3 Farklı Veri Setinin Dağılımlarının Görselleştirilmesi

Sahranur İnce

15 Kasım 2022

Özet

Bu raporda; Kaggle'dan alınan 3 farklı veri setinin dağılımlarının görselleştirme çalışmaları bulunmaktadır. Birinci veri seti Breaking Bad dizisi, ikinci veri seti The Big Bang Theory dizisi ve üçüncü veri seti 2009-2019 yılları arasında Amazon'da en çok satan 50 kitap 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("dplyr")
install.packages("ggridges")
install.packages("ggforce")
install.packages("MetBrewer")
library(ggplot2)
library(dplyr)
library(ggridges)
library(ggforce)
library(MetBrewer)
```

Uygulama 1: Breaking Bad Dizisi

Veri Setinin İncelenmesi

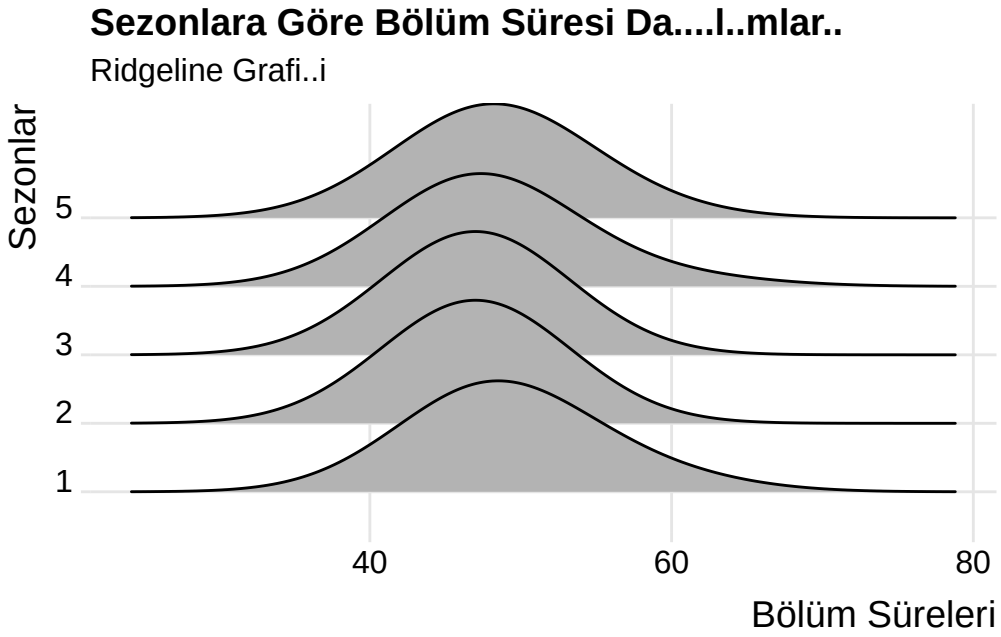
Bu veri seti Breaking Bad adlı dizi hakkında veriler içermektedir. Bu veri setinde 62 gözlem ve 10 değişken vardır. Bu değişkenler şunlardır: Gün, Sezon, Bölüm, Bölüm Adı,Yönetmen, Yazar, Bölüm Süreleri, Özet, Reyting Puanları, İzlenme Sayısı.

Sezonlara göre bölüm süresi dağılımlarını inceleyiniz. Bölüm sürelerindeki en yüksek değişimin gözlendiği sezonu belirleyiniz.

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

```
breaking_bad$Season <- factor(breaking_bad$Season, ordered = TRUE,
                              levels = c("1" , "2" , "3" , "4" , "5"))
```

