Veri Görselleştirme - Ödev 3

Sezai Ufuk Oral - 36664666542

Kullanılacak Kütüphanelerin Yüklenmesi

```
library("ggplot2")
library("dplyr")
library("priceR")
library("readr")
```

1) Data Science Salary Veri Seti

1.1 Veri Seti Hakkında

Bu veri seti Veri Bilimi alanında çalışan Veri Bilimcileri'nin çalıştıkları pozisyonları, çalışma düzenlerini, yıllık maaşlarını, çalıştıkları şirketlerin büyüklük skalalarını, şirketlerinin ve kendilerinin lokasyonlarını, uzaktan çalışma yüzdelerini ve çalıştıkları yılları içermektedir. https://www.kaggle.com/datasets/whenamancodes/data-science-fields-salary-categorization

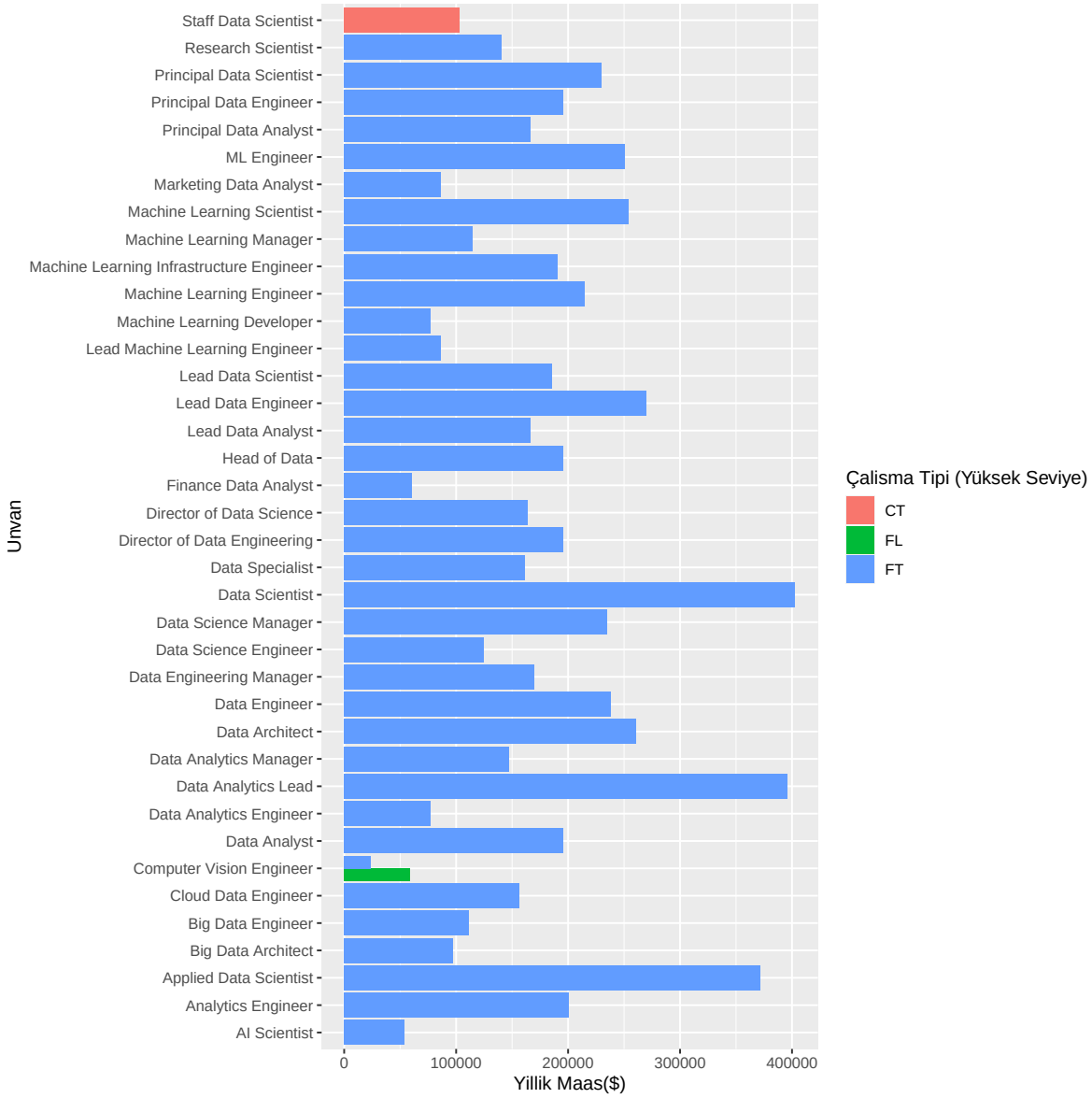
```
dataScienceDataset <- read.csv("Data_Science_Fields_Salary_Categorization.csv")
```

