Veri Bilimi Maaşları- En Çok Okunan Kitaplar- Harry Potter Karakterleri Veri Setleri ile Veri Görselleştirilmesi

Beste Ünal

27.11.2022

ÖZET

Veri Bilimi Maaşları- En Çok Okunan Kitaplar- Harry Potter Karakterleri Veri Setleri verilmiştir. Bu veri setleri ile ilgili 2-3 madde halinde belirli kriterler göz önünde bulundurularak uygun veri görselleştirme tekniklerine karar verilip, uygun grafiklendirmeler yapılmıştır.

Gerekli paketler yüklenip kütüphaneler çalıştırılmıştır.

```
install.packages("readr")
library(readr)
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)
install.packages("MetBrewer")
library(MetBrewer)
install.packages("tidyverse")
library(tidyverse)
install.packages("treemapify")
library(treemapify)
```

1 VERİ BİLİMCİ MAAŞLARI

VERİ TANIMI: Veri Bilimi Alanları Maaş Kategorizasyonu Veri Kümesi 9 sütun içermektedir.

Working Year: Maaşın ödendiği yıl. (2020, 2021, 2022) Designation: Yıl boyunca çalışılan rol. Experience: Yıl boyunca işteki deneyim düzeyidir. [EN - Giriş seviyesi / Junior, MI - Orta seviye / Orta, SE - Kıdemli seviye / Uzman, EX - Yönetici seviyesi / Direktör] Employment Status: Rol için istihdam türüdür. [PT - Yarı zamanlı, FT - Tam zamanlı, CT - Sözleşmeli, FL - Serbest] Salary In Rupees: Ödenen toplam brüt maaş tutarıdır. Employee Location: ISO 3166 ülke kodu olarak çalışanın çalışma yılı boyunca birincil ikamet ettiği ülkedir. (ISO 3166 ülke koduna PFB Bağlantısı) Company Location: İşverenin ana ofisi veya sözleşme şubesinin bulunduğu ülkedir. Company Size: Yıl boyunca şirket için çalışan ortalama kişi sayısıdır. [S(küçük) - 50'den az çalışan , M(orta) - 50 ila 250 çalışan , L(büyük) - 250'den fazla çalışan] Remote Working Ratio:Uzaktan yapılan toplam iş miktarıdır.

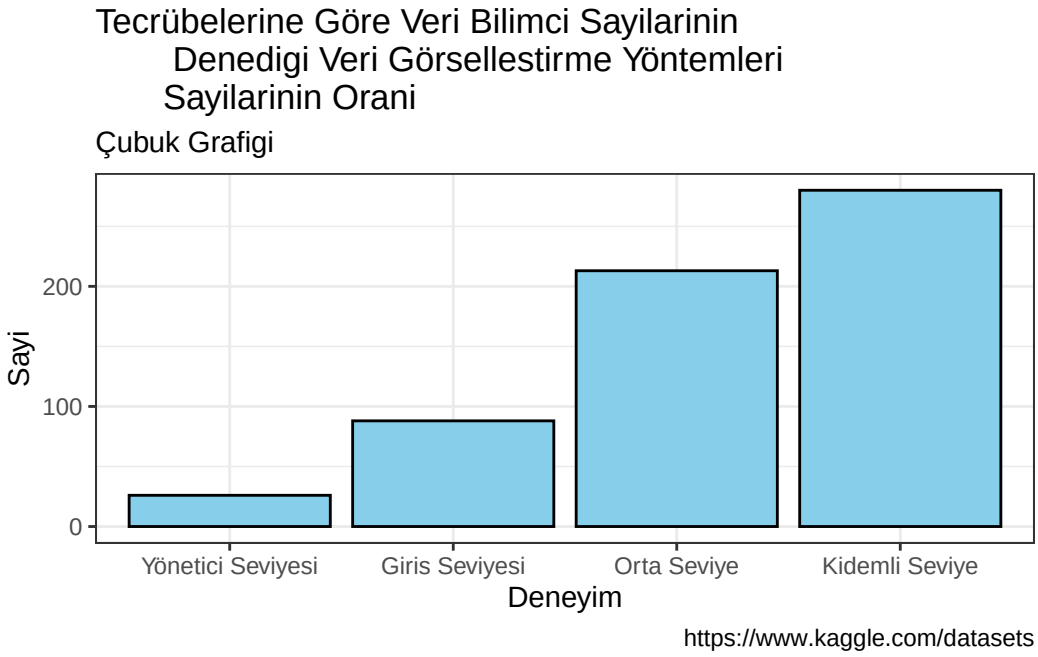
```
library(readr)
Data_Science_Fields_Salary_Categorization <- read_csv("Data_Science_Fields_Salary_Categori
```

