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

Sercan Öncü

18.11.2022

ÖZET

Bu raporda Breaking Bad dizisi, En çok satan kitaplar ve Veri bilimci maaşları veri setleri incelenmiştir. Breaking Bad dizisi sezonlara göre bölüm süresi, reyting punları ve izlenme sayısı dağılımları görselleştirilmiştir. En çok satan kitaplar yıllarına ve türlerine göre okuyu puanları ve fiyat dağılımları görselleştirilmiştir. Veri bilimci maaşları firma büyüklüğü, deneyim ve uzaktan çalışma yüzdesine göre yıllık ve aylık maaşları görselleştirilmiştir.

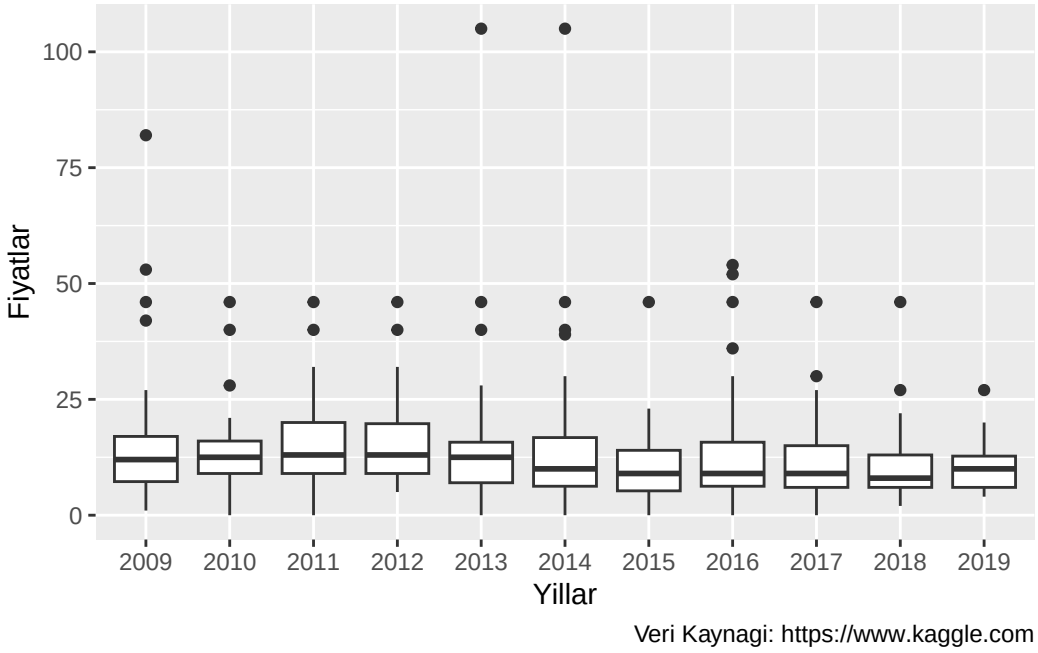
```r
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("palmerpenguins")
install.packages("dplyr")
install.packages("DALEX")
install.packages("gridExtra")
install.packages("ggforce")
install.packages("ggridges")
install.packages("reshape2")
install.packages("ggthemes")
install.packages("readr")
install.packages("readxl")
install.packages("forcats")
library(forcats)
library(reshape2)
library(ggplot2)
library(tidyverse)
library(palmerpenguins)
library(dplyr)
library(DALEX)
library(gridExtra)
library(ggforce)
library(ggridges)
library(readr)
library(readxl)
```

1. En çok satan kitaplar

Bu veri seti kullanıcı puanları, inceleme sayıları, fiyat, yıl, tür vb. başlıklar altında En çok satan kitaplar için veriler içermektedir.

