Breaking Bad, Veri Bilimci Maaşları ve En Çok Satan Kitaplar Veri Setlerinde Veri Görselleştirme

SERCAN ÖNCÜ

11/27/22

ÖZET

Bu raporda Breaking Bad dizisi, En çok satan kitaplar ve Veri bilimci maaşları veri setleri incelenmiştir. Breaking Bad dizisinin bölümlerin yazar sayısı ve reyting puanlarına göre oranları görselleştirilmiştir. En çok satan kitapların türlerine ve okuyucu puanlarına göre oranları görselleştirilmiştir. Veri bilimci sayıları tecrübe düzeyleri, çalışma sistemi ve firma büyüklüklerine göre oranları görselleştirilmiştir.

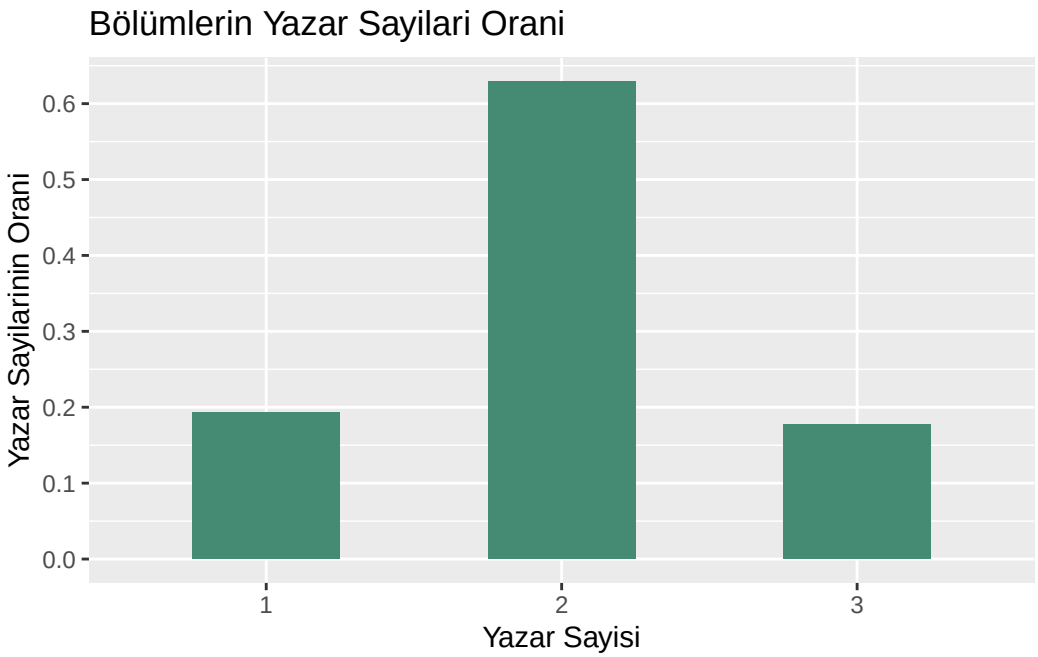
```r
options(repos = list(CRAN="http://cran.rstudio.com/"))
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("dplyr")
library(ggplot2)
library(tidyverse)
library(dplyr)
```

1. Breaking Bad

Bu veri seti sezon, bölüm, bölümlerin imdb puanı, bölümlerin süresi, bölümlerin yazar sayısı vb. başlıklar altında Breaking Bad dizisi için veriler içermektedir.