1.1 Tcrübelerine Göre Veri Bilimci Sayılarının Denediği Veri Görselleştirme Yöntemleri Grafiği

Data_Science_Fields_Salary_Categorization adlı veri setinden alınan deneyi(Experience) değişkeni grup haline getirilerek sayı ile özetlendi ve "data2"' ye atandı. Tecrübelerine göre veri bilimci sayılarının oranını gösteren Çubuk Grafiği çizilmiştir.

```r
data2 <- Data_Science_Fields_Salary_Categorization %>%
         group_by(Experience) %>%
         summarise(sayı = n())

ggplot(data2, aes(x =reorder(Experience, + sayı), y = sayı )) +
  geom_bar(stat = "identity",fill= "skyblue", color = "black" ) +
  labs(x = "Deneyim",
       y = "Sayi",
        title = "Tecrübelerine Göre Veri Bilimci Sayilarinin
        Denedigi Veri Görsellestirme Yöntemleri
       Sayilarinin Orani",
       subtitle = "Çubuk Grafigi" ,
  caption = "https://www.kaggle.com/datasets") +
  scale_x_discrete(labels = c("Yönetici Seviyesi", "Giris Seviyesi",
                              "Orta Seviye","Kidemli Seviye")) +
  theme_bw()
```



YORUM: Yukarıdaki grafikte görüldüğü üzere tecrübelerine göre veri bilimci oranını gösteren bir grafik mevcuttur. Tecrübe durumu arttıkça veri bilimi oranının arttığını söyleyemeyiz. Fakat en yüksek orana sahip olan tacrübe düzeyi "Kıdemli Seviye" iken en düşük tecrübe düzeyine sahip olan ise "Yönetici Seviye" sidir.

1.2 Veri bilimcilerinin Çalışma Sistemine Göre Oranlarının Grafiği

Data_Science_Fields_Salary_Categorization veri setindede bulunan Employment_Status(Çalışma Sistemi) değişkeni grup haline getirilmiştir ve sayı olarak özetlenip oranlanmıştır.

```
Data_Science_Fields_Salary_Categorization %>%
  group_by(Employment_Status) %>%
  summarise(sayi = n()) %>%
  mutate(oran = sayi/sum(sayi))
```

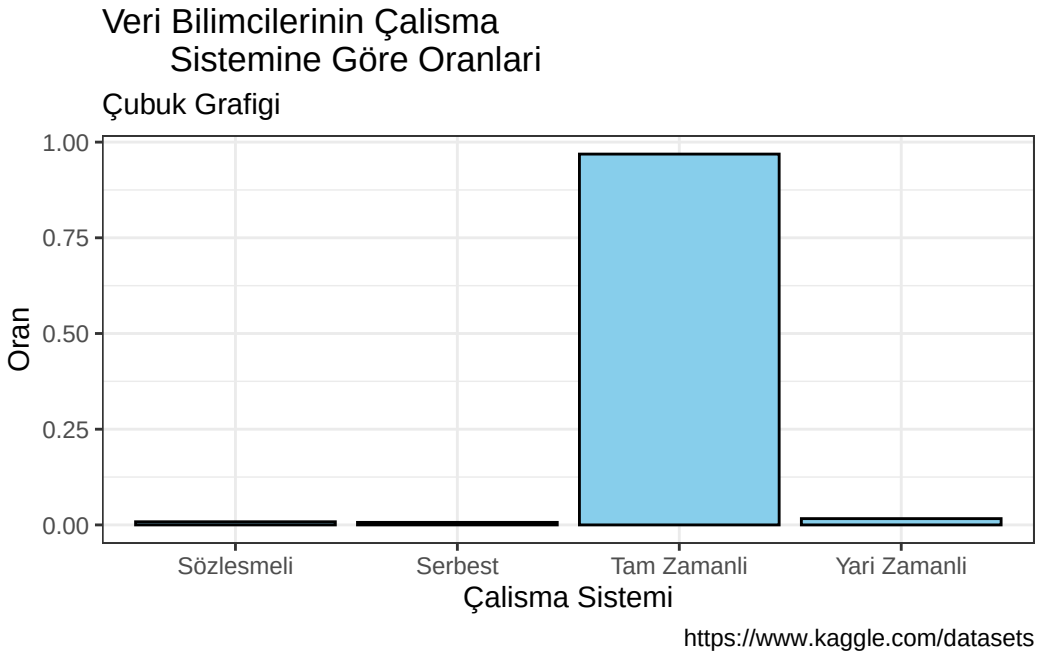
```
# A tibble: 4 x 3
  Employment_Status  sayi    oran
  <chr>             <int>   <dbl>
1 CT                    5 0.00824
2 FL                    4 0.00659
3 FT                  588 0.969
4 PT                   10 0.0165
```

Data_Science_Fields_Salary_Categorization veri setindede bulunan Employment_Status(Çalışma Sistemi) değişkeni grup haline getirilmiştir ve sayı olarak özetlenip oranlanmıştır. Ek olarak ise bunların hepsi data3 veri setine atanmıştır.

