
library(ggridges)
library(ggforce)
library("dplyr")
library("tidyr")
```

Veri Seti 3. En çok satan kitaplar

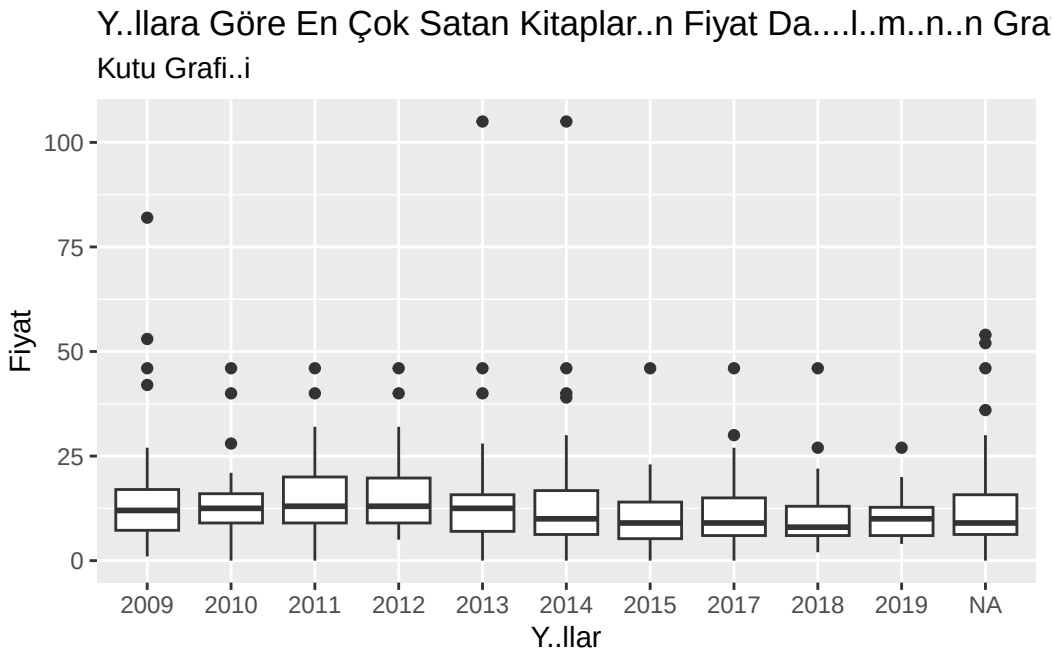
Veri seti 2009'dan 2019'a kadar Amazon'un en çok satan 50 kitabını içermektedir. Kitaplar kurgu ve kurgu olmayan olarak iki sınıfa ayrılmıştır. Veri setinde 7 değişken ve 550 gözlem olduğu görülmüştür. Veri setinde yer alan değişkenler şunlardır: Kitap adı, Yazar, Kullanıcı oyu, Yorumlar, Fiyat, Yıl, Tür.

Yıllara göre en çok satan kitapların fiyat dağılımının değişimini araştırınız.

```
library(readr)
bestsellers_with_categories <- read_csv("bestsellers with categories.csv")
```

```
bestsellers_with_categories$Year <- factor(bestsellers_with_categories$Year,
                                           ordered = TRUE,
                                    levels = c("2009", "2010", "2011",
                                               "2012","2013", "2014", "2015", "
                                               2016", "2017", "2018", "2019"))
```

