3 Farklı Veri Setinin Oranlarının Görselleştirilmesi

Sahranur İnce

22 Kasım 2022

Özet

Bu raporda; Kaggle'dan alınan 3 farklı veri setinin oranlarının görselleştirme çalışmaları bulunmaktadır. Birinci veri seti Breaking Bad dizisi, ikinci veri seti 2009-2019 yılları arasında Amazon'da en çok satan 50 kitap ve üçüncü veri seti Harry Potter karakterleri hakkındadır. Öncelikle her bir veri seti incelenip, yorumlanmıştır. Daha sonra her bir veri seti belirtilen durumlara göre veri setine uygun grafiklerle görselleştirilip, yorumlanmıştır.

Gerekli Paketlerin Yüklenmesi

```
install.packages("ggplot2")
install.packages("hrbrthemes")
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggmosaic")
install.packages("treemap")
install.packages("MetBrewer")
install.packages("tidyverse")
install.packages("treemapify")
library(ggplot2)
library(hrbrthemes)
library(dplyr)
library(tidyr)
library(ggmosaic)
library(treemap)
library(MetBrewer)
```

```
library(tidyverse)
library(treemapify)
```

Uygulama 1: Breaking Bad Dizisi

Veri Setinin İncelenmesi

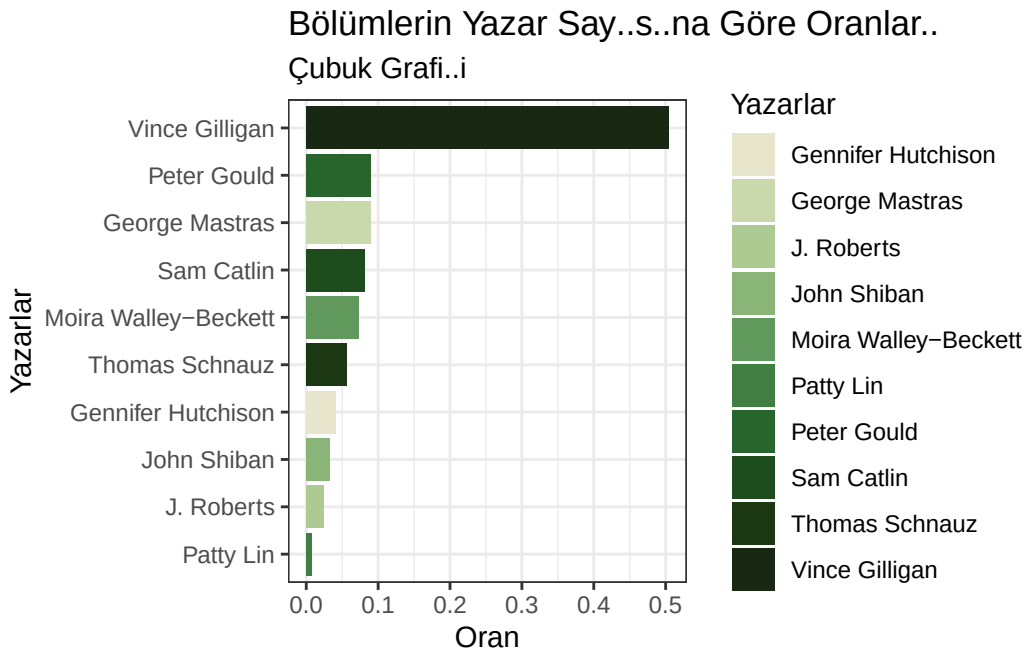
Bu veri seti Breaking Bad adlı dizi hakkında veriler içermektedir. Bu veri setinde 62 gözlem ve 10 değişken vardır. Bu değişkenler şunlardır: Gün, Sezon, Bölüm, Bölüm Adı,Yönetmen, Yazar, Bölüm Süreleri, Özet, Reyting Puanları, İzlenme Sayısı.

Bölümlerin yazar sayısına göre oranlarını veri görselleştirme yöntemleriyle araştırınız.

