Data Cleaning, Analysis and Visualization

Mehmet Emin Sahin

2023-12-09

Veri Analizi ve Görselleştirmenin Önemi

Günümüzde veri, her sektörde karar alma süreçlerinin temelini oluşturmakta ve iş dünyasından sağlık sektörüne kadar geniş bir yelpazede kullanılmaktadır. Veri analizi, büyük veri kümelerinden anlamlı bilgiler çıkarmak için kullanılırken, veri görselleştirme bu bilgileri anlaşılır ve etkili bir şekilde sunmanın en etkili yoludur. Bu çalışmada, Netflix platformundaki çeşitli filmler ve TV şovları hakkında bilgiler içeren bir veri seti kullanılarak, veri temizliği, analizi ve görselleştirmesi süreçlerini detaylı bir şekilde inceleyeceğiz. Amaç, veri biliminin iş dünyasındaki karar alma süreçlerine nasıl katkı sağladığını göstermek ve veri görselleştirmenin bilgi iletimindeki gücünü ortaya koymaktır.

Veri Seti Hakkında

Veri seti, Netflix platformundaki çeşitli filmler ve TV şovları hakkında bilgiler içermektedir.

- `show_id`: Her gösterinin benzersiz bir tanımlayıcısı.
- `type`: Gösterinin tipi (Film veya TV Şovu).
- `title`: Gösterinin adı.
- `director`: Gösterinin yönetmeni.
- `country`: Gösterinin üretildiği ülke.
- `date_added`: Netflix'e eklendiği tarih.
- `release_year`: Gösterinin yayınlandığı yıl.
- `rating`: Gösterinin derecelendirmesi.
- `duration`: Gösterinin süresi.
- `listed_in`: Gösterinin dahil olduğu kategoriler.

Veri Setinin İçeri Aktarılması

```
netflix <- read.csv("C:/Users/mehmet/Desktop/netflix1.csv")
head(netflix)
```

```
##   show_id    type                            title        director
## 1      s1   Movie             Dick Johnson Is Dead Kirsten Johnson
## 2      s3 TV Show                        Ganglands Julien Leclercq
## 3      s6 TV Show                    Midnight Mass   Mike Flanagan
## 4     s14   Movie Confessions of an Invisible Girl   Bruno Garotti
## 5      s8   Movie                          Sankofa    Haile Gerima
## 6      s9 TV Show    The Great British Baking Show Andy Devonshire
##          country date_added release_year rating  duration
## 1  United States  9/25/2021         2020  PG-13    90 min
```

```
## 2         France  9/24/2021         2021  TV-MA  1 Season
## 3  United States  9/24/2021         2021  TV-MA  1 Season
## 4         Brazil  9/22/2021         2021  TV-PG    91 min
## 5  United States  9/24/2021         1993  TV-MA   125 min
## 6 United Kingdom  9/24/2021         2021  TV-14 9 Seasons
##                                                      listed_in
## 1                                                  Documentaries
## 2 Crime TV Shows, International TV Shows, TV Action & Adventure
## 3                            TV Dramas, TV Horror, TV Mysteries
## 4                            Children & Family Movies, Comedies
## 5              Dramas, Independent Movies, International Movies
## 6                                  British TV Shows, Reality TV
```

Gerekli Kütüphanelerin Yüklenmesi

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.4      v purrr   1.0.2
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(rnaturalearth)
```

```
## Support for Spatial objects (`sp`) will be deprecated in {rnaturalearth} and will be removed in a fu
```

```
library(readr)
```

Veri Temizliği ve Dönüştürme Adımları

- Bu aşamada yapmamız gerekenler şunlar olacak:

1.Eksik Veriler: Eksik verilerin olup olmadığını kontrol edip, varsa bunlarla nasıl başa çıkacağımıza karar vereceğiz.

2.Veri Türleri: Her sütunun veri türünü kontrol edip, gerekiyorsa dönüştürme yapacağız.

3.Kategorik Veriler: Kategorik verileri düzgün bir şekilde işleyeceğiz.

4.Aykırı Değerler: Eğer varsa, aykırı değerleri tespit edip bunlarla başa çıkmak için strateji belirleyeceğiz.

```
# Veri setine genel bakış
summary(netflix)
```

```
##    show_id              type              title             director
##  Length:8790        Length:8790        Length:8790        Length:8790
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    country           date_added         release_year     rating
##  Length:8790        Length:8790        Min.   :1925   Length:8790
##  Class :character   Class :character   1st Qu.:2013   Class :character
##  Mode  :character   Mode  :character   Median :2017   Mode  :character
##                                        Mean   :2014
##                                        3rd Qu.:2019
##                                        Max.   :2021
##    duration          listed_in
##  Length:8790        Length:8790
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
# Eksik değerleri kontrol etme
sapply(netflix, function(x) sum(is.na(x)))
```

```
##      show_id         type        title     director      country   date_added
##            0            0            0            0            0            0
## release_year       rating     duration    listed_in
##            0            0            0            0
```