```
ggplot(breaking_bad, aes(x = Duration_mins, y = Season)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none") +
  labs(y = "Sezonlar",
       x = "Bölüm Süreleri",
       title = "Sezonlara Göre Bölüm Süresi Dağılımları",
       subtitle = "Ridgeline Grafiği")
```

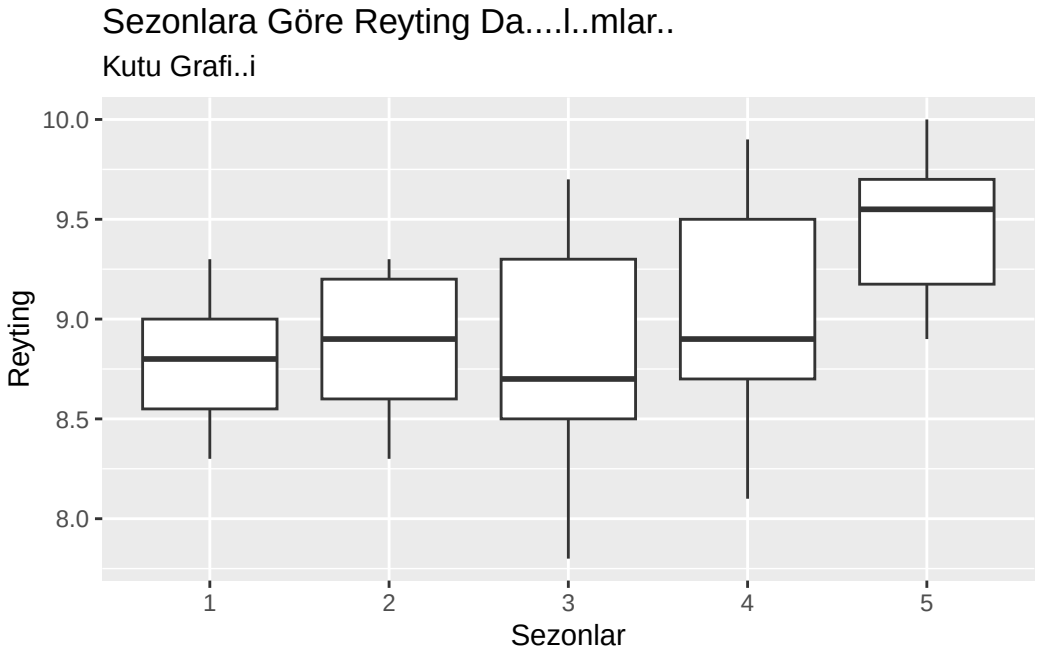


Grafikte elde edilen sonuçlara göre, bölüm sürelerinde en yüksek değişimin bulunduğu sezonu söylemek zordur. Çünkü, grafiğe genel olarak baktığımızda her beş sezon için de bölüm sürelerinin arasında çok fark olmadığını söyleyebiliriz. Tüm sezonlarda bölüm süreleri 40-60 dakika arasında değişmektedir.

Sezonlara göre reyting dağılımlarını araştırınız. Reyting değişiminin en düşük olduğu sezonu belirleyiniz.

```
breaking_bad$Season <- factor(breaking_bad$Season, ordered = TRUE,
                              levels = c("1" , "2" , "3" , "4" , "5"))
```