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

```
x <- breaking_bad %>%
  tidyr::separate_rows(`Written by`, sep = ", ") %>%
  group_by(`Written by`) %>%
  summarise(oran = n()/123)
```

```
ggplot(x, aes(fill = `Written by`,
              y= oran,
              reorder(x=`Written by`, +oran ))) +
  geom_bar(position = "dodge",
           stat = "identity") +
  labs(x = "Yazarlar",
       y = "Oran",
       fill = "Yazarlar",
       title = "Bölümlerin Yazar Sayısına Göre Oranları",
       subtitle = "Çubuk Grafiği") +
  theme_bw() +
  scale_fill_manual(values = met.brewer("VanGogh3",10)) +
   coord_flip()
```



Grafikte elde edilen sonuçlara göre, dizi için en çok senaryo yazan 0.5 oranında Vince Gilligan'dır. En az senaryo yazan ise Patty Lin'dir. Peter Gould ve George Mastras'ın neredeyse aynı oranda senaryo yazdıklarını görüyoruz. Grafiğe genel olarak baktığımızda ise, Vince Gilligan dışındaki yazarların oranları arasında çok büyük bir fark yoktur ancak Vince Gilligan diğer tüm yazarlardan oldukça fazla bir orana sahiptir.

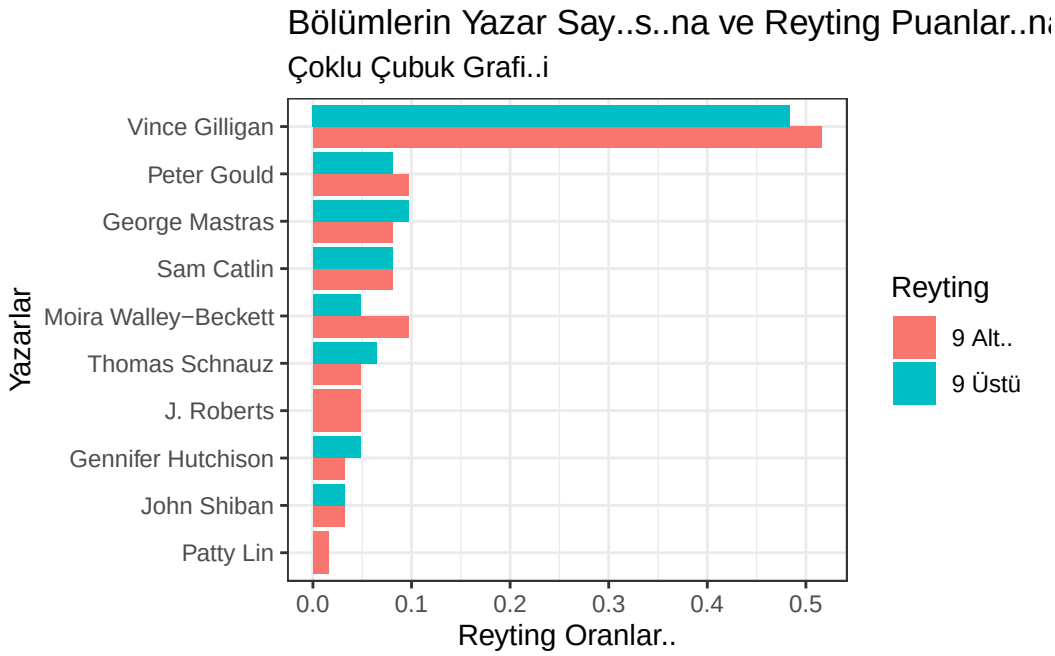
Bölümlerin yazar sayısına ve reyting puanlarına (iki gruba ayırınız: 9 puan altı ve üstü) göre oranlarını veri görselleştirme yöntemleriyle araştırınız.

```
breaking_bad <- breaking_bad %>%
  mutate(reyting = ifelse(Rating_IMDB >= 9, "9 Üstü","9 Altı"))
```

```
reyting <- breaking_bad %>%
  tidyr::separate_rows(`Written by`, sep = ", ") %>%
  group_by(reyting,`Written by`) %>%
  summarize(oran1 = n()/62, oran2 = n()/123)
```

```
ggplot(reyting, aes(fill = reyting,
                        y = oran1,
                        reorder(x = `Written by`, +oran1))) +
```

```
geom_bar(position = "dodge",
         stat = "identity") +
labs(x = "Yazarlar",
     y = "Reyting Oranları",
     fill = "Reyting",
     title = "Bölümlerin Yazar Sayısına ve Reyting Puanlarına Göre Oranları",
     subtitle = "Çoklu Çubuk Grafiği") +
theme_bw() +
coord_flip()
```