Çalışma Sistemine göre veri bilimcilerinin oranları gösterecek olan grafik, data3 veri seti kullanılarak çizdirilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```
data3 <- Data_Science_Fields_Salary_Categorization %>%
  group_by(Employment_Status) %>%
  summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))

ggplot(data3, aes(x = Employment_Status, y = oran)) +
    geom_bar(stat = "identity", fill = "skyblue",color = "black") +
  labs(x = "Çalisma Sistemi", y = "Oran",
       title = "Veri Bilimcilerinin Çalisma
       Sistemine Göre Oranlari",
       subtitle = "Çubuk Grafigi" ,
       caption = "https://www.kaggle.com/datasets") +
  scale_x_discrete(labels = c("Sözlesmeli", "Serbest", "Tam Zamanli", "Yari Zamanli")) +
  theme_bw()
```



YORUM: Yukarıdaki grafikte görüldüğü üzere veri bilimcilerin çalışma sistamine göre oranının en yüksek olduğu sistem Tam Zamanlı Sistemdir. Bu da Veri Bilimcilerin büyük bir çoğunluğunun tam zamanlı olarak çalıştığını analamına gelmektedir. Daha sonra Yarı Zamanlı çalışanlar, Sözleşmeli ve Serbest olarak çalışanlar da mevcuttur fakat bu oranlar oldukça düşüktür.

1.3 Firma Büyüklüğüne Ve Tecrübe Düzeyine Göre Veri Bilimci Sayılarının Oranını Gösteren Grafik

```
options(dplyr.summarise.inform = FALSE)
data3 <- Data_Science_Fields_Salary_Categorization %>%
  group_by(Company_Size, Experience) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / nrow(Data_Science_Fields_Salary_Categorization))
```

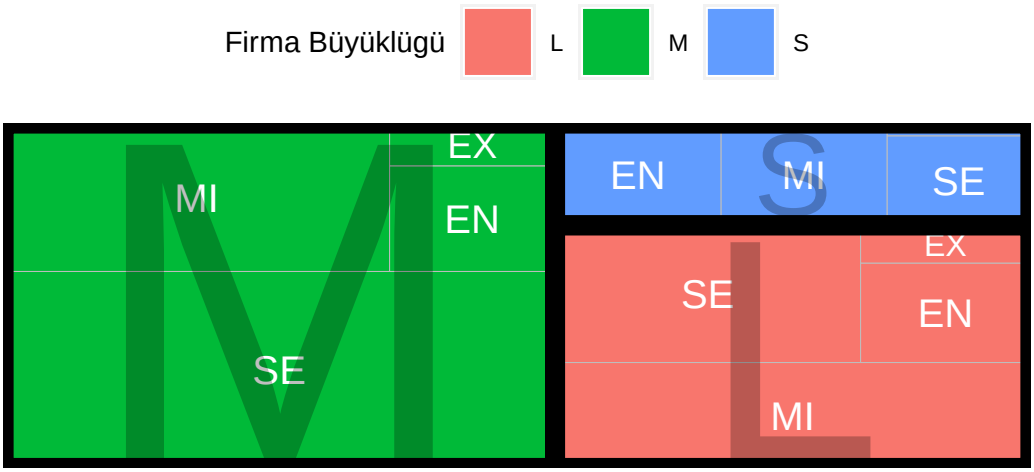
```
ggplot(data3, aes(area = data3$oran, label = data3$Experience,
                  subgroup = data3$Company_Size, fill = Company_Size)) +
  geom_treemap() +
  labs(title = "Firma Büyüklügüne Ve Tecrübe Düzeyine Göre
       Veri Bilimci Sayilarinin Oranlari Grafigi",
       fill = "Firma Büyüklügü",
       subtitle = "Agac Haritalari Grafigi",
       caption = "https://www.kaggle.com/datasets") +