1.1 Yıllara göre en çok satan kitapların fiyat dağılımı

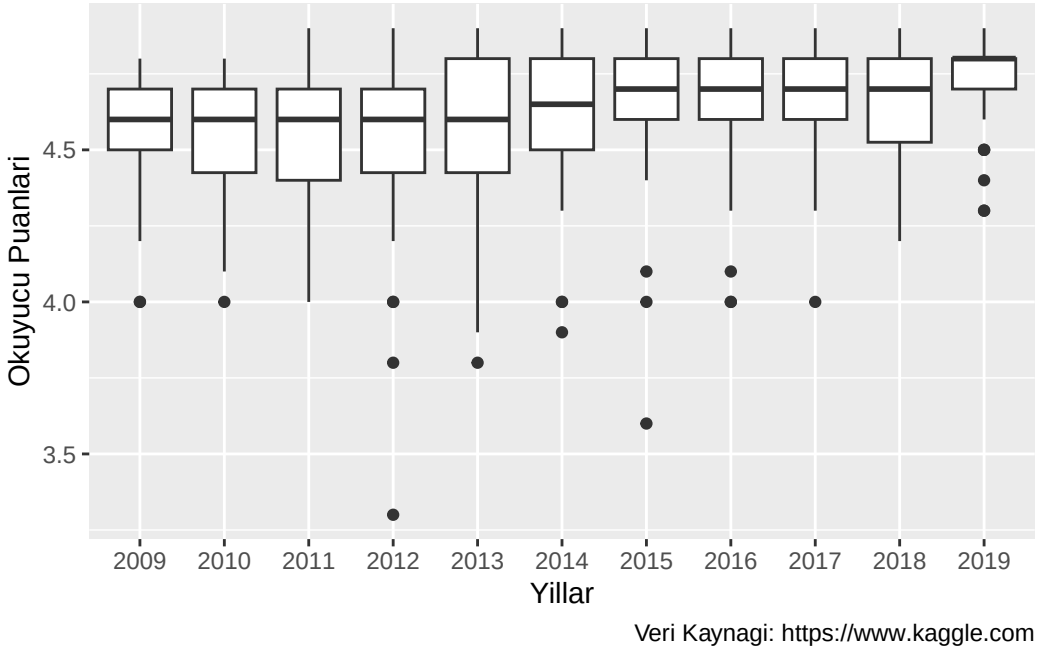
```
bestsellers_with_categories <- read_excel("bestsellers-with-categories.xlsx")
best <- bestsellers_with_categories
bestsellers_with_categories %>%
    ggplot(aes(x = as.factor(Year), y = Price))  +
    geom_boxplot() +
    labs(x="Yillar",
         y= "Fiyatlar",
       caption = "Veri Kaynagi: https://www.kaggle.com")
```



Yukarıdaki grafikte yıllara göre en çok satan kitapların fiyat dağılımı incelenmiştir. Yıllara göre kitap fiyatlarındaki en düşük 2019 yılında olduğu görülmüştür. En yüksek değişim ise 2011 yılında görülmüştür.

1.2 Yıllara göre en çok satan kitapların aldıkları okuyucu puanı dağılımı

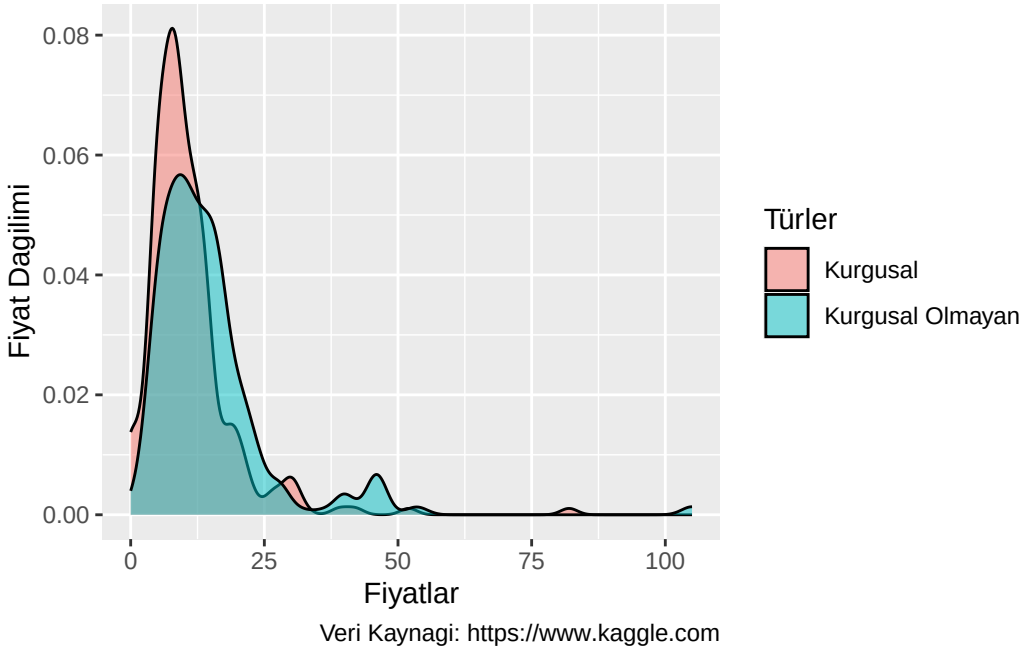
```
bestsellers_with_categories %>%
    ggplot(aes(x = as.factor(Year), y = `User Rating`))  +
    geom_boxplot() +
    labs(x="Yillar",
         y= "Okuyucu Puanlari",
       caption = "Veri Kaynagi: https://www.kaggle.com")
```



Yukarıdaki grafikte yıllara göre en çok satan kitapların aldıkları okuyucu puanı dağılımı incelenmiştir. Okuyu puanları yıllar ilerledikçe yükselmiştir ve en yüksek okuyu puanı 2019 yılında olduğu görülmüştür. 2013 yılında değişim en yüksek iken 2019 yılında değişim en azdır.