```
ggplot(bestsellers_with_categories, aes(x = Year, y = Price)) +
  geom_boxplot() +
  labs(y = "Fiyat",
       x = "Yıllar",
       title = "Yıllara Göre En Çok Satan Kitapların Fiyat Dağılımının Grafiği",
       subtitle = "Kutu Grafiği")
```

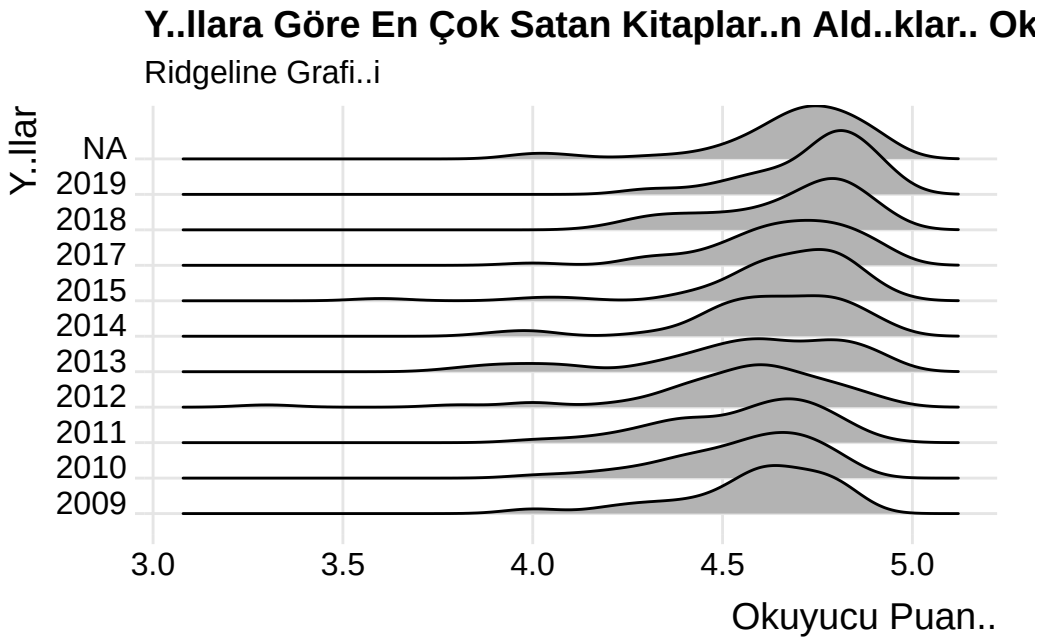


Yorum

Yıllara göre en çok satan kitapların fiyat dağılımı Kutu grafiği ile görselleştirilmiştir. x ekseninde 2009'dan 2019'a kadar yıllar, y ekseninde ise kitapların fiyatı yer almaktadır. Grafiğe bakarak her senede outlier(aykırı) değer olduğunu söyleyebiliriz. Kitap fiyatlarının 0 ile 100 küsür arasında olduğu görülmektedir. Fiyat dağılımı açısından en simetrik yıllar 2009 ve 2015'tir ama 2009 yılında 2015 yılına göre daha fazla aykrı değer bulunmaktadır.

Yıllara göre en çok satan kitapların aldıkları okuyucu puanı dağılımının değişimini araştırınız.

