3 tane veri setinin incelenmesi ve görselleştirilmesi

Ahmet Batuhan Özdoğan

25.11.2022

Özet

Bu ödevde Kaggle sitesinden alınmış 3 farklı veri setinin görselleştirme ve yorumlama çalışmaları bulunmaktadır. Birinci veri seti Breaking Bad dizisi ,İkinci veri seti En çok satan kitaplar ve üçüncü veri seti ise Marvel filmleri olarak kullanılmıştır

Gerekli Kütüphanlerin yüklenmesi

```
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("dplyr")
install.packages("ggridges")
install.packages("readxl")
library(ggridges)
library(ggplot2)
library(tidyverse)
library(dplyr)
```

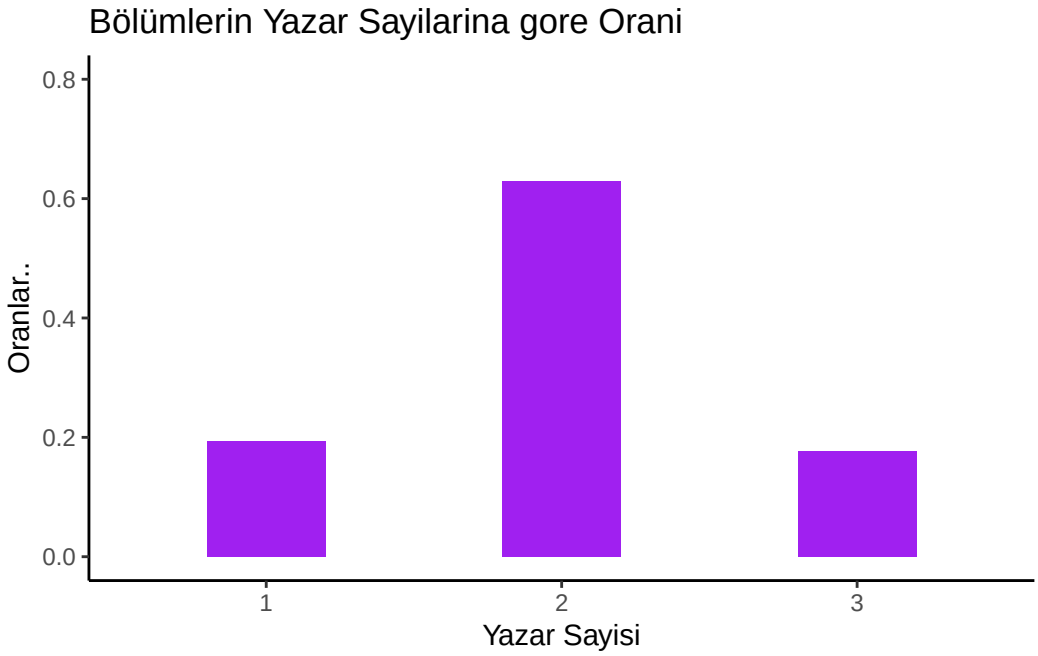
1. veri

Veri Setinin Hakkında

Bu veri setimizde Breaking bad adlı dizimizin incelemesi ve görselleştirmesi bulunmaktadır

Bölümlerin yazar sayısına göre oranları

```
breaking_bad <- read_csv("breaking_bad.csv")
breaking_bad <- breaking_bad %>%
add_column(imdb =if_else(breaking_bad$Rating_IMDB < 9, "1", "2"))
bbveri <- breaking_bad %>%
tidyr::separate_rows(`Written by`, sep = ", ") %>%
group_by(Season , Episode , Rating_IMDB , imdb) %>%
summarise(yazar = n())%>%
mutate(oran = frequency(yazar)/nrow(breaking_bad))
bbveri <- bbveri%>%
add_column(abc =
if_else(bbveri$imdb == "1", -bbveri$oran, bbveri$oran ))
ggplot(bbveri, aes(x = as.factor(yazar) ,
y=oran))+
geom_bar(stat = "identity",
fill = "purple",
width = 0.4)+
labs(x = "Yazar Sayisi",
y = "Oranları",
title = "Bölümlerin Yazar Sayilarina gore Orani")+
  ylim(0,0.8) +
theme_classic()
```

