Farklı Veri Setlerinin İncelenip Görselleştirilmesi ve Yorumlanması

Rümeysa Kurt

22.11.2022

ÖZET

Bu raporda 3 farklı veri seti kaggle.com'dan alınmıştır. İlk veri seti en çok satan kitaplar ile ilgilidir ve 2 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, 2 ayrı konuda görselleştirme yapılmıştır. Görelleştirme araçlarını kullanmak ve çeşitli işlemleri yapmak için install.packages("ggplot2"), install.packages("dplyr"), install.packages("tidyr") paketleri indirilmiştir. Oluşan grafikleri renklendirmek için ise install.packages("MetBrewer") paketi indirilip kullanılmıştır. Bu raporda istenilen oranlar hesaplanıp görselleştirilmiştir ve son olarak ortaya çıkan görseller yorumlanmıştır.

Gerekli paketlerin indirilmesi

```
install.packages("ggplot2")# ggplot görselleştirme araçlarını kullanmak için
install.packages("dplyr")# %>% operatörü kullanabilmek ve veri manipülasyonu yapabilmek iç
install.packages("tidyr")
install.packages("treemapify")
install.packages("ggmosaic")
library(ggplot2)
library(dplyr)
library(tidyr)
library(treemapify)
library(ggmosaic)
```

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

2009-2019 Yılları Arasında Amazon'da En Çok Satan 50 Kitabı İçeren Veri Setinin İncelenip İlgili Konular Dahilinde Görselleştirilmesi

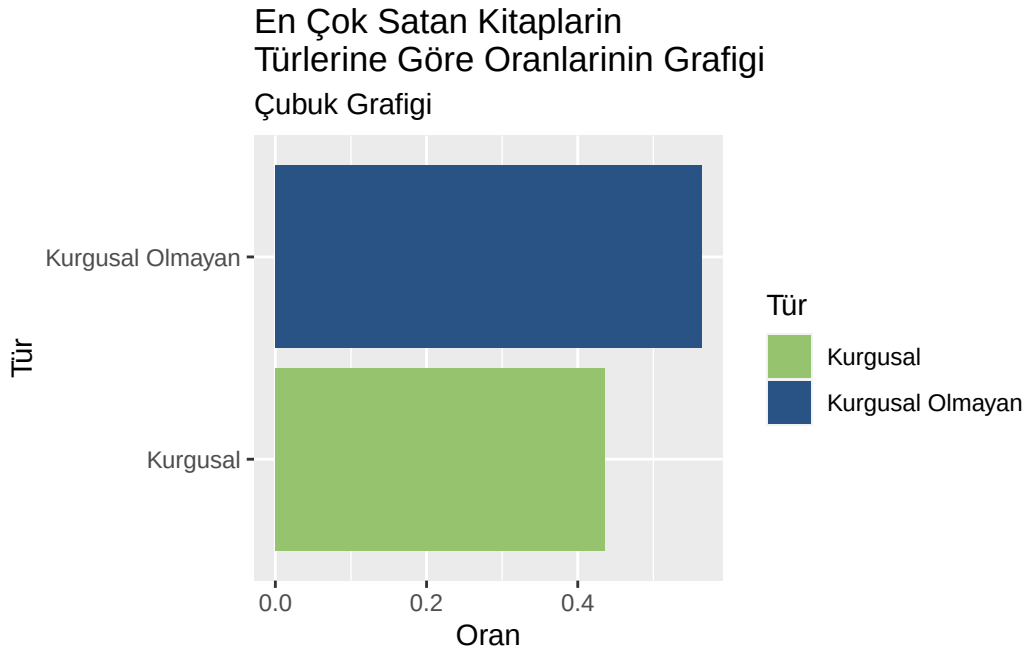
Veri seti 2009'dan 2019'a kadar Amazon'un en çok satan 50 kitabını içermektedir. Kitaplar kurgu ve kurgu olmayan olarak iki sınıfa ayrılmıştır. Veri setinde 7 değişken ve 550 gözlem olduğu görülmüştür. Veri setinde yer alan değişkenler şunlardır: Kitap adı, Yazar, Kullanıcı reytingi, Yorumlar, Fiyat,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

1.En çok satan kitapların türlerine göre oranlarını veri görselleştirme yöntemleriyle araştırınız.