```
ggplot(breaking_bad, aes(x = Season, y = Rating_IMDB)) +
  geom_boxplot() +
  labs(y = "Reyting",
       x = "Sezonlar",
       title = "Sezonlara Göre Reyting Dağılımları",
       subtitle = "Kutu Grafiği")
```



Grafikte elde edilen sonuçlara göre, reyting değişiminin en düşük olduğu sezon 1. sezondur. Reyting değişiminin en yüksek olduğu sezon ise 3. sezondur. Veri setinde herhangi bir aykırı değer bulunmamaktadır. En düşük reyting değerinin 3. sezonda verildiğini ve 8'den biraz daha düşük bir değer olduğunu, en yüksek reyting değerinin ise 5. sezonda verildiğini ve 10 olduğunu söyleyebiliriz.

Sezonlara göre izlenme sayısı dağılımlarını araştırınız. İzlenme sayısı değişiminin en düşük olduğu sezonu belirleyiniz.

Uygulama 2: Big Bang Theory Dizisi

Veri Setinin İncelenmesi

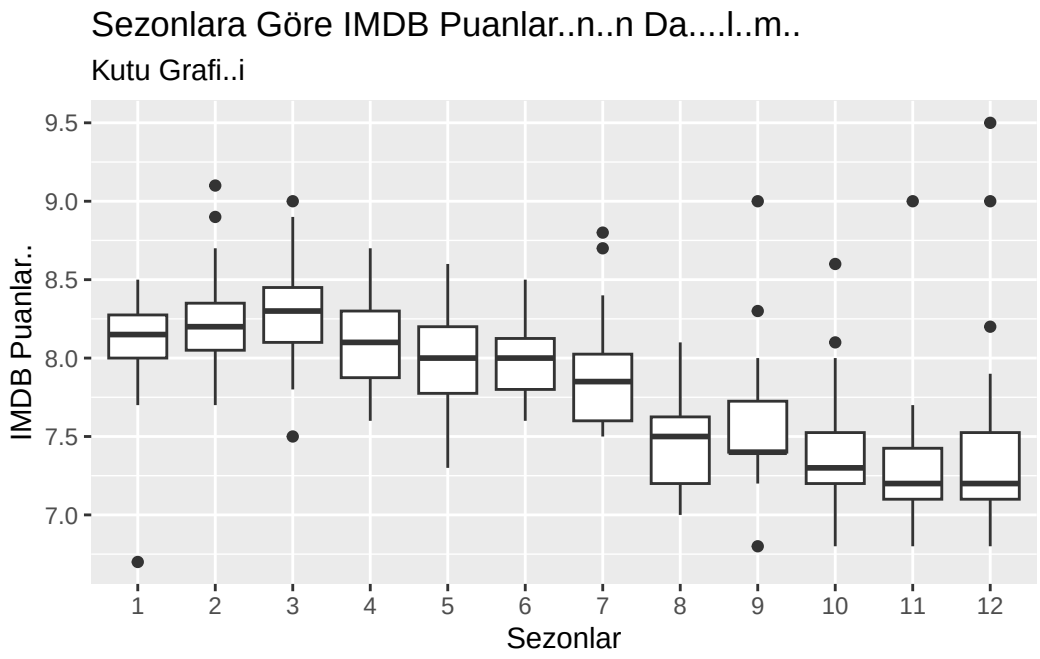
Bu veri seti The Big Bang Theory adlı dizi hakkında veriler içermektedir. Bu veri setinde 280 gözlem ve 7 değişken vardır. Bu değişkenler şunlardır: Sezon, Bölüm Numarası, Bölüm Adı, Yayınlandığı Gün, Rating Puanları, Oy Sayısı, Bölüm Özeti.

Sezonlara göre IMDB puanlarının dağılımını araştırınız. En asimetrik puan dağılımı gözlemlenen sezonu belirleyiniz.

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

```
big_bang_theory_imdb$season <- factor(big_bang_theory_imdb$season, ordered = TRUE,
                                      levels = c("1" , "2" , "3" , "4" , "5" ,
                                                 "6" , "7" , "8" , "9" , "10",
                                                 "11", "12"))
```

```
ggplot(big_bang_theory_imdb, aes(x = season, y = imdb_rating)) +
  geom_boxplot() +
  labs(y = "IMDB Puanları",
       x = "Sezonlar",
       title = "Sezonlara Göre IMDB Puanlarının Dağılımı",
       subtitle = "Kutu Grafiği")
```

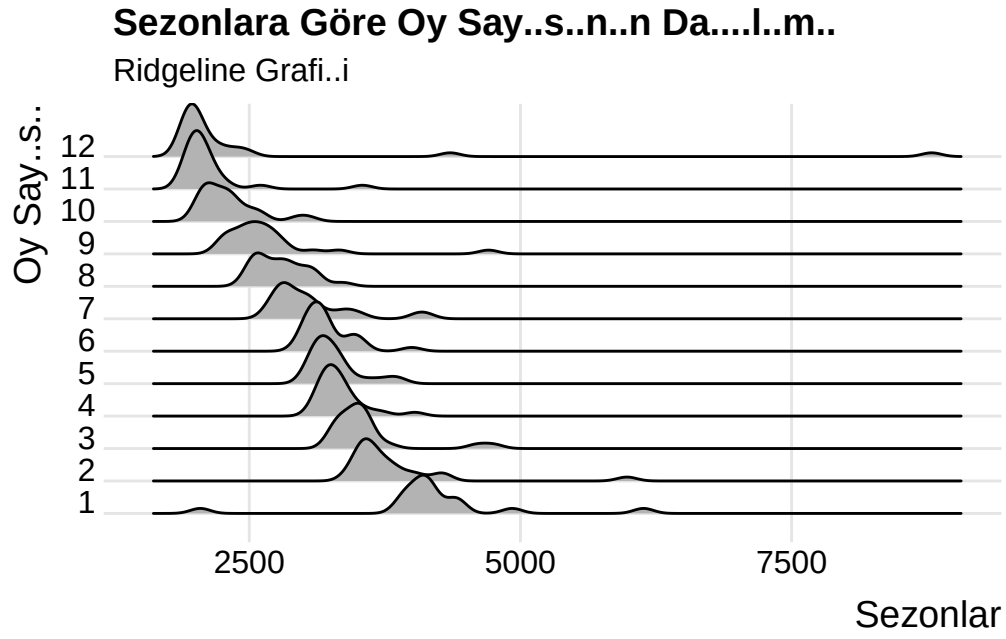


Grafikte elde edilen sonuçlara göre, en asimetrik puan dağılımı 9.sezondadır. En simetrik puan dağılımı ise 4. ve 5. sezonlarda görülmektedir. Her iki sezon için de medyan çizgisi kutuyu ortaladığı için, iki sezon arasında bir ayrım yapmak zordur. Bazı sezonlarda aykırı değerler bulunmaktadır. Bu sezonlar; 1,2,3,7,9,10,11 ve 12. sezonlardır. En düşük IMDB puanının 1.sezonda verildiğini ve 7'den biraz daha düşük bir değer olduğunu, en yüksek IMDB puanının ise 12. sezonda verildiğini ve 9.5 olduğunu görüyoruz.

Sezonlara göre oy sayısının dağılımını araştırınız. Puan sayıları arasında en yüksek farklılıkların gözlendiği sezonu belirleyiniz.