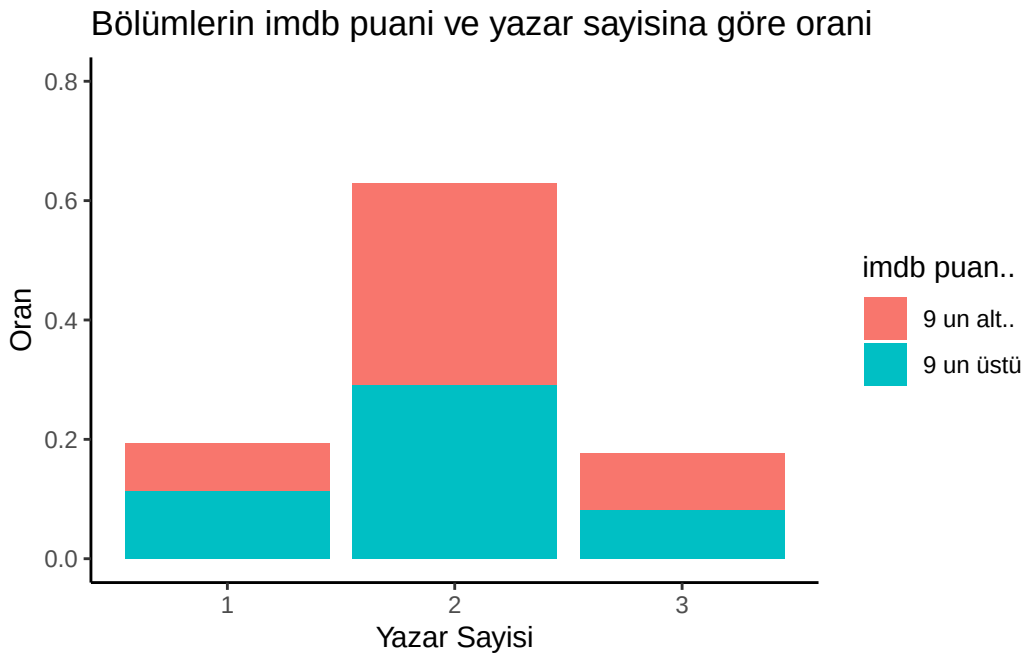


Rapor

Grafiğe Baktığımzda, genel olarak 2 kişi tarafından yazılan dizi bölümlerimizin tek ve 3 kişi tarafından yazılan bölümlerden daha fazla olduğu görülmektedir

Bölümlerin yazar sayısına ve IMDB puanlarına göre oranlarını

```
ggplot(bbveri, aes(x = as.character(yazar), fill = as.factor(imdb),
y=(frequency(yazar)/nrow(bbveri))))+
geom_bar(stat="identity") +
labs(x = "Yazar Sayisi",
y = "Oran",
title = "Bölümlerin imdb puani ve yazar sayisina göre orani",
fill="imdb puanı")+
ylim(0,0.8)+
scale_fill_discrete(labels = c("9 un altı", "9 un üstü")) +
  theme_classic()
```



Rapor

Grafikteki sonuçlara göre, imdb puanının 9 altında olan ve 9 üstünde olan bölümlerin oran farkının en çok 1 yazarı olan bölümlerde olduğunu söyleyebiliriz ve en düşük IMDB punlarının 3 yazarı olan bölümlerde olduğuda gözükmektedir