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

```
kitap <-bestsellers_with_categories %>%
group_by(Genre)%>%
summarise(oran=n()/550)
```

```
ggplot(kitap, aes(fill = Genre,
y = Genre,
x = oran)) +
geom_bar(position = "dodge",
stat = "identity") +
  labs(x = "Oran",
       y = "Tür" ,
       fill = "Tür",
       title = "En Çok Satan Kitaplarin
Türlerine Göre Oranlarinin Grafigi",
       subtitle = "Çubuk Grafigi") +
  scale_fill_manual(values = met.brewer("Hokusai3",2),
labels = c("Kurgusal", "Kurgusal Olmayan")) +
  scale_y_discrete(labels = c("Kurgusal", "Kurgusal Olmayan"))
```



Yorum

En çok satan kitapların türlerine göre oranları Çubuk gragiği ile görselleştirilmiştir. x eksenine oran, y eksenine ise kitap türleri konumlandırılmıştır. Renklendirmeler kitap türlerine göre yapılmıştır. Grafiğe bakıldığında 2009-2019 yılları arasında Amazon'da en çok satan kitapların türlerine göre oranlandığı ve kurgusal olmayan kitapların kurgusal olanlardan daha çok sattığı görülmektedir. Kurgusal olmayanların oranı 0.5 'i geçerken kurgusal olanlar ise 0.4 'ün biraz üzerine çıkmıştır.

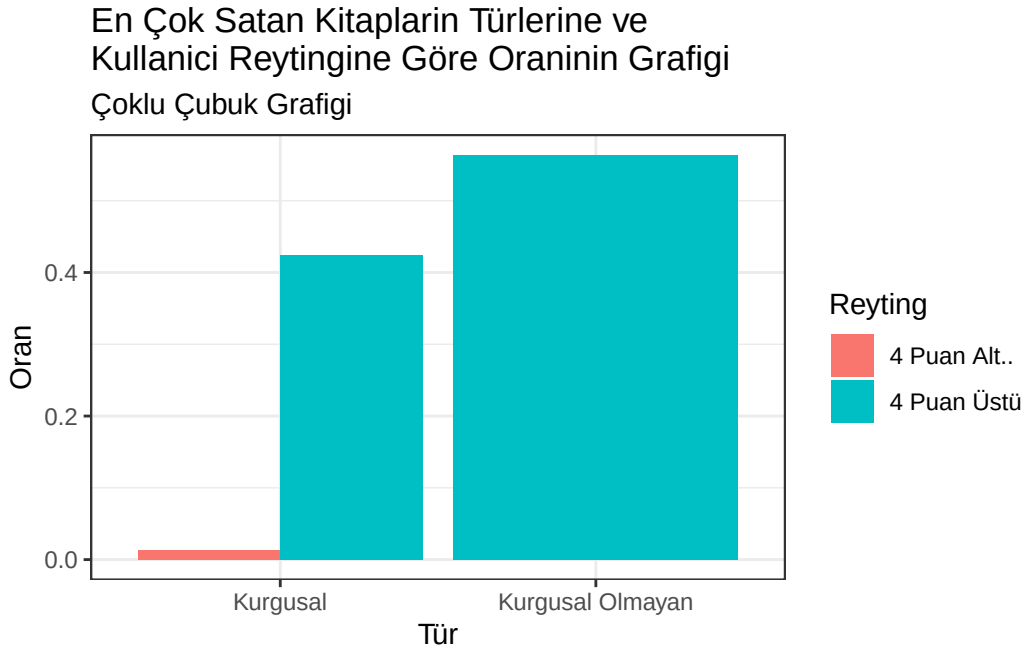
2.En çok satan kitapların türlerine ve kullanıcı reytingine (iki gruba ayırınız: 4 puan altı ve üstü) göre oranlarını veri görselleştirme yöntemleriyle araştırınız.

```
bestsellers_with_categories <- bestsellers_with_categories %>%
  mutate(reyting = ifelse(`User Rating` >=4, "4 Puan Üstü", "4 Puan Altı"))
```

```
reyting <- bestsellers_with_categories %>%
  group_by(Genre,reyting) %>%
  summarise(oran1 =n()/550, oran2 = n()/550)
```

```
ggplot(reyting, aes(fill = reyting,
                    y = oran2,
```

```
                          x = Genre)) +
  geom_bar(position = "dodge",
           stat = "identity") +
  labs(x = "Tür",
       y = "Oran",
       fill = "Reyting",
       title = "En Çok Satan Kitaplarin Türlerine ve
Kullanici Reytingine Göre Oraninin Grafigi",
       subtitle = "Çoklu Çubuk Grafigi") +
  scale_x_discrete(labels = c("Kurgusal", "Kurgusal Olmayan")) +
  theme_bw()
```