1.1 Bölümlerin yazar sayısına göre oranları

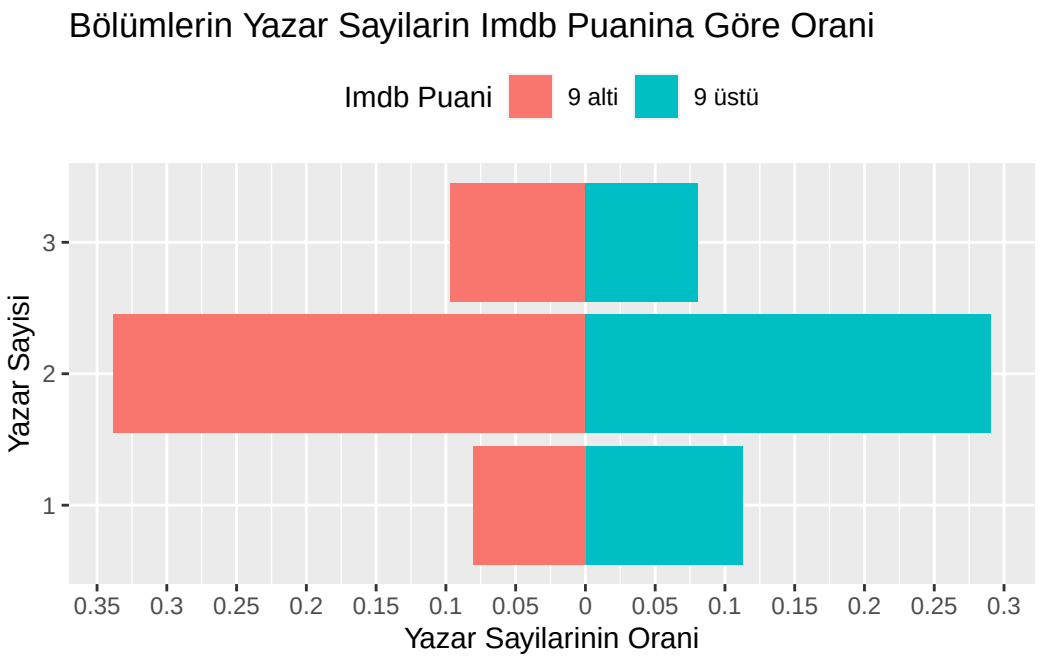
```
breaking_bad <- read_csv("C:/Users/serca/Desktop/SercanÖncü/breaking_bad.csv")
breaking_bad <- breaking_bad%>%
  add_column(imdb =
               if_else(breaking_bad$Rating_IMDB < 9, "1", "2"))
bb <- breaking_bad %>%
  tidyr::separate_rows(`Written by`,
                       sep = ", ") %>%
  group_by(Season,
           Episode,
           Rating_IMDB,
           imdb) %>%
  summarise(yazar = n())%>%
  mutate(oran = frequency(yazar)/nrow(breaking_bad))
bb <- bb%>%
  add_column(abc =
               if_else(bb$imdb == "1", -bb$oran, bb$oran ))
ggplot(bb, aes(x = as.factor(yazar) ,
               y=oran))+
  geom_bar(stat = "identity",
           fill = "aquamarine4",
           width = 0.5)+
  labs(x = "Yazar Sayisi",
       y = "Yazar Sayilarinin Orani",
       title = "Bölümlerin Yazar Sayilari Orani")+
  scale_y_continuous(breaks = seq(0,1,0.1))
```



Yukarıdaki grafikte Breaking Bad dizisinin bölümlerinin yazar sayılarına baktığımızda 2 yazarlı bölümler dizinin tüm bölümlerinin yarısından bile fazla olduğu görülmektedir. Yapımcılar 2 yazarlı bölümleri daha etkileyici bulmuş ve çoğunlukla 2 yazarlı şeklilde devam etmek istemiş olabilirler.