İkinci veri

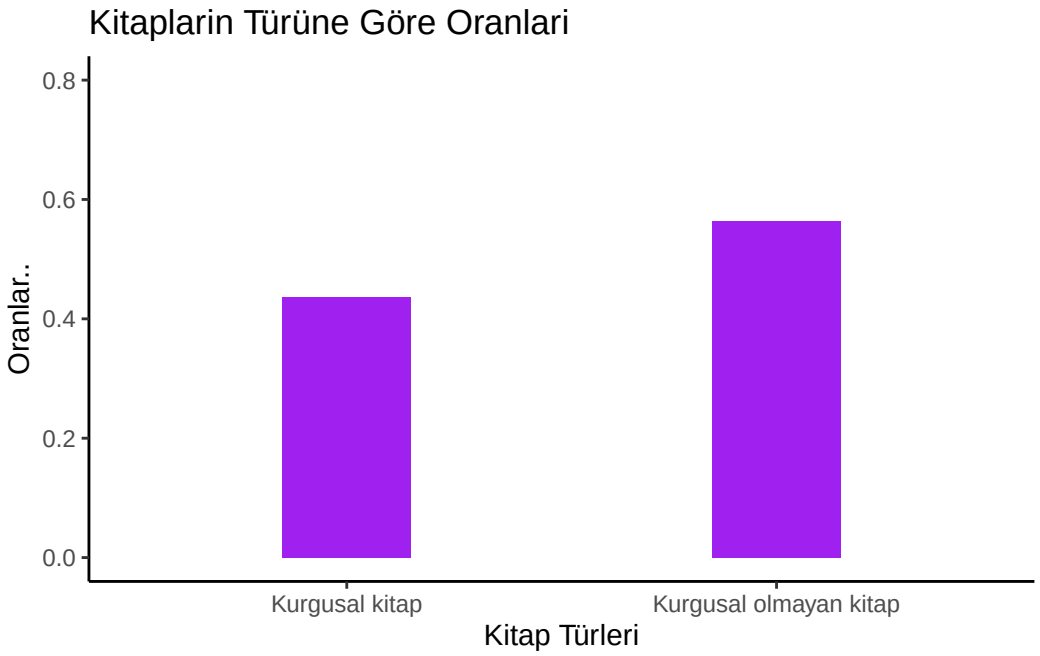
Veri Setinin Hakkında

Bu veri seti 2009-2019 yılları arasında Amazon'da en çok satan 50 kitabı içermektedir.

"En Çok Satan Kitaplarin Türlerine göre Oranları

```
library(readr)
kitap <- read_csv("bestsellers with categories.csv")
kitap1 <- kitap %>%
group_by(Genre) %>%
summarise(sayi = n()) %>%
mutate(oran = sayi / sum(sayi))
```

```r
ggplot(kitap1, aes(x = Genre, y = oran)) +
geom_bar(stat = "identity", width = 0.3,
fill = "purple") +
labs(x = "Kitap Türleri", y = "Oranları",
title = "Kitaplarin Türüne Göre Oranlari") +
scale_x_discrete(labels = c("Kurgusal kitap","Kurgusal olmayan kitap"))+
    ylim(0,0.8) +
theme_classic()
```