```
ggplot(bestsellers_with_categories, aes(x = `User Rating`, y = Year)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  labs(x = "Okuyucu Puanı",
       y = "Yıllar",
       title ="Yıllara Göre En Çok Satan Kitapların Aldıkları Okuyucu Puanı Dağılımı"
       ,subtitle = "Ridgeline Grafiği")
```



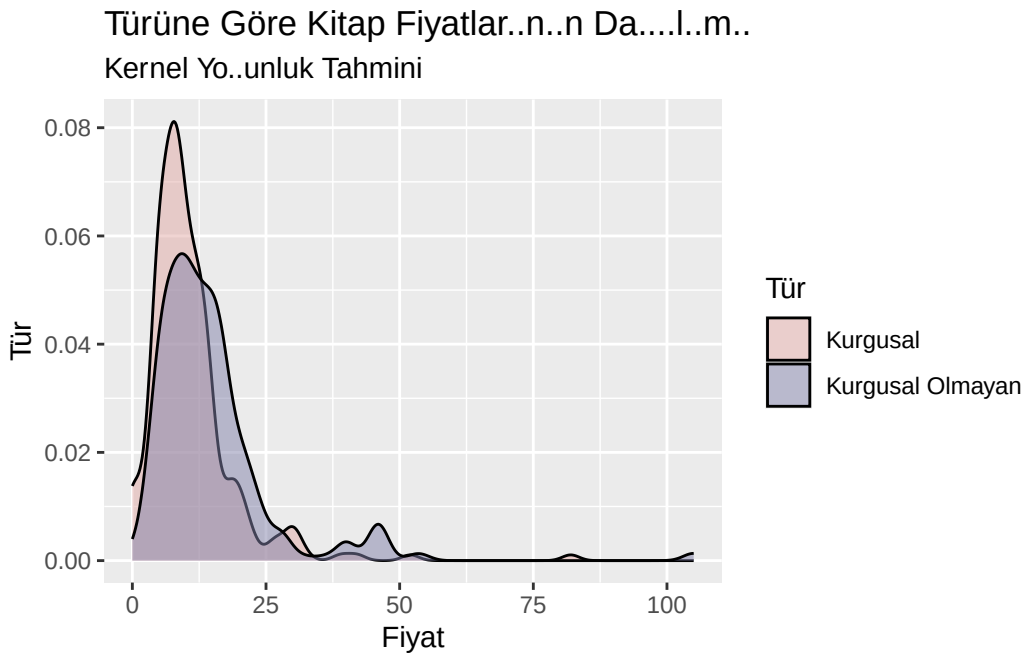
Yorum

Yıllara göre en çok satan kitapların aldıkları okuyucu puanı dağılımı Ridgeline grafiği ile görselleştirilmiştir. Grafiğe göre en yüksek okuyucu puanının 2019'da, en düşük okuyucu puanının ise 2013'te olduğu görülmektedir. Bu yorumu 2013'te olan dalgalanmalara bakarak yapabiliriz. 2019'da verilen okuyucu puanlarının 4.5 ile 5 arasında, 2013'te ise verilen puanların 4 ile 5 arasında olduğunu söyleyebiliriz.

Türüne göre kitap fiyatlarının dağılımını araştırınız.

```
install.packages("MetBrewer") # renklendirme yapmak için
library("MetBrewer")
```

```
ggplot(bestsellers_with_categories, aes(x = Price, fill = Genre)) +
  geom_density(alpha = 0.5) +
  labs(x = "Fiyat",
       y = "Tür",
       title = "Türüne Göre Kitap Fiyatlarının Dağılımı",
       subtitle = "Kernel Yoğunluk Tahmini",
       fill = "Tür") +
  scale_fill_manual(values = met.brewer("Cassatt1",2),
                    labels = c("Kurgusal", "Kurgusal Olmayan"))
```



Yorum

Türüne göre kitap fiyatlarının dağılımını görselleştirmek için Kernel yoğunluk grafiği oluşturulmuştur. x eksenine fiyat , ye eksenine kitap türleri konumlandırılmıştır. Grafiğe bakıldığında fiyat aralığının 0-25 olduğu kısımda kurgusal kitapların kurgusal olmayan kitaplara göre daha fazla olduğunu söyleyebiliriz. İki farklı kitap türü içinde outlier(aykırı) değerlerin olduğunu görilmektedir. Kurgusal olmayan kitaplar içinde fiyatı 100 liradan fazla kitapların olduğu görülmektedir. Kendi içinde kurgusal olmayan kitapları yorumlarsak, en fazla kitabın 0-25 arasında olduğunu söyleyebiliriz.

Veri Seti 6. Game of Thrones

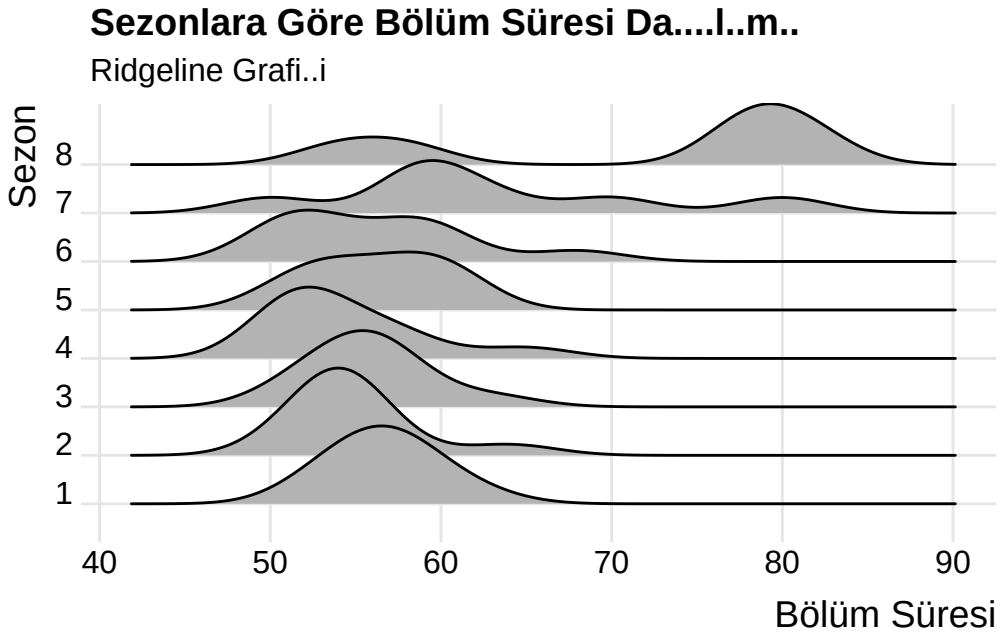
Bu veri seti Game of Thrones dizi ile ilgili verileri içermektedir. Game of Thrones veri setinde 18 değişken ve 73 gözlem yer almaktadır. Veri setinde bulunan değişkenler şunlardır:Sezon, Bölüm, Başlık, Yayın Tarihi, Rating, Oy, Özet, Yazar_1, Yazar_2, Başrol_1, Başrol_2, Başrol_3, Kullanıcı_yorumları, Eleştirmen_yorumları, US_yorumları, Süre, Yönetmen, Bütçe_tahmini.

Sezonlara göre bölüm süresi dağılımlarını inceleyiniz. Bölüm sürelerindeki en yüksek değişimin gözlendiği sezonu belirleyiniz.

