IMDb'nin En İyi 2000 Filmi Üzerine Puanları Etkileyen Faktörlerin Analizi

Mehmet Burak Sakallıoğlu

22 Haziran 2024

1. Veri Tanımı

Veri Kümesinin Tanımlanması

```
library(readxl)
dosya_yolu <- "C:\\Users\\Burak\\Desktop\\Veri Analizi\\IMDB'nin En İyi 2000 Filmi.xlsx"
filmler <- read_excel(dosya_yolu)
```

Veri Setine Genel Bakış

```
head(filmler)
```

```
# A tibble: 6 × 10
  FilmAdi       CikisYili  Sure IMDBPuan MetaPuan OySayisi Tur   Yonetmen Bas
rol
  <chr>             <dbl> <dbl>    <dbl>    <dbl>    <dbl> <chr> <chr>    <ch
r>
1 The Godfather      1972   175      9.2      100  2002655 Suç,… Francis… Mar
lo…
2 The Godfathe…      1974   202      9         90  1358608 Suç,… Francis… Al
Pa…
3 Ordiry People      1980   124      7.7       86    56476 Dram  Robert … Dol
d …
4 Lawrence of …      1962   218      8.3      100   313044 Mace… David L… Pet
er…
5 Straw Dogs         1971   113      7.4       73    64331 Suç,… Sam Pec… Dus
ti…
6 Close Encoun…      1977   138      7.6       90   216050 Dram… Steven … Ric
ha…
# ℹ 1 more variable: Hasilat <chr>
```

`head(filmler)` kodu, `filmler` isimli veri çerçevesinin ilk 6 satırını gösterir. Bu sayede veri setinin genel yapısı ve içerdiği bilgiler hakkında hızlı bir önizleme elde ederiz.

```
dim(filmler)
```

```
[1] 2000   10
```

dim fonksiyonu kullanarak, veri setinde 2000 satır ve 10 sütün olduğu görülüyor.

```
names(filmler)
```

```
 [1] "FilmAdi"   "CikisYili" "Sure"      "IMDBPuan"  "MetaPuan"  "OySayisi"
 [7] "Tur"       "Yonetmen"  "Basrol"    "Hasilat"
```

names fonksiyonu ile değişkenler inceleniyor.

Veri Kümesinin Yapısını İnceleme

```
str(filmler)
```

```
tibble [2,000 × 10] (S3: tbl_df/tbl/data.frame)
 $ FilmAdi  : chr [1:2000] "The Godfather" "The Godfather Part II" "Ordiry Pe
ople" "Lawrence of Arabia" ...
 $ CikisYili: num [1:2000] 1972 1974 1980 1962 1971 ...
 $ Sure     : num [1:2000] 175 202 124 218 113 138 166 150 137 124 ...
 $ IMDBPuan : num [1:2000] 9.2 9 7.7 8.3 7.4 7.6 8.5 7.7 8 7.8 ...
 $ MetaPuan : num [1:2000] 100 90 86 100 73 90 82 73 96 80 ...
 $ OySayisi : num [1:2000] 2002655 1358608 56476 313044 64331 ...
 $ Tur      : chr [1:2000] "Suç, Dram" "Suç, Dram" "Dram" "Macera, Biyografi,
Dram" ...
 $ Yonetmen : chr [1:2000] "Francis Ford Coppola" "Francis Ford Coppola" "Rob
ert Redford" "David Lean" ...
 $ Basrol   : chr [1:2000] "Marlon Brando" "Al Pacino" "Dold Sutherland" "Pet
er O'Toole" ...
 $ Hasilat  : chr [1:2000] "$134.97M" "$57.30M" "$54.80M" "$44.82M" ...
```

Veri seti, IMDB'nin en iyi 2000 filmi hakkında bilgiler içermektedir. Her bir film için film adı, çıkış yılı, süresi, IMDB puanı, Meta puanı, oy sayısı, türü, yönetmeni, başrol oyuncusu ve hasılat bilgileri bulunmaktadır.

Değişken tiplerine bakıldığında `CikisYili`, `Sure`, `IMDBPuan`, `MetaPuan`, ve `OySayisi` değişkenlerinin numerik olduğu, diğer değişkenlerin ise karakter tipinde olduğu görülmektedir. `Hasilat` değişkeni karakter tipinde olduğundan analiz öncesinde bu değişkenin numerik formata dönüştürülmesi gerekecektir.

Değişkenler

Veri setinde 2000 gözlem ve 10 değişken bulunmaktadır.

- `FilmAdi`: Filmin adı (karakter dizisi)
- `CikisYili`: Filmin çıkış yılı (sayısal)
- `Sure`: Filmin süresi (sayısal)
- `IMDBPuan`: Filmin IMDB puanı (sayısal)
- `MetaPuan`: Filmin Metascore puanı (sayısal)
- `OySayisi`: Film için oy kullanan kişi sayısı (sayısal)
- `Tur`: Filmin türü/türleri (karakter dizisi)
- `Yonetmen`: Filmin yönetmeni (karakter dizisi)
- `Basrol`: Filmin başrol oyuncusu (karakter dizisi)
- `Hasilat`: Filmin gişe hasılatı (karakter dizisi)

Veri Kümesindeki Değişkenlerin Özet İstatistikleri

```
summary(filmler)
```

```
   FilmAdi            CikisYili         Sure          IMDBPuan
 Length:2000        Min.   :1921   Min.   : 50.0   Min.   :1.500
 Class :character   1st Qu.:1992   1st Qu.: 98.0   1st Qu.:6.400
 Mode  :character   Median :2001   Median :110.0   Median :7.000
                    Mean   :1996   Mean   :113.9   Mean   :6.923
                    3rd Qu.:2006   3rd Qu.:125.0   3rd Qu.:7.600
                    Max.   :2010   Max.   :271.0   Max.   :9.300

    MetaPuan         OySayisi           Tur              Yonetmen
 Min.   :  9.00   Min.   :   1883   Length:2000        Length:2000
 1st Qu.: 48.00   1st Qu.:  79098   Class :character   Class :character
 Median : 61.00   Median : 135312   Mode  :character   Mode  :character
 Mean   : 61.04   Mean   : 223895
 3rd Qu.: 74.00   3rd Qu.: 252134
 Max.   :100.00   Max.   :2875249
 NA's   :81
    Basrol            Hasilat
 Length:2000        Length:2000
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Bu çıktı, `filmler` veri setindeki değişkenlerin özet istatistiklerini göstermektedir.

"Hasilat" Değişkenini Sayısal Hale Getirme

İleri aşamalardaki analizler için "Hasilat" değişkeni sayısal hale getirilmelidir.

```
library(dplyr)
library(stringr)
filmler <- filmler %>%
  mutate(Hasilat = gsub("M", "e6", Hasilat)) %>%
  mutate(Hasilat = gsub("\\$|\\,", "", Hasilat)) %>%
  mutate(Hasilat = as.numeric(Hasilat))
```

2. Problemin Tanımı ve Amaçlar

Problemin Tanımı

Film endüstrisi, dünya genelinde milyarlarca dolarlık bir sektördür ve filmlerin başarısı, gişe hasılatı, eleştirel beğeni ve izleyici puanları gibi çeşitli faktörlerle ölçülür. Bu faktörleri etkileyen unsurları anlamak, hem film yapımcıları hem de izleyiciler için büyük önem taşır.

Bu projede ele alınan problem, IMDB'nin En İyi 2000 Filmi veri seti üzerinden filmlerin IMDB puanlarını etkileyen faktörlerin belirlenmesidir. Bu problem, aşağıdaki sorulara yanıt aramayı amaçlar:

- Hangi film türleri daha yüksek IMDB puanları almaktadır?
- Belirli yönetmenlerin filmleri daha mı başarılıdır?
- Filmin süresi, çıkış yılı veya hasılatı IMDB puanını etkiliyor mu?
- IMDB puanı ile Metascore puanı arasında bir ilişki var mı?

Amaçlar

Bu projenin temel amacı, IMDB puanlarını etkileyen faktörleri belirleyerek film endüstrisi için değerli bilgiler sağlamaktır. Bu bilgiler, aşağıdaki amaçlar için kullanılabilir:

- **Film yapımcıları:** Hangi tür filmlerin, yönetmenlerin veya oyuncuların daha başarılı olduğunu anlayarak gelecekteki projelerini şekillendirebilirler.
- **İzleyiciler:** Beğenebilecekleri filmleri seçerken daha bilinçli kararlar verebilirler.
- **Eleştirmenler:** Film incelemelerini yaparken daha objektif kriterler kullanabilirler.
- **Akademisyenler:** Film endüstrisi ve izleyici davranışları üzerine daha derinlemesine araştırmalar yapabilirler.

3. Verilerin Toplanması

Veri seti Kaggle'dan alınmıştır.

Kaynakça: Sawhney, P. (2023, November 22). *IMDb Dataset - Top 2000 Movies*. Kaggle. https://www.kaggle.com/datasets/prishasawhney/imdb-dataset-top-2000-movies/data

Verilerin nasıl toplandığına dair bilgi bulunmamaktadır.

4. Verilerin Yapısı ve Niteliği

Analizde Kullanılmayacak Değişkenlerin Çıkarılması

```
filmler <- subset(filmler, select = -c(OySayisi, Basrol))
```

Analizde kullanılmayacak olan değişkenler çıkartılmıştır.

```
colnames(filmler)
```

```
[1] "FilmAdi"   "CikisYili" "Sure"      "IMDBPuan"  "MetaPuan"  "Tur"
[7] "Yonetmen"  "Hasilat"
```

colnames() fonksiyonu yardımıyla da gözlemlendiği üzere işlem başarılı olmuştur.

Aykırı/Uç Değerlerin İncelenmesi

Tanımlayıcı İstatistikler

```
summary(filmler[, c("CikisYili", "Sure", "IMDBPuan", "MetaPuan", "Hasilat")])
```

```
  CikisYili         Sure          IMDBPuan        MetaPuan
 Min.   :1921   Min.   : 50.0   Min.   :1.500   Min.   :  9.00
 1st Qu.:1992   1st Qu.: 98.0   1st Qu.:6.400   1st Qu.: 48.00
 Median :2001   Median :110.0   Median :7.000   Median : 61.00
 Mean   :1996   Mean   :113.9   Mean   :6.923   Mean   : 61.04
 3rd Qu.:2006   3rd Qu.:125.0   3rd Qu.:7.600   3rd Qu.: 74.00
 Max.   :2010   Max.   :271.0   Max.   :9.300   Max.   :100.00
                                                NA's   :81
    Hasilat
 Min.   :        0
 1st Qu.: 18220000
 Median : 44820000
 Mean   : 66186358
 3rd Qu.: 87070000
 Max.   :760510000
 NA's   :97
```

Sayısal değişkenlerin genel istatistikleri verilmiştir.

Sıklık Dağılımları

Çıkış yılı için sıklık tablosu

```
table(filmler$CikisYili)


1921 1925 1926 1927 1931 1932 1933 1934 1936 1937 1939 1940 1941 1944 1945 19
46
   1    1    1    1    2    1    1    1    1    1    4    3    2    2    1
3
1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 19
63
   3    1    3    3    3    4    6    3    4    6    4    5    8    7    7
7
1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 19
79
   9    4    6    6   11    5    4   10    6   13    8    7    9    6    9
13
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 19
95
  12   13   18   20   22   22   19   27   26   33   34   24   30   43   38
52
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
  43   58   62   69   67   78   78   90  108  109  125  131  127  130   66
```

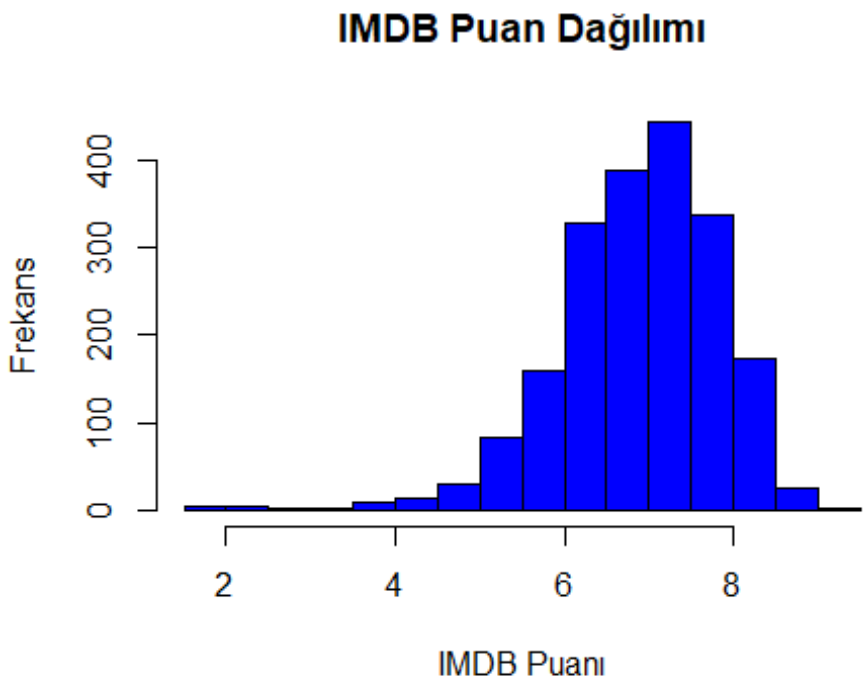
Bu sıklık dağılımı, film çıkış yıllarının zaman içindeki değişimini ve genel trendleri açıkça göstermektedir.

Artan Üretim: Genel olarak, 20. yüzyılın başlarından itibaren film çıkış sayılarında düzenli bir artış gözlenmektedir. Özellikle 1980'lerden sonra, film endüstrisinin büyük bir

genişleme yaşadığı söylenebilir. Değişkenlerdeki aykırı değerler incelendiğinde hatalı ölçümlere rastlanmadığı, bu yüzden ayıklama işlemine gerek olmadığı görülmüştür.

IMDB puanı için Histogram grafiği

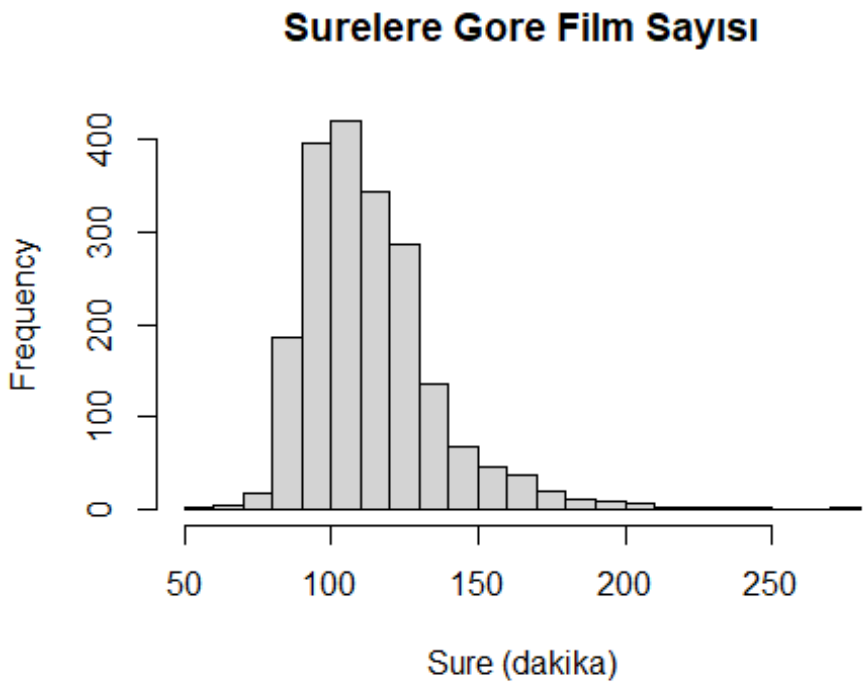
```
hist(filmler$IMDBPuan,
main = "IMDB Puan Dağılımı",
xlab = "IMDB Puanı",
ylab = "Frekans",
col = "blue",
border = "black")
```



Bu grafikte 2 ve 4 puan arasındaki düşük puanlar ile 9 ve üstü yüksek puanlar aykırı değerler olarak değerlendirilebilir. Bu puanlar, veri setinin genel dağılımından belirgin şekilde sapmaktadır ve nadir görülen puanlardır.

Süre için histogram