Grafikte reyting puanları 9 altı ve üstü olarak iki gruba ayrılmıştır. Elde edilen sonuçlara göre, her iki grupta da en çok reyting oranını alan yazar Vince Gilligan'dır. 9 altı grupta en az reyting oranını alan yazar Patty Lin iken, 9 üstü grupta en az reyting oranını alan yazar John Shiban'dır. J. Roberts ve Patty Lin adlı yazarların 9 üstü reyting oranı almadığını görüyoruz. Sam Catlin ve John Shiban'ın her iki grupta da aynı reyting oranına sahip olduğunu görüyoruz. Reyting oranları arasında en çok değişkenlik gösteren yazarın Moira Walley Beckett olduğunu söyleyebiliriz. Grafiğe genel olarak baktığımızda, Vince Gilligan dışındaki yazarların oranları arasında çok büyük bir fark yoktur ancak Vince Gilligan diğer tüm yazarlardan oldukça fazla bir orana sahiptir.

Uygulama 2: En Çok Satan Kitaplar

Veri Setinin İncelenmesi

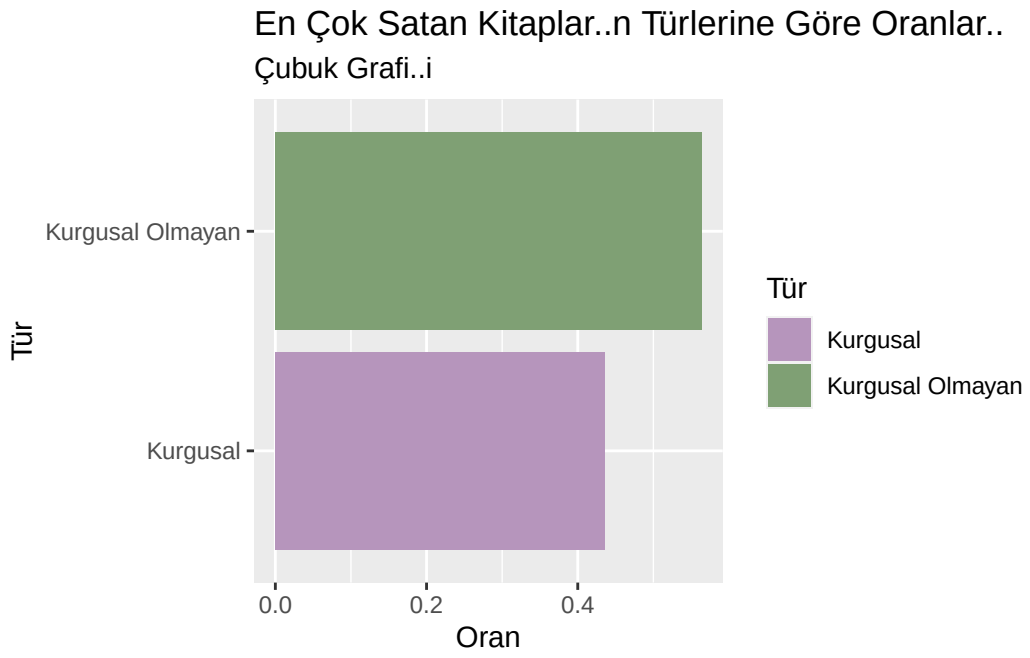
Bu veri seti 2009-2019 yılları arasında Amazon'da en çok satan 50 kitabı içermektedir. Bu veri setinde 550 gözlem ve 7 değişken vardır. Bu değişkenler şunlardır: Kitap Adı, Yazar, Okuyucu Puanı, Yorumlar, Fiyat, Yıl, Tür.