```
library(readr)
GOT_episodes_v4 <- read_csv("GOT_episodes_v4.csv")
```

```
GOT_episodes_v4$Season <- factor(GOT_episodes_v4$Season,
                                 ordered = TRUE,
                                                levels = c("1", "2", "3",
                                                           "4", "5","6",
                                                           "7", "8"))
```

```
ggplot(GOT_episodes_v4, aes(x = Duration, y = Season)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  labs(y = "Sezon",
       x = "Bölüm Süresi",
       title = "Sezonlara Göre Bölüm Süresi Dağılımı",
       subtitle = "Ridgeline Grafiği")
```

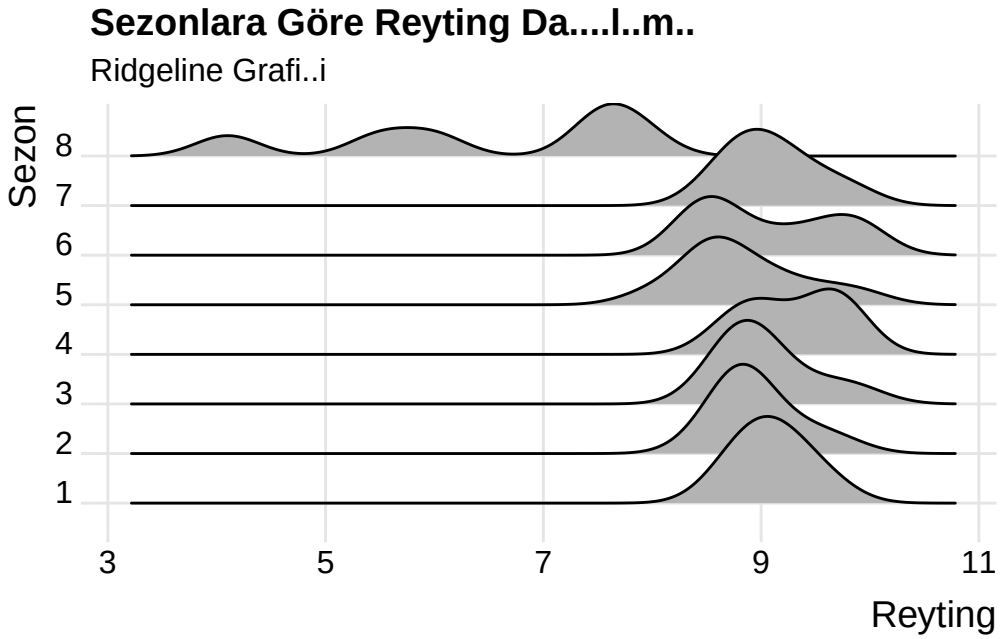


Yorum

Sezonlara göre bölüm süresi dağılımları Ridgeline grafiği ile görselleştirilmiştir. Bölüm süresi x eksininde, sezonlar ise y ekseninde yer almaktadır. Bölüm sürelerindeki en yüksek değişimin 7. sezonda olduğu gözlemlenmiştir. Bunu grafikteki dalgalanmalar sayesinde anlayabiliriz. 7. sezonu ise 8.sezon takip etmektedir. Grafikten bölüm sürelerinin 50 dakika ile 85 dakika arasında değişim gösterdiğini söyleyebiliriz. En düşük değişimin gözlendiği sezon ise 1.sezondur. Bölüm sürelerinin 50 dakika ile 60 dakika arasında değiştiğini söyleyebiliriz.

Sezonlara göre reyting dağılımlarını araştırınız. Reyting değişiminin en düşük olduğu sezonu belirleyiniz.