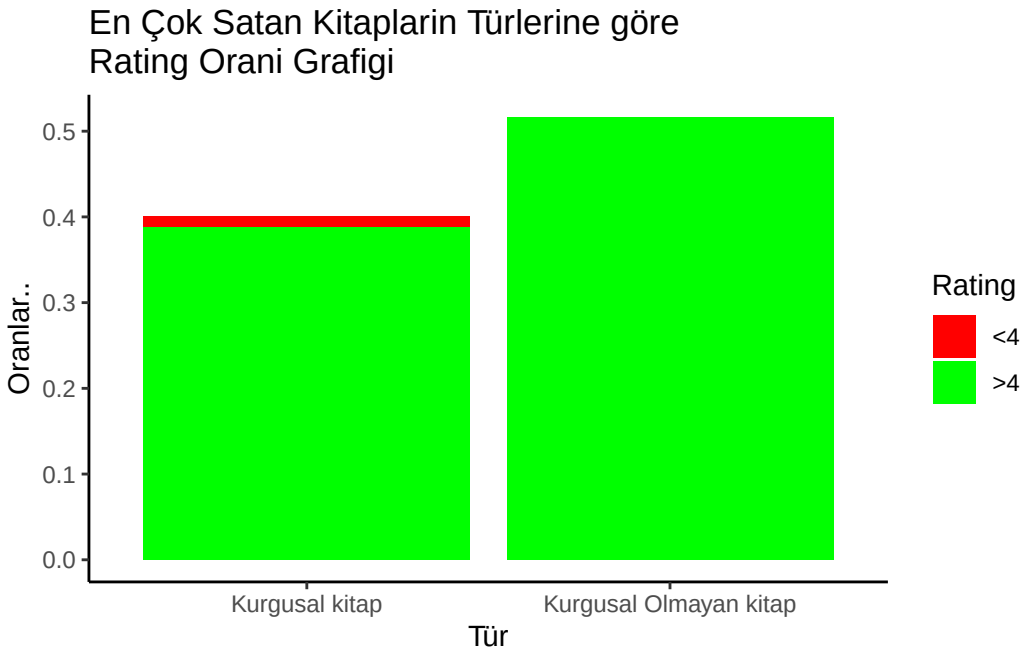
Rapor

Grafikteki sonuçlara göre, kurgusal olmayan kitapların kurgasal olan kitaplara nazaran daha fazla sayıda olduğu söylenebilir

En Çok Satan Kitaplarin Türlerine göre Rating Oranları

```r
library(readr)
kitap <- read_csv("bestsellers with categories.csv")
kitap <- kitap %>%
```

```r
mutate(kitap3 = ifelse(`User Rating` >=4, ">4", "<4"))
kitap2 <- kitap %>%
group_by(Genre,kitap3) %>%
summarise(oran1 =n()/600, oran2 = n()/600)
ggplot(kitap2, aes(fill = kitap3,
                   y = oran2,
                x = Genre)) +
geom_bar(position = "stack",
stat = "identity") + scale_fill_manual(values = c("red", "green")) +
labs(x = "Tür",
y = "Oranları",
fill = "Rating",
title = "En Çok Satan Kitaplarin Türlerine göre
Rating Orani Grafigi") +
scale_x_discrete(labels = c("Kurgusal kitap", "Kurgusal Olmayan kitap")) +
theme_classic()
```



Rapor

Grafikteki sonuçlarımıza göre, kurgusal olmayan kitapların rating oranı 4 ün altında olmayan bir eser yok ancak kurgusal olan bazı eserlerde 4 ün altında rating i olduğu gözlemlenebilmektedir