  geom_treemap_text(colour = "white", place = "centre", size = 15) +
  geom_treemap_subgroup_border(colour = "black", size = 10) +
  geom_treemap_subgroup_text(place = "centre", grow = TRUE, alpha = 0.25,
                             colour = "black", fontface = "plain") +


  theme(legend.position = "top")
```

Firma Büyüklügüne Ve Tecrübe Düzeyine Göre
Veri Bilimci Sayilarinin Oranlari Grafigi

Agac Haritalari Grafigi



https://www.kaggle.com/datasets

Ml: Orta seviye SE: Kıdemli seviye EN: Giriş seviyesi EX: Yönetici seviyesi

YORUM: Orta Düzey(M) firmda çalışanlar diğer iki düzeydeki firma çalışanlarına göre daha fazladır.Orta Düzey firma da çalışanların büyük çoğunluğu Kıdemli Seviyesidir(SE). Aynı zaman da Yönetici Seviyesinde (EX) çalışanların sayısı en azdır. Büyük Düzey(L) firmasında en fazla çalışan Orta Seviyedeki (Ml) çalışanlardır. Küçük Düzey(S) firma da Yönetici Seviyesinde (EX) çalışanlar bulunmamaktadır. Küçük Düzey firmada çalışan kişiler diğer firma düzeylerindeki çalışanlara göre en az çalışana sahiptir.

2 EN ÇOK SATAN KİTAPLAR

VERİ TANIMI: Amazonda 2009 - 2019 Yılları Arasında En Çok Satan 50 Kitap

Name: Kitap İsimleri Author: Kitap Yazarı User Rating: Amazon Kullanıcı Değerlendirmeleri Reviews: Amazon da yorum sayısı Price: Kitap fiyatı Year: Yıllar Genre: Kitapların Kurgusal-Kurgusal Olmama durumu

2.1 En Çok Satan Kitapların Türlerine Göre Oranlarını Gösteren Grafik

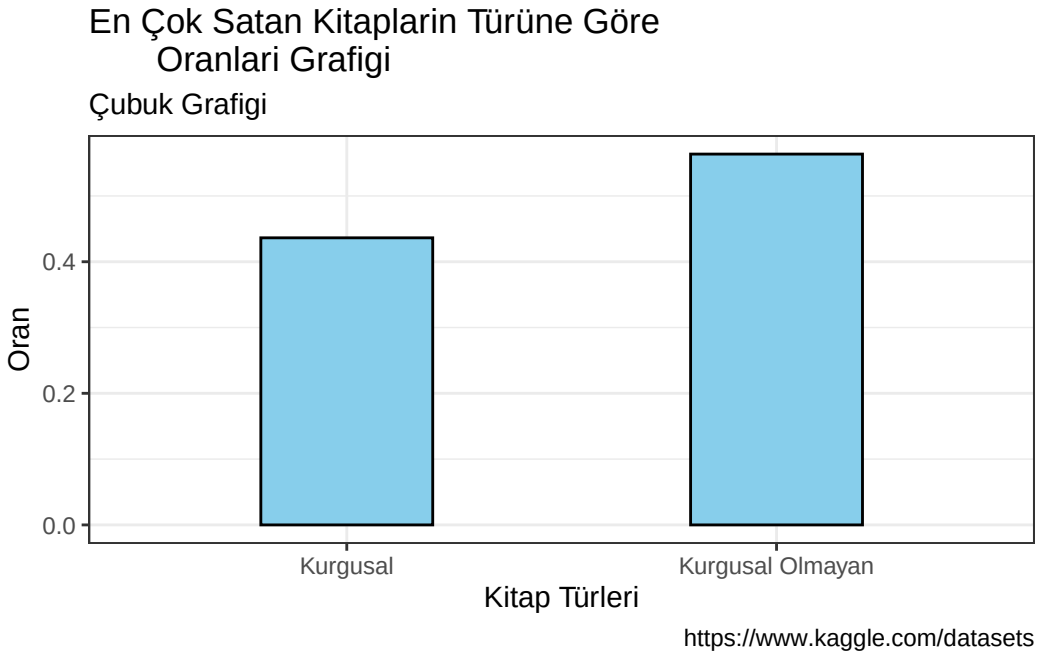
```
library(readr)
bestsellers_with_categories <- read_csv("bestsellers with categories.csv")
```

bestsellers_with_categories veri setindede bulunan Genre(Cinsiyet) değişkeni grup haline getirilmiştir ve sayı olarak özetlenip oranlanmıştır. Ek olarak "books" veri seti olarak yeniden atanmıştır.

En Çok Satan Kitapların Türüne Göre Oranlarlarını gösterecek olan grafik, books veri seti kullanılarak çizdirilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```
books <- bestsellers_with_categories %>%
  group_by(Genre) %>%
  summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))

ggplot(books, aes(x = Genre, y = oran)) +
    geom_bar(stat = "identity", width = 0.4,
             fill = "skyblue", color = "black") +
  labs(x = "Kitap Türleri", y = "Oran",
       title = "En Çok Satan Kitaplarin Türüne Göre
       Oranlari Grafigi",
       subtitle = "Çubuk Grafigi",
       caption = "https://www.kaggle.com/datasets") +
  scale_x_discrete(labels = c("Kurgusal", "Kurgusal Olmayan")) +
  theme_bw()
```