Yorum

Bu kısımda en çok satan kitaplar türlerine ve kullanıcı reytingine (4 puan altı ve üstü) göre gruplandırılmış ve oranları hesaplanmıştır. Görselleştirmek için çoklu çubuk grafiği kulllanılıp x eksenine kitap türleri, y eksenine oran konumlandırılmıştır. Renklendirmeler reyting gruplarına göre yapılmıştır. Grafiğe ilk bakıldığında kurgusal olmayan kitapların hepsinin 4 puan üstü reyting aldığı görülmüştür ve bu oran 0.5 'i geçmektedir. Kurgusal olan kitaplardan 4 puan üstü reyting alan kitapların oranı ise 0.4'ten biraz fazladır. Ayrıca 4 puan altı reyting alan kitapların da sadece kurgusal olduğunu söyleyebiliriz.

Game of Thrones Dizisi İle İlgili Verileri İçeren Veri Setinin İncelenip İlgili Konular Dahilinde Görselleştirilmesi

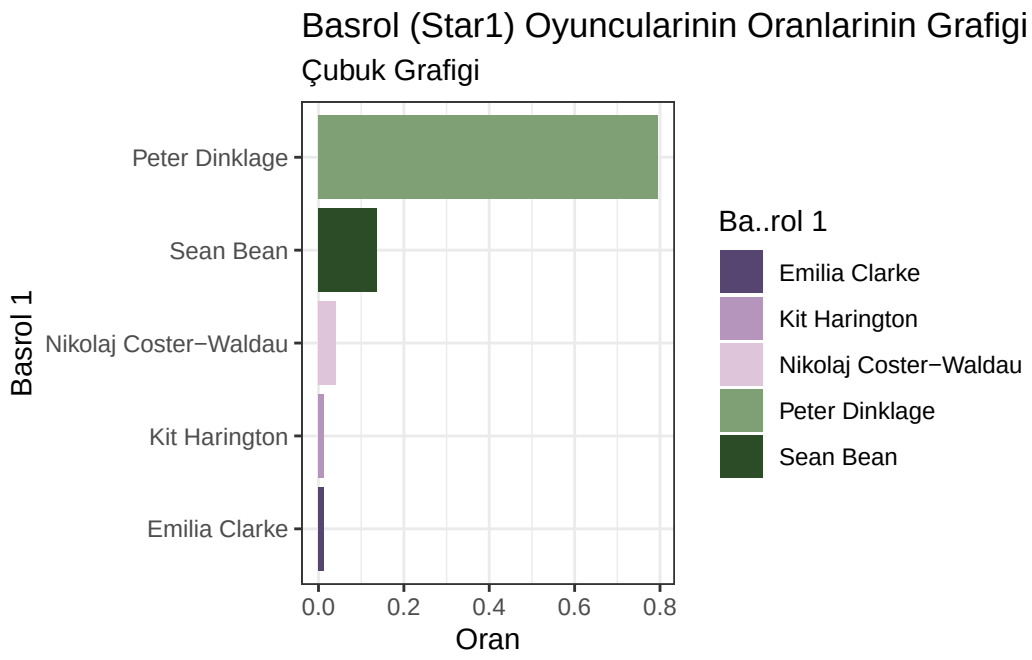
Bu veri seti Game of Thrones dizi ile ilgili verileri içermektedir. Game of Thrones veri setinde 18 değişken ve 73 gözlem yer almaktadır. Veri setinde bulunan değişkenler şunlardır:Sezon, Bölüm, Başlık, Yayın Tarihi, Rating, Oy, Özet, Yazar_1, Yazar_2, Başrol_1, Başrol_2, Başrol_3, Kullanıcı_yorumları,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