```
hist(filmler$Sure, breaks = 20, main = "Surelere Gore Film Sayısı", xlab = "S
ure (dakika)")
```

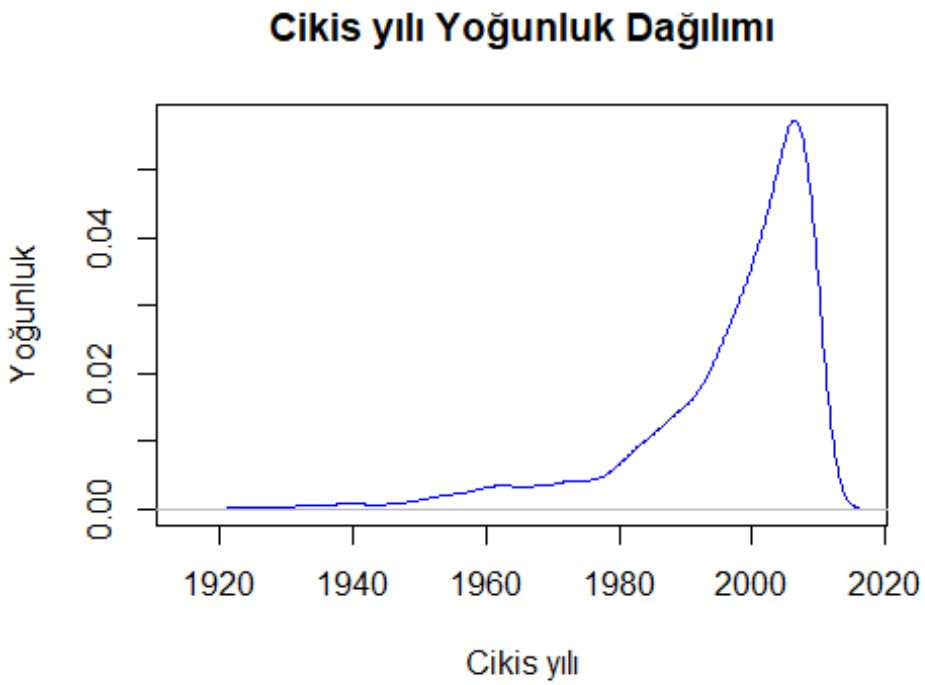


Grafiğe göre, filmlerin süreleri genellikle 100-150 dakika arasında yoğunlaşmaktadır. En yüksek frekans, yaklaşık 100-120 dakika arasındaki filmlerde görülmektedir. Ancak, grafikte 200 dakikayı geçen filmlerin sayısının oldukça az olduğu ve nadir görülen değerler olduğu anlaşılmaktadır.

Uç değerler (aykırı değerler), veri setinin genel eğiliminden belirgin şekilde farklı olan değerlerdir. Bu grafikte, özellikle 200 dakikayı aşan filmler uç değer olarak kabul edilebilir. Çünkü bu sürelerdeki film sayısı, diğer aralıklara göre oldukça düşük ve veri setinin büyük çoğunluğundan farklı bir konumdadır.

Çıkış yılı için yoğunluk grafiği

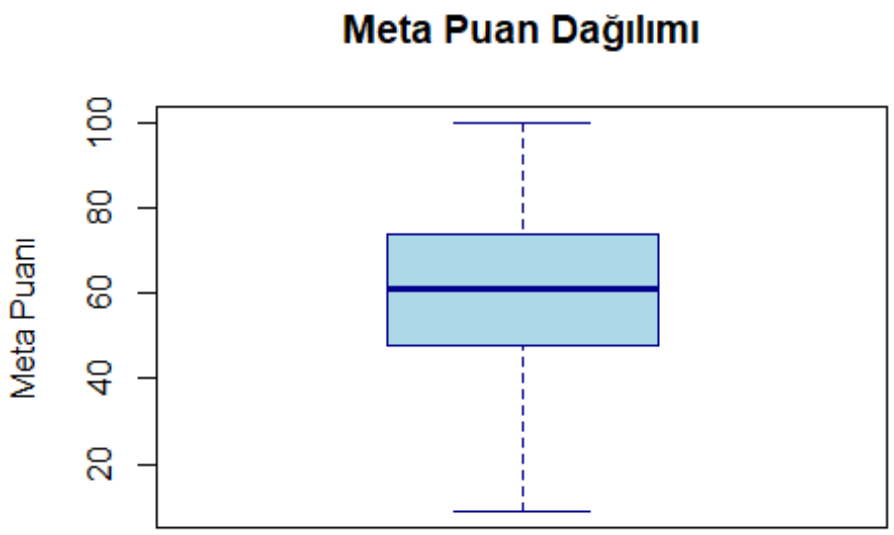
```
plot(density(filmler$CikisYili),
main = "Cikis yılı Yoğunluk Dağılımı",
xlab = "Cikis yılı",
 ylab = "Yoğunluk",
 col = "blue")
```



Grafikte, 1920'lerden 1980'lere kadar olan yıllarda, yoğunluk diğer yıllara göre oldukça düşük. Bu, bu dönemde daha az film yapıldığını veya veri setinde bu dönemdeki filmlerle ilgili eksiklik olduğunu gösteriyor olabilir.

Meta puanı için kutu grafiği

```
boxplot(filmler$MetaPuan,
 main = "Meta Puan Dağılımı",
 ylab = "Meta Puanı",
 col = "lightblue",
 border = "darkblue")
```

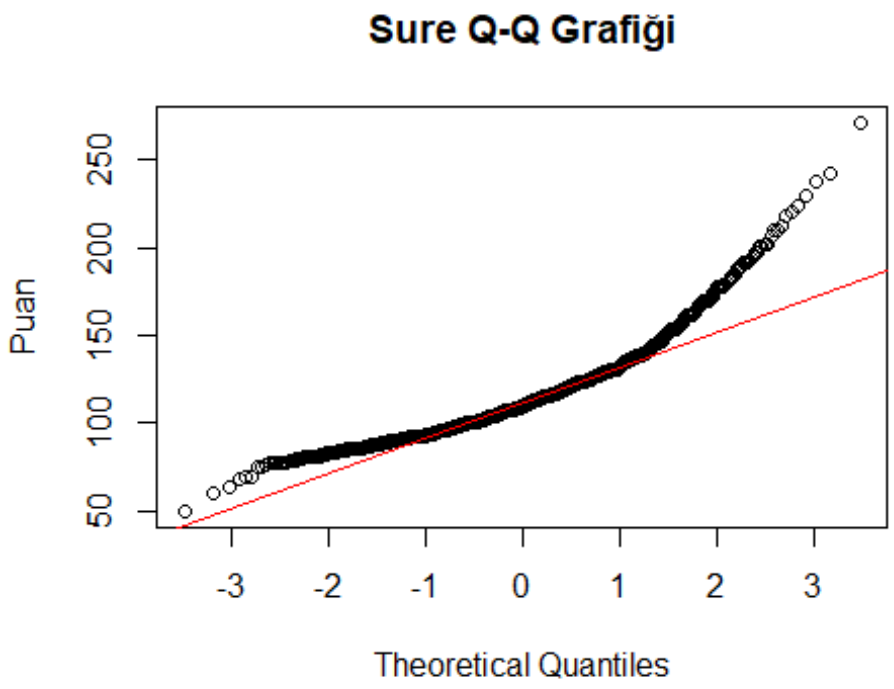


Bu grafikte aykırı (uç) değerler bulunmamaktadır çünkü kutunun dışındaki bölgelerin arasında herhangi bir nokta yoktur. Bu durum, verinin aykırı (uç) değer içermediğini veya tüm değerlerin kabul edilebilir bir aralıkta olduğunu gösterir.

Özetle, grafiğe göre Meta Puan Dağılımı'nda herhangi bir aykırı (uç) değer tespit edilmemiştir. Veriler, 20 ile 100 arasında düzenli bir şekilde dağılmıştır.

Süre için Q-Q grafiği

```
qqnorm(filmler$Sure,
 main = "Sure Q-Q Grafiği",
 ylab = "Puan")
 qqline(filmler$Sure, col = "red")
```



Q-Q grafiğinde, noktaların kırmızı çizgiye göre dağılımı, özellikle sağ uçta, normal dağılımdan sapıyor. Bu, verilerde bazı uç değerlerin olduğunu gösterebilir.

Eksik Gözlemlerin Varlığının Kontrolü

```
any(is.na(filmler))
```

```
[1] TRUE
```

Bu veride eksik değerler vardır.

Eksik Gözlemlerin Sayısı

```
sum(is.na(filmler))
```

```
[1] 178
```

Bu veride toplam 178 eksik değer vardır.

Eksik Gözlemlerin Hangi Sütunlarda Olduğunu Bulma

```
eksik_degerler <- colSums(is.na(filmler))
eksik_deger_yuzdeleri <- colMeans(is.na(filmler)) * 100
print(eksik_degerler)
```

```
  FilmAdi CikisYili      Sure  IMDBPuan  MetaPuan       Tur  Yonetmen   Hasil
at
        0         0         0         0        81         0         0
97

print(eksik_deger_yuzdeleri)
```

```
  FilmAdi CikisYili      Sure  IMDBPuan  MetaPuan       Tur  Yonetmen   Hasil
at
     0.00      0.00      0.00      0.00      4.05      0.00      0.00      4.
85
```

Bu çıktı, veri setindeki eksik değerlerin (NA) sayısını ve yüzdesini göstermektedir. İlk satırda her bir değişken için eksik değer sayısı, ikinci satırda ise eksik değer yüzdesi verilmiştir. Örneğin, `MetaPuan` değişkeninde 81 tane eksik değer bulunmaktadır ve bu veri setinin %4.05'ine denk gelmektedir. `Hasilat` değişkeninde ise 97 tane eksik değer bulunmaktadır ve bu veri setinin %4.85'ine denk gelmektedir.

Hangi Sütunlarda Eksik Gözlem Olduğunu Listeleme

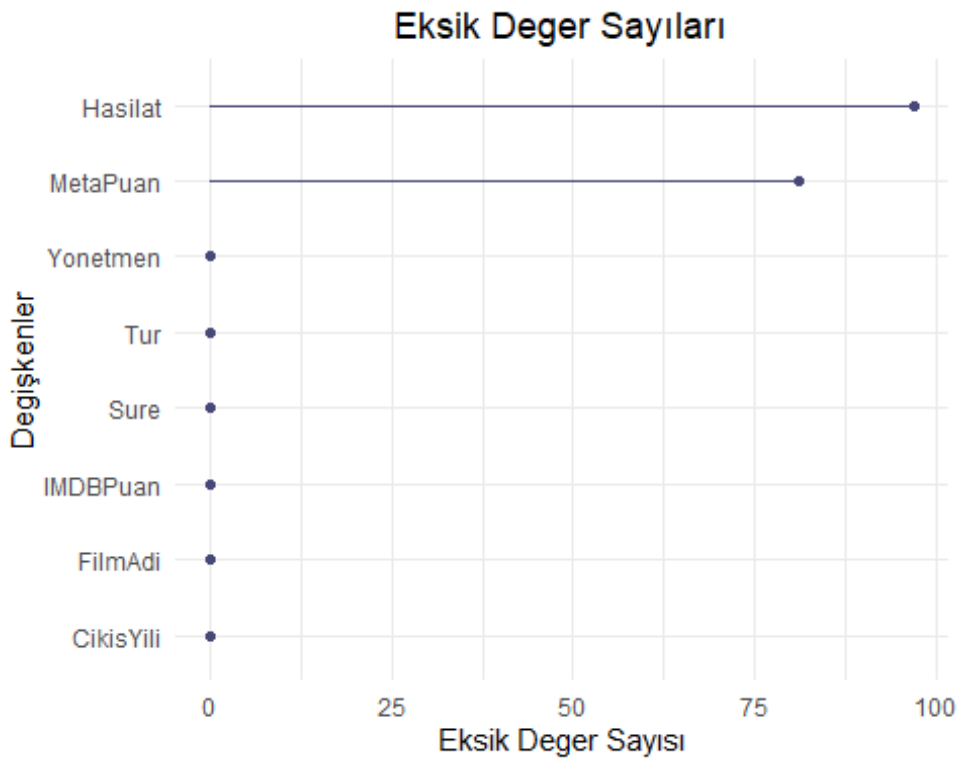
```
for (i in seq_along(filmler)) {
  if (any(is.na(filmler[[i]]))) {
    cat("Sutun", names(filmler)[i], "satırlarda eksik degerler var:", which(i
s.na(filmler[[i]])), "\n")
  }
}
```

```
Sutun MetaPuan satırlarda eksik degerler var: 29 46 72 91 114 163 170 172 175
180 181 208 210 211 213 471 545 553 570 577 604 652 657 671 683 686 697 713 7
49 753 863 876 919 921 941 947 953 967 1000 1148 1194 1362 1415 1419 1420 142
8 1437 1493 1496 1557 1596 1601 1603 1644 1658 1671 1672 1679 1682 1688 1691
1701 1726 1737 1738 1739 1818 1833 1857 1860 1891 1893 1895 1910 1928 1932 19
34 1941 1944 1947 1952
Sutun Hasilat satırlarda eksik degerler var: 5 9 17 18 29 43 44 49 54 55 67 7
2 80 81 91 99 100 103 113 136 137 141 145 147 154 164 167 171 173 177 180 185
194 196 200 201 204 210 219 230 231 234 237 240 245 464 570 667 753 919 935 9
36 941 947 953 1000 1148 1194 1362 1415 1428 1437 1481 1493 1496 1557 1572 16
01 1603 1644 1658 1671 1672 1679 1682 1688 1691 1701 1726 1737 1738 1739 1818
1833 1860 1869 1891 1893 1910 1928 1932 1934 1941 1944 1947 1982 1988
```

Bu çıktıda eksik gözlemlerin hangi sütunlarda olduğu bulunmuştur.

Eksik gözlemlerin görselleştirilmesi

```
library(naniar)
library(ggplot2)

gg_miss_var(filmler) +
  labs(title = "Eksik Deger Sayıları",
       x = "Degişkenler",
       y = "Eksik Deger Sayısı") +
  theme(plot.title = element_text(hjust = 0.5))
```



Eksik değerlerin hangi değişkenlerde kaç tane olduğu grafikte gösterilmiştir.

Hangi Değişkende Kaç Eksik Gözlem Var?

Metapuan değişkenindeki eksik gözlem sayısı

```
eksik_metapuan <- sum(is.na(filmler$MetaPuan))
print(eksik_metapuan)
```

```
[1] 81
```

"MetaPuan" değişkeninde toplam 81 eksik gözlem var.

Hasılat değişkenindeki eksik gözlem sayısı

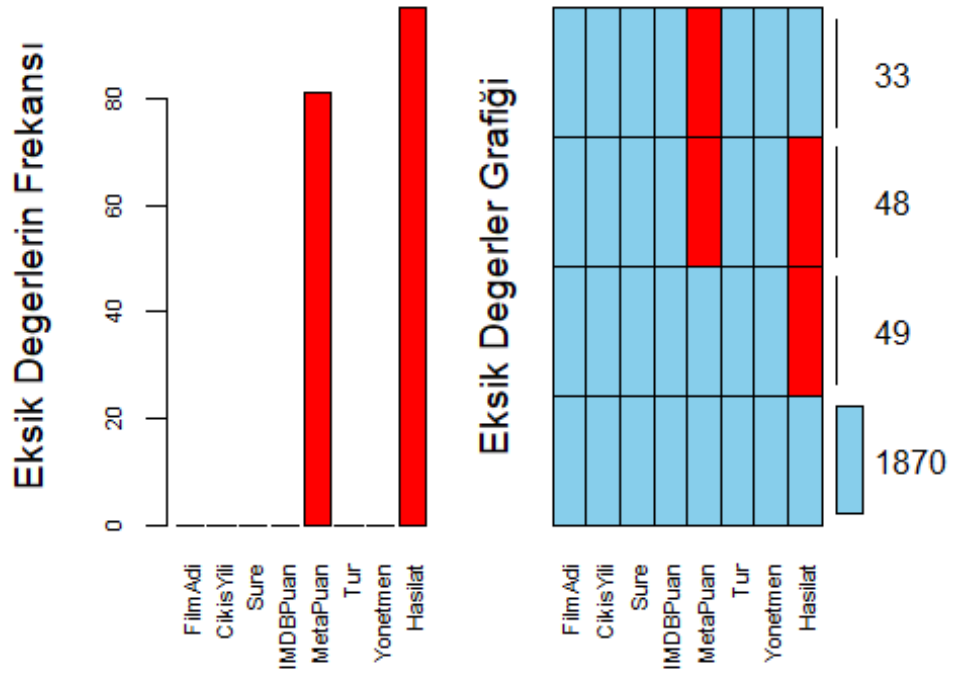
```
eksik_hasilat <- sum(is.na(filmler$Hasilat))
print(eksik_hasilat)
```

```
[1] 97
```

"Hasilat" değişkeninde toplam 97 eksik gözlem var.

Eksik Gözlemlerin Görselleştirilmesi 2

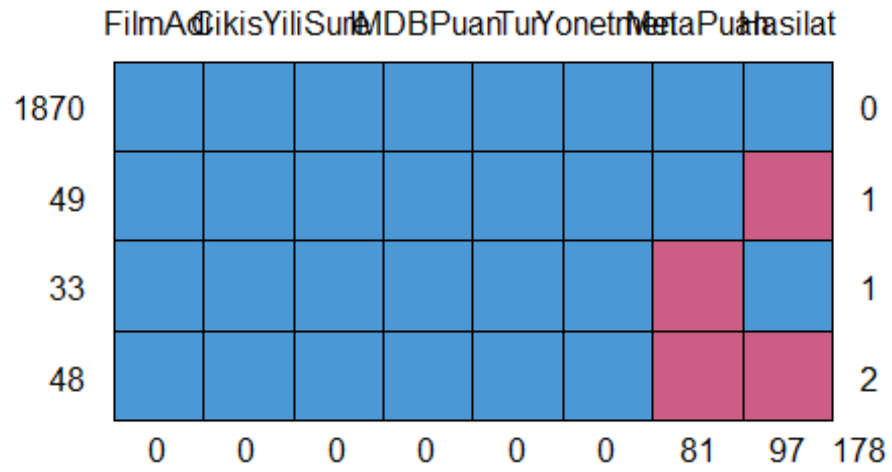
```
aggr(filmler,
     prop = FALSE,
     numbers = TRUE,
     cex.axis = 0.7,
     gap = 3,
     ylab = c("Eksik Degerlerin Frekansı", "Eksik Degerler Grafiği"))
```



Grafik, veri kümesindeki eksik değerlerin dağılımını göstermektedir. Sol taraftaki sütun grafiği, her bir değişkendeki eksik değer sayısını belirtirken, sağ taraftaki matris, farklı değişken kombinasyonlarında kaç eksik değer olduğunu gösterir. "Hasılat" değişkeninde en fazla eksik değer bulunmaktadır ve bu eksik değerler, diğer değişkenlerle kombinasyon halinde de mevcuttur.

Eksik Gözlemlerin Görselleştirilmesi 3