```
# Veri tiplerini kontrol etme
str(netflix)
```

```
## 'data.frame':    8790 obs. of  10 variables:
##  $ show_id     : chr  "s1" "s3" "s6" "s14" ...
##  $ type        : chr  "Movie" "TV Show" "TV Show" "Movie" ...
##  $ title       : chr  "Dick Johnson Is Dead" "Ganglands" "Midnight Mass" "Confessions of an Invisibl
##  $ director    : chr  "Kirsten Johnson" "Julien Leclercq" "Mike Flanagan" "Bruno Garotti" ...
##  $ country     : chr  "United States" "France" "United States" "Brazil" ...
##  $ date_added  : chr  "9/25/2021" "9/24/2021" "9/24/2021" "9/22/2021" ...
##  $ release_year: int  2020 2021 2021 2021 1993 2021 2021 2019 2021 2013 ...
##  $ rating      : chr  "PG-13" "TV-MA" "TV-MA" "TV-PG" ...
##  $ duration    : chr  "90 min" "1 Season" "1 Season" "91 min" ...
##  $ listed_in   : chr  "Documentaries" "Crime TV Shows, International TV Shows, TV Action & Adventure
```

- Veri setinde eksik veri bulunmuyor.
- Veri türleri de genel olarak uygun görünüyor.
- Ancak, "date_added" sütununun tarih olarak işlenmesi gerek, şu anda metin (string) formatında.

```
netflix$date_added <- as.Date(netflix$date_added, format="%m/%d/%Y")
```

Keşifsel Veri Analizi

```r
# Filmler ve TV Şovları Arasında Sayısal Karşılaştırma
table(netflix$type)
```

```
##
##   Movie TV Show
##    6126    2664
```

```r
# Yıllara Göre İçerik Ekleme Trendleri
netflix$year_added <- format(netflix$date_added, "%Y")
table(netflix$year_added)
```

```
##
## 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##    2    2    1   13    3   11   24   82  426 1185 1648 2016 1879 1498
```

```r
# En Popüler Türler ve Kategoriler
netflix$listed_in <- as.factor(netflix$listed_in)
top_genres <- sort(table(netflix$listed_in), decreasing = TRUE)
head(top_genres, 10)
```

```
##
##                     Dramas, International Movies
##                                              362
##                                    Documentaries
##                                              359
##                                  Stand-Up Comedy
##                                              334
##           Comedies, Dramas, International Movies
##                                              274
## Dramas, Independent Movies, International Movies
##                                              252
##                                        Kids' TV
##                                              219
##                          Children & Family Movies
##                                              215
##               Children & Family Movies, Comedies
##                                              201
##              Documentaries, International Movies
##                                              186
##    Dramas, International Movies, Romantic Movies
##                                              180
```

```r
# En Produktif Ülkeler ve Yönetmenler
top_countries <- sort(table(netflix$country), decreasing = TRUE)
head(top_countries, 10)
```

```
##
##  United States          India United Kingdom       Pakistan      Not Given
##           3240           1057            638            421            287
##         Canada          Japan    South Korea         France          Spain
##            271            259            214            213            182
```

```r
top_directors <- sort(table(netflix$director), decreasing = TRUE)
head(top_directors, 10)
```

```
##
##              Not Given          Rajiv Chilaka    Alastair Fothergill
##                   2588                     20                     18
## Raúl Campos, Jan Suter           Marcus Raboy            Suhas Kadav
##                     18                     16                     16
##              Jay Karas    Cathy Garcia-Molina            Jay Chapman
##                     14                     13                     12
##        Martin Scorsese
##                     12
```