```
big_bang_theory_imdb$season <- factor(big_bang_theory_imdb$season, ordered = TRUE,
                                levels = c("1" , "2" , "3" , "4" ,
                                           "5" , "6" , "7" , "8" ,
                                           "9" , "10", "11", "12"))
```

```
ggplot(big_bang_theory_imdb, aes(x = total_votes, y = season)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none") +
  labs(y = "Oy Sayısı",
       x = "Sezonlar",
       title = "Sezonlara Göre Oy Sayısının Dağılımı",
       subtitle = "Ridgeline Grafiği")
```



Grafikte elde edilen sonuçlara göre, oy sayıları arasındaki en yüksek farklılıkların gözlendiği sezon 12. sezondur. Oy sayıları arasındaki en düşük farklılıkların gözlendiği sezon ise 8.sezondur. Bunu grafikteki inişli çıkışlı dalgalanmalardan anlayabiliriz. 8. sezon dışındaki her sezonda aykırı değerler bulunmaktadır.

Uygulama 3: En Çok Satan Kitaplar

Veri Setinin İncelenmesi

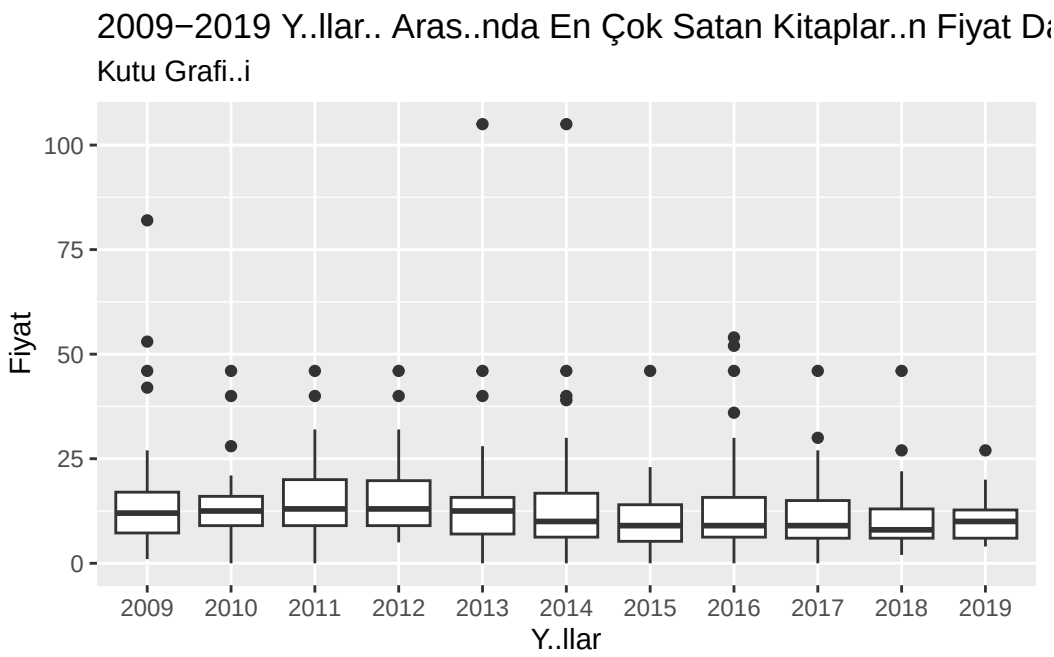
Bu veri seti 2009-2019 yılları arasında Amazon'da en çok satan 50 kitabı içermektedir. Bu veri setinde 550 gözlem ve 7 değişken vardır. Bu değişkenler şunlardır: Kitap Adı, Yazar, Okuyucu Puanı, Yorumlar, Fiyat, Yıl, Tür.

Yıllara göre en çok satan kitapların fiyat dağılımının değişimini araştırınız.

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