```
library(mice)
md.pattern(filmler)
```



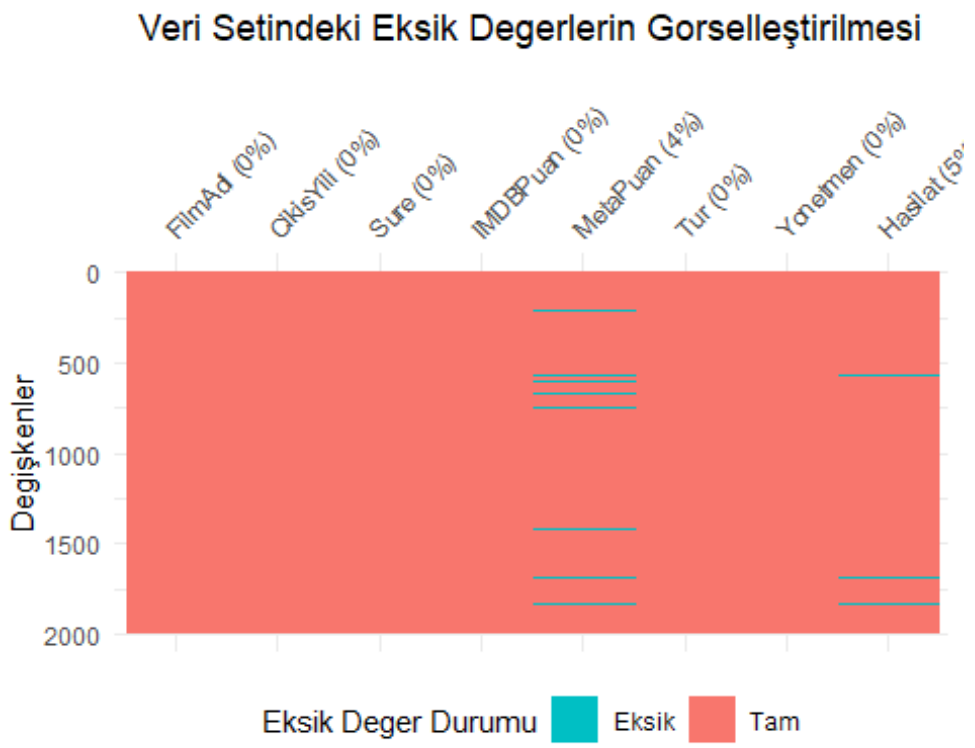

```
     FilmAdi CikisYili Sure IMDBPuan Tur Yonetmen MetaPuan Hasilat
1870       1         1    1        1   1        1        1       1   0
49         1         1    1        1   1        1        1       0   1
33         1         1    1        1   1        1        0       1   1
48         1         1    1        1   1        1        0       0   2
           0         0    0        0   0        0       81      97 178
```

Bu grafik, veri kümesindeki eksik değerlerin dağılımını gösteriyor:

- **"MetaPuan"**: 81 eksik değer var.
- **"Hasılat"**: 97 eksik değer var.
- **Diğer Değişkenler**: Eksik değer yok.

Eksik Gözlemlerin Görselleştirilmesi 4

```
library(naniar)
vis_miss(filmler) +
  labs(title = "Veri Setindeki Eksik Degerlerin Gorselleştirilmesi",
       y = "Degişkenler") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_discrete(name = "Eksik Deger Durumu", labels = c("Tam", "Eksik")
)

Scale for fill is already present.
Adding another scale for fill, which will replace the existing scale.
```



Burada eksik gözlemlerin oranlarının görselleştirilmesi görülmektedir. Buna göre "MetaPuan" değişkeninin %4'ü "Hasilat" değişkeninin ise %5'i eksik gözlemdir. Diğer değişkenlerde eksik gözlem yoktur.

Eksik Değerleri Doldurma

- **KNN (K En Yakın Komşu) İle Doldurma**

```
library(VIM)
filmler <- kNN(filmler, k = 5)
colSums(is.na(filmler))
```

```
      FilmAdi     CikisYili          Sure      IMDBPuan      MetaPuan 
            0             0             0             0             0 
          Tur      Yonetmen       Hasilat   FilmAdi_imp CikisYili_imp 
            0             0             0             0             0 
    Sure_imp  IMDBPuan_imp  MetaPuan_imp       Tur_imp  Yonetmen_imp 
            0             0             0             0             0 
 Hasilat_imp 
            0 
```

Eksik gözlemler KNN yöntemiyle doldurulmuştur. Kontrol yapınca hiçbir eksik gözlem kalmadığı görülmektedir. KNN, eksik değerlere en yakın komşularının değerlerine bakarak tahmin yapar. Veri setindeki benzerlik yapısını korur.

5. Eğitim ve Test Veri Kümelerinin Oluşturulması

Veri Setini Eğitim ve Test Kümelerine Ayırma (%80 eğitim, %20 test)

```
set.seed(123)
egitim_index <- sample(1:nrow(filmler), 0.8 * nrow(filmler))
egitim_veri <- filmler[egitim_index, ]
test_veri <- filmler[-egitim_index, ]
```

Veri seti %80 eğitim %20 test şekilde bölünmüştür. Bu adımdan sonra, geçerlilik adımına kadar yapılacak tüm işlemler, eğitim veri kümesi üzerinden gerçekleştirilecektir.

Gösterge Değişkenlerin Oluşturulması

2000 ve sonrası çıkan filmler için 1, diğerleri için 0 şeklinde gösterge değişken oluşturulacaktır.

```
filmler$yeni_film <- ifelse(filmler$CikisYili >= 2000, 1, 0)
```

Buna göre "yeni_film" değişkeni altında 2000 sonrası çıkan filmler 1, diğerleri 0 şeklinde kodlanarak gösterge değişken oluşturulmuştur.

Nicel Değişkeni Kategorize Etme

Hasilat değişkenini kategorilere ayırma

"Hasilat" değişkeninde çok büyük değerler olduğu için kategorilere ayrılması analiz için daha iyi olacaktır.

```
kategoriler <- c("Cok Dusuk", "Dusuk", "Orta", "Yuksek", "Cok Yuksek")

kesme_noktalari <- c(0, 20000000, 50000000, 80000000, 100000000, Inf)

egitim_veri$Hasilat <- cut(egitim_veri$Hasilat, breaks = kesme_noktalari, lab
els = kategoriler, include.lowest = TRUE)

table(egitim_veri$Hasilat)


 Cok Dusuk      Dusuk       Orta     Yuksek Cok Yuksek
       426        440        304         97        333
```

Bu şekilde "Hasilat" değişkeni 5 ayrı alt gruba bölünerek kategorize edilmiştir.

Türetilmiş Değişken Oluşturma

IMDBPuan ile MetaPuan arasındaki farkı hesaplama

Öncelikle MetaPuan 10'luk sistemde yazılmalı.

```
egitim_veri$MetaPuan_10 <- egitim_veri$MetaPuan / 10
```

Ardından IMDBPuan ile MetaPuan arasındaki farkı tanımlayan PuanFarki adlı bir değişken oluşturulur.

```
egitim_veri$PuanFarki <- egitim_veri$IMDBPuan - egitim_veri$MetaPuan_10
head(egitim_veri$PuanFarki)

[1]  1.6  0.1 -0.9  2.8  2.2  0.0
```

Bu sayede "IMDBPuan" ile "MetaPuan" arasındaki farkı "PuanFarki" adlı türetilmiş yeni bir değişkene yazdırdık.

MetaPuan ile IMDBPuan ortalamasını hesaplama

```
egitim_veri$OrtalamaPuan <- (egitim_veri$IMDBPuan*10 + egitim_veri$MetaPuan)
/ 2
```

filmlerin ortalama aldığı puanı yüz üzerinden değerlendirme yapmak için IMDBPuanını 10 ile çarpılmıştır.

```
head(egitim_veri$OrtalamaPuan)

[1] 58.0 60.5 85.5 56.0 69.0 72.0
```

görüldüğü üzere işlem başarılı bir şeklide gerçekleştirilmiş.

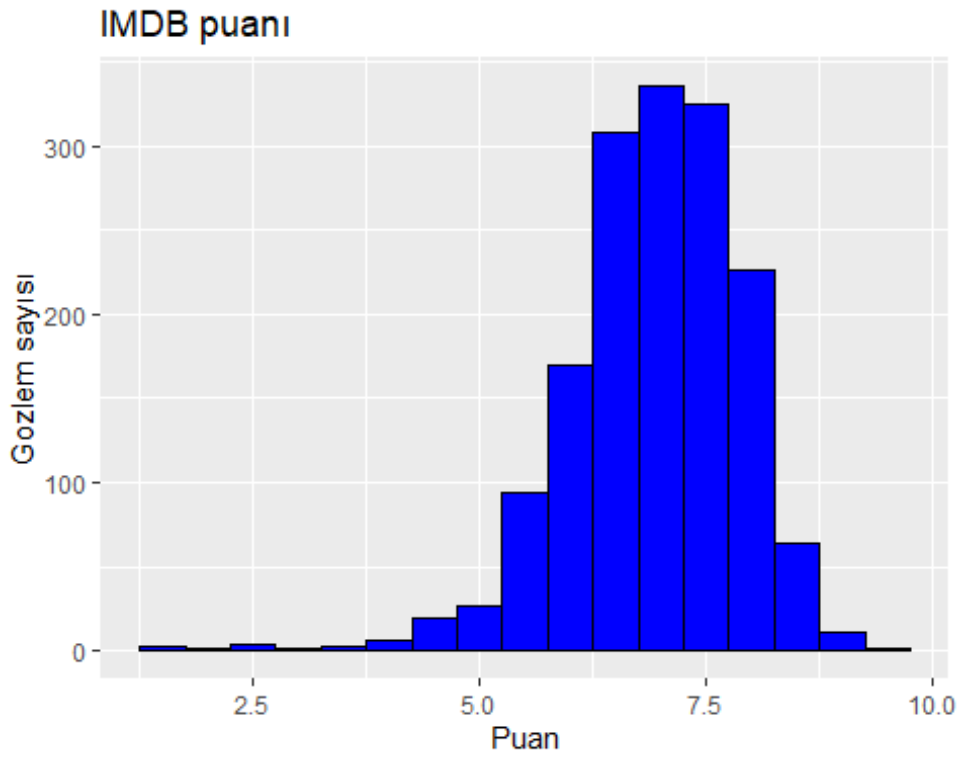
6. Verilerin Açıklayıcı/Keşfedici Çözümlemesi

Gerekli Kütüphanelerin İndirilmesi

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(gridExtra)
library(psych)
library(corrplot)

corrplot 0.92 loaded

library(sm)
library(MASS)
library(interactions)
library(aplpack)
library(DescTools)
```

IMDBPuanı İçin Histogram Grafiği Oluşturma

```
ggplot(egitim_veri, aes(x = IMDBPuan)) +
 geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
 labs(title = "IMDB puanı", x = "Puan", y = "Gozlem sayısı")
```



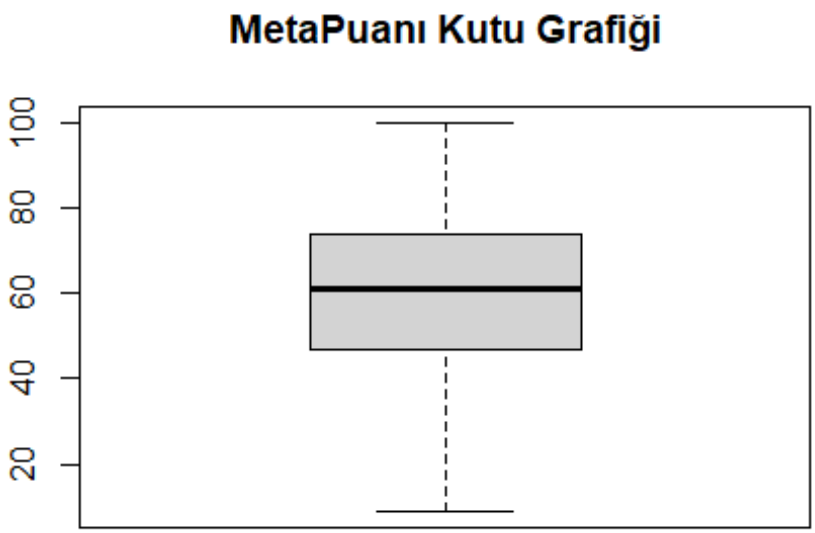
Histogram, puanların 7 civarında yoğunlaştığını ve 6 ile 8 arasında zirveye ulaştığını gösteriyor. Puanlar 6'dan düşük veya 8'den yüksek olduğunda gözlem sayısı hızla azalıyor. Bu, filmlerin çoğunun ortalama puana sahip olduğunu ve çok düşük veya çok yüksek puan alan filmlerin az olduğunu gösteriyor.

Sonuçlar:

- Ortalama: Puanlar 7 civarında yoğunlaşıyor.
- Varyans: Puan dağılımı düşük varyans gösteriyor, yani puanlar benzer.
- Çarpıklık: Histogram hafifçe sağa çarpık, yani daha fazla film 7'den düşük puan alıyor ve yüksek puan alan film sayısı az.

MetaPuanı için Kutu Grafiği Oluşturma

```
boxplot(egitim_veri$MetaPuan, main = "MetaPuanı Kutu Grafiği")
```



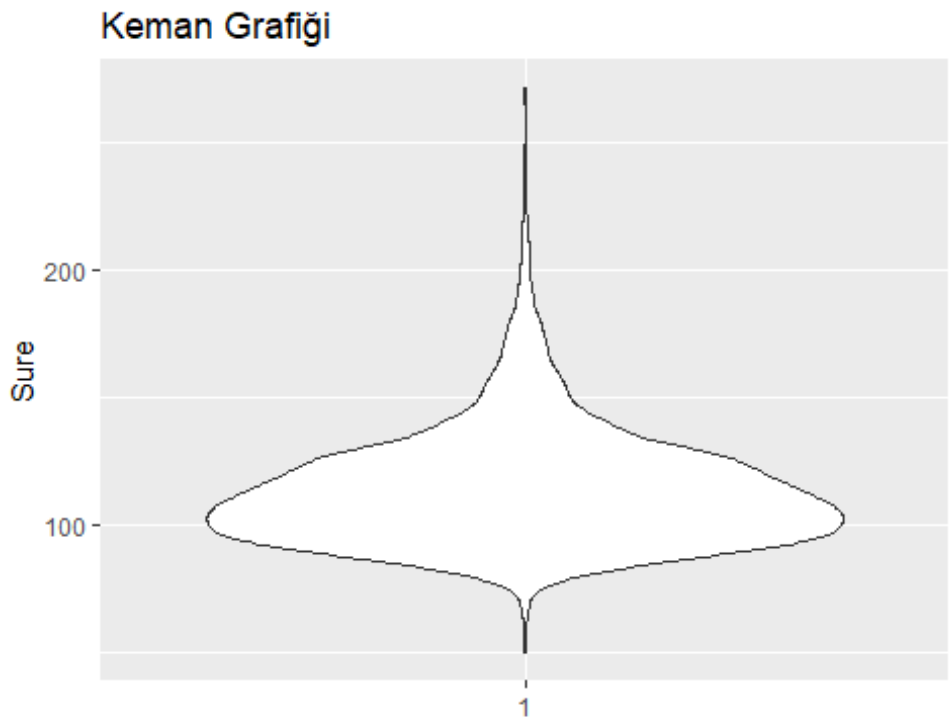
Kutu grafiğinden şu çıkarımlar yapılabilir:

- **Medyan:** Grafikteki ortadaki çizgi, medyanın yaklaşık 60 olduğunu gösterir.
- **Çeyreklikler:** Kutu kenarları, ilk ve üçüncü çeyrekleri gösterir; bunlar yaklaşık 50 ve 70'tir.
- **Çeyrekler Arası Aralık (IQR):** Kutunun yüksekliği yaklaşık 20'dir, verilerin ne kadar sıkı paketlendiğini gösterir.
- **Aykırı Değerler:** Grafikteki uç çizgiler, 10'un altındaki ve 100'ün üzerindeki aykırı değerleri gösterir.

Bu kutu grafiği, filmlerin çoğunun 50 ile 70 arasında puan aldığını, medyanın 60 olduğunu ve bazı aykırı değerlerin bulunduğunu gösterir.

Süre için Keman Grafiği Oluşturma

```
library(ggplot2)
 ggplot(egitim_veri, aes(x=factor(1), y=Sure)) +
 geom_violin() +
 labs(title="Keman Grafiği", x="", y="Sure")
```



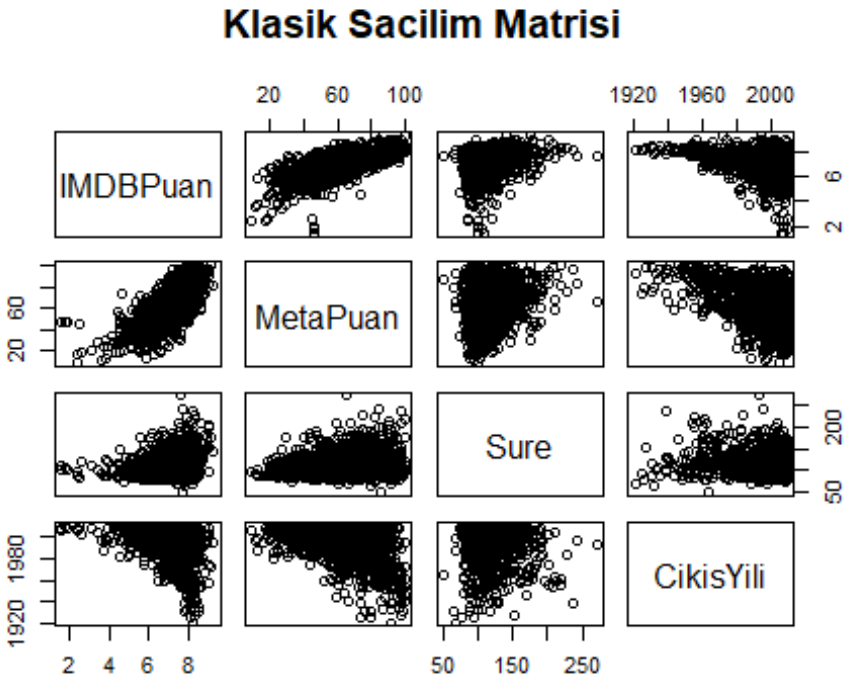
Keman grafiğinden şu çıkarımlar yapılabilir:

- **Ortalama:** Kemanın en geniş noktası, ortalama sürenin yaklaşık 100 dakika olduğunu gösterir. Bu, filmlerin çoğunun 100 dakika civarında sürdüğü anlamına gelir.
- **Varyans:** Kemanın şekli, sürede bir miktar varyans olduğunu gösterir. Kemanın dar kısmı, sürenin 100 dakikaya yakın yoğunlaştığını, daha geniş kısımlar ise filmlerin daha kısa veya daha uzun sürebildiğini belirtir.
- **Çarpıklık:** Keman hafifçe sağa çarpık görünmektedir, bu da bazı filmlerin daha uzun sürdüğünü gösterir.