1.Başrol (Star1) oyuncularının oranlarını veri görselleştirme yöntemleriyle araştırınız.

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

```
başrol <-GOT_episodes_v4 %>%
group_by(Star_1)%>%
summarise(oran=n()/73)
```

```
ggplot(başrol, aes(y = reorder(Star_1, +oran), x = oran, fill = Star_1)) +
 geom_bar(stat="identity")+
labs(x = "Oran",
y = "Basrol 1",
title = "Basrol (Star1) Oyuncularinin Oranlarinin Grafigi",
subtitle = "Çubuk Grafigi",
fill = "Başrol 1")+
  scale_fill_manual(values = met.brewer("Cassatt2",5)) +
  theme_bw()
```



Yorum

Başrol (Star1) oyuncularının oranları çubuk grafiği ile görselleştirilmiştir. x eksenine oran, y eksenine ise oyuncular konumlandırılmıştır. Renklendirmeler oyunculara göre yapılmıştır. 5 oyuncunun oranlarına bakıldığında Peter Dinklage açık ara farkla herkesi geçmiştir. En yakın rakibi Sean Bean ile arasında neredeyse 6 kat fark vardır. Emilia Clarke ve Kit Harington ise aynı oranlara sahiplerdir.(0.1 'in altında kalmışlardır.)