1.2 Veri Seti Hakkındaki Çıkarımlar

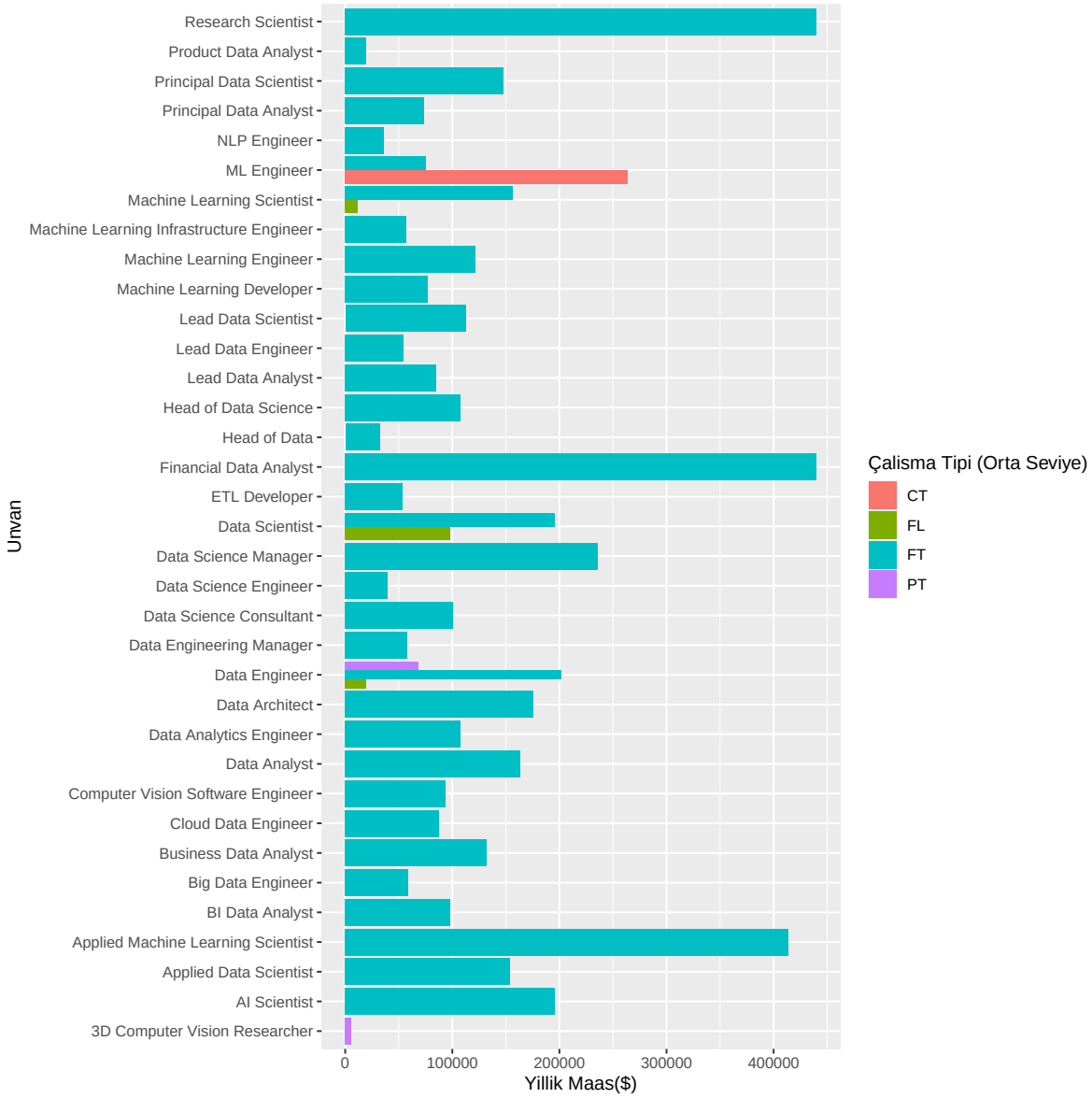
(1.2.1) Tecrübe düzeylerine göre veri bilimi pozisyonları aylık maaşlarının dolar karşılığı bazında dağılımını araştırınız.

```
salaryDistributionData <- select(dataScienceDataset, c('Designation','Experience','Employment_Statu

salaryDistributionData$Salary_In_Rupees <- round((parse_number(salaryDistributionData$Salary_In_Ru

colnames(salaryDistributionData)[4] <- "Salary_In_Dollars"

options("scipen"=100, "digits"= 2)

se <- ggplot(
```

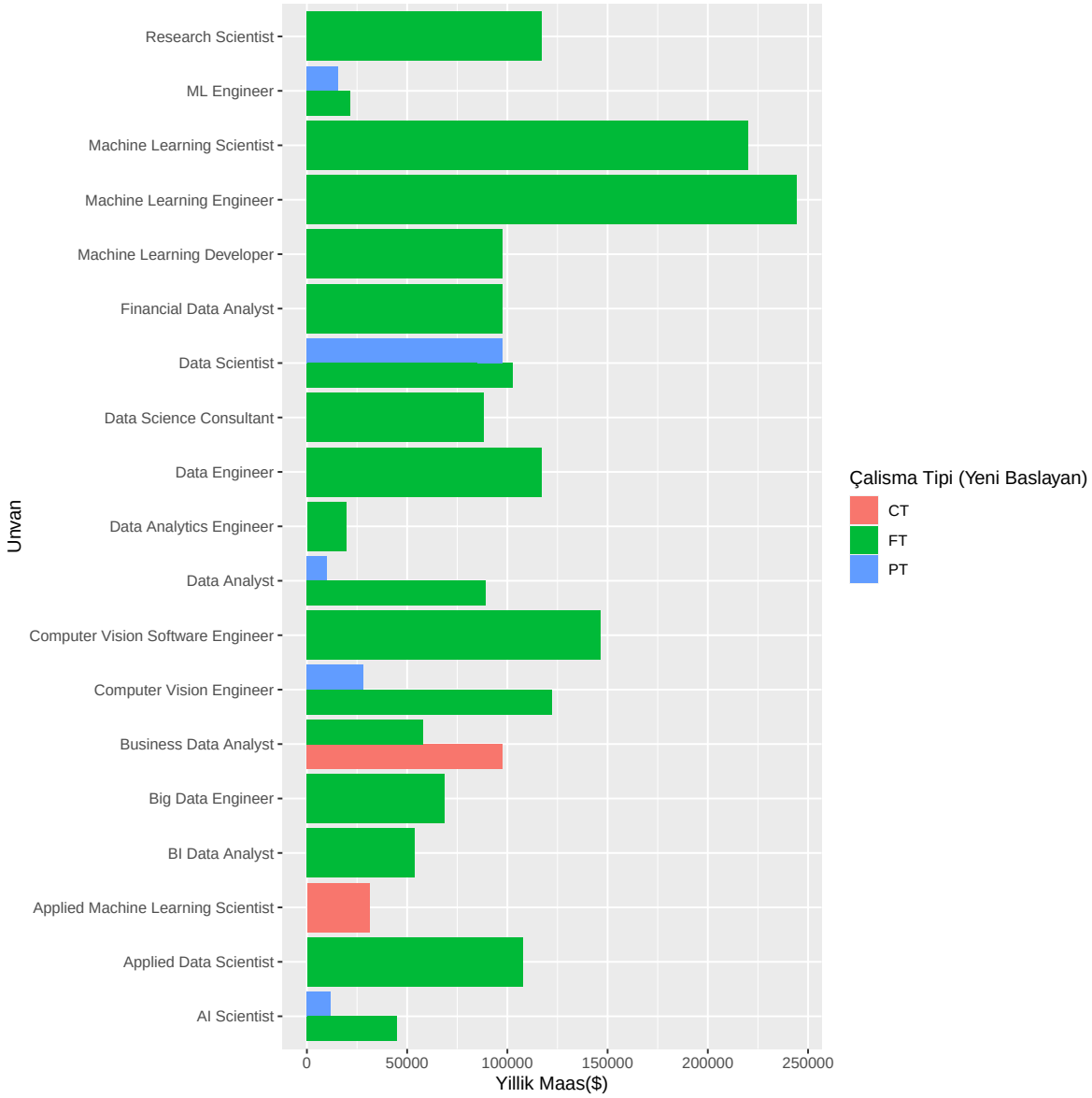
```
    salaryDistributionData[salaryDistributionData$Experience == "SE",],
    aes(x = Salary_In_Dollars, y = Designation, fill = Employment_Status)) +
    geom_col(position = 'dodge') + labs(y= "Unvan", x = "Yıllık Maaş($)") + scale_fill_discrete(nam

mi <- ggplot(
    salaryDistributionData[salaryDistributionData$Experience == "MI",],
    aes(x = Salary_In_Dollars, y = Designation, fill = Employment_Status)) +
    geom_col(position = 'dodge') + labs(y= "Unvan", x = "Yıllık Maaş($)") + scale_fill_discrete(nam

en <- ggplot(
    salaryDistributionData[salaryDistributionData$Experience == "EN",],
    aes(x = Salary_In_Dollars, y = Designation, fill = Employment_Status)) +
    geom_col(position = 'dodge') + labs(y= "Unvan", x = "Yıllık Maaş($)") + scale_fill_discrete(nam

ex <- ggplot(
    salaryDistributionData[salaryDistributionData$Experience == "EX",],
    aes(x = Salary_In_Dollars, y = Designation, fill = Employment_Status)) +
    geom_col(position = 'dodge') + labs(y= "Unvan", x = "Yıllık Maaş($)") + scale_fill_discrete(nam
```