Bu sonuçlar, filmlerin süresinin genellikle 100 dakika civarında yoğunlaştığını, ancak bazı filmlerin daha kısa veya daha uzun sürebildiğini ve birkaç filmin daha uzun sürdüğünü göstermektedir.

Klasik Saçılım Matrisi (ScatterplotMatrix)

```
sayisal_degiskenler<-c("IMDBPuan", "MetaPuan","Sure","CikisYili")
pairs(egitim_veri[, sayisal_degiskenler], main= "Klasik Sacilim Matrisi")
```



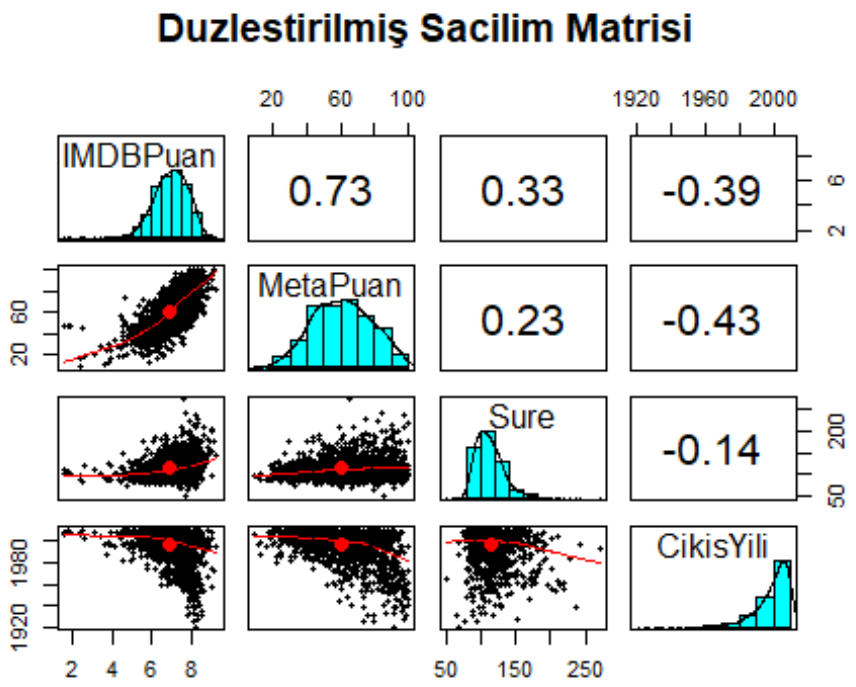
Saçılım matrisinden çıkarılan bazı gözlemler:

- **IMDb Puanı ve MetaPuan:** İki puan sistemi arasında güçlü pozitif bir ilişki var, yani yüksek IMDb puanı alan filmler genellikle Metapuan'da da yüksek puan alıyor.
- **IMDb Puanı ve Süre:** Aralarında belirgin bir ilişki yok, yani filmlerin süresi IMDb puanını etkilemiyor.
- **IMDb Puanı ve Çıkış Yılı:** Zayıf bir negatif ilişki var, yani eski filmler genellikle daha düşük IMDb puanına sahip olma eğiliminde, ancak bu ilişki çok zayıf.
- **MetaPuan ve Süre:** Aralarında belirgin bir ilişki yok, yani filmlerin süresi Metapuan'ı etkilemiyor.
- **MetaPuan ve Çıkış Yılı:** Belirgin bir ilişki yok, yani filmlerin çıkış yılı Metapuan'ı etkilemiyor.
- **Süre ve Çıkış Yılı:** Zayıf bir negatif ilişki var, yani eski filmler daha kısa olma eğiliminde, ancak bu ilişki çok zayıf.

Genel olarak, iki puan sistemi arasında güçlü bir ilişki bulunurken, diğer değişkenler ile puanlar arasında daha zayıf ilişkiler mevcut

Düzleştirilmiş Saçılım Matrisi (Smoothed Scatterplot Matrix)

```
pairs.panels(egitim_veri[, sayisal_degiskenler],
 main = "Duzlestirilmis Sacilim Matrisi",
 ellipses = TRUE)
```



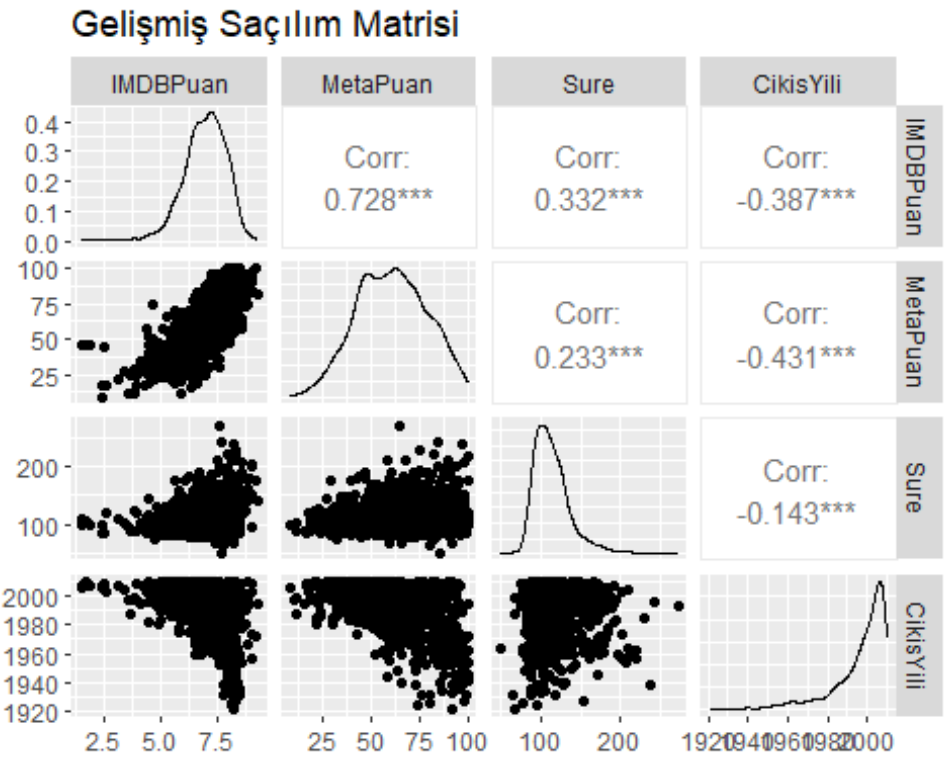
Düzleştirilmiş saçılım matrisinden çıkarılan bazı gözlemler:

- **IMDb Puanı ve Meta Puanı:** Güçlü pozitif doğrusal ilişki (korelasyon katsayısı 0,73). Yüksek IMDb puanı alan filmler genellikle yüksek Meta puan alıyor.
- **IMDb Puanı ve Süre:** Zayıf doğrusal ilişki (korelasyon katsayısı 0,23). Filmlerin süresi IMDb puanını fazla etkilemiyor.
- **IMDb Puanı ve Çıkış Yılı:** Zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,39). Eski filmler genellikle daha düşük IMDb puanına sahip.
- **Meta Puanı ve Süre:** Çok zayıf doğrusal ilişki (korelasyon katsayısı -0,14). Filmlerin süresi Meta puanını fazla etkilemiyor.
- **Meta Puanı ve Çıkış Yılı:** Zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,43). Eski filmler genellikle daha düşük Meta puanına sahip.
- **Süre ve Çıkış Yılı:** Çok zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,14). Eski filmler genellikle daha kısa süreli.

Genel olarak, iki puan sistemi arasında güçlü bir ilişki bulunurken, diğer değişkenler ile puanlar arasında daha zayıf ilişkiler var.

Gelişmiş Saçılım Matrisi

```
ggpairs(egitim_veri, columns = sayisal_degiskenler, title = "Gelişmiş Saçılım
Matrisi")
```



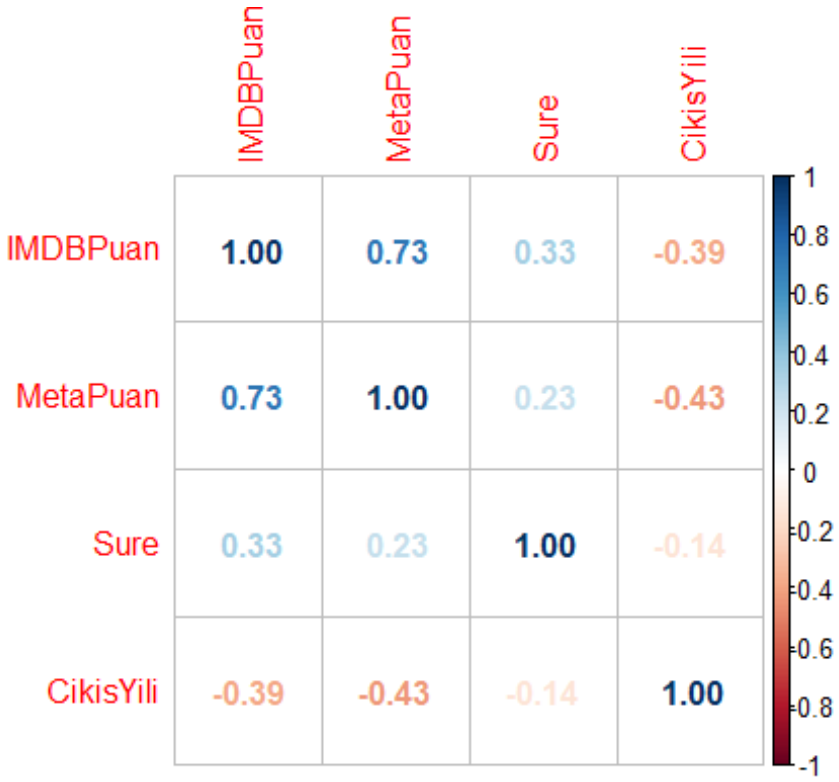
İşte gelişmiş saçılım matrisinden çıkarılan bazı gözlemler:

- **IMDb Puanı ve Meta Puanı:** Güçlü pozitif doğrusal ilişki (korelasyon katsayısı 0,729). Yüksek IMDb puanı alan filmler genellikle Meta puanında da yüksek puan alır. Saçılım diyagramında noktalar doğrusal bir eğriye yakın dağılmıştır.
- **IMDb Puanı ve Süre:** Zayıf doğrusal ilişki (korelasyon katsayısı 0,233). Filmlerin süresi IMDb puanını fazla etkilemez. Saçılım diyagramında noktalar dağınık dağılmıştır.
- **IMDb Puanı ve Çıkış Yılı:** Zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,387). Eski filmler genellikle daha düşük IMDb puanına sahiptir.
- **Meta Puanı ve Süre:** Çok zayıf doğrusal ilişki (korelasyon katsayısı -0,143). Filmlerin süresi Meta puanını etkilemez. Saçılım diyagramında noktalar dağınıktır.
- **Meta Puanı ve Çıkış Yılı:** Zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,434). Eski filmler genellikle daha düşük Meta puanına sahiptir.
- **Süre ve Çıkış Yılı:** Çok zayıf negatif doğrusal ilişki (korelasyon katsayısı -0,143). Eski filmler genellikle daha kısa sürelidir, ancak bu ilişki çok zayıftır.

Genel olarak, gelişmiş saçılım matrisi, değişkenler arasındaki ilişkilerin daha ayrıntılı bir resmini sunar ve korelasyon katsayıları bu ilişkilerin gücünü ve yönünü gösterir.

Korelasyon Matrisi

```
cor_matrix <- cor(egitim_veri[, sayisal_degiskenler], use = "complete.obs")
corrplot(cor_matrix, method = "number")
```



Bu çıktı, bir film veri setindeki Çıkış Yılı, Süre, IMDb Puanı ve Meta Puanı arasındaki ilişkileri gösteren bir korelasyon matrisidir.

- **Çıkış Yılı ve Süre:** Negatif korelasyon (-0.14). Zaman içinde filmler biraz daha kısalma eğilimindedir.
- **Çıkış Yılı ve IMDb/Meta Puan:** Negatif korelasyonlar (-0.39 ve -0.43). Daha eski filmler genellikle daha yüksek puanlara sahiptir.
- **Süre ve IMDb/Meta Puan:** Pozitif korelasyonlar (0.33 ve 0.23). Daha uzun filmler biraz daha yüksek puan alma eğilimindedir.
- **IMDb Puanı ve Meta Puanı:** Güçlü pozitif korelasyon (0.73). İki puan türü genellikle benzer eğilimlere sahiptir; biri yüksekse diğeri de yüksek olma olasılığı yüksektir.

Sonuç:

Bu korelasyon matrisi, film veri setindeki değişkenler arasındaki ilişkiler hakkında önemli bilgiler sağlar. Özellikle, IMDb ve Meta puanları arasındaki güçlü pozitif korelasyon, bu iki puan türünün benzer ölçütler olduğunu gösterir. Ayrıca, çıkış yılı ile puanlar arasındaki negatif korelasyon, zaman içinde film değerlendirmelerinde değişim olduğunu düşündürür.

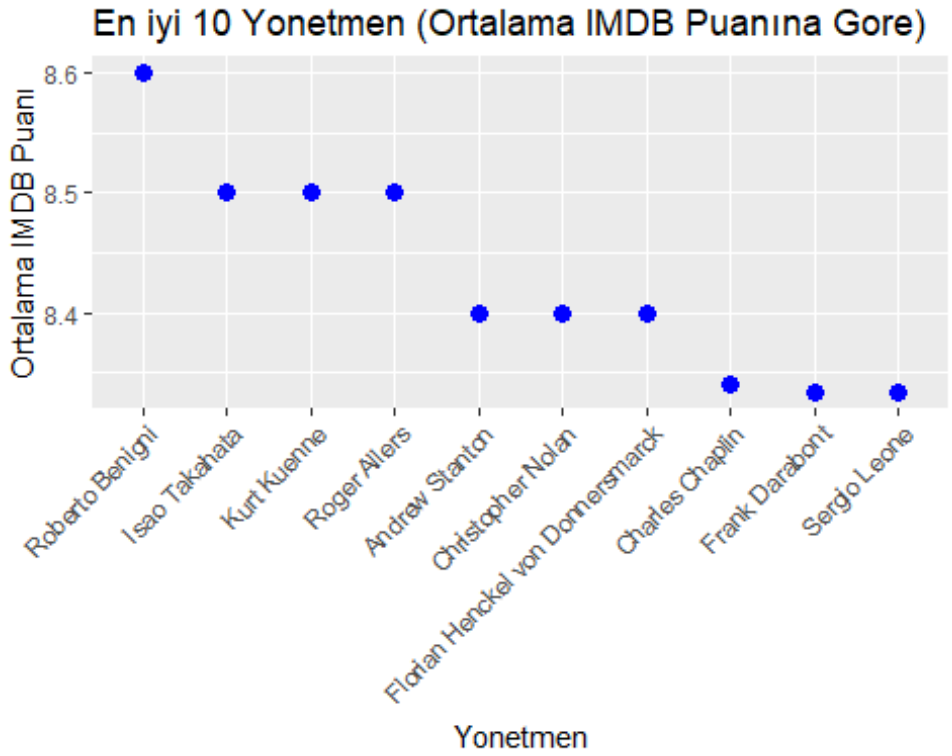
Belirli Yönetmenlerin Filmleri Daha Mı Başarılıdır?