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

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

```
kitaptürü <- bestsellers_with_categories %>%
  group_by(Genre) %>%
  summarize(oran = n()/550)
```

```
ggplot(kitaptürü, aes(fill = Genre,
                      x= oran,
                      y= Genre)) +
  geom_bar(position = "dodge",
           stat = "identity") +
  labs(x = "Oran",
       y= "Tür",
       fill = "Tür",
       title = "En Çok Satan Kitapların Türlerine Göre Oranları",
       subtitle = "Çubuk Grafiği") +
  scale_fill_manual(values = met.brewer("Cassatt2",2),
  labels = c("Kurgusal","Kurgusal Olmayan")) +
  scale_y_discrete(labels = c("Kurgusal", "Kurgusal Olmayan"))
```



Grafikte kitap türleri kurgusal ve kurgusal olmayan olarak iki gruba ayrılmıştır. Elde edilen sonuçlara göre, kurgusal olmayan kitapların kurgusal türündeki kitaplardan daha çok sattığını görüyoruz. Kurgusal olmayan kitapların satılma oranının yaklaşık 0.6, kurgusal kitapların satılma oranının ise 0.4'e yakın bir değer olduğunu söyleyebiliriz.

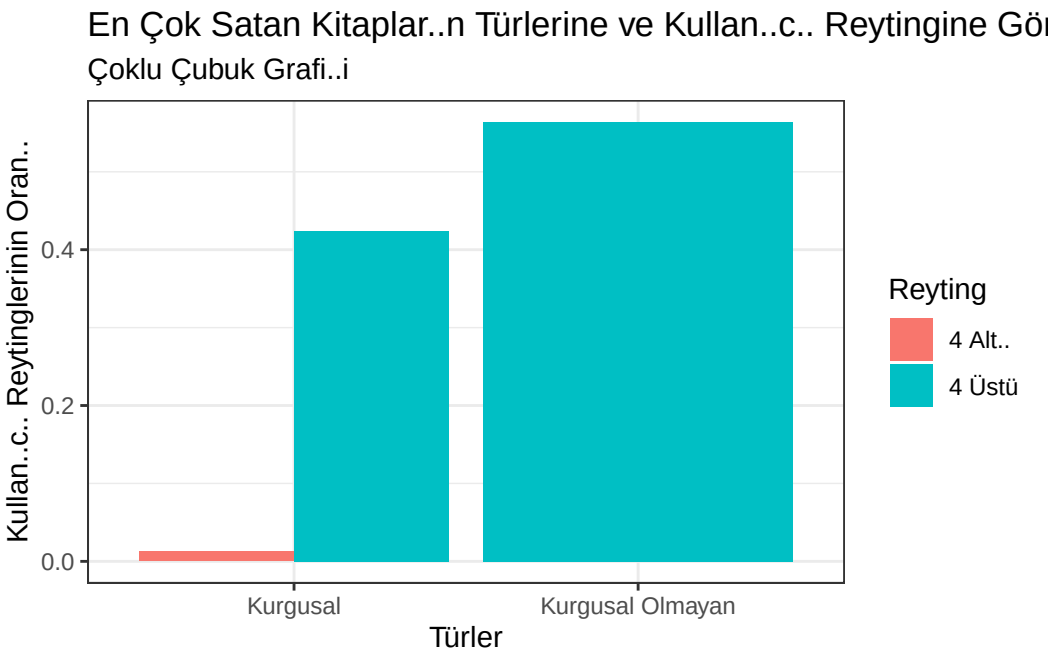
En çok satan kitapların türlerine ve kullanıcı reytingine (iki gruba ayırınız: 4 puan altı ve üstü) göre oranlarını veri görselleştirme yöntemleriyle araştırınız.

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

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

```
ggplot(kul_reyting, aes(fill = kul_reyting,
                        y = oran2,
                        x = Genre)) +
  geom_bar(position = "dodge",
           stat = "identity") +
```

```
labs(x = "Türler",
     y = "Kullanıcı Reytinglerinin Oranı",
     fill = "Reyting",
     title = "En Çok Satan Kitapların Türlerine ve Kullanıcı Reytingine Göre Oranı",
     subtitle = "Çoklu Çubuk Grafiği") +
scale_x_discrete(labels = c("Kurgusal", "Kurgusal Olmayan")) +
theme_bw()
```