1.3 Türüne göre kitap fiyatlarının dağılımı

```
bestsellers_with_categories %>%
    ggplot(aes(fill = Genre, x = Price))  +
    geom_density(alpha = 0.5)+
    labs(x="Fiyatlar",
         y= "Fiyat Dagilimi",
         fill = "Türler",
       caption = "Veri Kaynagi: https://www.kaggle.com")+
  scale_fill_discrete(labels= c("Kurgusal","Kurgusal Olmayan"))
```



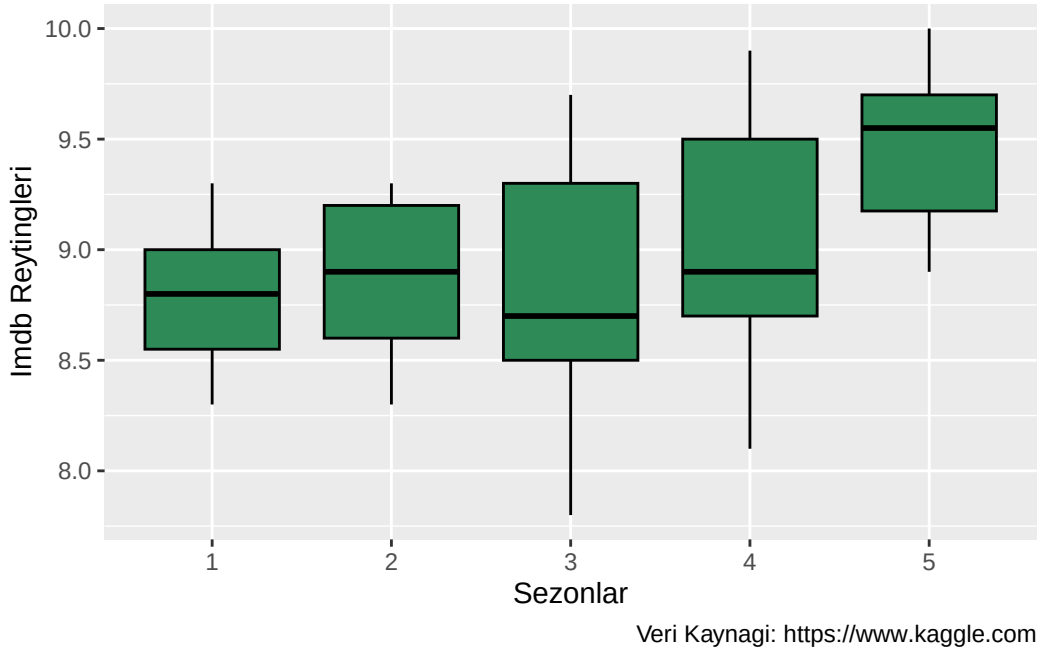
Yukarıdaki grafikte türüne göre kitap fiyatlarının dağılımı görselleştirilmiştir. Kurgusal olmayan türdeki kitapların fiyatının ortalmasının kurgusal türdeki kitapların fiyatından yüksek olduğu görülmüştür. Her iki türde de dağılım sağa çarpıktır. Her iki türde de fiyatlar 13-14 civarında çoğunluktadır.