En İyi 10 Yönetmenin Ortalama IMDB Puanları

```
top_yonetmenler <- egitim_veri %>%
  group_by(Yonetmen) %>%
  summarize(Ortalama_IMDB = mean(IMDBPuan, na.rm = TRUE), Film_Sayisi = n())
%>%
  arrange(desc(Ortalama_IMDB)) %>%
  head(10)
```

Nokta Grafiği

```
ggplot(top_yonetmenler, aes(x = reorder(Yonetmen, -Ortalama_IMDB), y = Ortala
ma_IMDB)) +
  geom_point(size = 3, color = "blue") +
  labs(title = "En iyi 10 Yonetmen (Ortalama IMDB Puanına Gore)", x = "Yonetm
en", y = "Ortalama IMDB Puanı") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



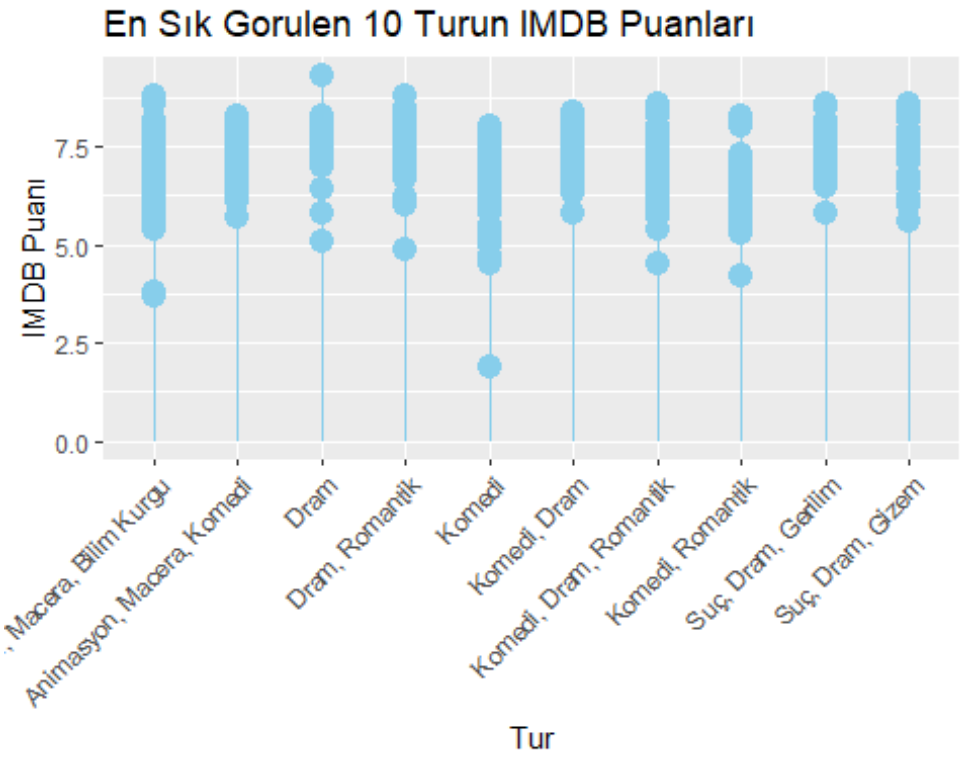
Bu grafik, en yüksek ortalama IMDb puanına sahip ilk 10 yönetmeni göstermektedir. Grafik, yatay eksende yönetmenlerin isimlerini, dikey eksende ise ortalama IMDb puanlarını içermektedir. Grafikte görülen noktalar, her bir yönetmenin ortalama IMDb puanını temsil ediyor.

Roberto Benigni en yüksek ortalama IMDb puanına sahip yönetmendir.

Lolipop Grafiği

En sık görülen 10 türün ve 10 yönetmenin IMDB puanları

```
top_turler <- names(sort(table(egitim_veri$Tur), decreasing = TRUE)[1:10])

top_yonetmenler <- names(sort(table(egitim_veri$Yonetmen), decreasing = TRUE)
[1:10])

ggplot(egitim_veri[egitim_veri$Tur %in% top_turler, ], aes(x = Tur, y = IMDBP
uan)) +
  geom_point(size = 4, color = "skyblue") +
  geom_segment(aes(xend = Tur, yend = 0), color = "skyblue") +
  labs(title = "En Sık Gorulen 10 Turun IMDB Puanları", x = "Tur", y = "IMDB
Puanı") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



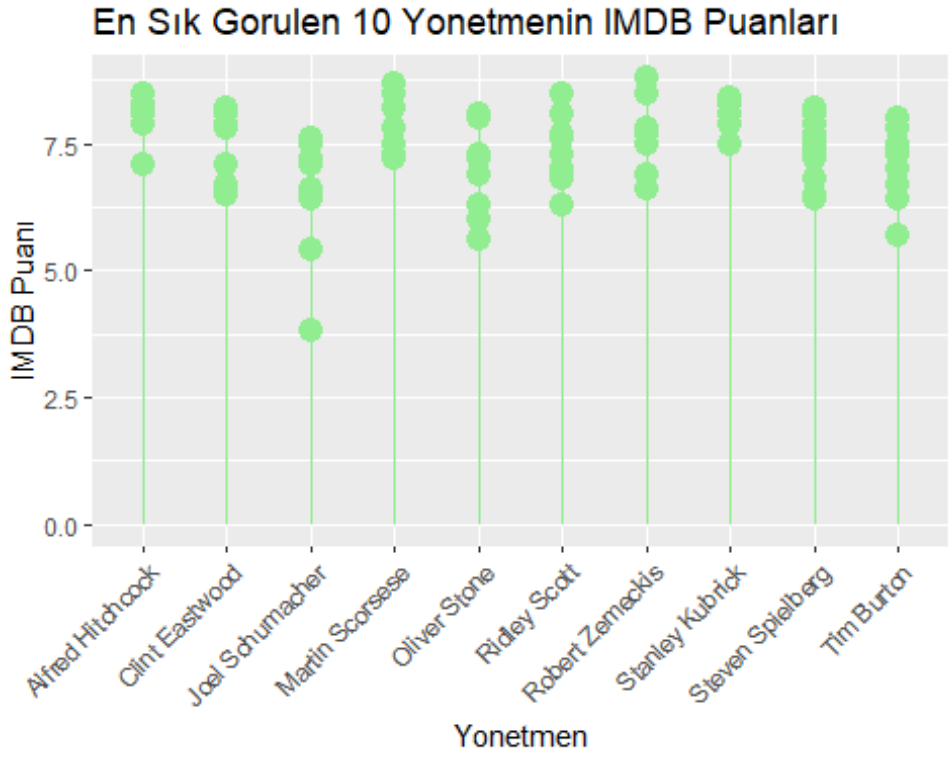
Grafik, en sık görülen 10 film türünün IMDB puanlarının dağılımını göstermektedir. Öne çıkan noktalar:

- **Aksiyon, Macera, Bilim Kurgu** ve **Dram** türlerinde puanlar geniş bir aralığa yayılmış, ancak genelde 5-8 arasında yoğunlaşmış.
- **Animasyon, Macera, Komedi** ve **Komedi, Romantik** türleri genellikle daha yüksek puanlar almış ve 6-8 aralığında yoğunlaşmış.

- **Dram, Romantik**, **Komedi**, **Komedi, Dram** ve **Suç, Dram, Gerilim** türleri de genelde 5-8 puan aralığında yoğunlaşmış.

Genel olarak, film türlerinin çoğunda puanların 5-8 aralığında yoğunlaştığı görülmektedir.

```
ggplot(egitim_veri[egitim_veri$Yonetmen %in% top_yonetmenler, ], aes(x = Yone
tmen, y = IMDBPuan)) +
  geom_point(size = 4, color = "lightgreen") +
  geom_segment(aes(xend = Yonetmen, yend = 0), color = "lightgreen") +
  labs(title = "En Sık Gorulen 10 Yonetmenin IMDB Puanları", x = "Yonetmen",
y = "IMDB Puanı") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Bu grafik, en sık görülen 10 yönetmenin IMDB puanlarını göstermektedir. Her yönetmenin filmlerinin IMDB puanları, yönetmenin isminin altında dikey olarak dağılmıştır.

Temel Gözlemler:

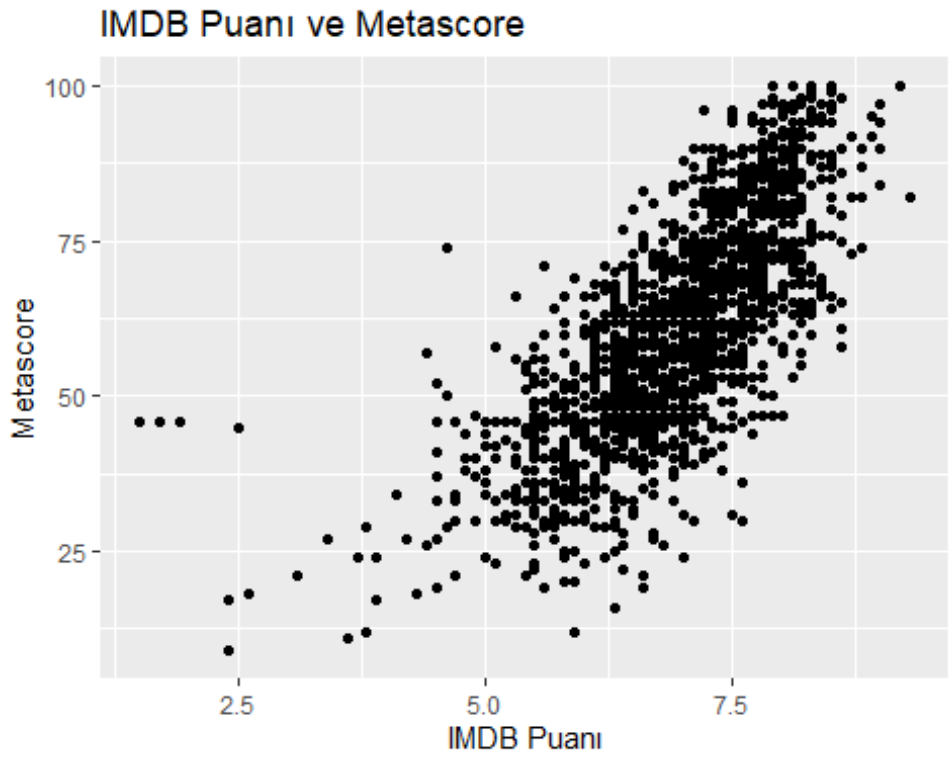
- **Puan Dağılımı:** Çoğu yönetmenin filmlerinin IMDB puanları 5.0 ile 8.0 arasında değişmektedir. Bu, genel olarak başarılı filmler yaptıklarını gösterir.
- **En Yüksek Puanlar:** Stanley Kubrick, Alfred Hitchcock ve Martin Scrorsese en yüksek puan ortalamasına sahip yönetmenler gibi görünmektedir.
- **En Düşük Puanlar:** Joel Schumacher, en düşük puan ortalamasına sahip yönetmen olarak öne çıkıyor.

- **Tutarlılık:** Stanley Kubrick daha tutarlı bir puan dağılımına sahiptir. Bu, genel olarak kaliteli filmler yapma eğiliminde olduğunu gösterir.
- **Değişkenlik:** Joel Schumacher adlı yönetmenin filmlerinin puanlarındaki değişim diğer yönetmenlere göre daha yüksektir.

IMDB Puanı ile Metascore Puanı Arasında İlişki Var Mı?

Saçılım Grafiği ve Korelasyon: IMDB Puanı ve Metascore

```
ggplot(egitim_veri, aes(x = IMDBPuan, y = MetaPuan)) +
  geom_point() +
  labs(title = "IMDB Puanı ve Metascore", x = "IMDB Puanı", y = "Metascore")
```



İlişkinin Yorumlanması:

- **Pozitif Korelasyon**: Genel olarak, IMDb puanı arttıkça Metascore puanı da artma eğilimindedir. Bu, iki puanlama sistemi arasında pozitif bir korelasyon olduğunu gösterir. Yani, yüksek IMDb puanına sahip filmler genellikle yüksek Metascore puanına da sahiptir.
- **Yüksek Yoğunluk**: 5 ile 8 arasındaki IMDb puanları, 50 ile 90 arasındaki Metascore puanlarıyla yoğun bir şekilde örtüşmektedir. Bu, birçok filmin bu aralıkta puanlandığını gösterir.

- **Düşük Puanlar**: Çok düşük IMDb puanlarına (2.5 civarı) sahip birkaç film Metascore puanı 50'nin altında yer almakta. Bu filmler hem IMDb hem de Metascore açısından düşük değerlendirilmiştir.
- **Dağınık Noktalar**: Her iki eksende de bazı filmler daha dağınık puanlara sahiptir. Yani, bazı filmler yüksek IMDb puanına sahipken düşük Metascore almış veya tam tersi duruma sahiptir.

Korelasyon

```
cor(egitim_veri$IMDBPuan, egitim_veri$MetaPuan, use = "complete.obs")
```

```
[1] 0.7282688
```

Bulunan korelasyon katsayısı, IMDB ve Metascore puanları arasında güçlü ve pozitif bir ilişki olduğunu gösterir. Bu, bir filmin IMDB puanı arttıkça Metascore puanının da artma eğiliminde olduğunu doğrular.

Hangi Film Türleri Daha Yüksek IMDB Puanı Almaktadır?

Türlere göre ortalama IMDB puanını hesaplama

```
tur_ortalamalari <- egitim_veri %>%
  group_by(Tur) %>%
  summarize(Ortalama_IMDB = mean(IMDBPuan, na.rm = TRUE)) %>%
  arrange(desc(Ortalama_IMDB))
```

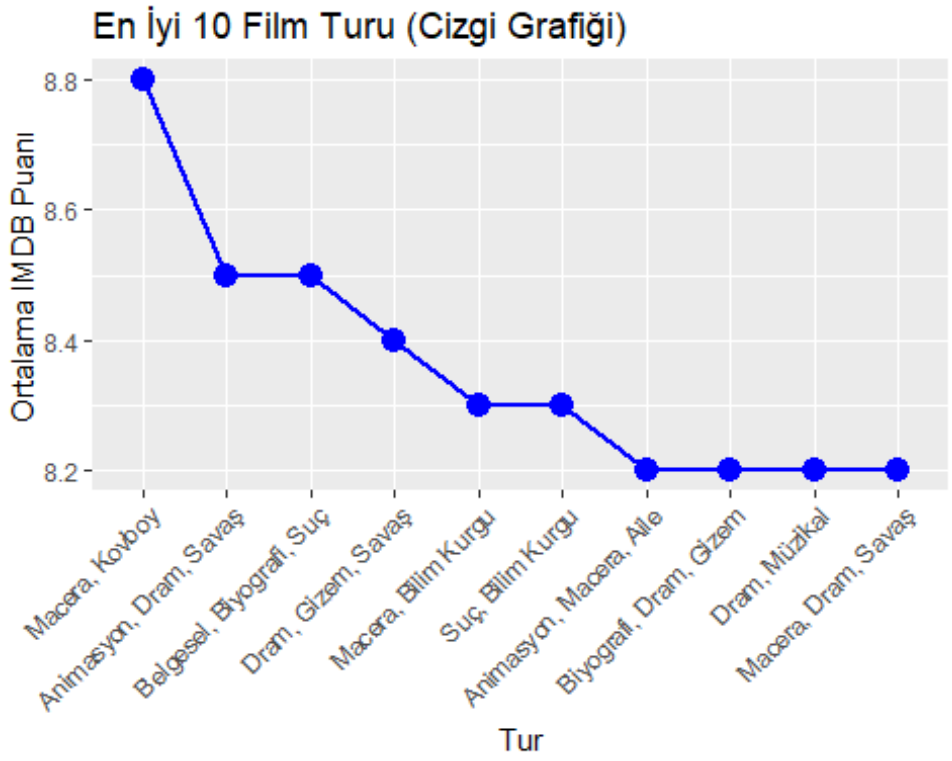
En iyi 10 türü seçme

```
en_iyi_10_tur <- head(tur_ortalamalari, 10)
```

En iyi türlerin ortalama IMDB puanlarını gösteren grafik

```
ggplot(en_iyi_10_tur, aes(x = reorder(Tur, -Ortalama_IMDB), y = Ortalama_IMDB
, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(size = 4, color = "blue") +
  labs(title = "En İyi 10 Film Turu (Cizgi Grafiği)", x = "Tur", y = "Ortalam
a IMDB Puanı") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



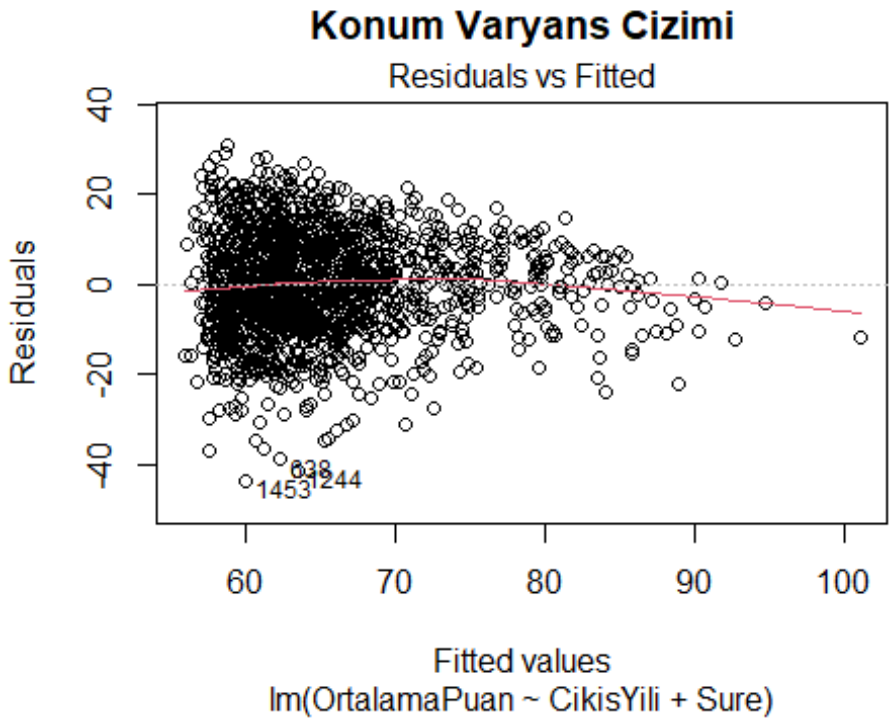
Grafik, En İyi 10 film türünün ortalama IMDB puanlarını göstermektedir.

Genel Yorum:

- En yüksek ortalama IMDB puanına sahip tür, 8.5 puanla “Macera, Kovboy” türüdür.
- Genel olarak, macera türü filmler yüksek puan alırken, dram ve savaş temalı filmler daha düşük puanlara sahiptir.

Konum Varyans Çizimleri (Location-Scale Plots)

```
#Konumvaryansçizimiiçinönceliklebirdoğrusalmodeloluşturuyoruz
model<-lm(OrtalamaPuan~ CikisYili +Sure, data=egitim_veri)
#Konumvaryansçizimi
plot(model,which=1,main="Konum Varyans Cizimi")
```



Bu grafikte gözlenen bulgular şunlardır:

1. **Homoskedastisite**: Noktaların yatay eksende eşit şekilde dağılmadığı gözlenmektedir. Bu durum, modelin homoskedastik olmadığını ve hataların varyansının tahmin edilen değerler arttıkça arttığını işaret eder.

2. **Ortalama Kalıntı**: Kalıntıların ortalamasının sıfır civarında olması beklenir. Grafikte bu durum gözlenmektedir, yani kalıntılar ortalaması sıfır civarında gibi görünmektedir.

3. **Kalıntı Dağılımı**: Kalıntılar normal dağılıma yakın bir şekilde olmalıdır. Ancak grafikte kalıntıların normal dağılıma yakın olmadığı gözlenmektedir.

4. **Aykırı Değerler**: Bazı noktalar diğerlerinden önemli ölçüde farklıdır, yani aykırı değerler mevcuttur.