```
ggplot(GOT_episodes_v4, aes(x = Rating, y = Season)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none") +
  labs(y = "Sezon",
       x = "Reyting",
       title = "Sezonlara Göre Reyting Dağılımı",
       subtitle = "Ridgeline Grafiği")
```

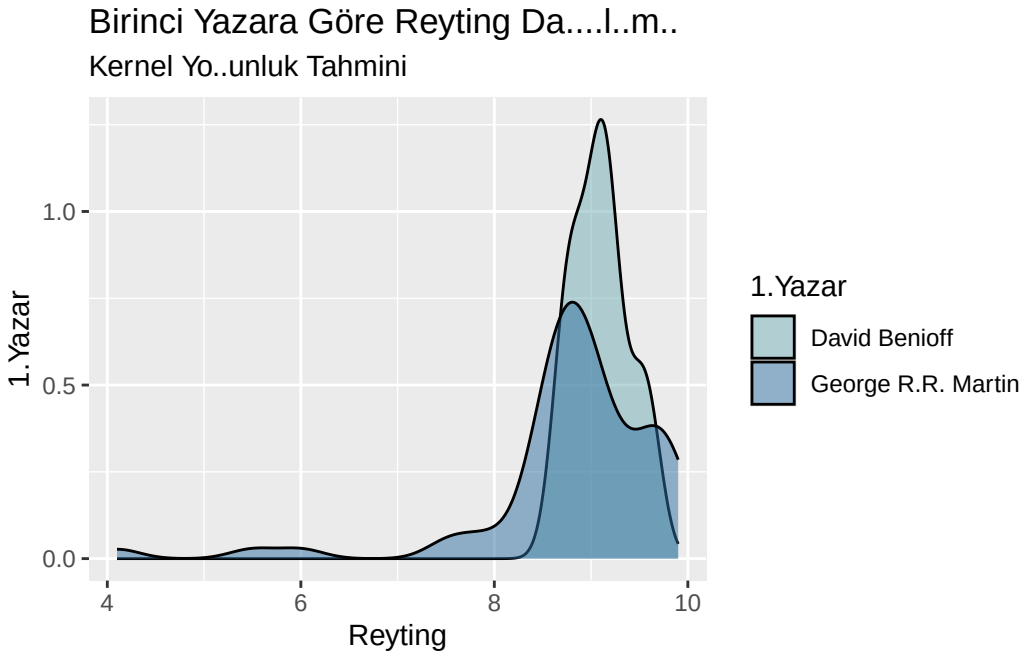


Yorum

Sezonlara göre rating dağılımları Ridgeline grafiği ile görselleştirilmiştir. x kesininde reyting değerleri, y ekseninde ise sezonlar yer almaktadır. Reyting değişiminin en düşük olduğu sezonun 1. sezon olduğu görülmektedir. 1.sezondaki reytinglerin 8.5 ile 9.5 arasında olduğunu söyleyebiliriz. Reyting değişiminin en fazla olduğu sezon ise 8.sezondur. Bunu oluşan girinti çıkıntılardan anlayabiliriz. Reytinglerin 4 ile 7.5 arasında olduğunu söyleyebiliriz.

Birinci yazara göre reyting dağılımını araştırınız. Hangi yazarın senaryosunun daha yüksek reyting aldığını belirleyiniz.