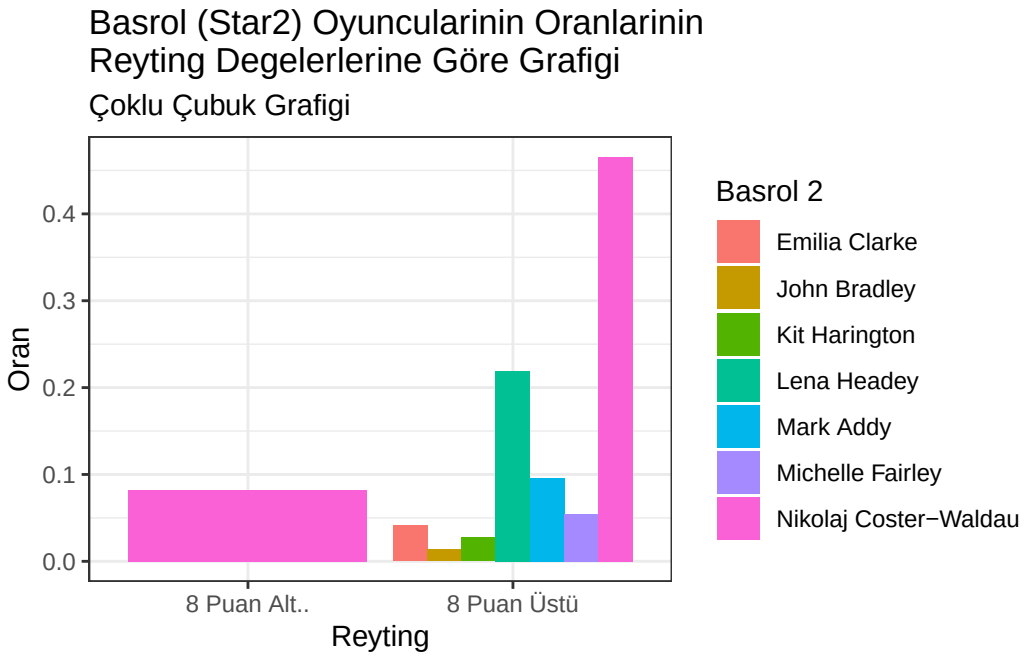
2.Başrol (Star2) oyuncularının oranlarını reyting değelerlerine göre (iki gruba ayırınız: 8 puan altı ve üstü) veri görselleştirme yöntemleriyle araştırınız.

```
GOT_episodes_v4 <- GOT_episodes_v4 %>%
  mutate(rating_got = ifelse(Rating >=8, "8 Puan Üstü", "8 Puan Altı"))
```

```
rating_got <- GOT_episodes_v4 %>%
  group_by(Star_2,rating_got) %>%
  summarise(oran1 =n()/73, oran2 = n()/73)
```

```
ggplot(rating_got, aes(x = rating_got,
                       y = oran2,
                       fill = Star_2)) +
```

```
  geom_bar(position = "dodge",
           stat = "identity") +
  labs(x = "Reyting",
       y = "Oran",
       fill = "Basrol 2",
       title = "Basrol (Star2) Oyuncularinin Oranlarinin
Reyting Degelerlerine Göre Grafigi",
       subtitle = "Çoklu Çubuk Grafigi") +
  theme_bw()
```



Yorum

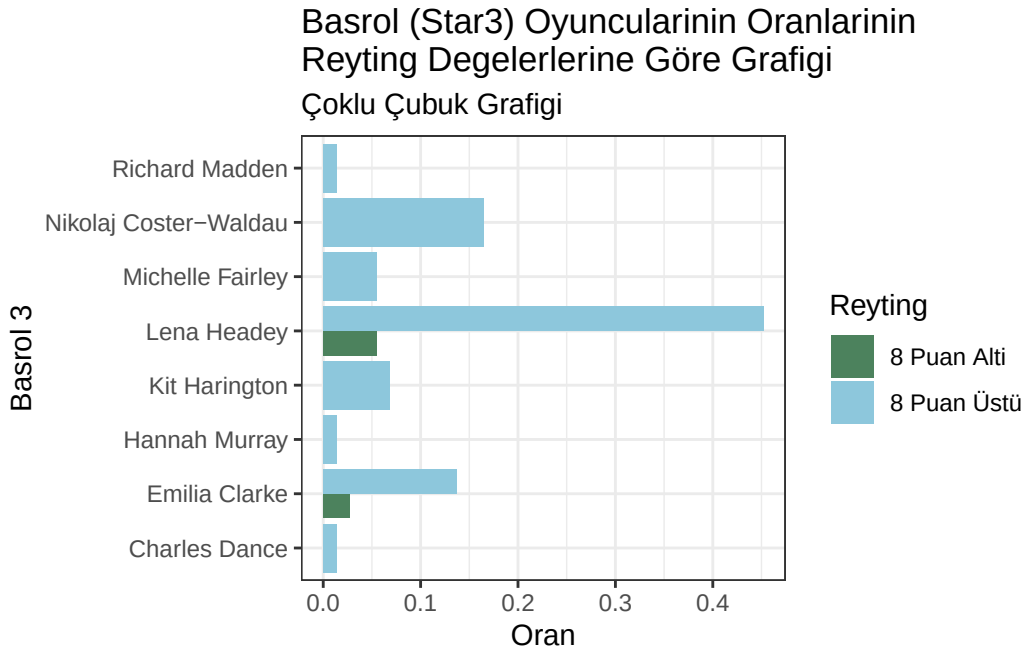
Başrol (Star2) oyuncularının oranları reyting değelerlerine göre "8 puan üstü", "8 puan altı" olmak üzere iki gruba ayrılmış ve çoklu çubuk grafiği ile görselleştirilmiştir. x ekseninde reyting değerleri, y ekseninde ise oranlar yer almaktadır. Renklendirmeler oyunculara göre yapılmıştır. Buradan reytingi 8 puan altında olan bölümlerde 2.başrol olan tek oyuncunun Nikolaj Coster-Waldau olduğu görülmektedir. Reytingi 8 puan üstünde olan bölümlerde yer alan 2.başrol oyuncularının oranına baktığımızda da Nikolaj Coster-Waldau'nun oranının 0.5'ten fazla olduğunu söyleyebiliriz. Onu Lena Headey takip etmektedir.