Grafikte kullanıcı reytinglerinin oranı 4 altı ve 4 üstü olmak üzere iki gruba ayrılmıştır. Elde edilen sonuçlara göre, kurgusal olmayan kitaplar kurgusal kitaplara göre daha fazla 4 üstünde kullanıcı reytingi almıştır. Kurgusal olmayan kitaplar 4 altında kullanıcı reytingi almamıştır. Kurgusal türündeki kitapların reyting oranları arasında büyük bir fark olduğunu söyleyebiliriz. Kurgusal olmayan kitapların 4 üstü kullanıcı reytinglerinin yaklaşık 0.6 oranında olduğunu söyleyebiliriz. Kurgusal kitapların ise 4 üstündeki kullanıcı reytinglerinin 0.4'den biraz daha fazla bir değer olduğunu, 4 altındaki kullanıcı reytinglerinin ise 0'dan biraz daha büyük bir değer olduğunu söyleyebiliriz.

Uygulama 3: Harry Potter Karakterleri

Veri Setinin İncelenmesi

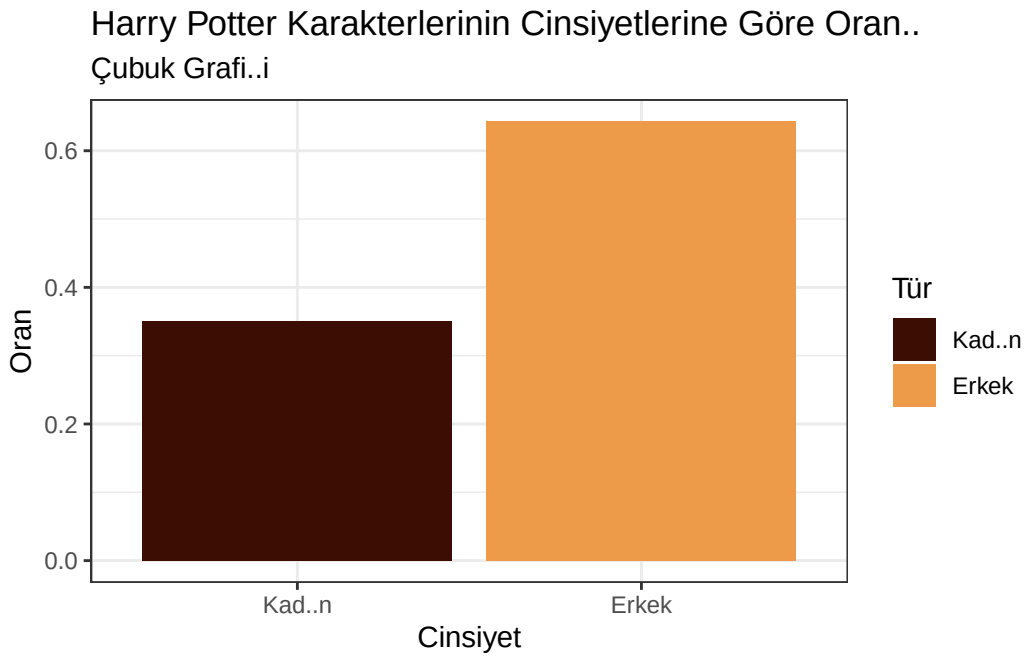
Bu veri seti Harry Potter adlı filmin karakterleri hakkında veriler içermektedir. Bu veri setinde 140 gözlem ve 15 değişken vardır. Bu değişkenler şunlardır: Numara, İsim, Cinsiyet, Meslek, Ev, Asa, Patronus, Türler, Safkan, Saç Rengi, Göz Rengi, Bağlılık, Yetenekler, Doğum Tarihi, Ölüm.