```
# Derecelendirmelere Göre Dağılım
table(netflix$rating)
```

```
##
##        G    NC-17       NR       PG    PG-13        R    TV-14     TV-G
##       41        3       79      287      490      799     2157      220
##    TV-MA    TV-PG     TV-Y    TV-Y7 TV-Y7-FV       UR
##     3205      861      306      333        6        3
```

Filmler ve TV Şovları Arasında Sayısal Karşılaştırma

- Filmler: 6126
- TV Şovları: 2664
- Netflix'teki içeriklerin büyük bir çoğunluğu filmlerden oluşuyor. Filmlerin sayısı, TV şovlarının yaklaşık iki katından fazla.

Yıllara Göre İçerik Ekleme Trendleri

- 2008-2021 yılları arasında, içerik ekleme sayıları her yıl artmış.
- 2019 yılı, en fazla içeriğin eklendiği yıl (2016 içerik).
- 2020 ve 2021'de de yüksek sayıda içerik eklenmiş, bu da Netflix'in içerik kütüphanesini sürekli genişlettiğini gösteriyor.

En Popüler Türler ve Kategoriler

- En popüler türler: "Dramas, International Movies" ve "Documentaries", her biri 362 ve 359 adetle.
- "Stand-Up Comedy" ve çeşitli drama kombinasyonları da oldukça popüler.