3.Başrol (Star3) oyuncularının oranlarını reyting değelerlerine göre (iki gruba ayırınız: 8 puan altı ve üstü) veri görselleştirme yöntemleriyle araştırınız.

```
GOT_episodes_v4 <- GOT_episodes_v4 %>%
  mutate(rating_got2 = ifelse(Rating >=8, "8 Puan Üstü", "8 Puan Alti"))
```

```
rating_got2 <- GOT_episodes_v4 %>%
  group_by(Star_3,rating_got2) %>%
  summarise(oran1 =n()/73, oran2 = n()/73)
```

```
ggplot(rating_got2, aes(x = oran2,
y = Star_3,
fill = rating_got2)) +
geom_bar(position = "dodge",
stat = "identity")+
labs(x = "Oran",
y = "Basrol 3",
fill = "Reyting",
title = "Basrol (Star3) Oyuncularinin Oranlarinin
Reyting Degelerlerine Göre Grafigi",
subtitle = "Çoklu Çubuk Grafigi") +
   scale_fill_manual(values = met.brewer("Pissaro",2)) +
theme_bw()
```



Yorum

Başrol (Star3) oyuncularının oranları reyting değelerlerine göre "8 puan üstü", "8 puan altı" olmak üzere iki gruba ayrılmış ve çoklu çubuk grafiği ile görselleştirilmiştir. x ekseninde reyting değerleri, y ekseninde ise oranlar yer almaktadır. Renklendirmeler reyting gruplarına göre yapılmıştır. Reytingi 8 puanın altında kalan bölümlerde 3.başrol olarak yer alan iki oyuncunun olduğu görülmektedir. Bunlar Emilia Clarke ve Lena Headey'dir. Kendi aralarında değerlendirmemiz gerekirse Lena Headey'in oranı Emilia Clarke'tan daha fazladır. Yani Lena Headey'in, reytingi 8'in altında olan ve 3. başrol olarak yer aldığı bölümlerin oranı Emilia Clarke 'a göre daha fazladır. Reytingi 8 puan üstünde olan bölümlerde yer alan 3.başrol oyuncularının oranına baktığımızda ise Lena Headey'in oranının 0.5'e yakın olduğunu söyleyebiliriz. Onu Nikolaj Coster-Waldau takip etmektedir.

Türk Televizyon Dizileri İle İlgili Verileri İçeren Veri Setinin İncelenip İlgili Konular Dahilinde Görselleştirilmesi