```
bestsellers_with_categories$Year <- factor(bestsellers_with_categories$Year,
                                           ordered = TRUE,
                                  levels = c("2009" , "2010" , "2011" ,
                                             "2012" , "2013" , "2014" ,
                                             "2015" , "2016" , "2017" ,
                                             "2018" , "2019"))
```

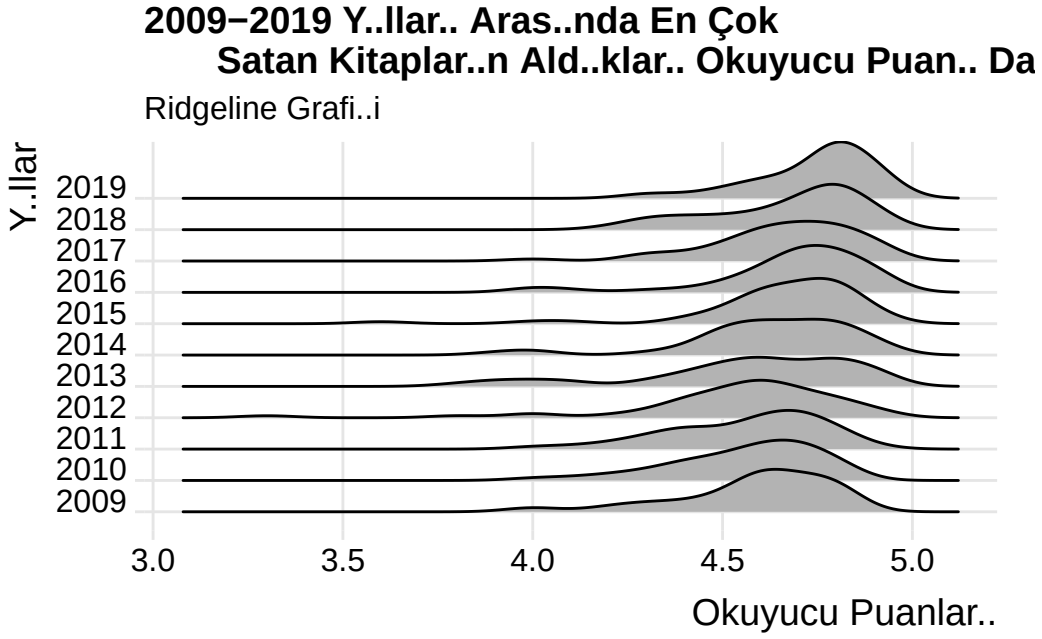
```
ggplot(bestsellers_with_categories, aes(x = Year, y = Price)) +
  geom_boxplot() +
  labs(y = "Fiyat",
       x = "Yıllar",
       title = "2009-2019 Yılları Arasında En Çok Satan Kitapların Fiyat Dağılımı",
       subtitle = "Kutu Grafiği")
```



Grafikte elde edilen sonuçlara göre, en düşük kitap fiyatının 0, en yüksek kitap fiyatının ise 100'ün üzerinde bir değer olduğunu görülmektedir. Her yılda aykırı değerler bulunmaktadır. Bu da demek oluyor ki her yıl, ortalama kitap fiyatının üstünde yani ortalama kitap fiyatından daha pahalı kitaplar bulunmaktadır. Fiyat dağılımı açısından en simetrik olan yılların 2009 ve 2015 olduğunu söyleyebiliriz.

Yıllara göre en çok satan kitapların aldıkları okuyucu puanı dağılımının değişimini araştırınız.