```
ggplot(GOT_episodes_v4, aes(x = Rating, fill =Writer_1 )) +
  geom_density(alpha = 0.5) +
  labs(x = "Reyting",
       y = "1.Yazar",
       title = "Birinci Yazara Göre Reyting Dağılımı",
       subtitle = "Kernel Yoğunluk Tahmini",
       fill = "1.Yazar") +
  scale_fill_manual(values = met.brewer("Hokusai2",2),
                    labels = c("David Benioff", "George R.R. Martin"))
```



Yorum

Bu kısımda birinci yazara göre reyting dağılımını görselleştirmek için Ridgeline ve Kernel yoğunluk grafiği oluşturulmuştur. 1.Yazarlar iki kişiden oluşmaktadır. Bunlar David Benioff ve George R.R. Martin'dir. Oluşturulan iki grafiktede David Benioff'un yer aldığı bölümlerin George R.R. Martin' e göre daha yüksek reyting aldığı görülmektedir.

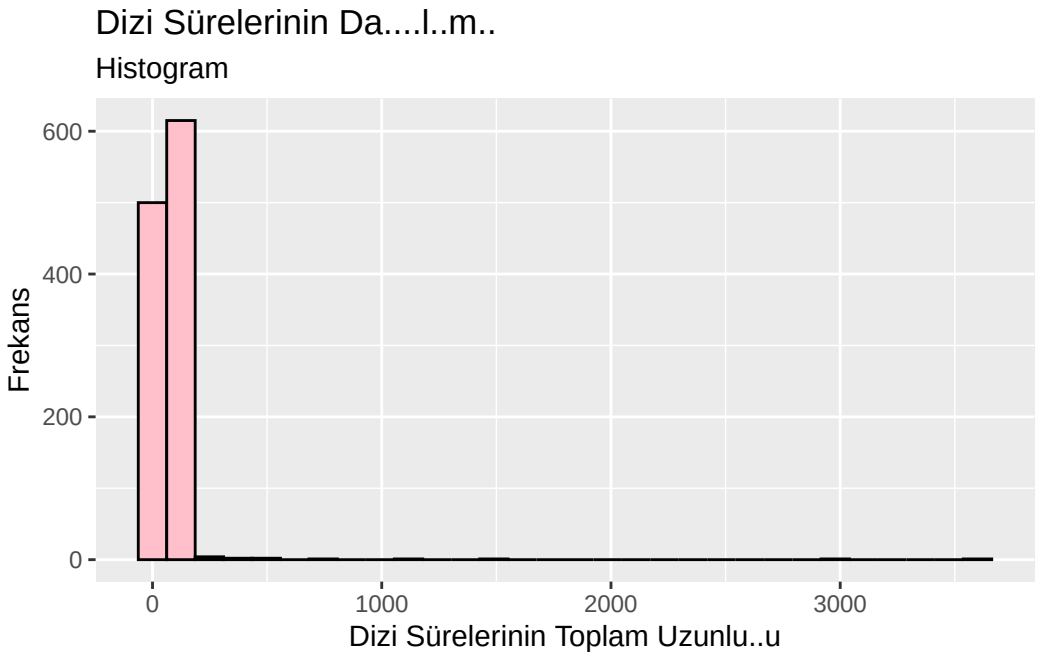
Veri Seti 11.Türk TV dizileri

Ver seti kaggle.com dan alınmıştır. Geçmişten günümüze yayınlanmış televizyon dizilerini, dizilerin yayınlandıkları tarihleri, oyuncuları, sezon sayılarını vb. içermektedir. Bu veri setinde 103 değişken ve 2128 tane gözlem yer almaktadır.

Dizi sürelerinin dağılımını araştırınız.