1.2 Bölümlerin yazar sayısına ve reyting puanlarına göre oranları

```
ggplot(bb, aes(x = as.character(yazar),
               fill = as.factor(imdb),
               y=abc))+
  geom_bar(stat = "identity" )+
  labs(x = "Yazar Sayisi",
       y = "Yazar Sayilarinin Orani",
       fill = "Imdb Puani",
       title = "Bölümlerin Yazar Sayilarin Imdb Puanina Göre Orani")+
  scale_fill_discrete(labels = c("9 altı", "9 üstü"))+
  scale_y_continuous(labels = abs, breaks = seq(-0.5, 0.5, 0.05))+
  coord_flip()+
  theme(legend.position = "top")
```



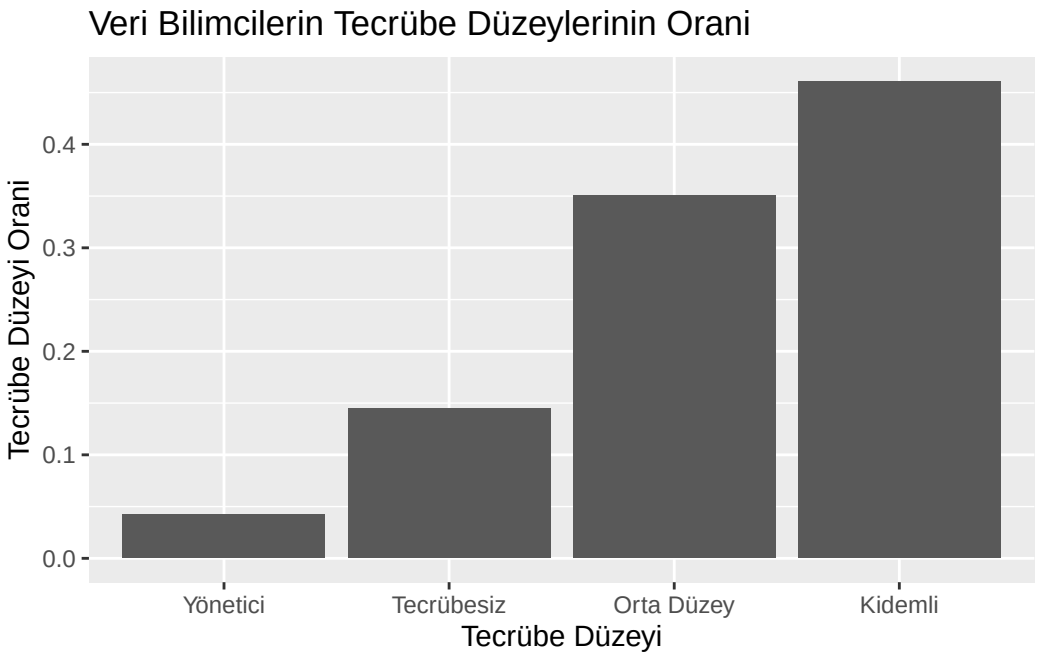
Yukarıdaki grafikte yazar sayiları kendi içlerinde tek tek incelendiğinde imdb puanının 9 altı ve 9 üstü sayıları oran farkı 1 yazarlı bölümlerde daha çok olduğu görülmektedir. 1 yazarlı bölümler çoğunlukla 9 üstü puan alırken 2 ve 3 yazarlı bölümlerin 9 puan altı bölümleri sayıca fazla çıkmıştır. Belki de tek yazarla devam edilseydi çok daha farklı bir Breaking Bad izleyebilirdik. Yukarıdaki oranlara bakıldığında en kötü sonuçlar 3 yazarlı bölümlerde alınmıştır. 3 yazarlı bölümlerde fikir ayrılıkları biraz fazla yaşanmış olabilir.

2. Veri bilimci maaşları

Bu veri seti çalışma sistemi, deneyim ve şirket büyüklüğü vb. başlıklar altında Veri bilimci maaşları için veriler içermektedir.