Karakterlerin cinsiyetlerine göre oranını veri görselleştirme yöntemleriyle araştırınız.

```
Characters <- read_delim("Characters.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
Cinsiyet <- Characters %>%
  group_by(Gender) %>%
  drop_na(Gender) %>%
  summarize(oran = n()/140)
```

```
ggplot(Cinsiyet, aes(fill= Gender,
                y= oran,
                x = Gender)) +
  geom_bar(position = "dodge",
           stat = "identity") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  labs(x = "Cinsiyet",
       y = "Oran",
       fill = "Tür",
       title = "Harry Potter Karakterlerinin Cinsiyetlerine Göre Oranı",
       subtitle = "Çubuk Grafiği") +
  scale_fill_manual(values = met.brewer("Greek",2),
  labels = c("Kadın","Erkek")) +
  scale_x_discrete(labels = c("Kadın", "Erkek")) +
  theme_bw()
```



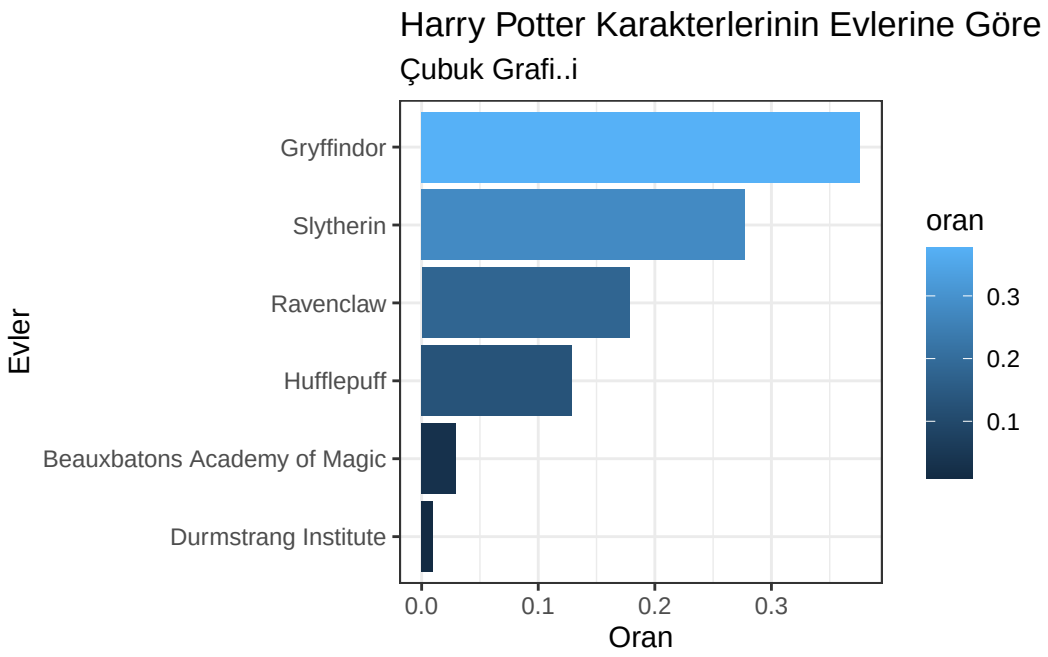
Grafikte elde edilen sonuçlara göre, Harry Potter karakterlerinin daha çok erkek olduğunu görüyoruz. Erkek karakterlerin oranı 0.6'dan biraz daha fazla iken, kadın karakterlerin oranı yaklaşık 0,35'dir. Bu durumda, kadın karakterlerin sayısının neredeyse erkek karakterlerin yarısı kadar olduğunu söyleyebiliriz.

Karakterlerin evlerine göre oranını veri görselleştirme yöntemleriyle araştırınız.

```
EvOranları <- Characters %>%
  group_by(House) %>%
  drop_na(House) %>%
  summarize(oran = n()/101)
```

```
ggplot(EvOranları, aes(fill= oran,
              y= oran,
              reorder(x = House, +oran))) +
  geom_bar(position = "dodge",
           stat = "identity") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  labs(x = "Evler",
       y = "Oran",
       title = "Harry Potter Karakterlerinin Evlerine Göre Oranı",
```

```
      subtitle = "Çubuk Grafiği") +
  theme_bw() +
   coord_flip()
```