- Bu, Netflix'in uluslararası ve belgesel içeriklere önem verdiğini ve komedi türünün de popüler olduğunu gösteriyor.

En Produktif Ülkeler ve Yönetmenler

- En fazla içerik üreten ülkeler: ABD (3240 içerik) Hindistan (1057 içerik) Birleşik Krallık (638 içerik)
- En aktif yönetmenler: Rajiv Chilaka, Alastair Fothergill, Raúl Campos, Jan Suter vb. ABD'nin içerik üretimindeki liderliği ve Hindistan'ın da önemli bir katkısı olduğunu gösteriyor.

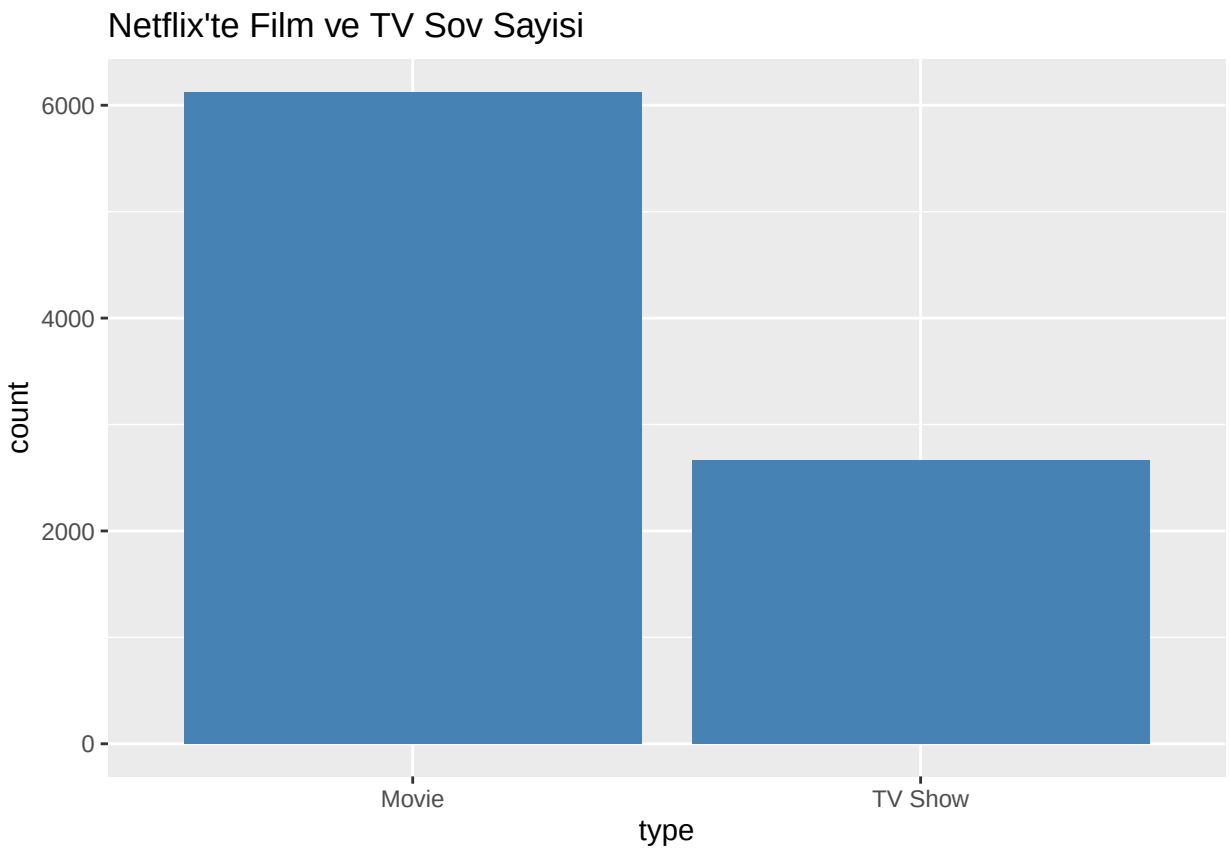
Derecelendirmelere Göre Dağılım

- En yaygın derecelendirmeler: "TV-MA" (3205 içerik) ve "TV-14" (2157 içerik).
- "R" ve "PG-13" derecelendirmesi de oldukça sık kullanılmış.

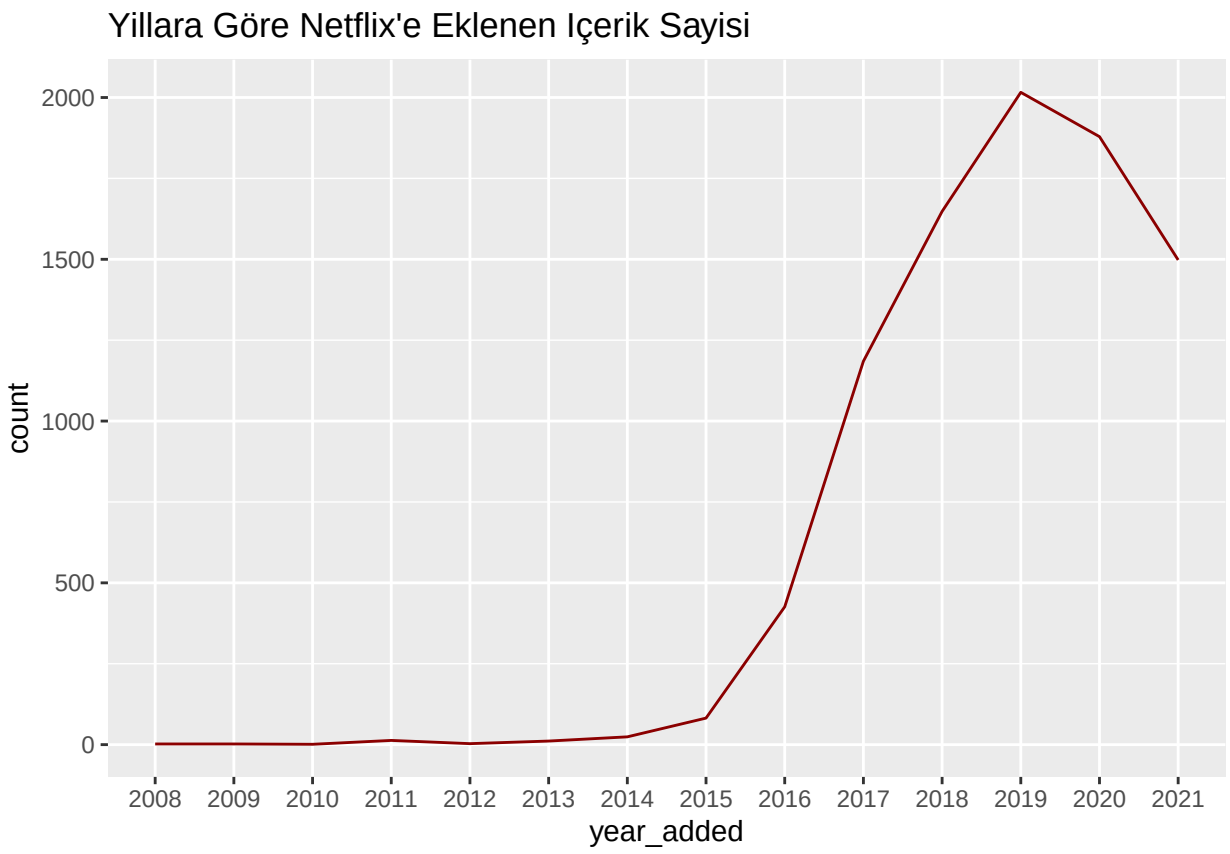
- Bu dağılım, Netflix'in genellikle yetişkinlere ve genç yetişkinlere yönelik içeriklere odaklandığını gösteriyor.

Veri Görselleştirme

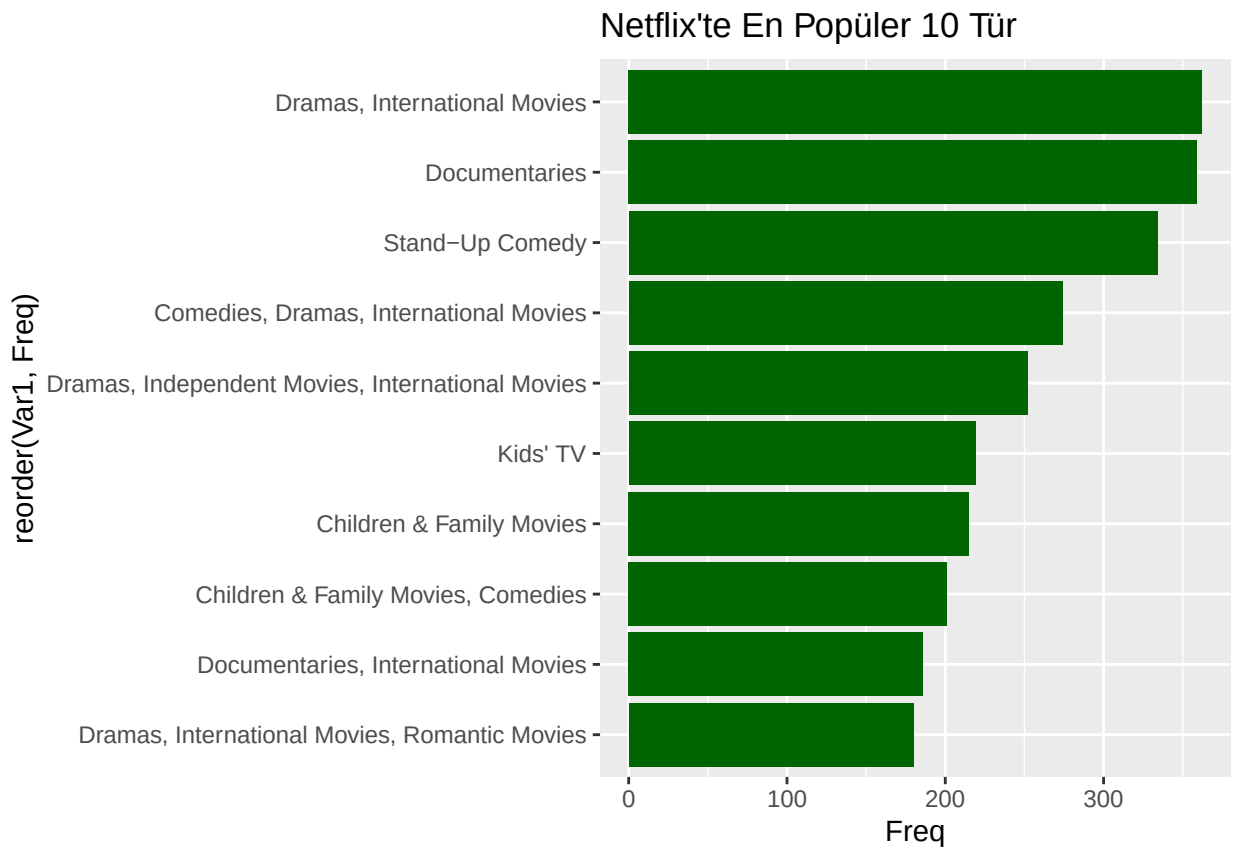
```
# Filmler ve TV Şovları Arasındaki Oranın Görselleştirilmesi
ggplot(netflix, aes(x = type)) +
  geom_bar(fill = "steelblue") +
  ggtitle("Netflix'te Film ve TV Şov Sayısı")
```



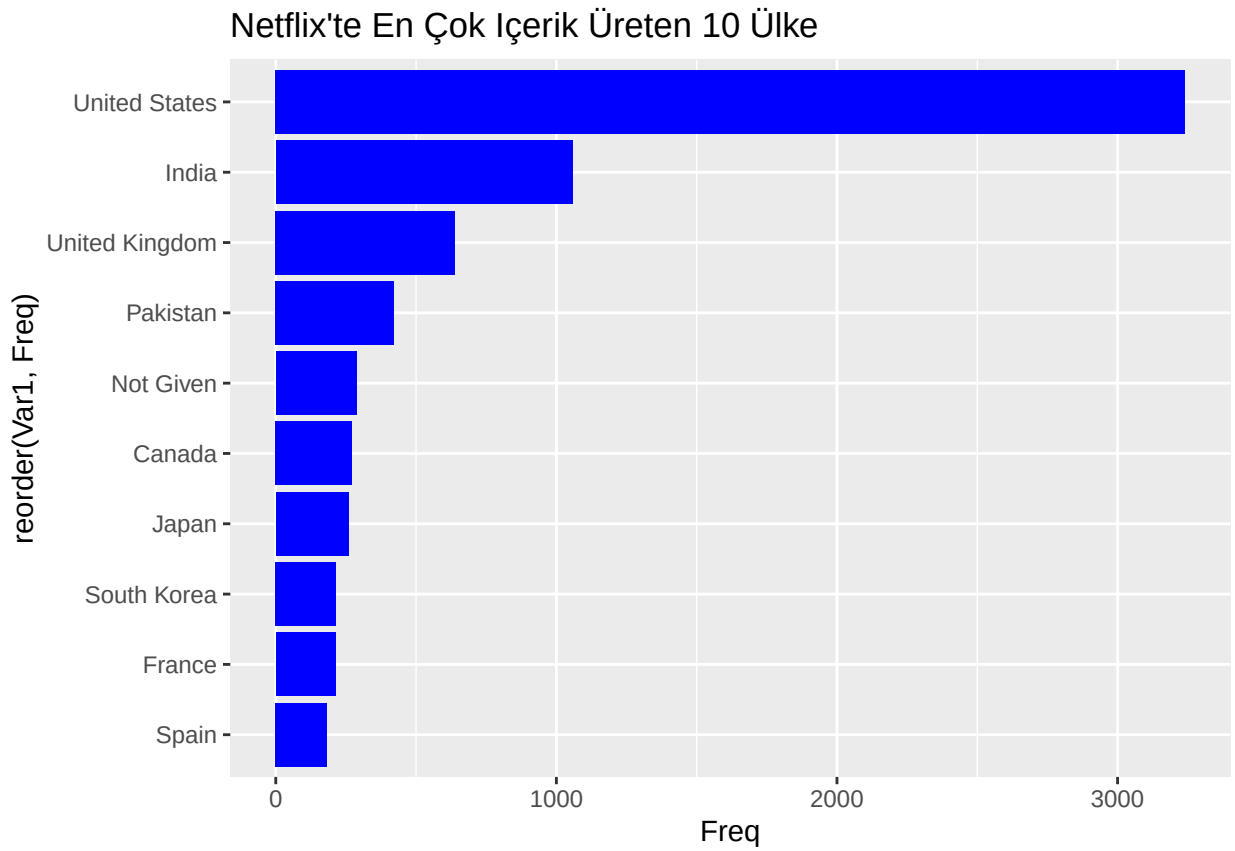