```
ggplot(bestsellers_with_categories, aes(x = `User Rating`, y = Year)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none") +
  labs(y = "Yıllar",
       x = "Okuyucu Puanları",
       title = "2009-2019 Yılları Arasında En Çok
       Satan Kitapların Aldıkları Okuyucu Puanı Dağılımı",
       subtitle = "Ridgeline Grafiği")
```

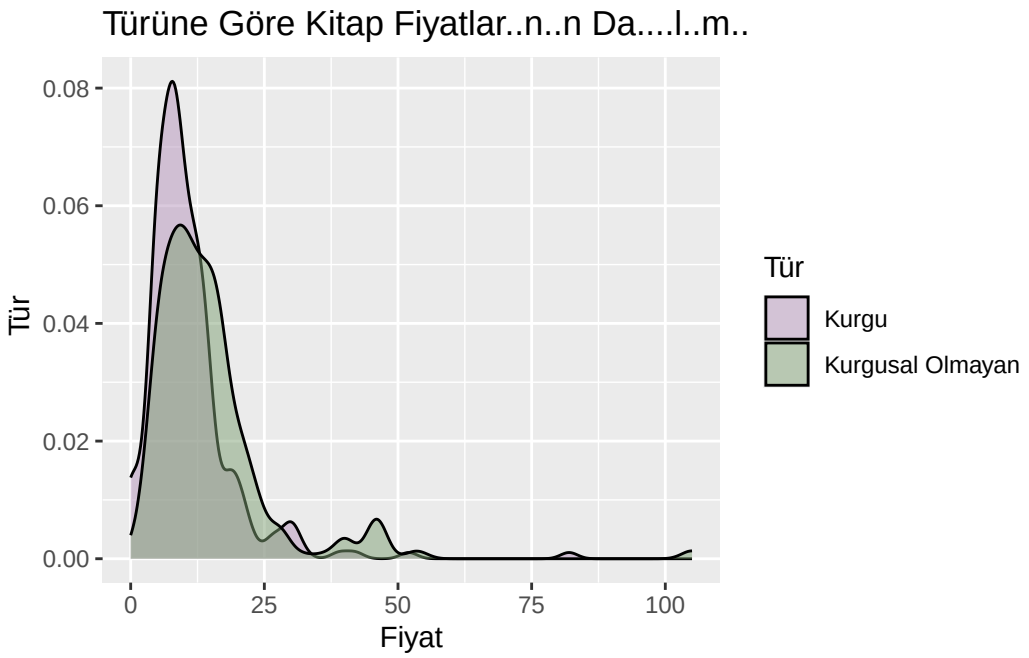


Grafikte elde edilen sonuçlara göre, en yüksek okuyucu puanı 2019 yılında, en düşük okuyucu puanı ise 2012 yılında bulunmaktadır. 2019 yılındaki okuyucu puanları yaklaşık 4.5-5 arasında

değerler alırken, 2012 yılındaki okuyucu puanları yaklaşık 3.4-5 arasında değerler almıştır. Verilen okuyucu puanlarının genellikle 4.5-5 arasında değerler aldığını söyleyebiliriz.

Türüne göre kitap fiyatlarının dağılımını araştırınız.

```
ggplot(bestsellers_with_categories, aes(x = Price, fill = Genre)) +
  geom_density(alpha = 0.5) +
  labs( x = "Fiyat" ,
        y = " Tür",
        title = "Türüne Göre Kitap Fiyatlarının Dağılımı",
        fill = "Tür") +
  scale_fill_manual(values = met.brewer("Cassatt2",2),
                    labels = c("Kurgu","Kurgusal Olmayan"))
```



Grafiğe baktığımızda, kitap türlerinin "Kurgu" ve "Kurgu Olmayan" olarak ikiye ayrıldığını görüyoruz. Elde edilen sonuçlara göre 0-25 fiyat aralığında kurgu türündeki kitaplar daha fazladır, kurgusal olmayan kitaplar daha azdır. 40-50 fiyat aralığında kurgusal olmayan kitaplar kurgu kitaplarından daha fazladır. İki farklı kitap türü içinde aykırı değerler bulunmaktadır. Kurgu türündeki kitaplar için yaklaşık 80 değerinde, kurgusal olmayan kitaplar için ise 100 ve üzerindeki değerlerde kitaplar bulunmaktadır. Her iki kitap türününde en çok 0-25 fiyat aralığında kitapları vardır.