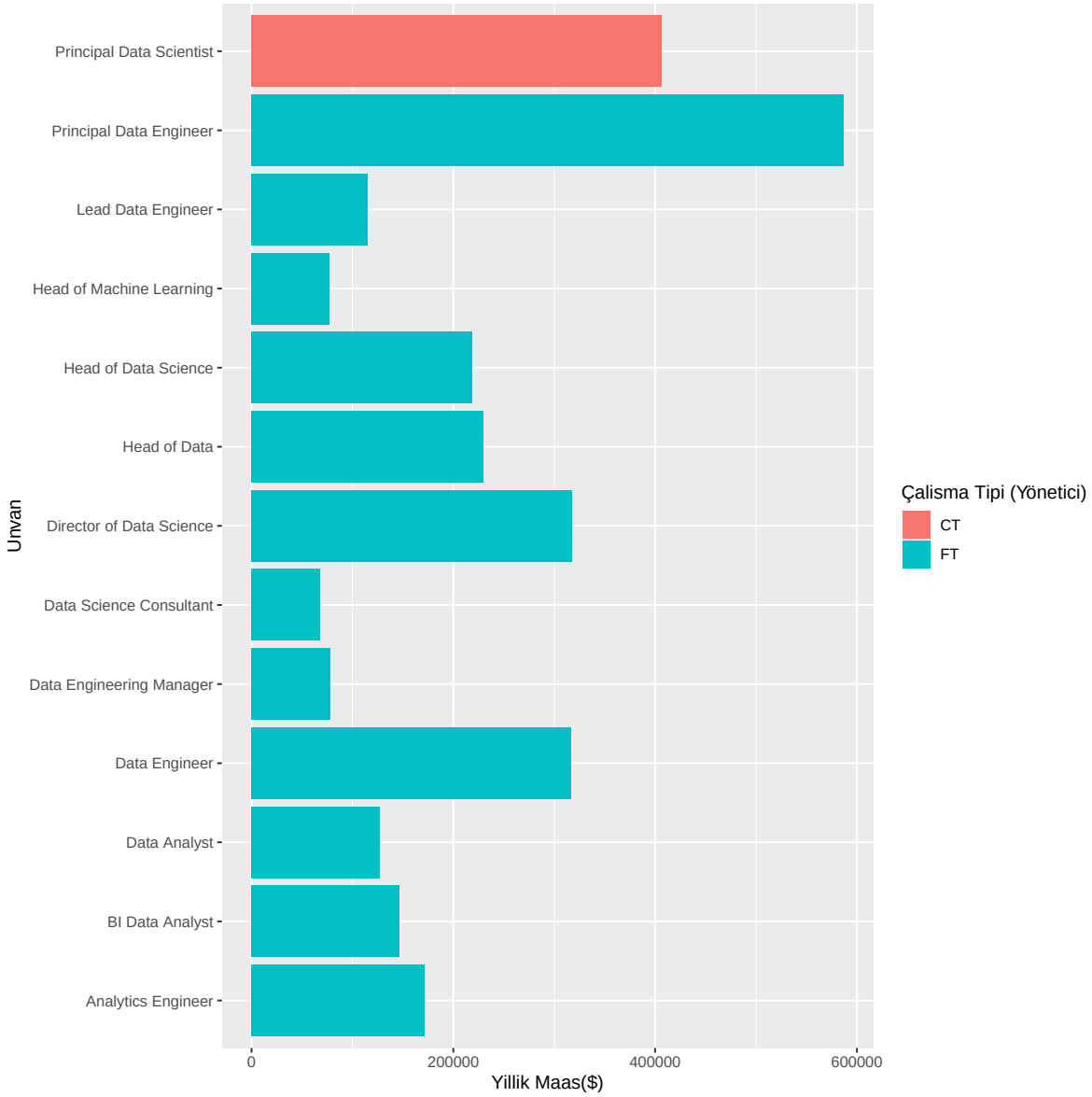
Yukarıdaki grafikte Veri Bilimi alanında çalışan Yüksek Seviye(Senior) çalışanların Unvanlarını, Çalışma Tiplerini (CT = "Kontratlı", FL="Freelance", FT= "Tam Zamanlı") ve aldıkları yıllık maaş seviyelerini görebilmekteyiz. Bu grafikte Yüksek Seviye çalışanlarda Freelance ve Kontratlı çalışma tiplerinin yaygın olmadığı çıkarımını yapabilir, en yüksek maaş getirisi olan Unvanların Data Scientist, Data Analytics ve Applied Data Scientist olduğunu söyleyebiliriz. Çoğunluk olarak maaşlar 250.000$ civarından düşük seviyede seyir etmektedir.



Yukarıdaki grafikte Veri Bilimi alanında çalışan Orta Seviye(Middle) çalışanların Unvanlarını, Çalışma Tiplerini (CT = "Kontratlı", FL="Freelance", FT= "Tam Zamanlı",PT="Yarı Zamanlı") ve aldıkları yıllık maaş seviyelerini görebilmekteyiz. Bu grafikte Orta Seviye çalışanlarda Freelance ve Kontratlı çalışma tiplerinin yaygın olmadığı çıkarımını yapabilir, en yüksek maaş getirisi olan Unvanların Financial Data Analyst, Research Scientist ve Applied Machine Learning Scientist olduğunu söyleyebiliriz. Çoğunluk olarak maaşlar 200.000$ civarından düşük seviyede seyir etmektedir. ML Engineer (Makine Öğrenmesi Mühendisliği) alanında diğer Unvanlardan farklı olarak Kontratlı çalışma modelinin hakimiyet göstermekte olduğu görülmektedir.