Bu bulgular, modelimizin temel varsayımlarının ihlal edildiğini ve hata terimlerinin tutarlı bir şekilde dağılmadığını gösteriyor. Hata terimlerinin tutarsız dağılması, modelin tahminlerinin güvenilirliğini azaltır. Ayrıca, modelin hata terimlerinin normal dağılım göstermemesi, modelin doğruluğunu ve güvenilirliğini olumsuz etkiler.

Düzleştirme

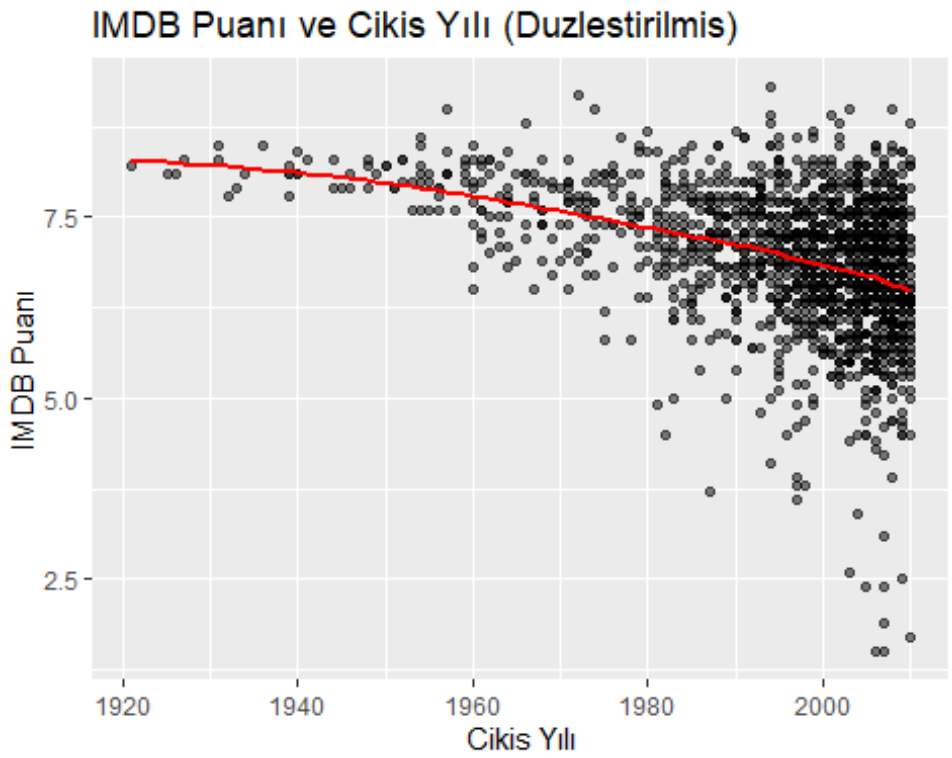
LOESS düzleştirme

Düzleştirilmiş değerleri tahmin etme

```
loess_model <- loess(IMDBPuan ~ CikisYili, data = egitim_veri)
egitim_veri$IMDBPuan_Duzeltilmis <- predict(loess_model)
```

Düzleştirme Grafiği

```
ggplot(egitim_veri, aes(x = CikisYili, y = IMDBPuan)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = IMDBPuan_Duzeltilmis), color = "red", size = 1) +
  labs(title = "IMDB Puanı ve Cikis Yılı (Duzlestirilmis)", x = "Cikis Yılı",
y = "IMDB Puanı")
```



Grafik üzerindeki kırmızı çizgi, IMDB puanı ile Çıkış Yılı arasındaki ilişkiyi LOESS yöntemiyle düzleştirilmiş olarak göstermektedir. Bu grafikten çıkarılabilecek bazı yorumlar şunlar olabilir:

1. **Genel IMDB Puanı Eğilimi**: Grafik genel olarak zaman içinde IMDB puanlarının hafifçe düşme eğiliminde olduğunu göstermektedir. Bu durum, daha yeni filmlerin genellikle daha düşük IMDB puanları aldığına işaret edebilir.

2. **Erken Dönemdeki Yüksek Puanlı Filmler**: Özellikle 1920'ler ve 1930'lar gibi erken dönemlerde bazı çok yüksek puanlı filmler bulunmaktadır. Bu filmler, sinemanın erken dönemlerindeki klasikler olabilir ve yüksek puanlarıyla dikkat çekmektedirler.

3. **Yoğunlaşma ve Homojenlik**: Yakın dönemde, yani 2000'li yıllardan sonra IMDB puanlarının daha dar bir aralıkta yoğunlaştığı görülmektedir. Bu durum, yeni

filmlerin puanlarının daha homojen dağıldığını ve genel bir standardizasyon eğilimi gösterdiğini işaret edebilir.

4. **Değişim Dalgalanmaları**: Grafikte belirgin bir zirve veya düşüş noktası olmaması, IMDB puanının zaman içinde sürekli bir değişim göstermediğini, daha çok dalgalanmalar şeklinde seyrettiğini düşündürmektedir. Bu, IMDB puanlarının zamanla istikrarlı bir eğilim göstermediğini ancak belirli dönemlerde dalgalanmalar yaşadığını gösterebilir.

Nitel/kategorik Değişkenler İçin Gini ve Entropi

Gini

```
kategorik_degiskenler <- c("Tur", "Hasilat")
gini_degerleri <- sapply(egitim_veri[, kategorik_degiskenler], function(x) {
  Gini(table(x), na.rm = TRUE)
})
print(gini_degerleri)
```

```
      Tur   Hasilat
0.6265574 0.2525000
```

Sonuçlar:

- **Tur değişkeni için Gini katsayısı: 0.6265574**
- **Hasilat değişkeni için Gini katsayısı: 0.2493750**

Bu sonuçlar şunları ifade eder:

- **Tur** değişkeninin Gini katsayısı daha yüksek (0.6265574), bu da bu değişkenin daha fazla çeşitlilik veya düzensizlik içerdiğini gösterir. Yani, "Tur" değişkeni birçok farklı kategoriye sahip ve bu kategoriler arasında daha eşit bir dağılım söz konusu.
- **Hasilat** değişkeninin Gini katsayısı daha düşük (0.2493750), bu da bu değişkenin daha az çeşitlilik veya düzensizlik içerdiğini gösterir. Yani, "Hasilat" değişkeni daha az sayıda kategoriye sahip olabilir ve/veya bu kategoriler arasında daha dengesiz bir dağılım olabilir.

Özetle, "Tur" değişkeni "Hasilat" değişkenine göre daha çeşitlidir.

Entropi

```
entropi_degerleri <- sapply(egitim_veri[, kategorik_degiskenler], function(x)
{
  frekanslar <- table(x) / length(x)
  -sum(frekanslar * log2(frekanslar))
})
print(entropi_degerleri)
```

```
     Tur  Hasilat
6.840856 2.192182
```

Sonuçlar:

- **Tur** değişkeninin entropisi daha yüksek (6.840856), bu da bu değişkenin daha fazla çeşitlilik veya belirsizlik içerdiğini gösterir. Yani, "Tur" değişkeni birçok farklı kategoriye sahip ve bu kategoriler arasında daha eşit bir dağılım söz konusu.
- **Hasilat** değişkeninin entropisi daha düşük (2.192203), bu da bu değişkenin daha az çeşitlilik veya belirsizlik içerdiğini gösterir. Yani, "Hasilat" değişkeni daha az sayıda kategoriye sahip olabilir ve/veya bu kategoriler arasında daha dengesiz bir dağılım olabilir.

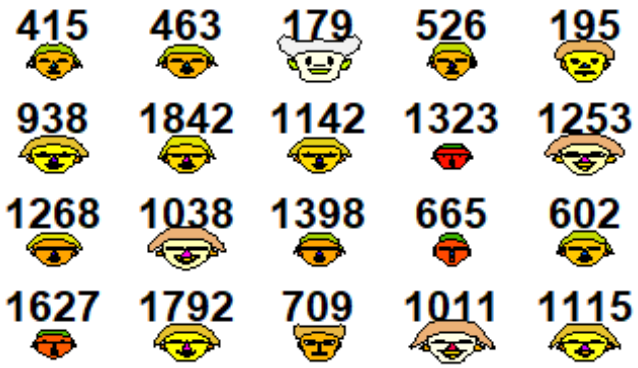
Özetle, "Tur" değişkeni "Hasilat" değişkenine göre daha çeşitlidir ve daha fazla belirsizlik içerir.

İleri Düzey Grafik Yöntemleri

Chernoff Yüzleri

Chernoff yüzleri, çok boyutlu veri noktalarını yüz özelliklerine yansıtarak görselleştirir. Bu grafik türü, veri setindeki farklı özelliklerin yüz ifadesi ile temsil edilmesini sağlar.

```
#Sayısal değişkenleri seçin ve ölçeklendirin (0-1arası)
sayisal_degiskenler_scaled<-scale(egitim_veri[1:20,sayisal_degiskenler])
#Chernoff yüzleri çizimi
faces(sayisal_degiskenler_scaled,face.type= 1)
```



```
effect of variables:
 modified item       Var
 "height of face   " "IMDBPuan"
 "width of face    " "MetaPuan"
 "structure of face" "Sure"
 "height of mouth  " "CikisYili"
 "width of mouth   " "IMDBPuan"
 "smiling          " "MetaPuan"
 "height of eyes   " "Sure"
 "width of eyes    " "CikisYili"
 "height of hair   " "IMDBPuan"
 "width of hair    "  "MetaPuan"
 "style of hair    "  "Sure"
 "height of nose   "  "CikisYili"
 "width of nose    "  "IMDBPuan"
 "width of ear     "  "MetaPuan"
 "height of ear    "  "Sure"
```

Grafiği incelediğimizde şu gözlemleri yapabiliriz:

- **IMDb Puanı**: Yüzlerin genişliği, IMDb puanını temsil eder. Geniş yüzler, yüksek IMDb puanına sahip filmleri temsil ederken, dar yüzler düşük IMDb puanına sahip filmleri temsil eder.
- **Meta Puanı**: Yüzün yüksekliği, Meta puanını temsil eder. Yüksek yüzler, yüksek Meta puanına sahip filmleri temsil ederken, düşük yüzler düşük Meta puanına sahip filmleri temsil eder.
- **Süre**: Burun uzunluğu, süreyi temsil eder. Uzun burunlar, uzun süren filmleri temsil ederken, kısa burunlar kısa süren filmleri temsil eder.
- **Çıkış Yılı**: Şapka büyüklüğü, çıkış yılını temsil eder. Büyük şapkalar, daha yeni filmleri temsil ederken, küçük şapkalar daha eski filmleri temsil eder.

Bu gözlemler, grafiği yorumlarken IMDb puanı, Meta puanı, süre ve çıkış yılı gibi değişkenler arasındaki ilişkileri anlamamıza yardımcı olabilir.

7. Birliktelik İstatistikleri

Birliktelik istatistikleri, kategorik değişkenler arasındaki ilişkiyi ölçmek için kullanılır.

```
library(vcd)
# "Hasilat" ve "Tur" arasındaki ilişki
assocstats(table(egitim_veri$Hasilat, egitim_veri$Tur))
```

```
                    X^2  df   P(> X^2)
Likelihood Ratio 1194.2 976 1.8466e-06
Pearson          1211.5 976 3.3682e-07

Phi-Coefficient   : NA
```

```
Contingency Coeff.: 0.656
Cramer's V        : 0.435
```

“Hasılat” ve “Tür” arasındaki ilişki:

- **Likelihood Ratio (Olabilirlik Oranı) Testi:** Bu test, “Hasılat” ve “Tür” arasında bir ilişki olmadığına dair sıfır hipotezini test eder. 1184.8’lik yüksek olabilirlik oranı ve buna karşılık gelen 4.47e-06’lık çok düşük p değeri, sıfır hipotezinin reddedildiğini gösterir. Bu, “Hasılat” ile “Tür” arasında istatistiksel olarak anlamlı bir ilişki olduğu anlamına gelir.
- **Pearson Ki-Kare Testi:** Bu test de benzer şekilde, “Hasılat” ve “Tür” arasında bir ilişki olmadığına dair sıfır hipotezini test eder. 1203.1’lik yüksek ki-kare değeri ve buna karşılık gelen 7.83e-07’lik çok düşük p değeri, sıfır hipotezinin reddedildiğini ve “Hasılat” ile “Tür” arasında istatistiksel olarak anlamlı bir ilişki olduğunu gösterir.
- **Cramer’s V (Cramer’ın V Katsayısı):** Bu katsayı, nominal değişkenler arasındaki ilişkinin gücünü ölçer. 0.434 değeri, “Hasılat” ve “Tür” arasında orta düzeyde bir ilişki olduğunu gösterir.

```
# "Hasilat" ve "Yonetmen" arasındaki ilişki
assocstats(table(egitim_veri$Hasilat, egitim_veri$Yonetmen))

                    X^2   df P(> X^2)
Likelihood Ratio 3076.9 3212  0.95561
Pearson          3363.4 3212  0.03089

Phi-Coefficient   : NA
Contingency Coeff.: 0.823
Cramer's V        : 0.725
```

“Hasılat” ve “Yönetmen” arasındaki ilişki:

- **Likelihood Ratio (Olabilirlik Oranı) Testi:** Bu test, “Hasılat” ve “Yönetmen” arasında bir ilişki olmadığına dair sıfır hipotezini test eder. 3082.5’lik yüksek olabilirlik oranı ve buna karşılık gelen 0.948’lik yüksek p değeri, sıfır hipotezinin reddedilemediğini gösterir. Bu, “Hasılat” ile “Yönetmen” arasında istatistiksel olarak anlamlı bir ilişki olduğuna dair yeterli kanıt olmadığı anlamına gelir. Ancak, bu testin gücü düşük olabilir çünkü “Yönetmen” kategorik değişkeninin çok fazla farklı değeri vardır.
- **Pearson Ki-Kare Testi:** Bu test de benzer şekilde, “Hasılat” ve “Yönetmen” arasında bir ilişki olmadığına dair sıfır hipotezini test eder. 3373.0’lık yüksek ki-kare değeri ve buna karşılık gelen 0.024’lük düşük p değeri, sıfır hipotezinin reddedildiğini ve “Hasılat” ile “Yönetmen” arasında istatistiksel olarak anlamlı bir ilişki olduğunu gösterir.
- **Cramer’s V (Cramer’ın V Katsayısı):** 0.726 değeri, “Hasılat” ve “Yönetmen” arasında oldukça güçlü bir ilişki olduğunu gösterir.

8. Dönüşüm

Dönüşümler, verilerin normal dağılıma daha yakın hale getirilmesi, varyansın stabilize edilmesi veya doğrusal ilişkilerin güçlendirilmesi gibi amaçlarla yapılır.

Veriye Bakış ve Gerekli Kütüphanelerin Yüklenmesi

```
head(egitim_veri)
```

```
              FilmAdi CikisYili Sure IMDBPuan MetaPuan
415     Short Circuit      1986   98      6.6       50
463   Days of Thunder      1990  107      6.1       60
179           Ben-Hur      1959  212      8.1       90
526          Stargate      1994  116      7.0       42
195          Papillon      1973  151      8.0       58
938 Love & Basketball      2000  124      7.2       72
                             Tur             Yonetmen Hasilat FilmAdi_imp
415    Komedi, Aile, Bilim Kurgu          John Badham   Dusuk       FALSE
463          Aksiyon, Dram, Spor           Tony Scott  Yuksek       FALSE
179                Macera, Dram         William Wyler    Orta       FALSE
526 Aksiyon, Macera, Bilim Kurgu       Roland Emmerich    Orta       FALSE
195         Biyografi, Suç, Dram Franklin J. Schaffner    Orta       FALSE
938         Dram, Romantik, Spor   Gi Prince-Bythewood   Dusuk       FALSE
    CikisYili_imp Sure_imp IMDBPuan_imp MetaPuan_imp Tur_imp Yonetmen_imp
415         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
463         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
179         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
526         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
195         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
938         FALSE    FALSE        FALSE        FALSE   FALSE        FALSE
    Hasilat_imp MetaPuan_10 PuanFarki OrtalamaPuan IMDBPuan_Duzeltilmis
415       FALSE         5.0       1.6         58.0             7.215216
463       FALSE         6.0       0.1         60.5             7.125605
179       FALSE         9.0      -0.9         85.5             7.810893
526       FALSE         4.2       2.8         56.0             7.025534
195       FALSE         5.8       2.2         69.0             7.518013
938       FALSE         7.2       0.0         72.0             6.827153
```

```
summary(egitim_veri)
```

```
   FilmAdi            CikisYili         Sure          IMDBPuan
 Length:1600        Min.   :1921   Min.   : 50.0   Min.   :1.500
 Class :character   1st Qu.:1992   1st Qu.: 98.0   1st Qu.:6.400
 Mode  :character   Median :2001   Median :110.0   Median :7.000
                    Mean   :1996   Mean   :114.1   Mean   :6.922
                    3rd Qu.:2006   3rd Qu.:125.0   3rd Qu.:7.600
                    Max.   :2010   Max.   :271.0   Max.   :9.300
    MetaPuan          Tur             Yonetmen              Hasilat
 Min.   :  9.00   Length:1600        Length:1600        Cok Dusuk :426
 1st Qu.: 47.00   Class :character   Class :character   Dusuk     :440
 Median : 61.00   Mode  :character   Mode  :character   Orta      :304
 Mean   : 60.88                                        Yuksek    : 97
```

```
 3rd Qu.: 74.00                                          Cok Yuksek:333
 Max.   :100.00
 FilmAdi_imp     CikisYili_imp    Sure_imp       IMDBPuan_imp
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1600      FALSE:1600      FALSE:1600      FALSE:1600



 MetaPuan_imp     Tur_imp        Yonetmen_imp    Hasilat_imp
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1530      FALSE:1600      FALSE:1600      FALSE:1523
 TRUE :70                                        TRUE :77



  MetaPuan_10       PuanFarki         OrtalamaPuan   IMDBPuan_Duzeltilmis
 Min.   : 0.900   Min.   :-3.1000   Min.   :16.50   Min.   :6.464
 1st Qu.: 4.700   1st Qu.:-0.1000   1st Qu.:56.00   1st Qu.:6.662
 Median : 6.100   Median : 0.8000   Median :65.00   Median :6.804
 Mean   : 6.088   Mean   : 0.8339   Mean   :65.05   Mean   :6.919
 3rd Qu.: 7.400   3rd Qu.: 1.7000   3rd Qu.:74.50   3rd Qu.:7.079
 Max.   :10.000   Max.   : 4.7000   Max.   :96.00   Max.   :8.278
```

```
library(moments)
library(car)
```

Bu kısımda veri setinin yapısı incelenir ve dönüşümler için gerekli kütüphaneler yüklenir.