2.1 Tecrübelerine göre veri bilimci sayılarının oranı

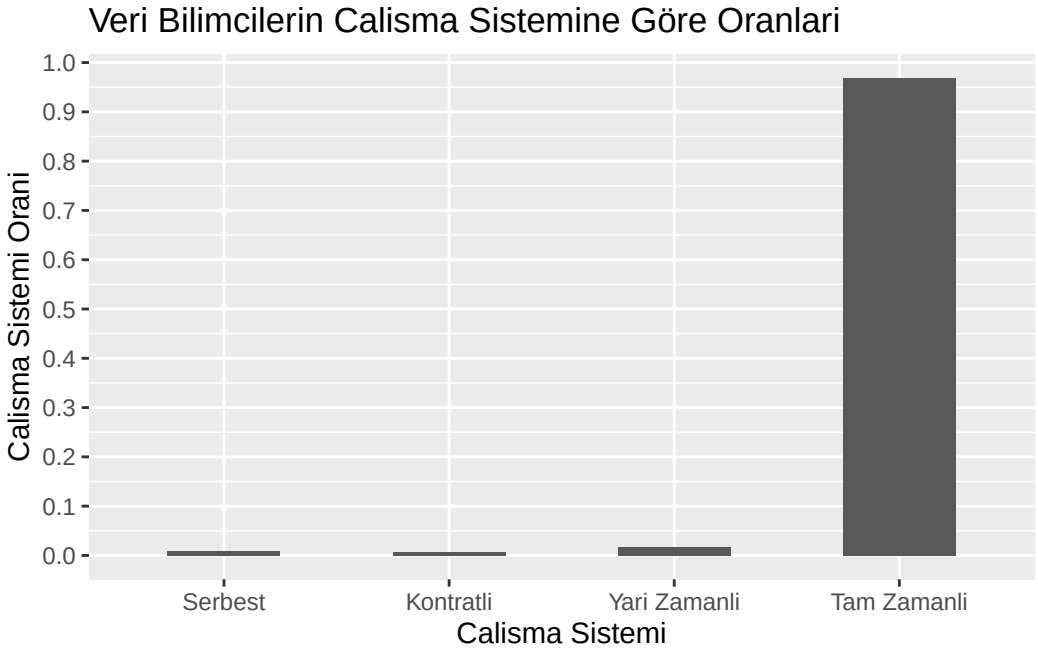
```
Data <- read_csv("C:/Users/serca/Desktop/SercanÖncü/Data_Science_Fields_Salary_Categorizat
data2 <- Data %>%
  group_by(Experience,
           Company_Size,
           Employment_Status) %>%
  summarise(exp = n())%>%
  mutate( oran2 = exp / nrow(Data))
ggplot(data2, aes(x = reorder(Experience,+oran2),
                  y = oran2))+
  geom_bar(stat = "identity")+
  labs(x = "Tecrübe Düzeyi",
       y = "Tecrübe Düzeyi Orani",
       title = "Veri Bilimcilerin Tecrübe Düzeylerinin Orani")+
scale_x_discrete(labels = c("Yönetici", "Tecrübesiz", "Orta Düzey", "Kidemli"))
```



Yukarıdaki grafiğe bakıldığında yönetici bir veri bilimci olmak zor olabilir diyebiliriz çünkü tüm veri bilimciler arasında %5'i bile yönetici değil. Kıdemli veri bilimci oranına bakıldığında ise yakın zaman diliminde çok fazla veri bilimci çıkmamış olabilir çünkü uzun süredir sektörde olup kendini geliştirmiş olan veri bilimcilerin sayıları epey bir fazla gözükmektedir.