```
# Yıllara Göre İçerik Ekleme Trendlerinin Görselleştirilmesi
ggplot(netflix, aes(x = year_added)) +
  geom_line(stat = "count", group = 1, color = "darkred") +
  ggtitle("Yıllara Göre Netflix'e Eklenen İçerik Sayısı")
```



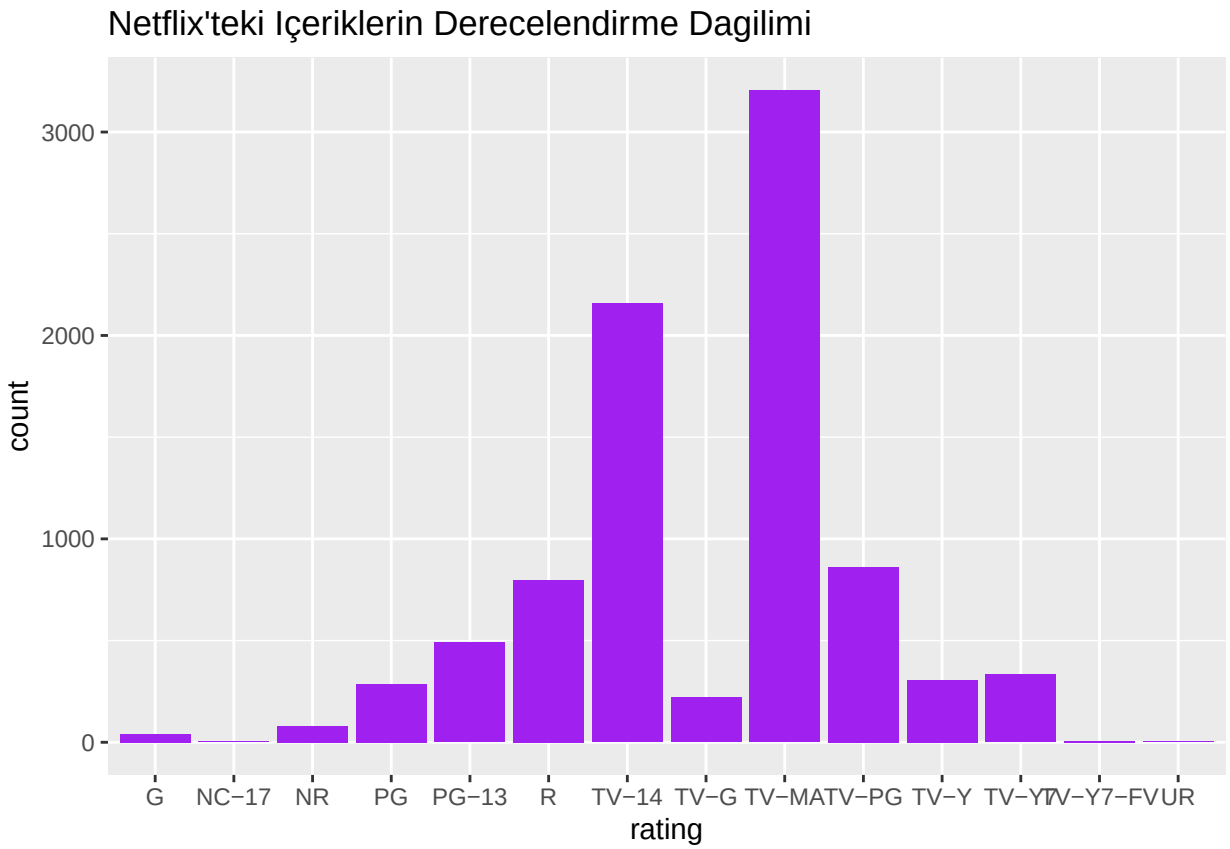