YORUM: Grafiktede görüldüğü üzere En Çok Okunan Kitapların arasıdan Kurgusal Olmayan türüne sahip olanların oranı daha fazladır. Bu da bu tür kitapların daha çok tercih edildiği ablamına gelmektedir. Kurgusal türündeki kitapların oranı 0.4' ün biraz üzerindeyken Kurgusal Olmayan türündeki kitapların oranı ise 0.5' e yaklaşık oluğunu söyleyebiliriz.

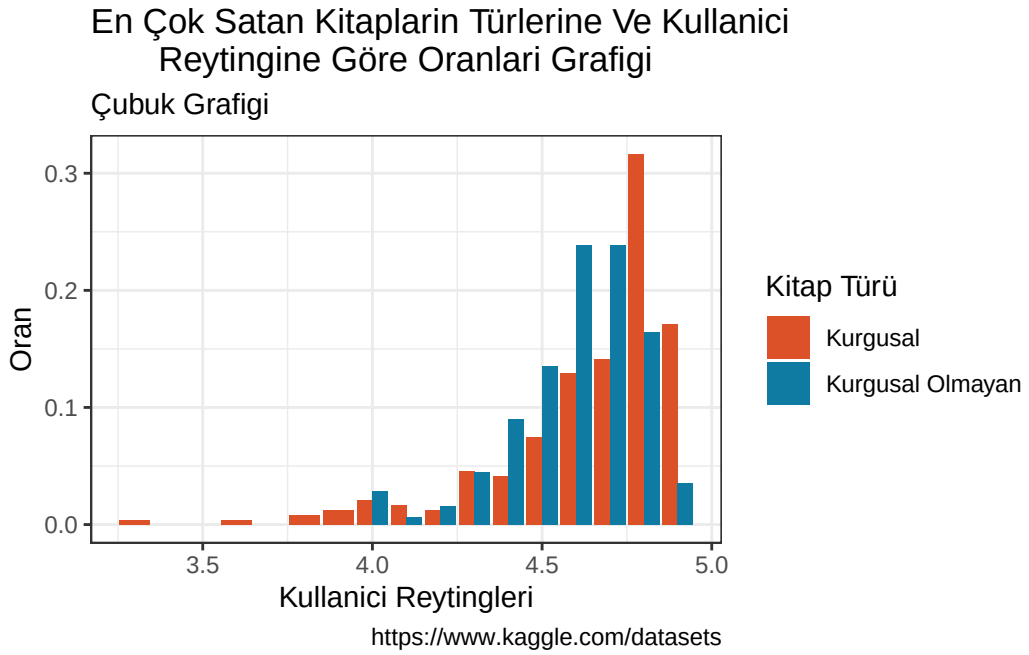
2.2 En Çok Satan Kitapların Türlerine Ve Kullanıcı Reytingine Göre Oranları Grafiği

bestsellers_with_categories veri setindede bulunan Genre(Cinsiyet) ve User Rating(Kullanıcı Reytinleri) değişkenleri grup haline getirilmiştir ve sayı olarak özetlenip oranlanmıştır. Ek olarak "books2" veri seti olarak yeniden atanmıştır.

En Çok Satan Kitapların Türlerine Ve Kullanıcı Reytingine Göre Oranlarını gösterecek olan grafik, books2 veri seti kullanılarak çizdirilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```r
books2 <- bestsellers_with_categories %>%
  group_by(Genre,`User Rating`) %>%
  summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))