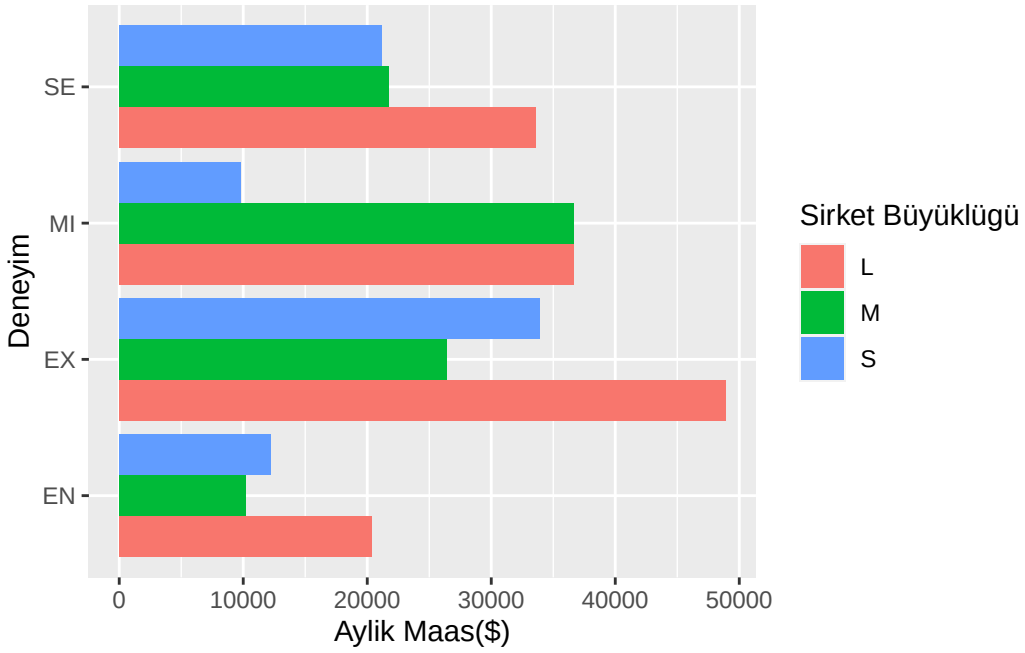
Yukarıdaki grafikte Veri Bilimi alanında çalışan Yeni Başlayan(Entry Level) çalışanların Unvanlarını, Çalışma Tiplerini (CT = "Kontratlı", FT= "Tam Zamanlı",PT="Yarı Zamanlı") ve aldıkları yıllık maaş seviyelerini görebilmekteyiz. Bu grafikte yeni başlayan çalışanlarda Yarı Zamanlı ve Kontratlı çalışma tiplerinin yaygın olmadığı çıkarımını yapabilir, en yüksek maaş getirisi olan Unvanların Machine Learning Engineer ve Machine Learning Scientist olduğunu söyleyebiliriz. Çoğunluk olarak maaşlar 150.000$ civarından düşük seviyede seyir etmektedir. Yeni başlayan bir Veri Bilimci'nin Makine Öğrenmesi alanı ile ilişkisi arttıkça maaşında da artış gözlemlenmektedir, buradan da çalışma piyasasının Makine Öğrenmesi üzerine bir trende geçmesi, talep olması yönünde bir yorum yapılabilir.



Yukarıdaki grafikte Veri Bilimi alanında çalışan Yönetici(Executive Level) çalışanların Unvanlarını, Çalışma Tiplerini (CT = "Kontratlı", FT= "Tam Zamanlı") ve aldıkları yıllık maaş seviyelerini görebilmekteyiz. Bu grafikte yöneticilikte Yarı Zamanlı ve Freelance gibi çalışma tiplerinin yaygın olmadığı çıkarımını yapabilir, en yüksek maaş getirisi olan Unvanların Principal Data Engineer ve Principal Data Scientist olduğunu söyleyebiliriz. Çoğunluk olarak maaşlar 300.000$ civarından düşük seviyede seyir etmektedir.

(1.2.2) Firma büyüklüğüne ve tecrübe düzeyine göre aylık maaş tutarı dağılımlarını araştırınız. Tecrübesiz adaylar için çalışmaya başlamak için en uygun firma ölçeği hangisidir?

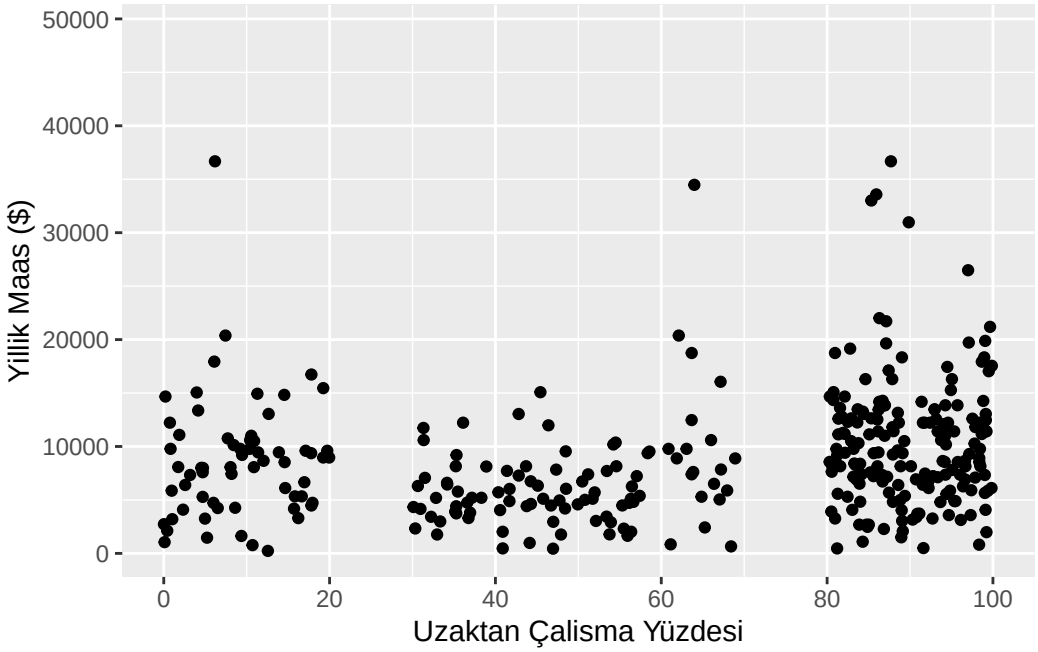
```
salaryDistributionData$Salary_In_Dollars <- salaryDistributionData$Salary_In_Dollars/12
ggplot(
    salaryDistributionData,
    aes(x = Salary_In_Dollars, y = Experience, fill = Company_Size)) +
    geom_col(position = 'dodge') + labs(y= "Deneyim", x = "Aylık Maaş($)") + scale_fill_discrete(na
```



Yukarıdaki grafikte deneyim faktörünün aylık maaş üzerindeki artışa etkisini görebilmekteyiz. İşe yeni başlayan tecrübesiz bir Veri Bilimci için en mantıklı seçimin büyük ölçekli bir şirkete girmesi olduğu yorumunu yapabiliriz.