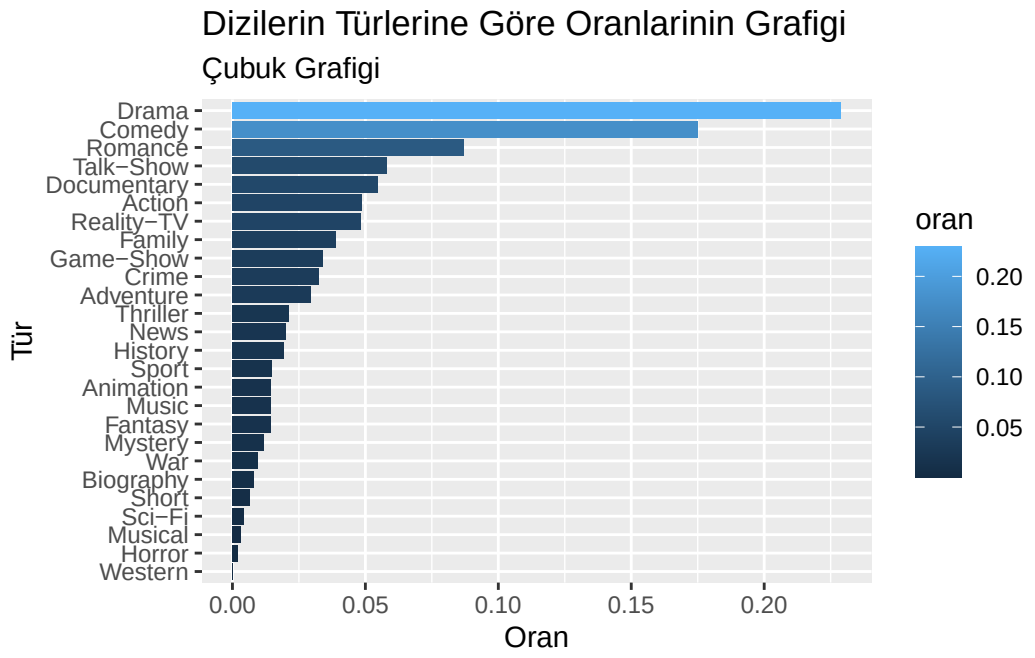
Ver seti kaggle.com dan alınmıştır. Geçmişten günümüze yayınlanmış televizyon dizilerini,dizilerin yayınlandıkları tarihleri, oyuncuları, sezon sayılarını vb. içermektedir. Bu veri setinde 103 değişken ve 2128 tane gözlem yer almaktadır.

1.Dizilerin türlerine göre oranlarına veri görselleştirme yöntemleriyle araştırınız. (Her bir türü grafikte ayrı ayrı kullanınız, oranları iki veya daha fazla tür üzerinden etiketlemeyiniz.)

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

```
türkdizi2 <- turkish_tvseries %>%
  tidyr::separate_rows(genres, sep = ", ") %>%
  drop_na(genres) %>%
  group_by(genres) %>%
  summarise(oran=n()/3276)
```

```
ggplot(türkdizi2, aes(x = reorder(genres, +oran),
                      y = oran,
                      fill = oran)) +
  geom_bar(stat="identity") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  labs(x = "Tür",
       y = "Oran",
       title = "Dizilerin Türlerine Göre Oranlarinin Grafigi",
       subtitle = "Çubuk Grafigi") +
  coord_flip()
```