ggplot(books2, aes(x = `User Rating`, y = oran, fill = Genre)) +
    geom_bar(stat = "identity", position = "dodge" ) +
  labs(x = "Kullanici Reytingleri", y = "Oran", fill = "Kitap Türü",
       title = "En Çok Satan Kitaplarin Türlerine Ve Kullanici
       Reytingine Göre Oranlari Grafigi",
        subtitle = "Çubuk Grafigi",
       caption = "https://www.kaggle.com/datasets" ) +
  scale_fill_manual(labels = c("Kurgusal","Kurgusal Olmayan"), values = met.brewer("Egypt"
  theme_bw()
```



YORUM: Grafiğe ilk baktığımızda gözümüze çarpan şey 4.5-5.0 kullanıcı reytingleri arasında Kurgusal Kitap türüne verilen oranın en çok olduğu bir nokta vardır. Diğer aralılarda genellikle Kurgusal Olmayan kitap türüne verilen Kullanıcı reytingleri daha fazladır. 3.5-4.0 Kullanıcı Reytingleri arasında Kurgusal Olmayan kitap türüne hiç ya da çok az tercih edilmiştir diyebiliriz.

3 HARRY POTTER KARAKTERLERİ

VERİ TANIMI: Harry Potter Filmindeki Karakterlerin veri Kümesi

Name: Karakter İsimleri Gender: Karakter Cinsiyetleri Job: Karakterlerin İşleri House: Karakterlerin Evleri Wand: Karakterlerin Asaları Patronus:
Species:
Blood Status: Karakterlerin saf kan ya da Muggle olma durumu Hair Colour: Karakterlerin Saç Renkleri

3.1 Karakterlerin Cinsiyete Göre Veri Görselleştirilmesi

Harry Potter Karakterleri adlı kaggle dosyasında bulunun "Characters.csv" veri seti aktif hale getirildi ve gerekli olan kütüphane çalıştırılarak Characters adına atandı.

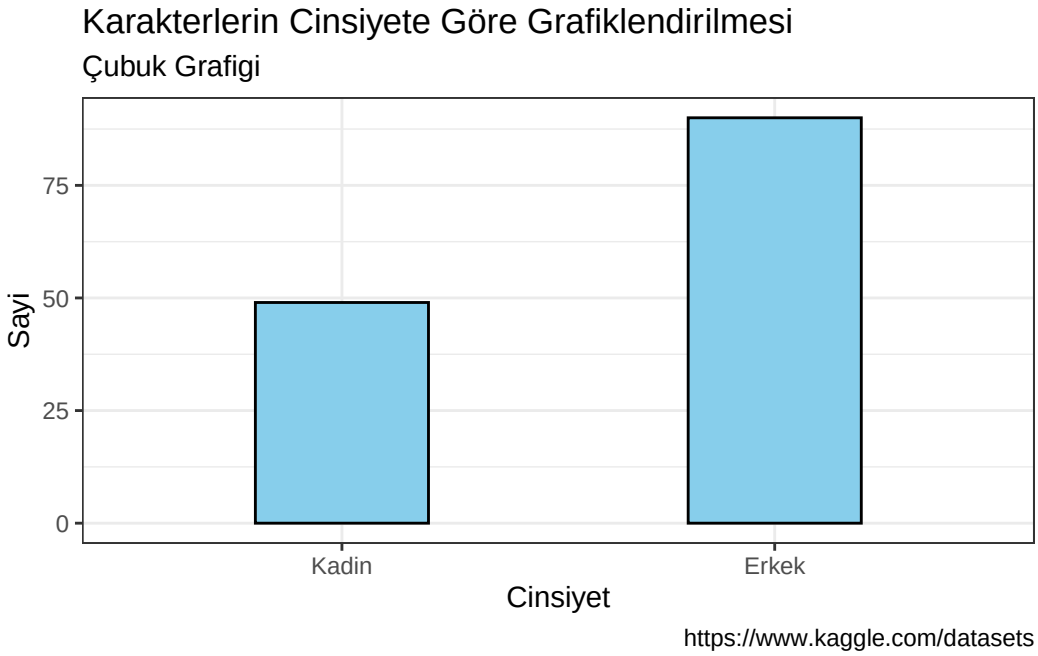
```
library(readr)
Characters <- read_delim("Characters.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

Characters veri setinde bulunan Cinsiyetier(Gender) grup haline getirildi, Cinsiyetlerin(Gender) içerisinde bulunan NA değerleri çıkarıldı, isimler ile özetlendi ve hepsi harry1 olarak atandı.