```
# En Popüler Türlerin ve Kategorilerin Görselleştirilmesi
top_genres_df <- as.data.frame(head(top_genres, 10))
ggplot(top_genres_df, aes(x = reorder(Var1, Freq), y = Freq)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  coord_flip() +
  ggtitle("Netflix'te En Popüler 10 Tür")
```




```r
# En Produktif Ülkelerin ve Yönetmenlerin Görselleştirilmesi
top_countries_df <- as.data.frame(head(top_countries, 10))
ggplot(top_countries_df, aes(x = reorder(Var1, Freq), y = Freq)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() +
  ggtitle("Netflix'te En Çok İçerik Üreten 10 Ülke")
```




```
# Derecelendirmelere Göre İçerik Dağılımının Görselleştirilmesi
ggplot(netflix, aes(x = rating)) +
  geom_bar(fill = "purple") +
  ggtitle("Netflix'teki İçeriklerin Derecelendirme Dağılımı")
```




```
# Hangi yılda daha fazla Film ve TV Şovu yayınlandı
netflix_years <- netflix %>%
  filter(release_year >= 2010) %>%
  count(release_year, type) %>%
  arrange(release_year, type)

head(netflix_years)
```

```
##   release_year    type   n
## 1         2010   Movie 153
## 2         2010 TV Show  39
## 3         2011   Movie 145
## 4         2011 TV Show  40
## 5         2012   Movie 173
## 6         2012 TV Show  63
```