2.2 Veri bilimcilerinin çalışma sistemine göre oranları

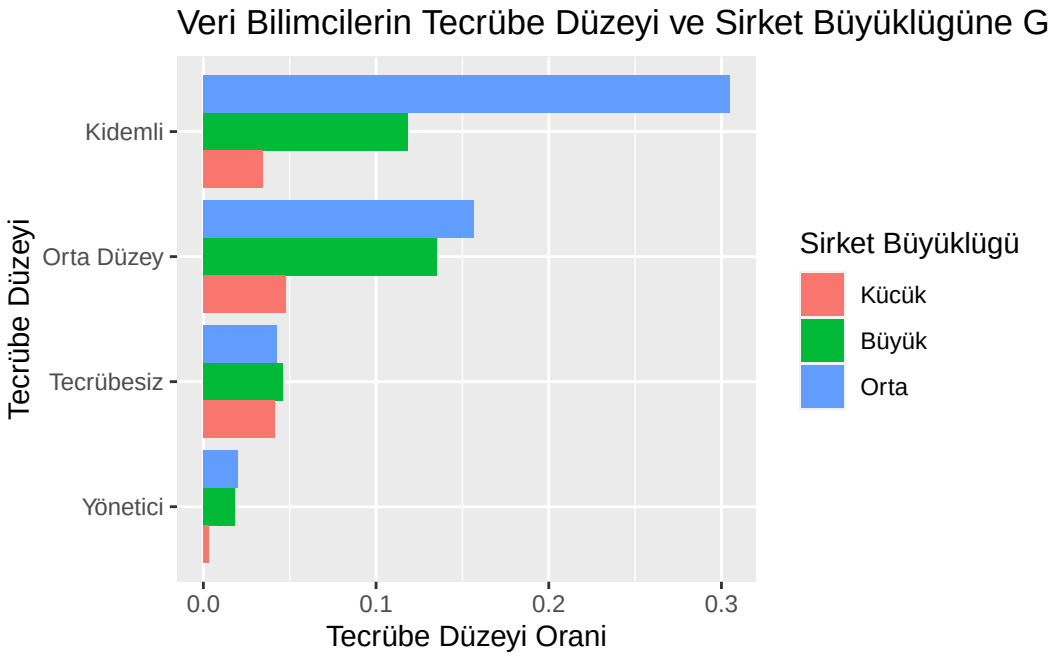
```
ggplot(data2, aes(x = reorder(Employment_Status, +oran2),
                  y = oran2))+
  geom_bar(stat = "identity", width = 0.5)+
  labs(x = "Calisma Sistemi",
       y = "Calisma Sistemi Orani",
       title = "Veri Bilimcilerin Calisma Sistemine Göre Oranlari")+
scale_x_discrete(labels = c("Serbest",
                            "Kontratli", "Yari Zamanli", "Tam Zamanli"))+
  scale_y_continuous(minor_breaks = seq(0,1,0.05), breaks = seq(0,1,0.1))
```



Yukarıdaki grafiğe bakıldığında veri bilimi sektörü için vazgeçilmez olan çalışma sistemi tam zamanlı diyebiliriz. Çünkü tüm çalışma sistemlerinin %95'inden fazlasını tam zamanlı çalışma sistemi kapsıyor.

2.3 Firma büyüklüğüne ve tecrübe düzeyine göre veri bilimci sayılarının oranı

```
ggplot(data2, aes(x = reorder(Experience,+oran2),
                  y = oran2,
                  fill = reorder(Company_Size, +oran2)))+
  geom_bar(stat = "identity", position = "dodge")+
  labs(x = "Tecrübe Düzeyi",
       y = "Tecrübe Düzeyi Orani",
       fill = "Sirket Büyüklügü",
       title = "Veri Bilimcilerin Tecrübe Düzeyi ve Sirket Büyüklügüne Göre Oranlari")+
  scale_fill_discrete(labels = c("Kücük", "Büyük", "Orta"))+
  scale_x_discrete(labels = c("Yönetici", "Tecrübesiz", "Orta Düzey", "Kidemli"))+
  coord_flip()
```