Karakterlerin Cinsiyete Göre Grafiği, harry1 veri seti kullanılarak çizdirilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```
harry1 <- Characters %>%
  group_by(Gender) %>%
  drop_na(Gender) %>%
  summarise(sayi = n())

ggplot(harry1, aes(x = Gender, y = sayi)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black",
           width = 0.4) +
  labs(x = "Cinsiyet",
       y = "Sayi",
        title = "Karakterlerin Cinsiyete Göre Grafiklendirilmesi",
       subtitle = "Çubuk Grafigi" ,
  caption = "https://www.kaggle.com/datasets") +
  scale_x_discrete(labels = c("Kadin", "Erkek")) +
  theme_bw()
```



YORUM: Harry Potter Karakterlerinin Cinsiyet Sayısını veren bir grafik çizişlmiştir. Burada görüldüğü üzere Harry Potter dizinde yer alan karakterlerin çoğunluğu "Erkek" cinsiyetine aittir. Kadın karakter sayısı 49 veya 50 dir diyebiliriz, aynı şekilde Erkek karakter sayısı 80 üzeridir demek yanlış olmayacaktır.

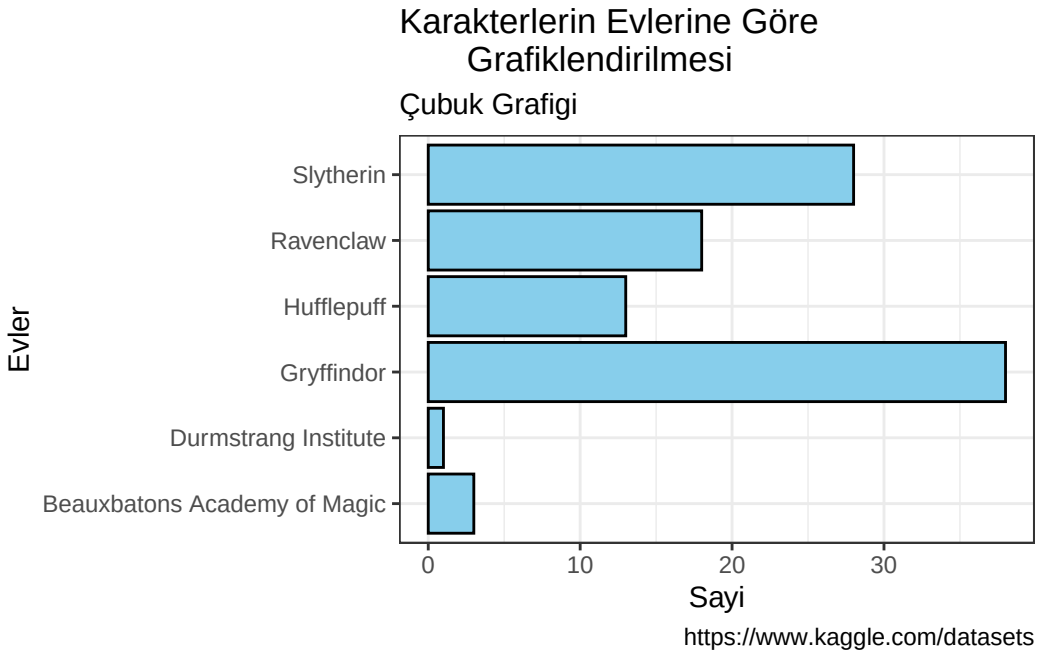
3.2 Karakterlerin Evlerine Göre Veri Görselleştririlmesi

Characters veri setinde bulunan Karakterlerin Evleri(House) grup haline getirildi, Karakterlerin Evleri(House) içerisinde bulunan NA değerleri çıkarıldı, isimler ile özetlendi ve hepsi harry2 olarak atandı.

Karakterlerin Evlerine Göre Grafiği, harry2 veri seti kullanılarak çizdirilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```r
harry2 <- Characters %>%
  group_by(House) %>%
  drop_na(House) %>%
  summarise(sayi = n())

ggplot(harry2, aes(x = sayi, y = House)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(x = "Sayi",
       y = "Evler",
        title = "Karakterlerin Evlerine Göre
       Grafiklendirilmesi",
       subtitle = "Çubuk Grafigi" ,
  caption = "https://www.kaggle.com/datasets") +
   theme_bw()
```



YORUM: Harry Potter karakterlerinin Evlerine göre sayılarını gösteren bir grafik çizilmiştir. En çok üyeye sahip olan ev "Gryffindor" iken en az üyeye sahip olan ise "Durmstrang Institule" olmuştur.

3.3 Karakterlerin Evlerine Ve Muggle Olması (iki gruba ayırınız: muggle ve diğerleri) Durumlarına Göre Oranını Veri Görselleştirmesi

Characters veri setinde bulunan "Blood status" değişkeni, Muggle Olanlar ve Olmayanlar olarak if-else kalıbı kullanılanılarak bir sütun haline getirilmiştir. Bu sutüns Muggle ismi verilmiştir ve bu durum yine Characters olarak atanmıştır.