(1.2.3) Uzaktan çalışma yüzdesi sistemine göre yıllık maaş ücreti dağılımlarının değişimini araştırınız.

```
salaryDistributionData$Remote_Working_Ratio <- dataScienceDataset$Remote_Working_Ratio

ggplot(salaryDistributionData, aes(x = Remote_Working_Ratio, y = Salary_In_Dollars)) + geom_jitter
  labs(y = "Yıllık Maaş ($)",
       x = "Uzaktan Çalışma Yüzdesi")
```



Yukarıdaki grafik üzerindeki yoğunlaşmaları yorumlar isek yüksek uzaktan çalışma yüzdelerinde bir çalışan yığılması olduğunu ve birçok kişinin birbirine yakın maaşlar aldığı söylenebilir.

2) Breaking Bad Veri Seti

2.1 Veri Seti Hakkında

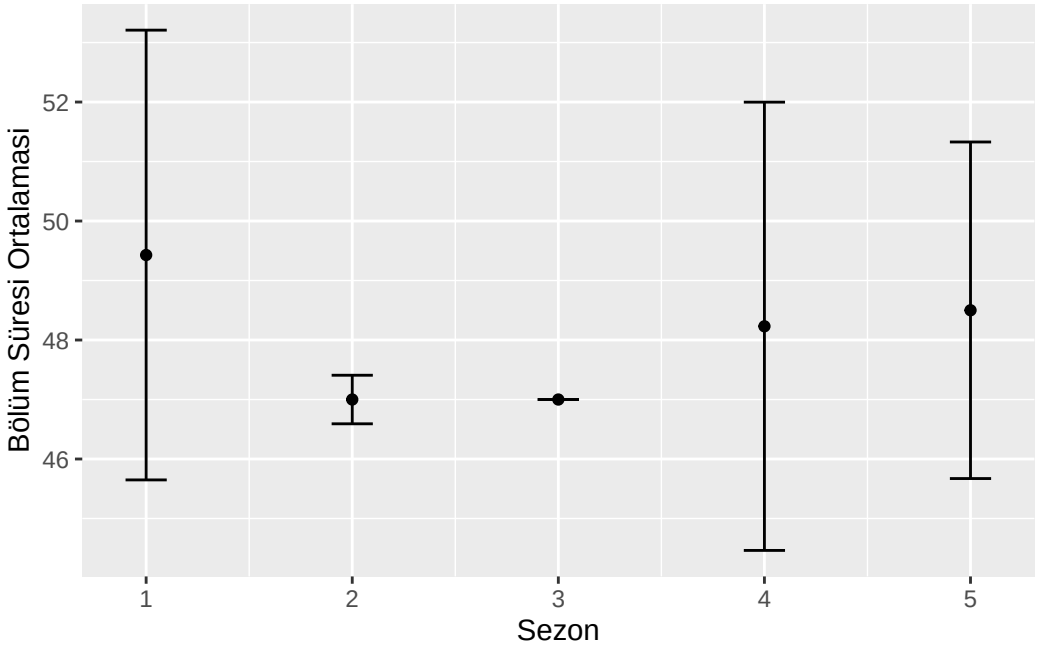
Bu veri seti Breaking Bad isimli Amerikan Televizyon Dizisinin bölümleri hakkında bilgiler içermektedir (IMDB Puanı, Yayınlanma Tarihi, Bölüm Adı, Yönetmen Adı vb.) https://www.kaggle.com/datasets/varpit94/breaking-bad-tv-show-all-seasons-episodes-data

```
breakingBadDataset <- read.csv("breaking_bad.csv")
```

2.2 Veri Seti Hakkındaki Çıkarımlar

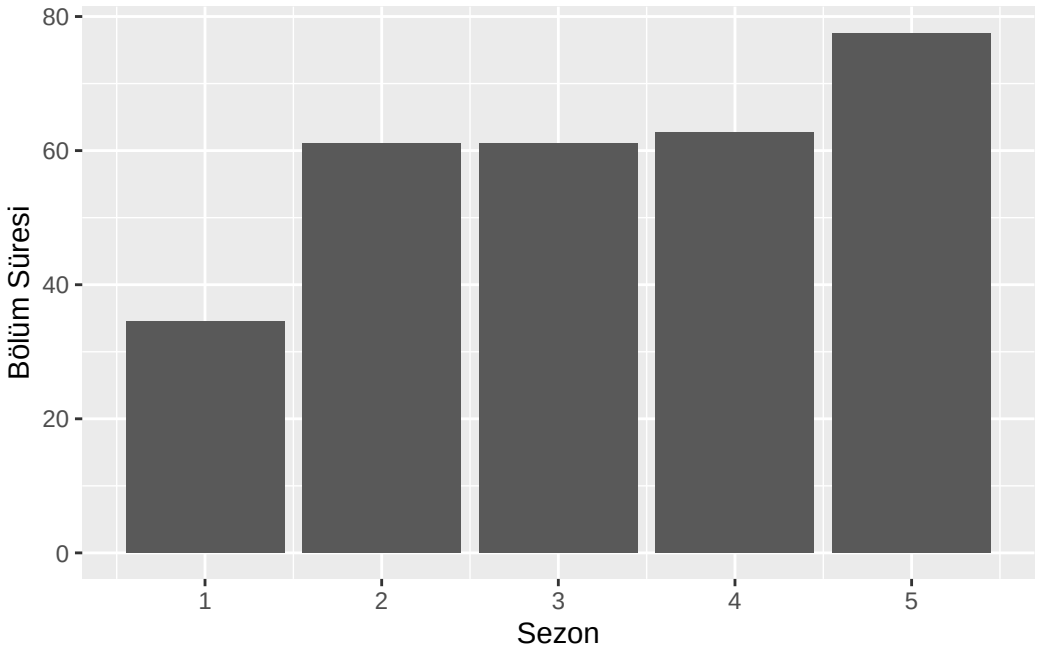
(2.2.1) Sezonlara göre bölüm süresi dağılımlarını inceleyiniz. Bölüm sürelerindeki en yüksek değişimin gözlendiği sezonu belirleyiniz.

```
by_season <- breakingBadDataset %>% group_by(Season)

by_seasonSummary <- by_season %>%
  group_by(Season) %>%
  summarise(
    sd = sd(Duration_mins, na.rm = TRUE),
    mean = mean(Duration_mins)
  )

ggplot(
  by_seasonSummary,
  aes(x = Season, y = mean, ymin = mean-sd, ymax = mean+sd)
  ) + geom_errorbar(width = 0.2) + geom_point() + labs(x="Sezon",y="Bölüm Süresi Ortalaması")
```



Yukarıdaki grafik incelenerek bölüm sürelerindeki en yüksek değişimin 4. sezonda yaşandığı çıkarımını yapabiliriz. Hata çubukları bizlere ortalamadan olan uzaklığı, çeşitlilik sınırlarını vermektedir.