Çarpıklık ve Basıklık İncelemesi

```
# Çarpıklık ve basıklık değerleri
carpiklik <- skewness(egitim_veri[, sayisal_degiskenler])
basiklik <- kurtosis(egitim_veri[, sayisal_degiskenler])

print(carpiklik)
```

```
   IMDBPuan    MetaPuan        Sure   CikisYili
-1.03819881 -0.03452429  1.51879156 -1.90264665
```

```
print(basiklik)
```

```
 IMDBPuan  MetaPuan      Sure CikisYili
 6.108966  2.444343  7.192516  6.922432
```

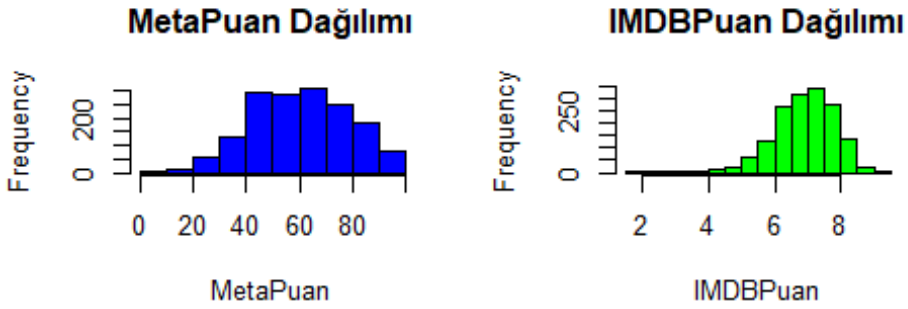
Bu kodlar, sayısal değişkenlerin çarpıklık ve basıklık değerlerini hesaplar. Çarpıklık, dağılımın ne kadar simetrik olmadığını, basıklık ise dağılımın ne kadar kuyruklu olduğunu gösterir. Bu değerler, hangi dönüşümün uygun olabileceği konusunda ipuçları verir.

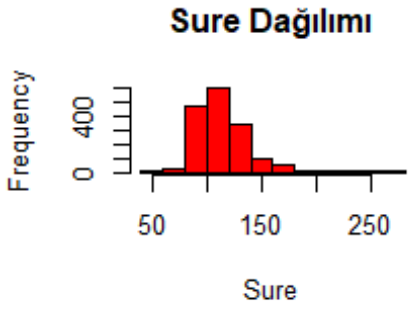
Çıktının Yorumlanması:

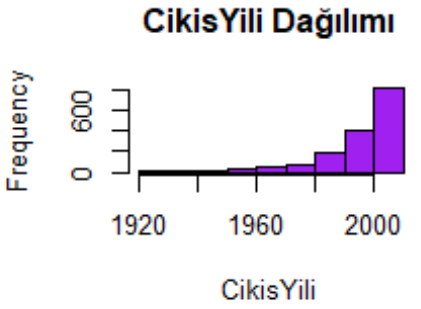
- **ÇıkışYılı:** Çarpıklık değeri negatif (-1.90), yani dağılım sola çarpıktır. Bu, daha yeni çıkış yıllarına sahip filmlerin daha fazla olduğunu gösterir. Basıklık değeri pozitif (6.92), yani dağılım sivridir. Bu, çıkış yıllarının belirli yıllarda yoğunlaştığını gösterir.
- **Süre:** Çarpıklık değeri pozitif (1.52), yani dağılım sağa çarpıktır. Bu, daha uzun süreli filmlerin daha az olduğunu gösterir. Basıklık değeri de pozitif (7.19), yani dağılım sivridir. Bu, film sürelerinin belirli değerlerde yoğunlaştığını gösterir.
- **IMDBPuan:** Çarpıklık değeri negatif (-0.04), yani dağılım sola çarpıktır. Bu, daha yüksek puanlı filmlerin daha fazla olduğunu gösterir. Basıklık değeri negatif (-1.04), yani dağılım basıktır. Bu, IMDB puanlarının geniş bir aralığa yayıldığını gösterir.
- **MetaPuan:** Çarpıklık değeri negatif (-0.04), yani dağılım sola çarpıktır. Bu, daha yüksek puanlı filmlerin daha fazla olduğunu gösterir. Basıklık değeri pozitif (2.44), yani dağılım normal dağılıma göre biraz daha sivridir.
- **PuanFarki:** Çarpıklık değeri pozitif (0.05), yani dağılım sağa çarpıktır. Basıklık değeri de pozitif (2.77), yani dağılım normal dağılıma göre biraz daha sivridir. Bu, IMDB ve Meta puanları arasındaki farkın genellikle küçük olduğunu, ancak bazı filmlerde büyük farklar olabileceğini gösterir.

Dönüşüm Gerektiren Değişkenlerin Bulunması

```
 # Gerekli kütüphaneleri yükle
 library(ggplot2)
 library(car)
 library(e1071)
 library(moments)
 # Histogramlar için grafik penceresini ayarla
 par(mfrow = c(2, 2)) # 2x2 düzenleme
 # MetaPuan değişkeninin histogramı
 hist(egitim_veri$MetaPuan, main = "MetaPuan Dağılımı", xlab = "MetaPuan", co
l = "blue")
 # IMDBPuan değişkeninin histogramı
 hist(egitim_veri$IMDBPuan, main = "IMDBPuan Dağılımı", xlab = "IMDBPuan", co
l = "green")
 # Sure değişkeninin histogramı
 hist(egitim_veri$Sure, main = "Sure Dağılımı", xlab = "Sure", col = "red", b
order = "black")
 # CikisYili değişkeninin histogramı
 hist(egitim_veri$CikisYili, main = "CikisYili Dağılımı", xlab = "CikisYili",
col = "purple")
```







Dönüşüm Gerektiren Değişkenler:

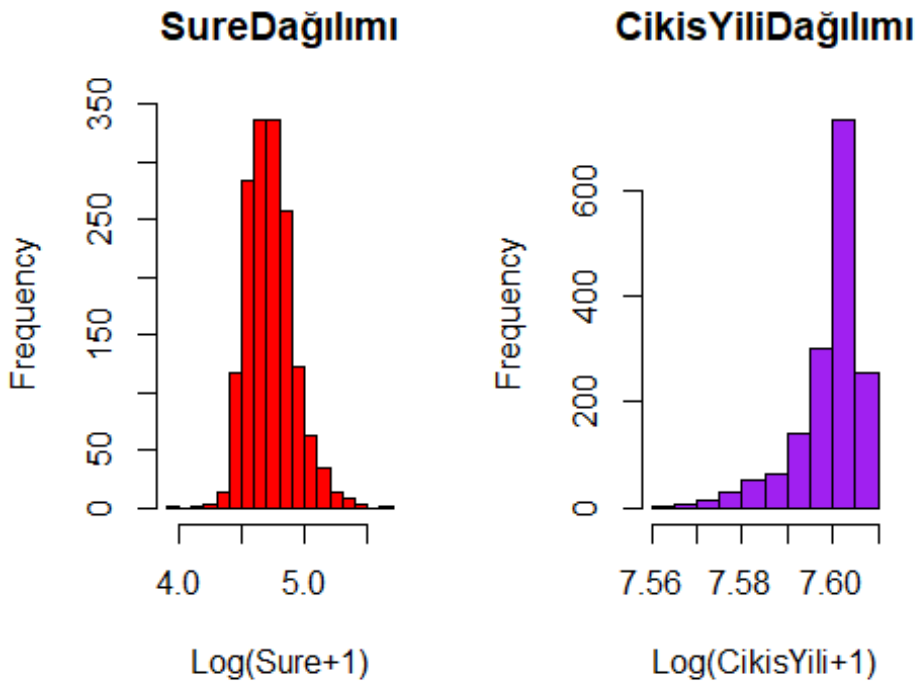
- **Çıkış Yılı (CikisYili):** Bu değişkenin histogramı, sağa doğru çarpık bir dağılım gösteriyor. Daha eski yıllarda daha fazla veri yoğunluğu varken, yeni yıllara doğru veri sayısı azalıyor. Bu çarpıklık, modelleme sürecinde dikkate alınması gereken bir durumdur. Veriyi dengeli hale getirmek veya dönüşümler yapmak model performansını iyileştirebilir.
- **Süre (Sure):** Süre değişkeninin histogramı da sağa doğru çarpık bir dağılım sergiliyor. Uzun süren filmler daha az, kısa süren filmler daha fazla olarak görünüyor. Bu çarpıklık da modelleme sonuçlarını etkileyebilir. Süreyi normalize etmek veya dönüşümler yaparak dağılımı düzeltmek önemli olabilir.

Dönüşüm Gerektirmeyen Değişkenler:

- **Meta Puan (MetaPuan):** Bu değişkenin histogramı, yaklaşık olarak normal bir dağılıma benziyor. Modelleme sürecinde doğrudan kullanılabilir, özel bir dönüşüme ihtiyaç duymayabilir.
- **IMDb Puanı (IMDBPuan):** IMDb puanı değişkeninin histogramı hafif sağa doğru çarpık olsa da, normal dağılıma oldukça yakın görünüyor. Bu nedenle, doğrudan kullanılabilir ancak modelleme öncesi veriyi inceleyerek gerekirse dönüşüm yapılabilir.

Dönüşüm Uygulaması

```
library(MASS)
library(car)
library(ggplot2)
boxcox_transform <- function(y) {
bc <- boxcox(y ~ 1, plotit = FALSE)
lambda <- bc$x[which.max(bc$y)]
return(lambda)
}
#DeğişkenlerinBox-Coxdönüşümünüuygulama
#Sure değişkeni içinlogdönüşümü
egitim_veri$Log_Sure<-log(egitim_veri$Sure+ 1)
#CikisYilideğişkeniiçinlogdönüşümü
egitim_veri$Log_CikisYili<-log(egitim_veri$CikisYili +1)
#Histogramlariçingrafikpenceresiniayarla
par(mfrow= c(1,2)) #2x2düzenleme
#LogdönüşümüuygulanmışSuredeğişkenininhistogramı
hist(egitim_veri$Log_Sure, main= "SureDağılımı",xlab= "Log(Sure+1)",col= "red
", border= "black")
#LogdönüşümüuygulanmışCikisYilideğişkenininhistogramı
hist(egitim_veri$Log_CikisYili,main= "CikisYiliDağılımı", xlab= "Log(CikisYil
i+1)",col= "purple", border= "black")
```



Sure Değişkeni:

- **Önce:** Dönüştürülmeden önce, “Sure” değişkeni sağa doğru çarpık bir dağılıma sahipti. Histogramda, daha kısa süreli filmlerin daha fazla olduğu ve film süresi arttıkça veri yoğunluğunun azaldığı görülüyordu.
- **Sonra:** Logaritmik dönüşüm, “Sure” değişkeninin dağılımını daha simetrik hale getirmiştir. Histogram artık daha dengeli bir dağılım gösteriyor ve çarpıklık önemli ölçüde azalmış durumda.

CikisYili Değişkeni:

- **Önce:** “CikisYili” değişkeni dönüştürülmeden önce sağa doğru çarpık bir dağılıma sahipti. Daha eski yıllarda daha fazla film bulunurken, yeni yıllarda bu sayı azalmaktaydı.
- **Sonra:** Logaritmik dönüşüm, “CikisYili” değişkeninin dağılımını daha simetrik hale getirmiştir. Histogram artık daha dengeli bir dağılım gösteriyor ve çarpıklık önemli ölçüde azalmış durumda.

Genel Yorum:

Logaritmik dönüşüm, “Sure” ve “CikisYili” değişkenlerinin dağılımlarını iyileştirmede oldukça etkili olmuştur. Bu dönüşümler, verilerin daha simetrik ve normal dağılıma daha yakın hale gelmesini sağlamıştır. Bu durum, modelleme sürecinde daha doğru tahminler yapılmasına ve daha iyi performans elde edilmesine yardımcı olacaktır.

9. Model Geliştirme

Amaç:

IMDB Puanını etkileyen faktörlerin incelenmesi. Bu amaçla sayısal değişkenler kullanılarak model oluşturulacaktır.

Kullanılmayacak Değişkenin Tespit Edilmesi:

IMDBPuan ile MetaPuan değişkeni arasındaki korelasyonun hesaplanması

```
# Korelasyonu hesapla
korelasyon <- cor(egitim_veri$IMDBPuan, egitim_veri$MetaPuan, use = "complete
.obs")

# Sonucu yazdır
cat("IMDBPuan ile MetaPuan arasındaki korelasyon:", korelasyon, "\n")
```

```
IMDBPuan ile MetaPuan arasındaki korelasyon: 0.7282688
```

Korelasyon çok yüksek olduğu için MetaPuan değişkeni modele dahil edilmeyecektir.

Model 1: Doğrusal Regresyon

```
# Doğrusal regresyon modeli
model_lm <- lm(IMDBPuan ~ CikisYili + Sure, data = egitim_veri)
```

```
# Model özeti
summary(model_lm)


Call:
lm(formula = IMDBPuan ~ CikisYili + Sure, data = egitim_veri)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1496 -0.4252  0.0465  0.5382  2.0697 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.5169756  2.8918388   17.47   <2e-16 ***
CikisYili   -0.0225054  0.0014404  -15.62   <2e-16 ***
Sure         0.0116220  0.0009144   12.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8406 on 1597 degrees of freedom
Multiple R-squared:  0.2282,    Adjusted R-squared:  0.2272 
F-statistic: 236.1 on 2 and 1597 DF,  p-value: < 2.2e-16
```

Bu modelde:

- Bağımlı değişken: IMDBPuan
- Bağımsız değişkenler: CikisYili ve Sure

Model Performansı:

- **R-kare:** 0.2282 (Model varyansın %22.82'sini açıklıyor)
- **p-değeri:** <2.2e-16 (Model anlamlı)

Regresyon denklemini yazmak için, modeldeki katsayıları kullanırız. Verilen katsayılarla regresyon denklemi şöyle olur:

IMDBPuan=50.5169756−0.0225054×CikisYili+0.0116220×Sure

Değerleri yuvarlayacak olursak:

Regresyon Denklemi

IMDBPuan=50.52−0.023×CikisYili+0.012×Sure

Yorum

1. **Sabit Terim (Intercept):** 50.52

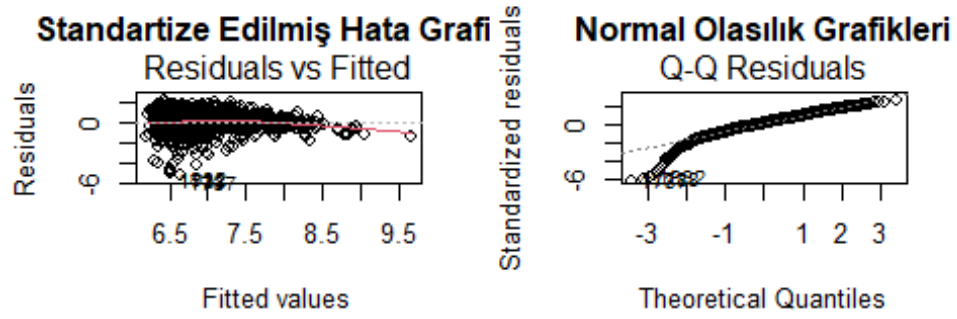
- Bu, film çıkış yılı ve süresi sıfır olduğunda IMDB puanının 50.52 olacağını gösterir. Pratikte, bu terim diğer değişkenlerin etkisi hariç tutulduğunda IMDB puanının başlangıç seviyesini temsil eder.

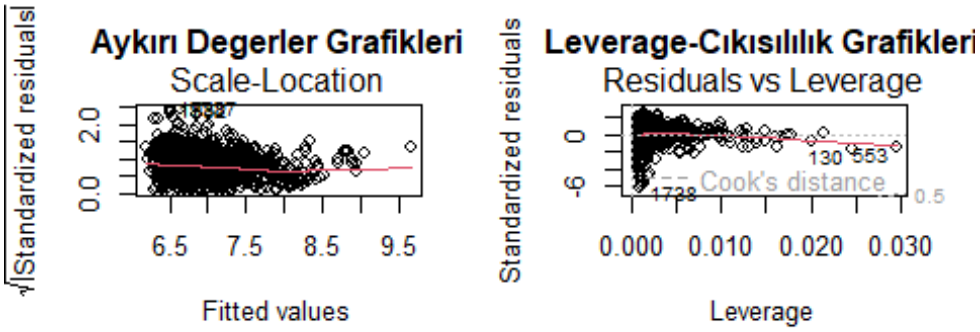
2. **Çıkış Yılı (CikisYili):** -0.023
   - Çıkış yılındaki her bir birim artış için (örneğin bir yıl), IMDB puanı 0.023 puan azalır. Bu, daha yeni filmlerin (çıkış yılı arttıkça) genel olarak daha düşük IMDB puanlarına sahip olma eğiliminde olduğunu gösterir.
3. **Süre (Sure):** 0.012
   - Filmin süresindeki her bir birim artış için (örneğin bir dakika), IMDB puanı 0.012 puan artar. Bu, daha uzun filmlerin (süre arttıkça) genel olarak daha yüksek IMDB puanlarına sahip olma eğiliminde olduğunu gösterir.