Oluşturduğumuz bu yeni Charachters veri seti kullanılarak, Karakterlerin Evleri(House) ve Muggle değişkenleri kullanılarak grup haline getrilip içerisindeki NA değerleri çıkartılmıştır. Sayı olarak özetlemdikten sonra oranlanmıştır ve "harry3" olarak atanmıştır.

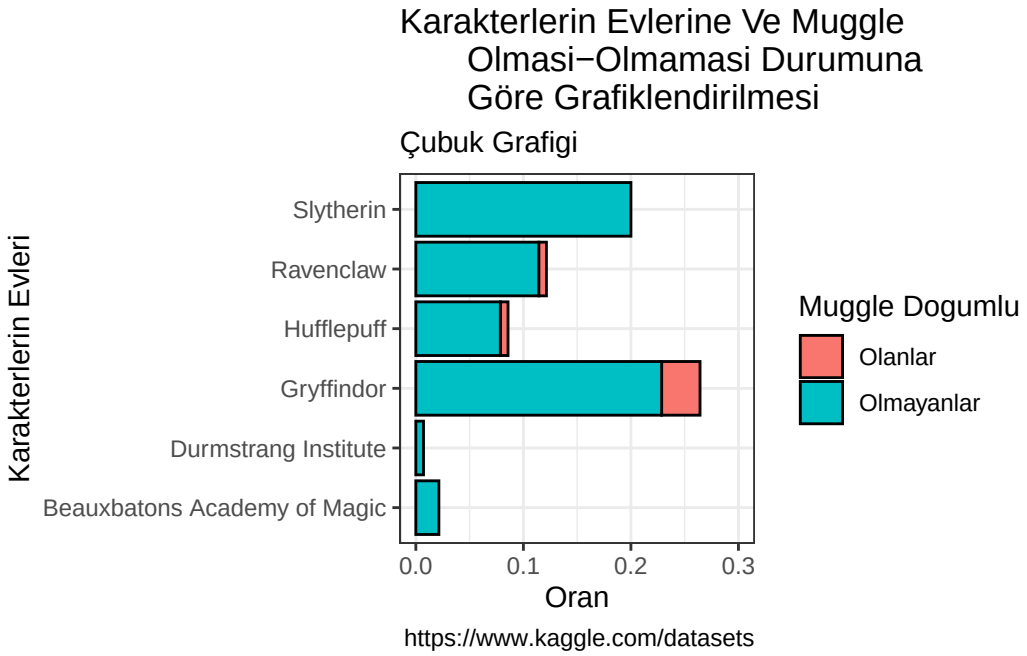
```
Characters <- Characters %>%
  add_column(Muggle = if_else(Characters$`Blood status` ==
                                "Muggle-born", "Olanlar", "Olmayanlar"))

harry3 <- Characters %>%
  group_by(House, Muggle) %>%
  drop_na(House, Muggle) %>%
  summarise(sayi = n()) %>%
mutate(oran = sayi / nrow(Characters))
```

Yeni oluşturduğumuz veri seti kullanırak Karakterlerin Evlerine Ve Muggle Olması-Olmaması Durumuna Göre çubuk grafiği çizilmiştir. Grafiğin okunurluğunu arttırmak ve kolaylaştırmak için gerekli estetikler ve lejantlar düzenlenmiştir.

```
harry3 %>%

  ggplot() +
  geom_bar(aes(x = House, y = oran, fill = Muggle),
    stat = "identity", color = "black") +
  labs(x = "Karakterlerin Evleri",
       y = "Oran",
        title = "Karakterlerin Evlerine Ve Muggle
       Olmasi-Olmamasi Durumuna
       Göre Grafiklendirilmesi",
       subtitle = "Çubuk Grafigi" ,
       fill = "Muggle Dogumlu",
  caption = "https://www.kaggle.com/datasets") +
   coord_flip() +
  ylim(0,0.3) +
  theme_bw()
```



YORUM: Harry Potter Karakterlerinin Evlerine Göre Muggle Doğumlu Olanlar Ve Olmayanların oranlarının verildiği grafikte , Mugggle Doğumlu olanların çoğunun Gryffindor evinde bulunduğunu görebiliyoruz. Aynı zamanda Muggle Doğumlu olmayanların da çoğunluğunun bu eve ait olduğunu söylebiliyoruz. “Slytherin”, “Durmstang Institute”ve “Beauxbatons Academy of Magic” evlerine ait olan karakterlerin tümü Muggle Doğumlu olmayanlardan oluşmaktadır.