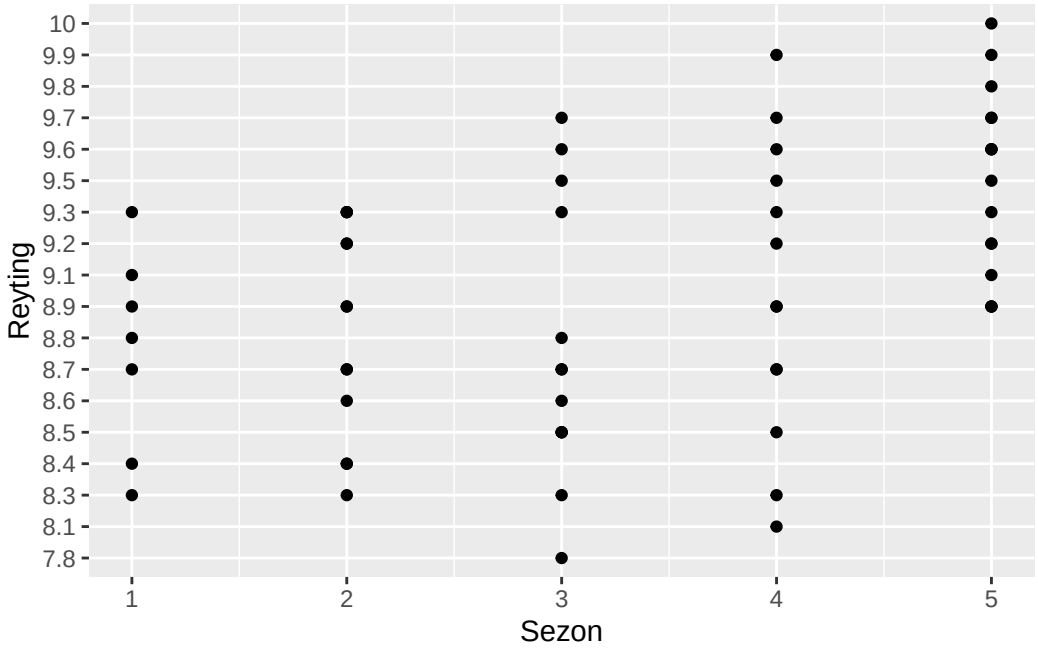
```
ggplot(data=by_season, aes(x=Season, y=Duration_mins/10)) + geom_col() + labs(x = "Sezon", y = "Bö
```



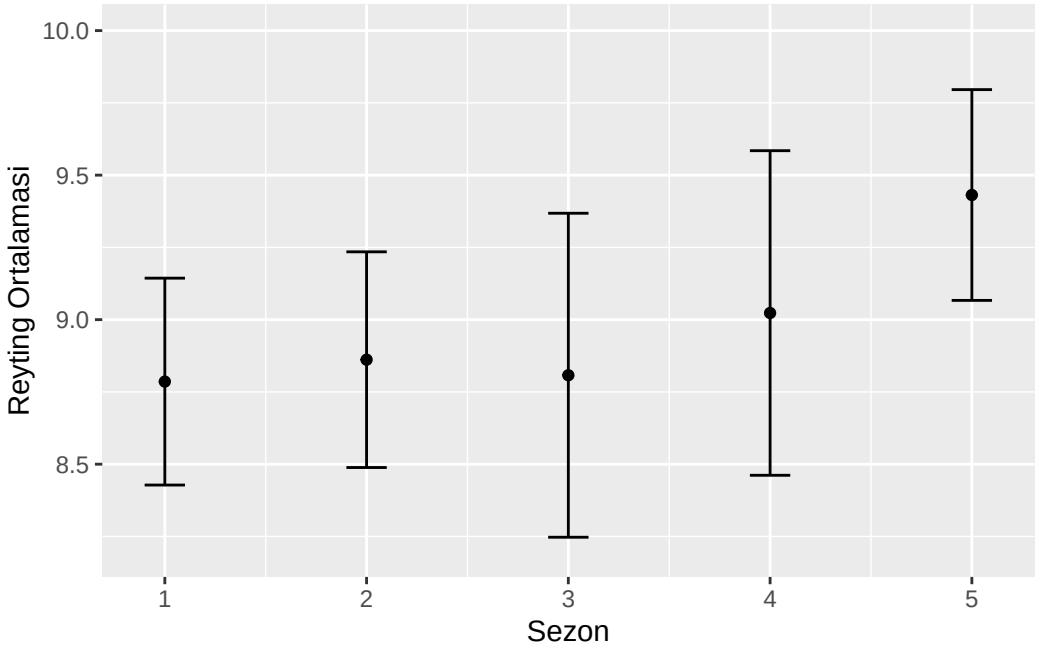
Yukarıdaki grafik incelendiğinde ise en yüksek bölüm sürelerine son sezon itibari ile ulaşıldığı görünmektedir.

(2.2.2) Sezonlara göre rating dağılımlarını araştırınız. Rating değişiminin en düşük olduğu sezonu belirleyiniz.

```
ggplot(data=by_season, aes(x=Season, y=as.factor(Rating_IMDB),ymax=5)) + geom_point() + labs(x = "S
```



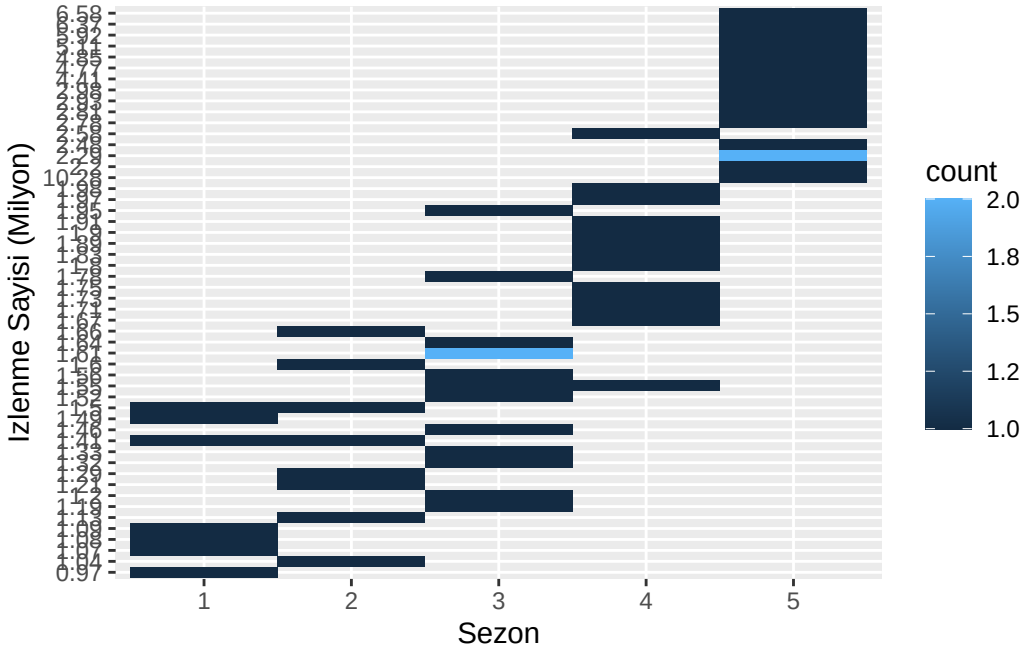

```
by_seasonRatingSummary <- by_season %>%
  group_by(Season) %>%
  summarise(
    sd = sd(Rating_IMDB, na.rm = TRUE),
    mean = mean(Rating_IMDB)
  )

ggplot(
  by_seasonRatingSummary,
  aes(x = Season, y = mean, ymin = mean-sd, ymax = mean+sd)
  ) + geom_errorbar(width = 0.2) + geom_point() + labs(x="Sezon",y="Reyting Ortalaması") + scale_y
```