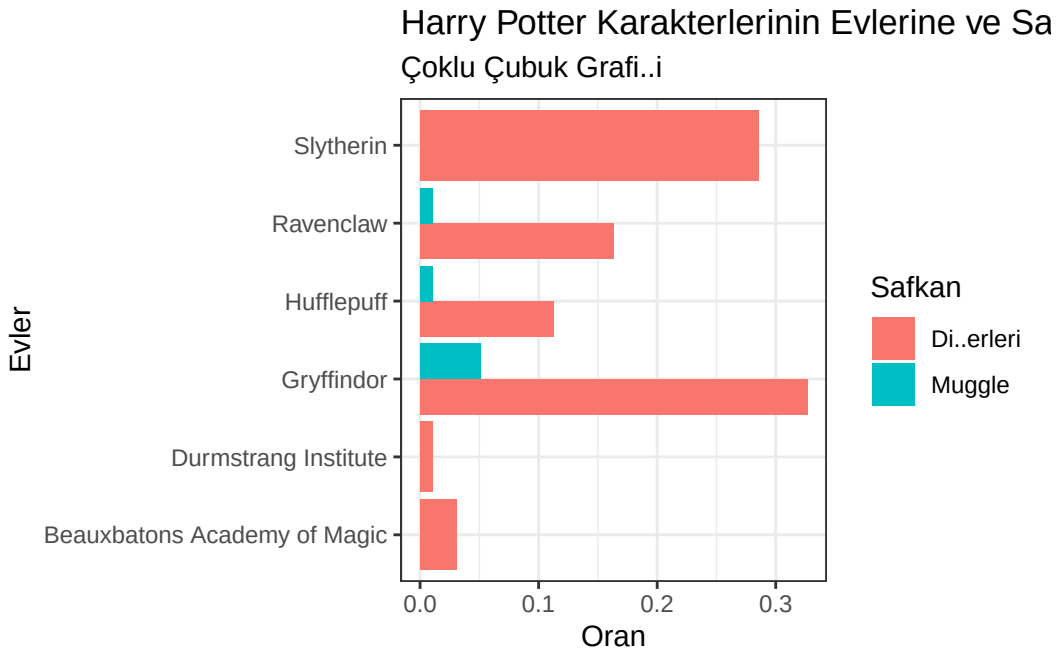
Grafikte elde edilen sonuçlara göre, Harry Potter karakterlerinin en çok bulunduğu ev Gryffindor, en az bulunduğu ev ise Durmstrang Institude'dur. Gryffindor evinde bulunan karakterlerin oranı yaklaşık 0.4 iken Durmstrang Institute evinde bulunan karakterlerin oranı 0'a çok yakın bir değerdir.

Karakterlerin evlerine ve muggle olması (iki gruba ayırınız: muggle ve diğerleri) durumlarına göre oranını veri görselleştirme yöntemleriyle araştırınız.

```
Characters <- Characters %>%
  mutate(muggle = ifelse(`Blood status` == "Muggle-born", "Muggle","Diğerleri"))
```

```
harry <- Characters %>%
  group_by(House, muggle) %>%
  drop_na(House, muggle) %>%
  summarize(oran1 = n()/98, oran2 = n()/98)
```

```r
ggplot(harry, aes(fill = muggle,
                  y = oran2,
                  x = House)) +
  geom_bar(position = "dodge",
           stat = "identity") +
  labs(x = "Evler",
       y = "Oran",
       fill = "Safkan",
       title = "Harry Potter Karakterlerinin Evlerine ve Safkanlıklarına Göre Oranı",
       subtitle = "Çoklu Çubuk Grafiği") +
  theme_bw() +
  coord_flip()
```



Grafikte Harry Potter karakterlerinin safkanlığı muggle ve diğerleri olmak üzere iki gruba ayrılmıştır. Elde edilen sonuçlara göre en çok muggle bulunan ev Gryffindor, en az muggle bulunan ev ise Hufflepuff ve Ravenclaw'dır. Slytherin, Durmstrang Institute ve Beauxbatons Academy of Magic evlerinde hiç muggle bulunmamaktadır. Muggle dışındaki safkanların en çok bulunduğu ev Gryffindor, en az bulunduğu ev ise Durmstrang Institute'dır. Gryffindor ve Slytherin evlerinde bulunan muggle dışındaki safkanların oranları arasında çok büyük bir fark olmadığını söyleyebiliriz. Grafiğe genel olarak baktığımızda ise Harry Potter karakterlerinin sadece küçük bir kısmının muggle olduğunu söyleyebiliriz.