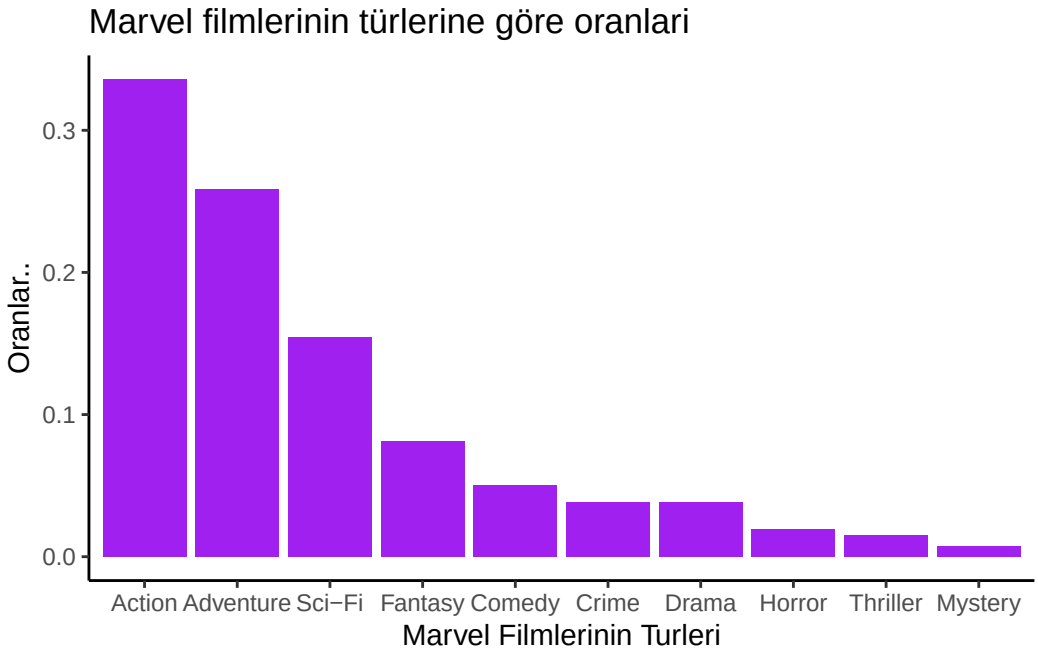
Üçüncü veri

Veri Setinin Hakkında

Bu veri seti Marvel Filmleri hakkında bilgiler içermektedir

Marvel filmlerinin türlerine göre oranları

```
marv = read.csv("mdc.csv")
marvel = marv %>%
tidyr::separate_rows(genre, sep = ", ") %>%
group_by(genre)
marvel = marvel %>%
group_by(genre) %>%
summarise(number = n()) %>%
mutate(oran = number / sum(number))
ggplot(marvel) +
aes(x =reorder(genre, -oran), weight = oran) +
geom_bar(fill = "purple") +
labs(
x = "Marvel Filmlerinin Turleri",
y = "Oranları",
title = "Marvel filmlerinin türlerine göre oranlari"
) +
theme_classic()
```



Rapor

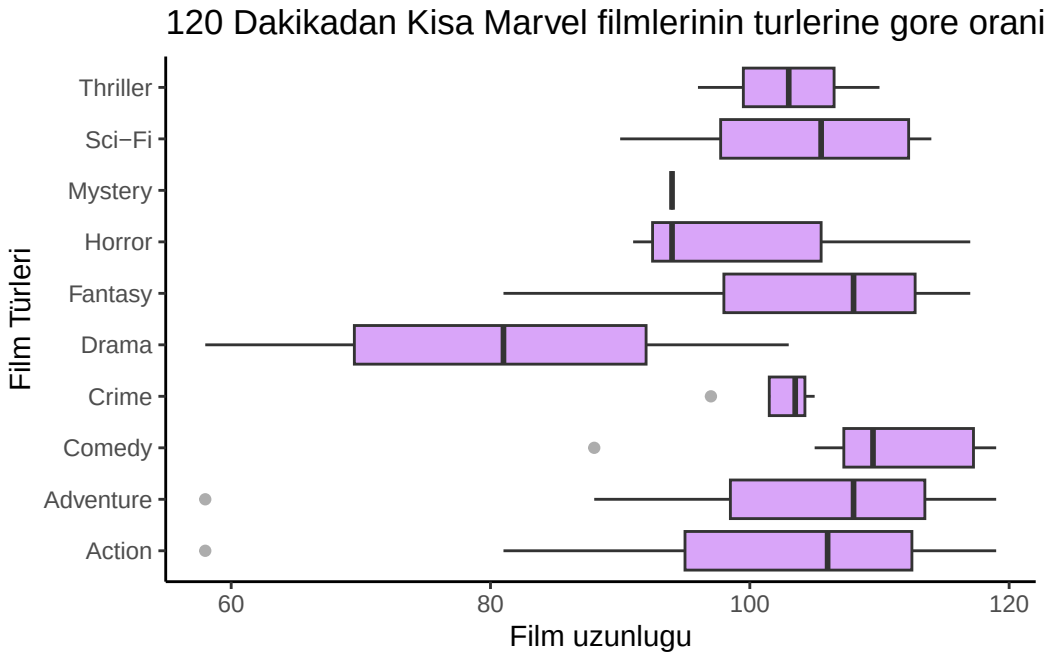
Grafikteki sonuçlarımıza göre, en çok filmin bulunduğu kategori aksiyon ve an az filmin bulunduğu kategori ise Gizem olarak gözükmektedir genel olarak baktığımızda ise aksiyon ve macera kategorisi filmlerinin bütün filmlere oranla daha fazla olduğu görülmektedir