Yukarıdaki grafikler incelendiğinde 1,2 ve 5. sezonların Rating değişimlerine neredeyse aynı derecede rastlandığı, diğer reytinglerden daha az değişime sahip oldukları, en az değişimin son sezonda gözlendiği ve son sezonda ratinglerin ortalamasının diğerlerinin üzerinde olduğu görülebilir. Bu grafikten halk tarafından ortalama bölüm kalitesi olarak en çok 5. sezonun beğenildiği çıkarımını yapabiliriz, ancak dizinin final sezonu olması sebebi ile seyircilerin duygusal yaklaşımını da göz ardı etmemeliyiz.

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

```
by_season <- by_season[which(by_season$U.S..viewers_million != "N/A"),]
by_season <- by_season[order(by_season$U.S..viewers_million),]

ggplot(by_season, aes(x = as.factor(Season),
                      y = U.S..viewers_million,
                      ymin = min(by_season$U.S..viewers_million),
                      ymax = max(by_season$U.S..viewers_million))) +
  geom_bin2d() + labs(y = "İzlenme Sayısı (Milyon)", x = "Sezon")
```



Yukarıdaki grafiğe bakarak en az izlenme sayısı değişiminin aralık sıklığına bakarak son sezon olduğunu söyleyebiliriz. En fazla değişime ise 2. Sezon'da rastlamaktayız.

3. Best Seller Books Veri Seti

3.1 Veri Seti Hakkında

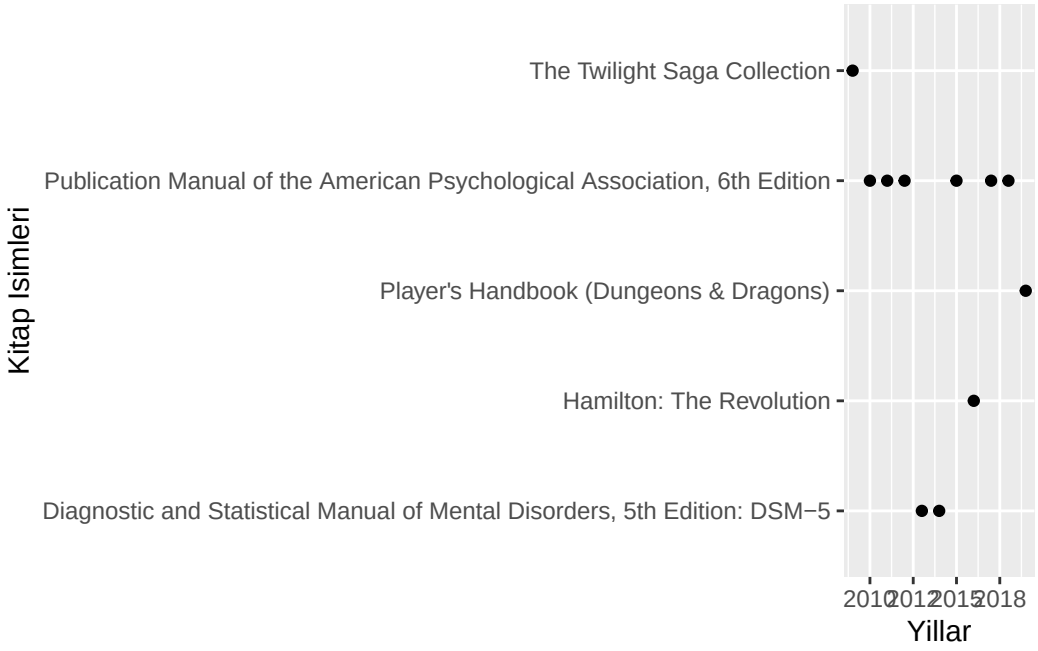
Bu veri seti 2009-2019 yılları arasında Amazon'un "En Çok Satan 50 Kitap" sıralamasına girebilmiş kitapları bulundurmaktadır. Bu veri setinde 550 adet kitap bulunmaktadır.

https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019

3.2 Veri Seti Hakkındaki Çıkarımlar

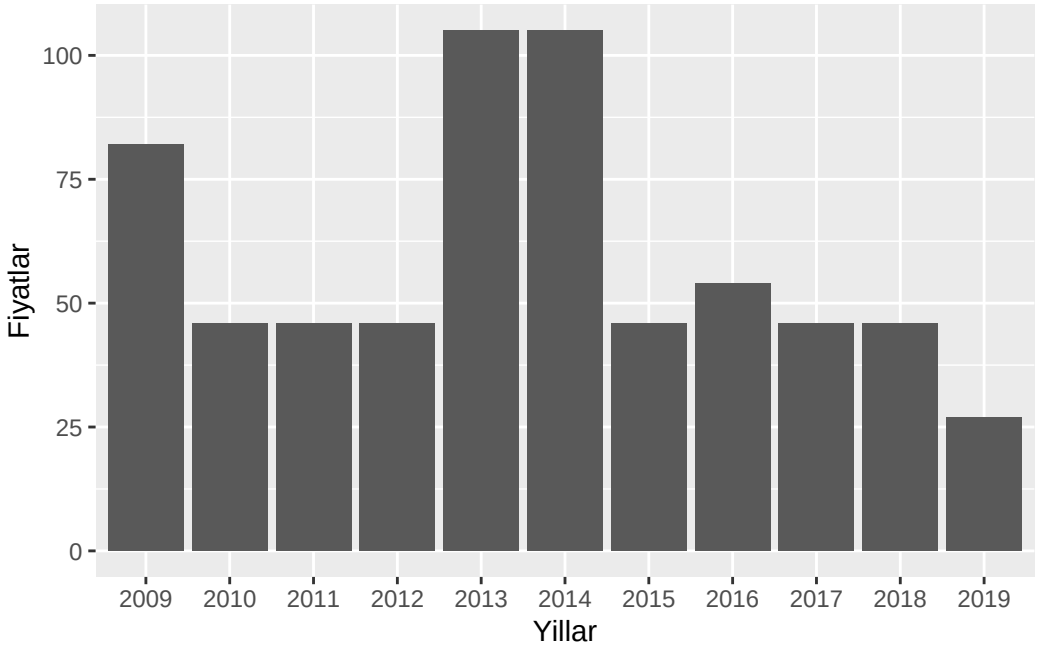
3.2.1 Yıllara göre en çok satan kitapların fiyat dağılımının değişimini araştırınız.

```
bestsellers <- read.csv("bestsellers.csv")
bestSellersOfYears <- bestsellers %>%
  group_by(Year) %>%
  filter(Price == max(Price))

ggplot(bestSellersOfYears,aes(y=Name,x=Year)) + geom_point()+labs(x="Yıllar",y="Kitap İsimleri")
```



Yukarıdaki grafikte görüldüğü üzere yıllara göre en çok satan kitap "Publication Manual of the American Psychological Association 6th Edition" kitabı olmuştur (6 kez).