```
ggplot(data = netflix_years, aes(x = release_year, y = n, fill = type)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Hangi yılda daha fazla Film ve TV Şovu yayınlandı.",
       x = "Yayın Yılı",
       y = "Sayı") +
  theme(panel.background = element_blank(),
        plot.title = element_text(size = 20, family = "sans"),
        axis.title.x = element_text(size = 14, family = "sans"),
        axis.title.y = element_text(size = 14, family = "sans"),
        axis.text.x = element_text(size = 12, family = "sans", angle = 90, hjust = 1),
        axis.text.y = element_text(size = 12, family = "sans")) +
```

```
  scale_fill_manual(breaks = c("Movie", "TV Show"),
                    values = c("navy blue", "light blue"))
```

- En çok film yayınının olduğu yıllar 2017 ve 2018 iken, 2020 en fazla televizyon şovu başlatan yıldır.

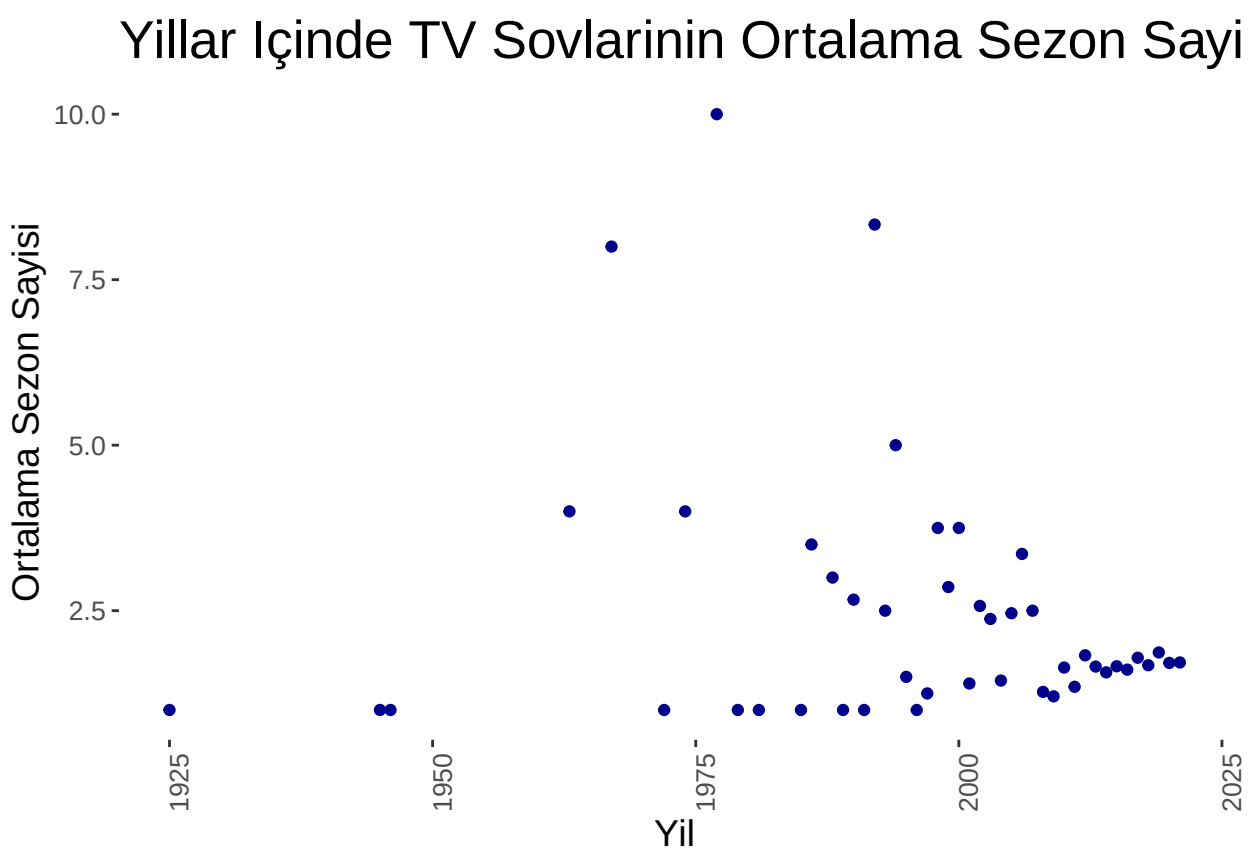
```
# Filmlerin ve TV şovlarının Sürelerinin Yıllar İçinde Artıp Artmadığı

# 'duration' sütunundaki sezon sayılarını sayısal bir değere dönüştürme
netflix <- netflix %>%
  mutate(duration_num = str_extract(duration, "\\d+")) %>%
  mutate(duration_num = as.numeric(duration_num))

# TV şovlarını filtreleme ve yıllara göre ortalama sezon sayısını hesaplama
duration_tv_shows <- netflix %>%
  filter(type == 'TV Show') %>%
  group_by(release_year) %>%
  summarise(average_seasons = mean(duration_num, na.rm = TRUE))