Yorum

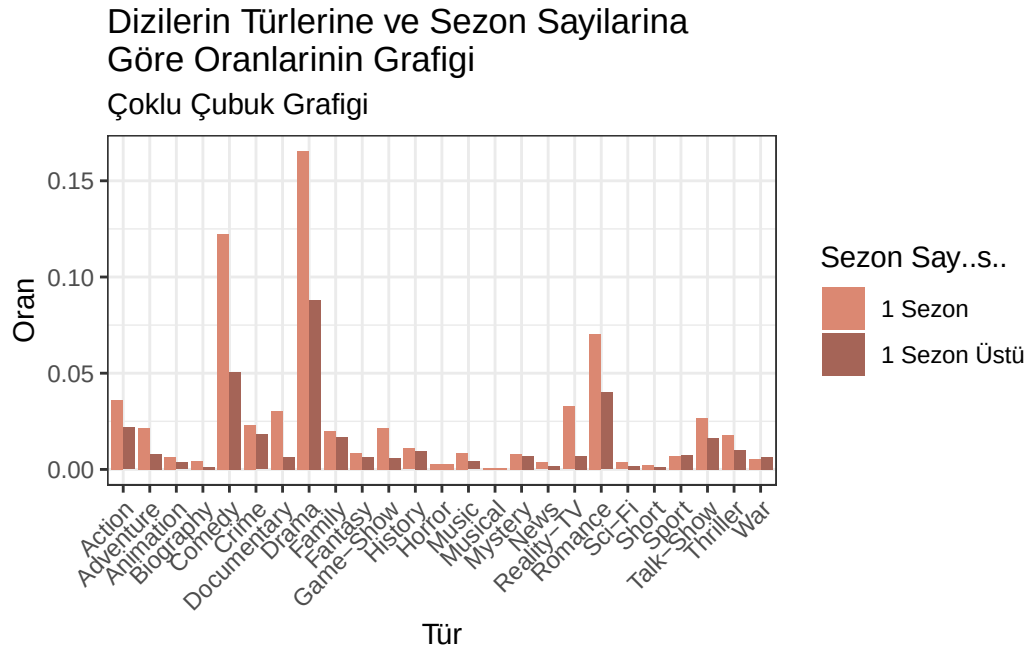
Dizilerin türlerine göre oranları çubuk grafiği ile görselleştirilmiştir. x ekseninde oran, y ekseninde ise dizilerin türleri yer almaktadır. Grafiğin daha rahat anlaşılması için azalan-oran şeklinde sıralanmıştır. Grafiğe bakarsak Dram türündeki dizilerin oranının 0.2'den fazla olduğunu ve en çok çekilen tür olduğunu söyleyebiliriz. Dram türünü ise Komedi ve Romantik diziler takip etmektedir. Oranı en az olan dizi türü ise Westerndir.

2.Dizilerin türlerine ve sezon sayılarına (iki gruba ayırınız: 1 sezon ve 1 sezondan fazla) göre oranlarına veri görselleştirme yöntemleriyle araştırınız. (Her bir türü grafikte ayrı ayrı kullanınız, oranları iki veya daha fazla tür üzerinden etiketlemeyiniz.)

```
turkish_tvseries <- turkish_tvseries %>%
  mutate(sezon_sayısı = ifelse(seasons == 1, "1 Sezon ", "1 Sezon Üstü"))
```

```
sezon_sayısı <- turkish_tvseries %>%
  tidyr::separate_rows(genres, sep = ", ") %>%
  group_by(genres,sezon_sayısı) %>%
   drop_na(genres,sezon_sayısı) %>%
  summarise(oran1 =n()/2200, oran2 = n()/2200)
```

```r
ggplot(sezon_sayısı, aes(x = oran2,
                         y = genres,
                         fill = sezon_sayısı)) +
  geom_bar(position = "dodge",
           stat = "identity")+
  labs(x = "Oran",
       y = "Tür",
       fill = "Sezon Sayısı",
       title = "Dizilerin Türlerine ve Sezon Sayilarina
Göre Oranlarinin Grafigi",
       subtitle = "Çoklu Çubuk Grafigi") +
  coord_flip()+
   theme_bw() +
  scale_fill_manual(values = met.brewer("Morgenstern",2)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



Yorum

Dizilerin türlerine ve sezon sayılarına (1 sezon ve 1 sezon üstü) göre oranları çoklu çubuk grafiği ile görselleştiirlmiştir. x eksenine dizi türleri, y eksenine ise oran konumlandırılmıştır. Renklendirmeler ise sezon sayısına göre yapılmıştır. Grafiğe bakıldığında her iki sezon grubunda da en fazla orana sahip dizi türünün Dram olduğu görülmektedir. Korku ve Müzikal türündeki

dizilerin ise sadece 1 sezon sürdüğünü söyleyebiliriz. İkisini karşılaştırmak gerekirse, Korku türündeki dizilerinin oranı Müzikal türündeki dizilerinin oranından fazladır.