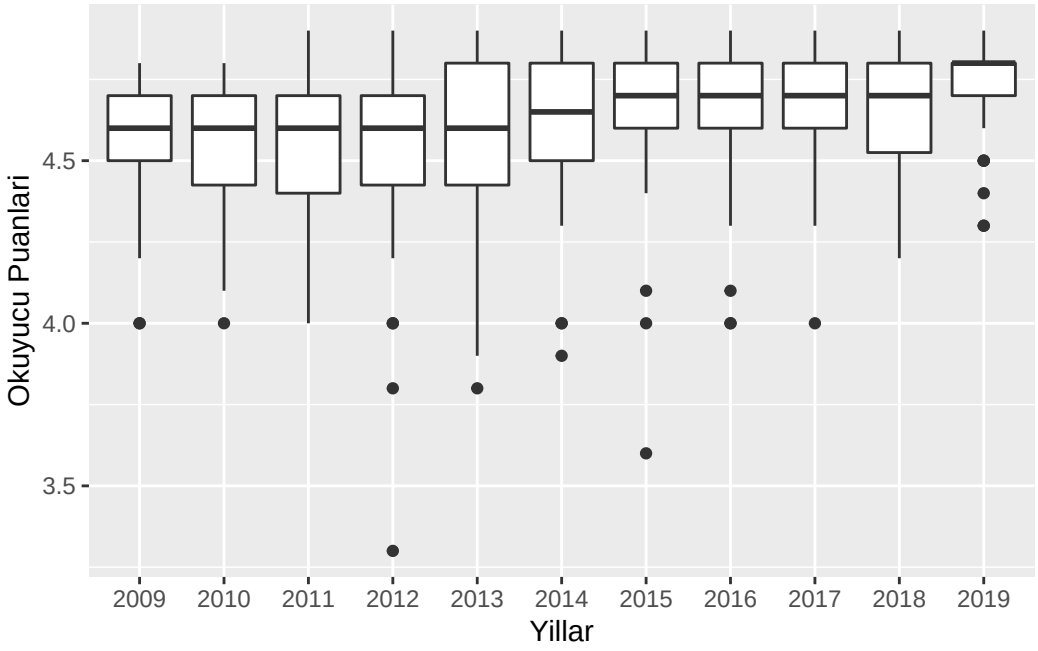
```
ggplot(bestSellersOfYears, aes(x = as.factor(Year), y = Price))+geom_col() + scale_y_continuous(lim
```



Yukarıdaki grafikte ise fiyatların 2010-2012 ve 2015-2018 aralığında genel olarak sabit olduğunu, fiyatların 2013-2014 yıllarında 105 ile zirve yaptığını söyleyebiliriz.

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

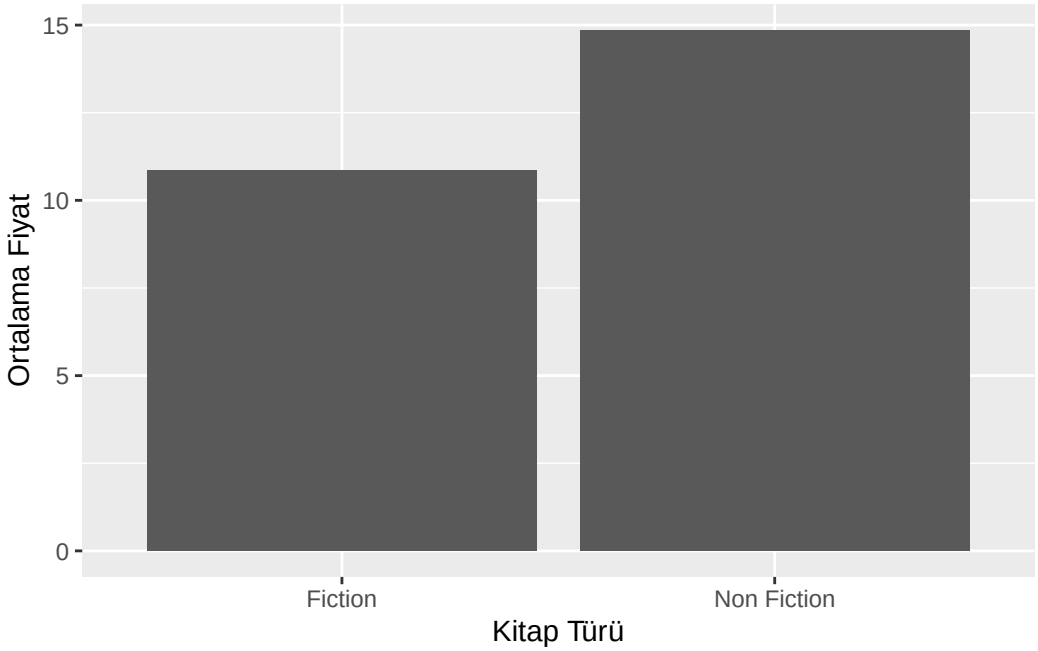
```
ggplot(bestsellers, aes(x = as.factor(Year), y = User.Rating)) +geom_boxplot(show.legend = FALSE) +
```



Yukarıdaki grafikte görüleceği üzere, okuyucular tarafından en çok puan değişimi (kutu uzunluğu) 2013 yılında elde edilmiştir, kutuların çizgileri ötesinde bulunan noktalar diğer oylara nazaran çok daha düşük veya yüksek olan oylar olmaktadır, çizgilerin sonları maksimum ve minimum oy limitlerini gösterir, ötesindeki oylar beklenenin dışındadır. Ayrıca bu grafikte genel olarak okuyucuların düşükten ziyade daha yüksek puanlar vermeye eğilimli olmuş oldukları çıkarımı yapılabilir (Kutuların ortasındaki çizgilerin yaklaştığı taraflar).

3.2.3 Türüne göre kitap fiyatlarının dağılımını araştırınız.

```
bestSellersOfYearsByGroups <- bestsellers %>%
  group_by(Genre) %>%
  summarise(Price=mean(Price))

ggplot(bestSellersOfYearsByGroups, aes(x = Genre, y = Price)) + geom_col(show.legend = FALSE) + lab
```



Yukarıdaki grafikten yapabileceğimiz çıkarım ortalama olarak Roman türündeki kitapların Roman türünde olmayan kitaplardan daha ucuz olduğudur.