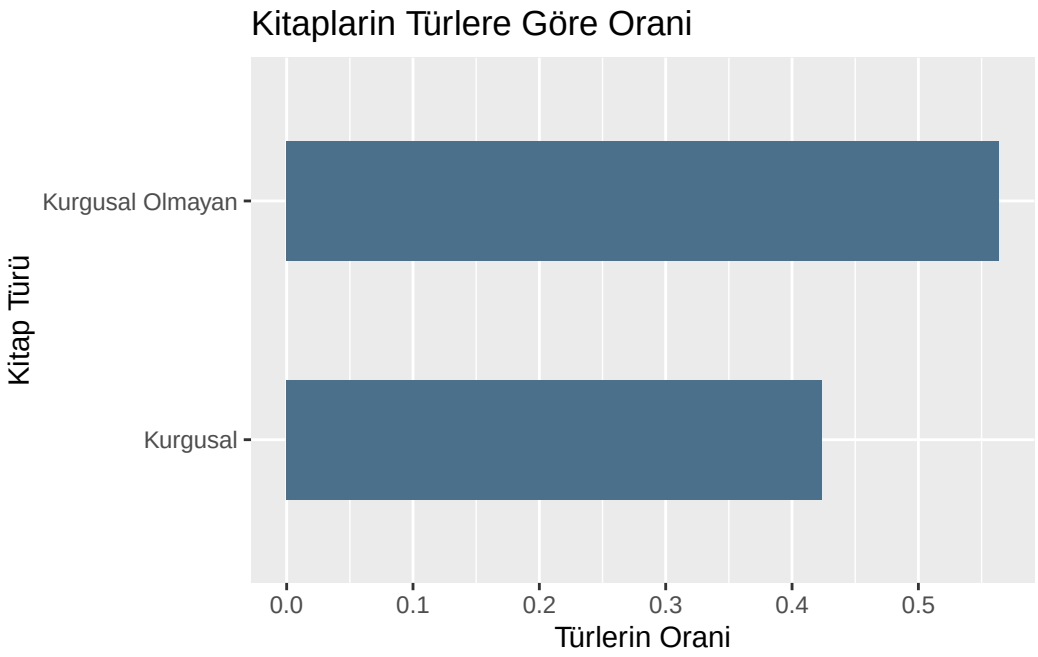
Yukarıdaki grafiğe bakıldığında tecrübeler farketmeksizin en çok veri bilimci orta büyüklükte şirketlerde çalışmaktadır. Bunun sebebi belki de orta büyüklükteki şirketlerin veri bilimcilere daha yüksek ücret vermesinden kaynaklanıyor olabilir. Orta büyüklükteki şirketlerin veri bilimcileri ise çoğunlukla kıdemli veri bilimcilerden oluşmaktadır. Küçük ve büyük şirketlerde çoğunlukla orta düzey tecrübeye sahip veri bilimci çalıştırmaktadır. Bunun sebebi büyük şirketler için belkide sayıca fazla veri bilimci çalıştırıyor olmasından kaynaklı olabilir, küçük şirketler için ise kıdemli veri bilimcilerin istedikleri ücretler bütçelerini aşıyor olabilir.

3. En çok satan kitaplar

Bu veri seti kullanıcı puanları ve kitap türleri vb. başlıklar altında En çok satan kitaplar için veriler içermektedir.