```
library(readr)
turkish_tvseries <- read_csv("turkish_tvseries.csv")
```

```
ggplot(turkish_tvseries, aes(x = runtimes)) +
  geom_histogram(color = "black", fill = "pink") +
  labs(x = "Dizi Sürelerinin Toplam Uzunluğu",
       y = "Frekans",
       title = "Dizi Sürelerinin Dağılımı",
       subtitle = "Histogram")
```



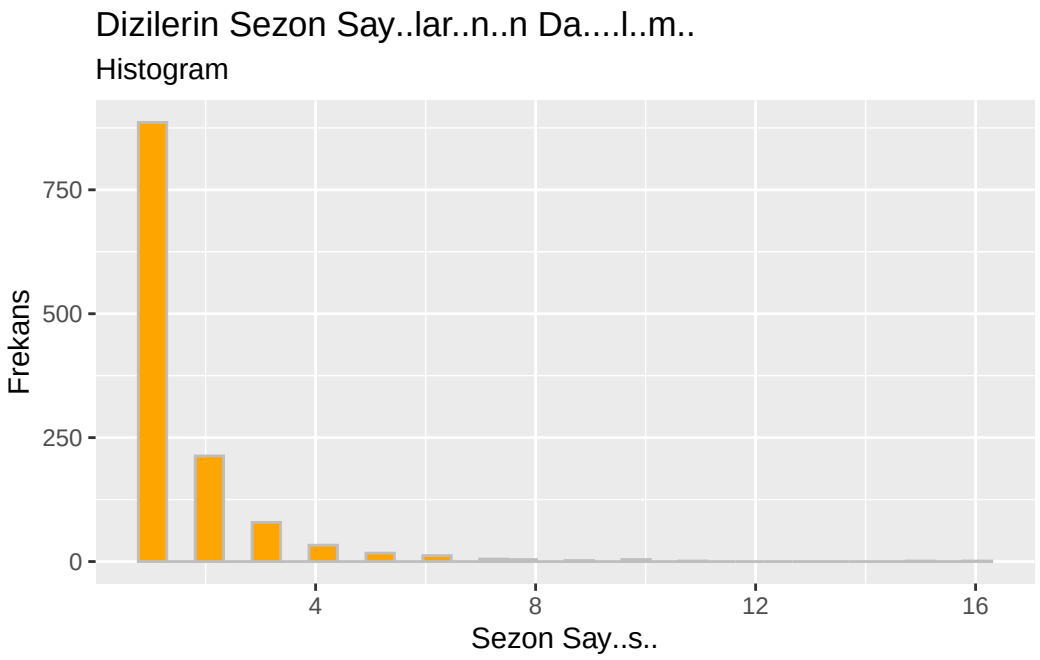
Yorum

Dizi sürelerinin dağılımını görselleştirmek için Histogram grafiği kullanılmıştır. Grafiğe bakıldığında yaklaşık 1200 dizinin süresinin birbirine yakın olduğu görülmüştür. 1000 saatten fazla olan 4 dizi olduğunu söyeleyebiliriz.

Dizi sürelerinin 2000 yılı öncesi ve sonrası dağılımını araştırınız.

Dizilerin sezon sayılarının dağılımını araştırınız.

```
ggplot(turkish_tvseries, aes(x = seasons)) +
  geom_histogram(color = "gray", fill = "orange") +
  labs(x = "Sezon Sayısı",
       y = "Frekans",
       title = "Dizilerin Sezon Sayılarının Dağılımı",
       subtitle = "Histogram")
```



Yorum

Dizilerin sezon sayılarının dağılımını görselleştirmek için Histogram grafiği çizdirilmiştir. Histogram grafiğine bakıldığında diziler en çok 1 sezon, en az 16 sezon sürmüştür. Grafiğe genel olarak bakıldığında dizilerin sezon sayılarının 1 ile 4 arasında olduğu görülmektedir.