Marvel filmlerinin sürelerine ve türlerine göre oranları

120 dakikadan kısa olan filmler ve türleri

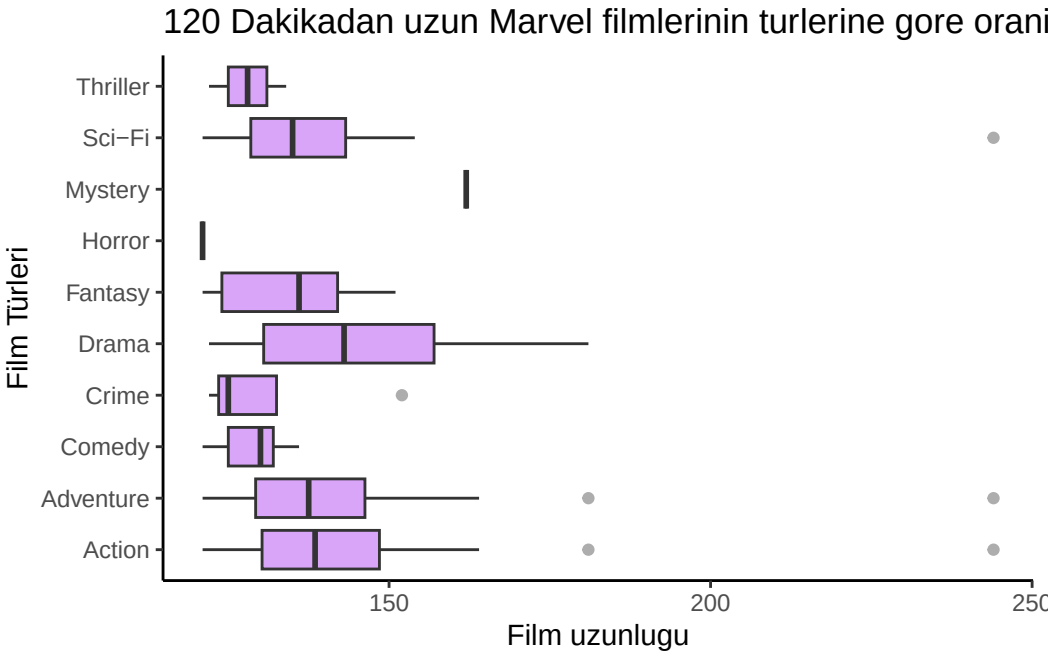
```
marvelsüre = marv %>%
tidyr::separate_rows(genre, sep = ", ") %>%
group_by(genre,runtime) %>%
summarise(n())
marvelsüre =filter(marvelsüre, runtime <120)
ggplot(marvelsüre) +
aes(x = runtime, y = genre) +
geom_boxplot(alpha=0.4 , fill="purple")+
labs(
x = "Film uzunlugu",
```

```
y = "Film Türleri",
title = "120 Dakikadan Kisa Marvel filmlerinin turlerine gore orani") +
theme_classic()
```



120 dakikadan uzun olan filmler ve türleri

```
Marvelsüre2 = marv %>%
tidyr::separate_rows(genre, sep = ", ") %>%
group_by(genre,runtime) %>%
summarise(n())
marvelsüre2 =filter(Marvelsüre2, runtime >120)
ggplot(marvelsüre2) +
aes(x = runtime, y = genre) +
geom_boxplot(alpha=0.4 , fill="purple")+
labs(
x = "Film uzunlugu",
y = "Film Türleri",
title = "120 Dakikadan uzun Marvel filmlerinin turlerine gore orani") +
theme_classic()
```



Rapor

Grafikteki sonuçlarımıza bakıcak olur ise genel olarak marvel filmlerinin büyük bir kısmının 120 dakinanın üstünde olduğu görülmektedir Action, Adventure ve Sci-Fi,türündeki filmlerin süreleri genellikle 120 dakikanın üstündedir onun aksine marvelin yapmış olduğu Comedy, crime vede horror türlerinde de filmlerin süresi büyük oranda 120 dakikanın altındadır.