3.1 En çok satan kitapların türlerine göre oranları

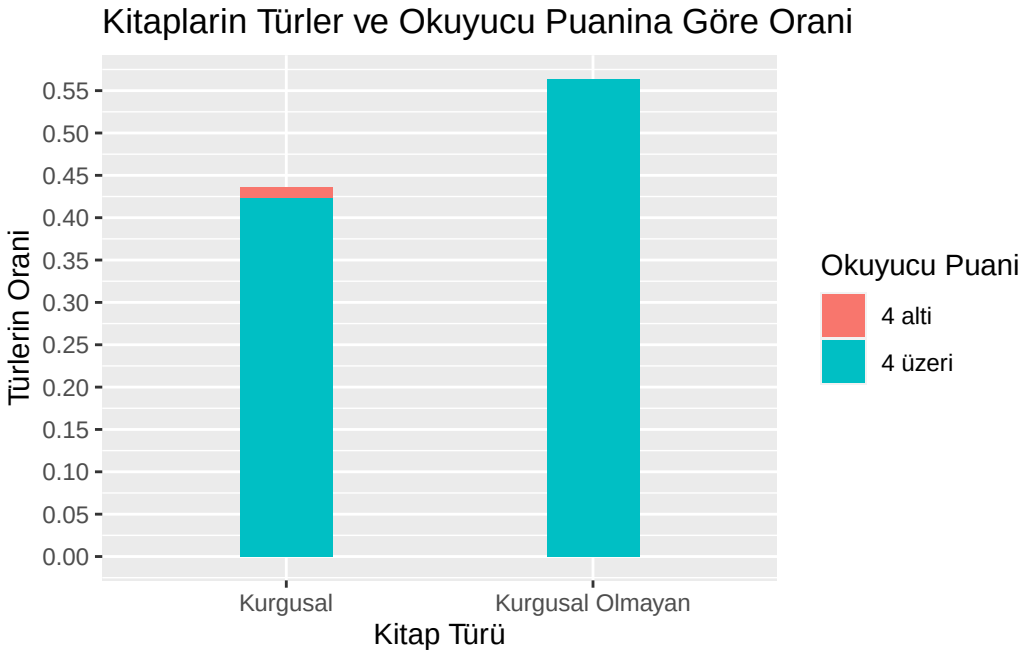
```
bestsellers <- read_csv("C:/Users/serca/Downloads/archive (1)/bestsellers.csv")
bestsellers <- bestsellers %>%
  add_column(rating =
               if_else(bestsellers$`User Rating`< 4, "1", "2"))
best <- bestsellers %>%
  group_by(Genre,
           rating) %>%
  summarise(tür = n())%>%
  mutate( oran2 = tür / nrow(bestsellers))
ggplot(best, aes(x = Genre,
                 y = oran2))+
  geom_bar(stat = "identity",
           position = "dodge",
           fill = "skyblue4",
           width = 0.5)+
scale_y_continuous(breaks = seq(0,1,0.1))+
  labs(x = "Kitap Türü",
       y = "Türlerin Orani",
       title = "Kitaplarin Türlere Göre Orani")+
  scale_x_discrete(labels = c("Kurgusal", "Kurgusal Olmayan"))+
  coord_flip()
```



Yukarıdaki grafiğe bakıldığında kurgusal olmayan kitapların en çok satan kitaplar arasında daha fazla olduğunu görülmektedir. Okuyucuların kurgusal olmayan kitaplara daha çok yöneldiğini söyleyebiliriz. Kurgusal kitaplar yazan yazarların hikayelerini okuyucuların isteklerine göre şekillendirmesi belkide ileride bu oranı değiştirebilir.

3.2 En çok satan kitapların türlerine ve kullanıcı reytingine göre oranları

```
ggplot(best, aes(x = Genre,
                 y = oran2,
                 fill = as.factor(rating)))+
  geom_bar(stat = "identity", width = 0.3)+
  labs(x = "Kitap Türü",
       y = "Türlerin Orani",
       title = "Kitaplarin Türler ve Okuyucu Puanina Göre Orani",
       fill = "Okuyucu Puani")+
  scale_x_discrete(labels = c("Kurgusal", "Kurgusal Olmayan"))+
  scale_y_continuous(breaks = seq(0,0.55,0.05))+
  scale_fill_discrete(labels = c("4 alti", "4 üzeri"))
```



Yukarıdaki grafiğe bakıldığında en çok satan kitapların çok büyük bir bölümü okuyucular tarafından 4 puan üzerinde puanlanmıştır. Kurgusal olmayan kitapların tümü 4 puan üzerindeyken kurgusal kitaplarda çok az da olsa 4 puan altında okuyucu puanı görülmektedir. Yine burada kurgusal kitap yazarlarına hikayelerinizi okuyuculara göre şekillendirin gibi bir çağrıda bulunabiliriz.