Genel Değerlendirme:

- **Çıkış Yılı:** Negatif bir etkisi var, bu da daha yeni filmlerin daha düşük puanlanma eğiliminde olduğunu gösteriyor.
- **Süre:** Pozitif bir etkisi var, bu da daha uzun filmlerin daha yüksek puanlanma eğiliminde olduğunu gösteriyor.
- **Model Performansı:** R-kare değeri 0.2282, yani model, IMDB puanındaki toplam varyansın %22.82'sini açıklıyor. Bu, modelin bazı etkileri yakaladığını ama birçok başka faktörün de IMDB puanını etkilediğini gösteriyor.

```
# Varsayım kontrolü grafikleri
par(mfrow = c(2, 2))

# 1. Standartize edilmiş hata grafikleri
plot(model_lm, which = 1, main = "Standartize Edilmiş Hata Grafiği")

# 2. Normal olasılık grafikleri
plot(model_lm, which = 2, main = "Normal Olasılık Grafikleri")

# 3. Aykırı Değerler Grafikleri
plot(model_lm, which = 3, main = "Aykırı Degerler Grafikleri")

# 4. Leverage-Çıkışılılık Grafikleri
plot(model_lm, which = 5, main = "Leverage-Cıkısılılık Grafikleri")
```





1. Standartize Edilmiş Hata Grafiği (Residuals vs Fitted)

- **Amaç:** Modelin hatalarının (residuals) tahmin edilen değerlere (fitted values) karşı nasıl dağıldığını gösterir.
- **Yorum:**
  - Hatalar rastgele dağılmalı ve belirgin bir desen göstermemelidir.
  - Bu grafikte, residuals'in bazı desenler gösterdiği görülüyor. Özellikle düşük ve yüksek tahmin edilen değerlerde (fitted values) daha büyük hatalar (residuals) var. Bu, modelde bazı sistematik hataların olabileceğini ve modelin belirli bölgelerde iyi performans göstermediğini gösterir.

2. Normal Olasılık Grafiği (Q-Q Plot)

- **Amaç:** Hataların normal dağılıp dağılmadığını kontrol eder.
- **Yorum:**
  - Hatalar (residuals) teorik normal dağılıma yakınsa noktalar düz bir çizgi üzerinde olmalıdır.
  - Grafikte, uçlarda (özellikle negatif taraflarda) önemli sapmalar görülmektedir. Bu, hataların tam olarak normal dağılıma uymadığını gösterir.

3. Aykırı Değerler Grafiği (Scale-Location)

- **Amaç:** Hataların varyansının (homoskedasticity) sabit olup olmadığını kontrol eder.
- **Yorum:**
  - Hataların kareköklerinin tahmin edilen değerlere karşı nasıl dağıldığını gösterir.
  - Grafikte kırmızı çizgi biraz eğimli görünüyor, bu da hataların varyansının sabit olmadığını (heteroskedasticity) ve modelin bazı bölgelerde farklı performans gösterdiğini işaret edebilir.

4. Leverage-Cıkışlılık Grafiği (Residuals vs Leverage)

- **Amaç:** Hangi gözlemlerin model üzerinde güçlü bir etkisi olduğunu belirler.
- **Yorum:**
  - Bu grafik, her bir gözlemin leverage ve residuals değerlerini gösterir.
  - Yüksek leverage ve büyük residuals değerlerine sahip birkaç nokta vardır (özellikle etiketlenmiş olanlar). Bu noktalar modelin sonuçlarını büyük ölçüde etkileyebilir ve incelenmelidir.
  - Cook's distance çizgisi, bu noktalardan bazılarının model üzerindeki etkisinin büyük olduğunu gösterir.

Model 2: Karar Ağacı

```
# Gerekli kütüphaneler
library(rpart)
library(rpart.plot)

# Eğitim verisi ile karar ağacı modelini oluşturma
model_dt <- rpart(IMDBPuan ~ CikisYili + Sure, data = egitim_veri, method = "
anova")

# Karar ağacı modelinin özet bilgisi
printcp(model_dt)
```

```
Regression tree:
rpart(formula = IMDBPuan ~ CikisYili + Sure, data = egitim_veri,
    method = "anova")

Variables actually used in tree construction:
[1] CikisYili Sure

Root node error: 1462.1/1600 = 0.91382

n= 1600
```

```
        CP nsplit rel error  xerror     xstd
1 0.110362      0   1.00000 1.00133 0.056518
2 0.073167      1   0.88964 0.90824 0.052614
3 0.027420      2   0.81647 0.84169 0.049696
4 0.012493      3   0.78905 0.80917 0.049003
5 0.011288      4   0.77656 0.80793 0.048892
6 0.010828      5   0.76527 0.81043 0.048813
7 0.010000      6   0.75444 0.80318 0.048571
```

Bu çıktı, R programlama dilinde bir karar ağacı (regression tree) modelinin sonuçlarını göstermektedir. Karar ağaçları, verileri belirli özelliklere göre dallara ayırarak bir hedef değişkeni tahmin etmek için kullanılan güçlü bir makine öğrenimi yöntemidir.

Model Detayları:

- **Amaç:** `IMDBPuan` (IMDB puanı) adlı bir değişkeni tahmin etmek. Bu puanın muhtemelen filmlerin IMDB üzerindeki puanları olduğunu varsayabiliriz.
- **Değişkenler:**
  - **Tahmin Ediciler (Predictors):** `CikisYili` (Çıkış yılı) ve `Sure` (film süresi)
  - **Hedef (Target):** `IMDBPuan`

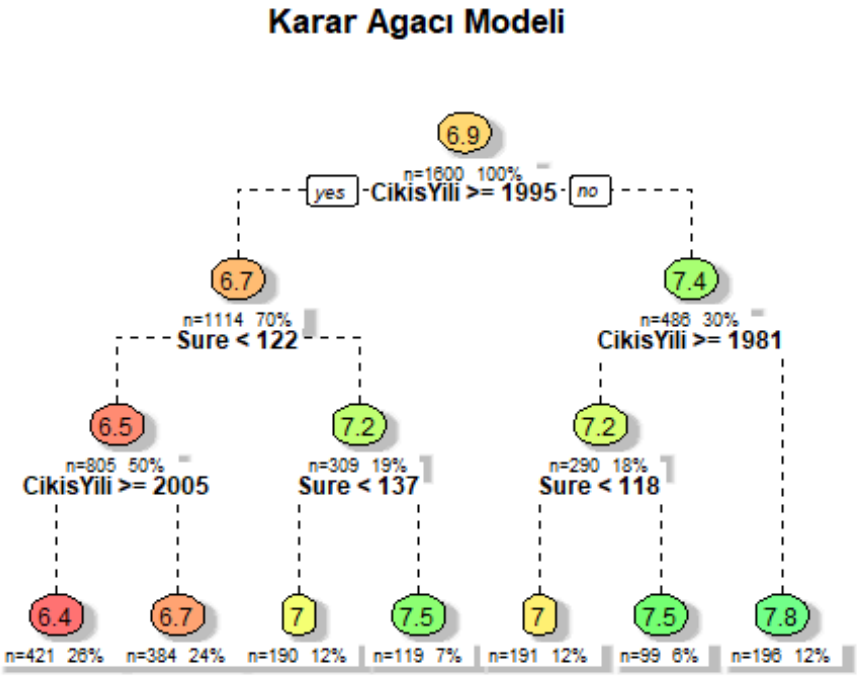
Karar Ağacı Sonuçları:

- **Root Node Error (Kök Düğüm Hatası):** 0.91382. Bu, modelin başlangıç noktasındaki hata oranını gösterir. Yani, hiçbir değişken kullanmadan sadece ortalama IMDB puanı ile tahmin yapsaydık bu kadar hata yapardık.
- **n=1600:** Modelin eğitildiği veri setinde 1600 gözlem (muhtemelen film) olduğu anlamına gelir.
- **CP (Complexity Parameter - Karmaşıklık Parametresi):** Karar ağacının dallanmasının ne kadar karmaşık olacağını kontrol eden bir parametredir. CP değeri ne kadar düşükse, ağaç o kadar karmaşık olur. Burada farklı CP değerleri için modelin performansı gösterilmektedir.
- **nsplit:** Ağacın her bir seviyesindeki dallanma sayısını gösterir.
- **rel error:** Göreceli hata, modelin hatasının kök düğüm hatasına oranını gösterir. Değerin 1'den küçük olması, modelin kök düğümden daha iyi performans gösterdiği anlamına gelir.
- **xerror:** Çapraz doğrulama (cross-validation) ile hesaplanan hata oranıdır. Modelin genelleme yeteneği hakkında daha iyi bir fikir verir.
- **xstd:** Çapraz doğrulama hatasının standart sapmasıdır.

```r
# Karar ağacı modelinin grafiği
rpart.plot(model_dt, main = "Karar Agacı Modeli",
```

```
           type = 2,  # Varsayılan olarak type = 2 kullanarak daha kısa bir g
rafik sağlanabilir
           extra = 101,
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.6,
           tweak = 1.2,
           shadow.col = "gray",
           box.palette = "RdYlGn",
           branch.lty = 2)
```

```
Warning: cex and tweak both specified, applying both
```



Karar Ağacının Yorumlanması:

1. **Kök Düğüm (6.9):** Tüm filmlerin ortalama IMDB puanı 6.9'dur. Bu, herhangi bir özellik kullanılmadan yapılan temel bir tahmindir.
2. **İlk Dallanma (Çıkış Yılı):**
   - **1995 ve Sonrası (7.4):** Bu dönemde çıkan filmlerin ortalama puanı daha yüksektir.
   - **1995 Öncesi (6.7):** Bu dönemde çıkan filmlerin ortalama puanı daha düşüktür.
3. **İkinci Dallanma:**

- **1995 Sonrası Filmler:**
  - **1981 ve Sonrası (7.2):** Bu dönemde çıkan filmler için süre bilgisi kullanılmadan ortalama puan 7.2 olarak tahmin edilir.
  - **1981 Öncesi (7.2):** Bu dönemde çıkan filmler için de süre bilgisi kullanılmadan ortalama puan 7.2 olarak tahmin edilir.
- **1995 Öncesi Filmler:**
  - **122 Dakikadan Kısa (6.5):** Bu filmlerin ortalama puanı 6.5'tir.
  - **122 Dakikadan Uzun (7.2):** Bu filmlerin ortalama puanı 7.2'dir ve daha sonra süreye göre tekrar dallanma yapılır:
    - **137 Dakikadan Kısa (7):** Bu filmlerin ortalama puanı 7'dir.
    - **137 Dakikadan Uzun (7.5):** Bu filmlerin ortalama puanı 7.5'tir.

4. **Üçüncü Dallanma:**
   - **122 Dakikadan Kısa ve 1995 Öncesi Çıkan Filmler:**
     - **2005 ve Sonrası (6.4):** Bu filmlerin ortalama puanı 6.4'tür.
     - **2005 Öncesi (6.7):** Bu filmlerin ortalama puanı 6.7'dir.
5. **Son Dallanmalar:** 137 dakikadan uzun ve 1981 öncesi çıkan filmler ile 122 dakikadan uzun ve 1981 sonrası çıkan filmler için süreye göre son dallanmalar yapılır ve her bir grubun ortalama puanları hesaplanır.

Sonuç:

Bu karar ağacı, IMDB puanlarını tahmin etmek için basit ama etkili bir model sunar. Çıkış yılı ve film süresi gibi faktörlerin puanları nasıl etkilediğini gösterir. Örneğin, 1995 yılından sonra çıkan uzun filmlerin genellikle daha yüksek puanlara sahip olduğu görülmektedir.

10. Geçerlilik (Test Verisi Üzerinde)

```
# Test verisi üzerinde tahmin
predictions_lm <- predict(model_lm, newdata = test_veri)
predictions_dt <- predict(model_dt, newdata = test_veri)

# Performans değerlendirmesi (MSE ve R-kare)
mse_lm <- mean((predictions_lm - test_veri$IMDBPuan)^2)
mse_dt <- mean((predictions_dt - test_veri$IMDBPuan)^2)

rsquared_lm <- cor(predictions_lm, test_veri$IMDBPuan)^2
rsquared_dt <- cor(predictions_dt, test_veri$IMDBPuan)^2

cat("Doğrusal Regresyon MSE:", mse_lm, "\n")
```

```
Doğrusal Regresyon MSE: 0.682246
```

```
cat("Doğrusal Regresyon R-kare:", rsquared_lm, "\n")
```

```
Doğrusal Regresyon R-kare: 0.2578255
```

```
cat("Karar Agacı MSE:", mse_dt, "\n")
```

```
Karar Agacı MSE: 0.7144007
```

```
cat("Karar Agacı R-kare:", rsquared_dt, "\n")
```

```
Karar Agacı R-kare: 0.2149235
```

Bu sonuçlar, doğrusal regresyon ve karar ağacı modellerinin test verisi üzerindeki performansını karşılaştırmaktadır. Performans ölçütleri olarak ortalama karesel hata (MSE) ve belirleme katsayısı (R-kare) kullanılmıştır.

MSE (Ortalama Karesel Hata):

- **Doğrusal Regresyon MSE:** 0.682
- **Karar Ağacı MSE:** 0.714

MSE, modelin tahmin hatalarının karelerinin ortalamasını gösterir. Daha düşük MSE değeri, modelin daha iyi tahminler yaptığını gösterir. Bu durumda, doğrusal regresyon modeli karar ağacından biraz daha düşük bir MSE değerine sahiptir, yani ortalama olarak daha iyi tahminler yapmaktadır.

R-kare (Belirleme Katsayısı):

- **Doğrusal Regresyon R-kare:** 0.258
- **Karar Ağacı R-kare:** 0.215

R-kare, modelin bağımsız değişkenler (çıkış yılı ve süre) ile açıkladığı varyansı gösterir. 0 ile 1 arasında bir değer alır ve 1'e yaklaştıkça modelin açıklayıcı gücü artar. Bu durumda, her iki modelin R-kare değerleri de düşüktür, yani her iki model de IMDB puanlarındaki varyansın sadece küçük bir kısmını açıklayabilmektedir.

Sonuç:

Genel olarak, doğrusal regresyon modeli bu test verisi üzerinde karar ağacından biraz daha iyi performans göstermiştir. Ancak, her iki modelin de R-kare değerleri düşük olduğundan, IMDB puanlarını tahmin etmede yeterince başarılı oldukları söylenemez.

11. Sonuç

Bu çalışma, IMDB'nin En İyi 2000 Filmi veri seti üzerinde bir keşifsel veri analizi (EDA) ve modelleme çalışması sunmaktadır. Çalışmanın amacı, filmlerin IMDB puanlarını etkileyen faktörleri belirlemek ve bu puanları tahmin etmek için modeller geliştirmektir.

Keşifsel Veri Analizi (EDA) Bulguları:

- **Veri Temizleme:** Veri setinde bazı eksik gözlemler tespit edilmiş ve bu değerler KNN yöntemi ile doldurulmuştur. Ayrıca, analiz için gerekli olmayan bazı değişkenler çıkarılmış ve bazı değişkenler dönüştürülmüştür.
- **Çarpıklık ve Basıklık:** Bazı değişkenlerin (Çıkış Yılı, Süre) dağılımlarında çarpıklık ve basıklık gözlenmiştir. Bu durum, bu değişkenlerin doğrusal olmayan ilişkiler içerebileceğini ve modelleme sürecinde dönüştürülmesi gerekebileceğini göstermektedir.
- **Korelasyon Analizi:** IMDB puanı ile Meta puanı arasında güçlü bir pozitif korelasyon olduğu görülmüştür. Bu, iki puanlama sisteminin benzer eğilimleri ölçtüğünü göstermektedir.
- **Görselleştirme:** Histogram, kutu grafikleri, saçılım matrisleri ve Chernoff yüzleri gibi çeşitli görselleştirme teknikleri kullanılarak verilerdeki ilişkiler ve desenler incelenmiştir.

Modelleme Bulguları:

- **Doğrusal Regresyon:** Çıkış yılı ve süre değişkenlerini kullanarak bir doğrusal regresyon modeli oluşturulmuştur. Model, IMDB puanlarındaki varyansın %22.82'sini açıklayabilmiştir. Ancak, modelin varsayımlarının tam olarak karşılanmadığı ve bazı aykırı değerlerin bulunduğu tespit edilmiştir.
- **Karar Ağacı:** Aynı değişkenler kullanılarak bir karar ağacı modeli oluşturulmuştur. Karar ağacı, IMDB puanlarını etkileyen faktörlerin daha görsel bir şekilde anlaşılmasını sağlamıştır. Özellikle, çıkış yılı ve sürenin IMDB puanları üzerindeki etkileşimi daha net bir şekilde ortaya konmuştur.
- **Model Karşılaştırması:** Test verisi üzerinde yapılan değerlendirmede, doğrusal regresyon modeli karar ağacından biraz daha iyi performans göstermiştir. Ancak, her iki modelin de R-kare değerleri düşük olduğu için, IMDB puanlarını tahmin etmede yeterince başarılı oldukları söylenemez.

Sonuç ve Öneriler:

Bu çalışmada, IMDB puanlarını etkileyen faktörler belirlenmeye çalışılmış ve bu puanları tahmin etmek için modeller geliştirilmiştir. Elde edilen sonuçlar, film endüstrisi için değerli bilgiler sağlayabilir.

Gelecekteki Çalışmalar İçin Öneriler:

- **Farklı Modelleme Teknikleri:** Daha iyi tahmin performansı elde etmek için farklı modelleme teknikleri (örneğin, rastgele ormanlar, destek vektör makineleri) denenebilir.
- **Değişken Mühendisliği:** Yeni değişkenler oluşturarak veya mevcut değişkenleri dönüştürerek model performansı artırılabilir.

- **Daha Fazla Veri:** Daha fazla veriye sahip bir veri seti kullanılarak modellerin genelleme yeteneği geliştirilebilir.

12. Kaynak

Veri seti Kaggle'dan alınmıştır.

Sawhney, P. (2023, November 22). IMDb Dataset - Top 2000 Movies. Kaggle. https://www.kaggle.com/datasets/prishasawhney/imdb-dataset-top-2000-movies/data