# Grafik çizimi
ggplot(duration_tv_shows, aes(x = release_year, y = average_seasons)) +
  geom_point(color = 'dark blue') +
  labs(title = 'Yıllar İçinde TV Şovlarının Ortalama Sezon Sayısı',
       x = 'Yıl',
       y = 'Ortalama Sezon Sayısı') +
  theme(panel.background = element_blank(),
        plot.title = element_text(size = 20),
```

```
        axis.text.x = element_text(size = 10, angle = 90),
        axis.text.y = element_text(size = 10),
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14))
```



- Grafik, TV şovlarının son yıllarda bir sezon olarak yoğunlaştığını göstermektedir.

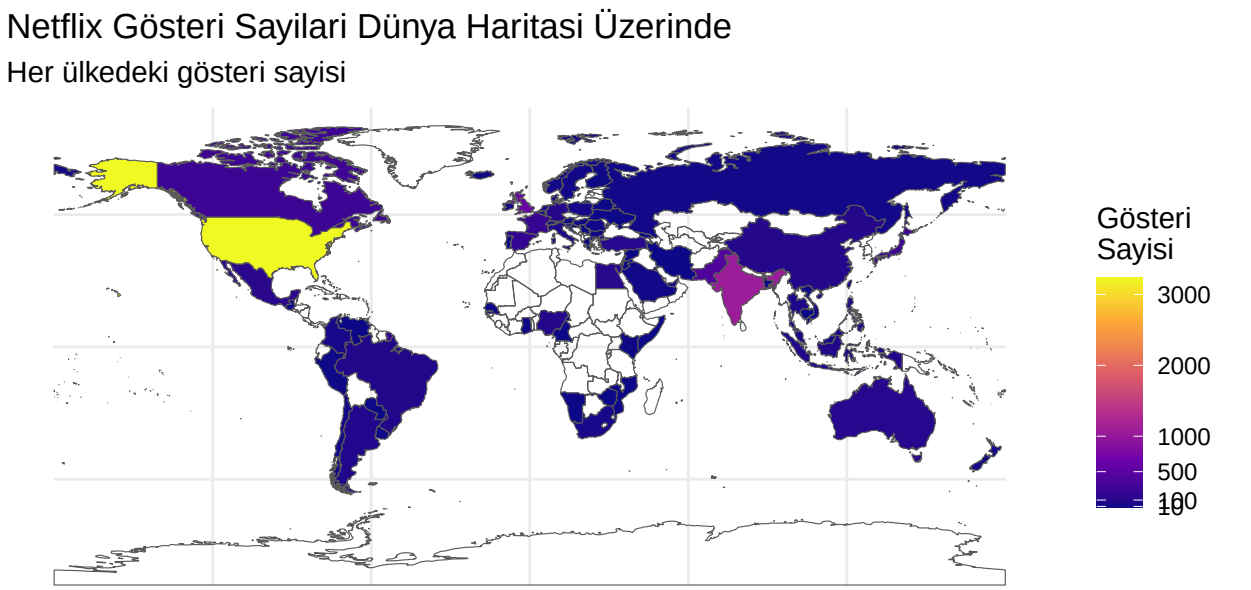
```
# Ülke bazında gösteri sayısını hesaplama
country_counts <- netflix %>%
  count(country) %>%
  rename(show_count = n)

# Dünya ülkelerinin coğrafi verilerini yükleme
world <- ne_countries(scale = "medium", returnclass = "sf")

# Ülke isimlerini eşleştirme ve veri birleştirme
world_data <- left_join(world, country_counts, by = c("name" = "country"))

# Haritayı ve veriyi birleştirme ve görselleştirme
ggplot(data = world_data) +
  geom_sf(aes(fill = show_count)) +
  scale_fill_viridis_c(option = "plasma", na.value = NA, guide = "colorbar",
                       breaks = c(10, 100, 500, 1000, 2000, 3000),
                       labels = c("10", "100", "500", "1000", "2000", "3000")) +
  labs(title = "Netflix Gösteri Sayıları Dünya Haritası Üzerinde",
       subtitle = "Her ülkedeki gösteri sayısı",
```

```
      fill = "Gösteri\nSayısı") +
 theme_minimal()
```



Sonuç ve Değerlendirme

Bu analizde, Netflix veri seti üzerinden veri temizleme, analiz ve görselleştirme süreçlerini uyguladım. Elde edilen sonuçlar, Netflix'in içerik stratejilerini, popüler türleri ve izleyici kitlesinin tercihlerini yansıtmaktadır. Özellikle filmler ve TV şovları arasındaki sayısal farklılıklar, içerik ekleme trendleri ve en popüler türlerin analizi, platformun içerik çeşitliliğine ve kalitesine dair önemli bilgiler sunmaktadır. Görselleştirme teknikleri, bu verileri etkili bir şekilde sunmamızı sağladı, böylece karmaşık veri kümelerini daha anlaşılır hale getirdik. Bu çalışma, veri biliminin ve görselleştirmenin, kompleks veri kümelerinden anlamlı bilgiler çıkarmak ve bu bilgileri etkili bir şekilde sunmak için ne kadar önemli olduğunu göstermektedir.