2. Breaking Bad

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

2.1 Sezonlara göre reyting dağılımları

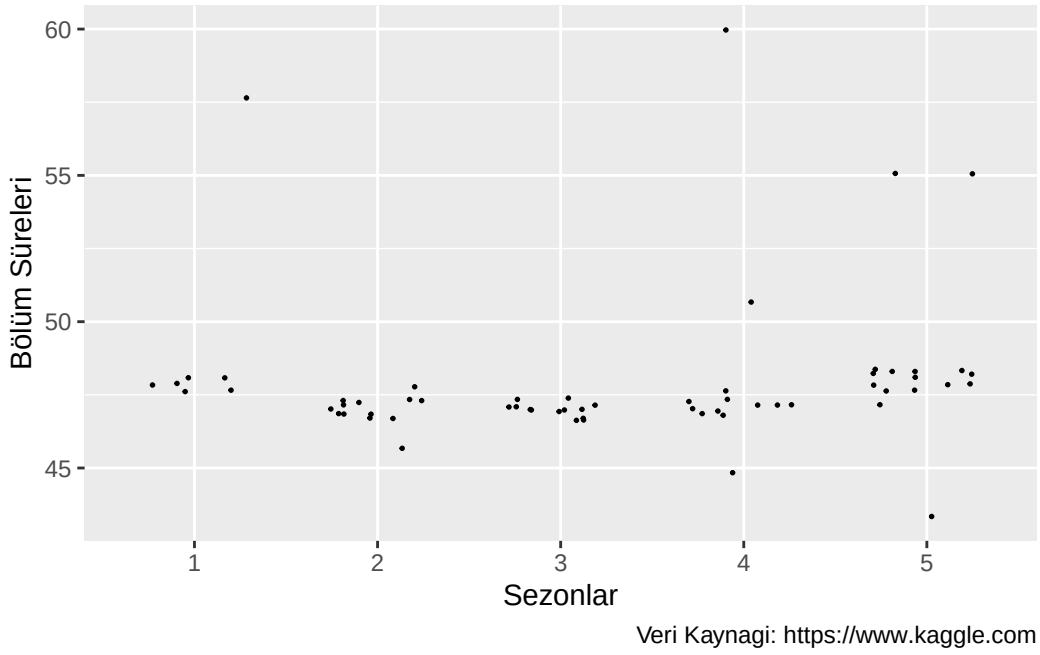
```
breaking_bad <- read_excel("breaking_bad.xlsx")
breaking_bad$Season <- factor(breaking_bad$Season, ordered = TRUE, levels = c("1", "2", "3
breaking_bad$Episode <- factor(breaking_bad$Episode, ordered = TRUE)
breaking_bad %>%
    ggplot(aes(x = Season, y = Rating_IMDB))  +
    geom_boxplot(color = "black", fill = "seagreen" ) +
   labs(x="Sezonlar",
         y= "Imdb Reytingleri",
       caption = "Veri Kaynagi: https://www.kaggle.com")
```



Yukarıdaki grafikte sezonlara göre rating dağılımları incelenmiştir. Bölümlerin imdb puanları sezonlar ilerledikçe yükselmiştir. Bölümlerin imdb puanlarındaki değişim en yüksek 3. sezonda ve en düşük değişim ise 1. sezondadır.

2.2 Sezonlara göre bölüm süresi dağılımları

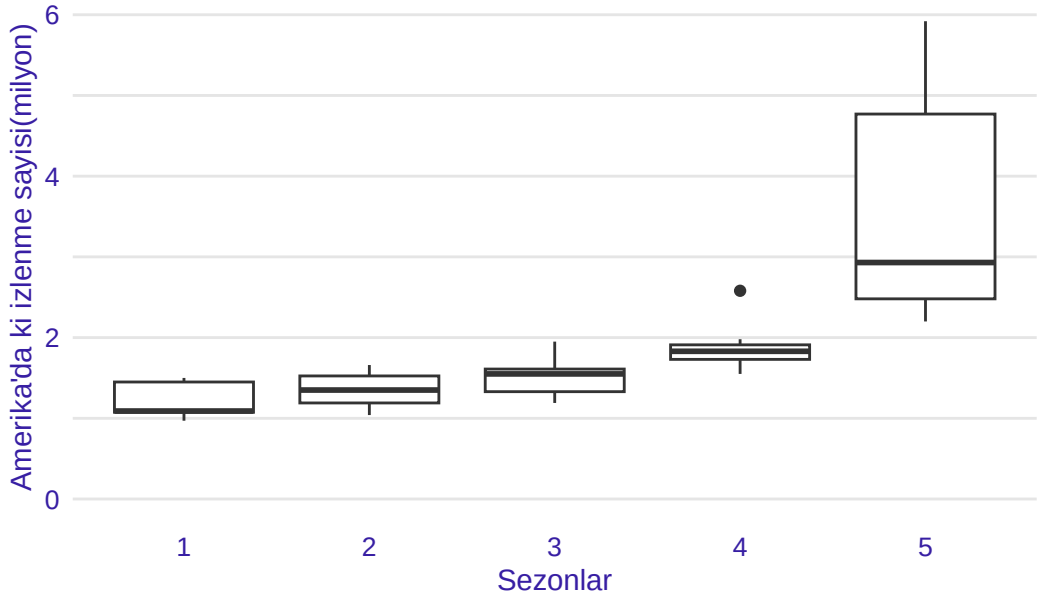
```
breaking_bad %>%
    ggplot(aes(x = Season,  y = Duration_mins))  +
    geom_jitter(cex = 0.3, width = 0.3, stat = "identity") +
    labs(x="Sezonlar",
         y= "Bölüm Süreleri",
       caption = "Veri Kaynagi: https://www.kaggle.com")
```



Yukarıdaki grafikte sezonlara göre bölüm süresi dağılımları incelenmiştir. Bölüm sürelerindeki en düşük değişim 3. sezonda ve en yüksek değişim 4. sezondadır. En uzun bölüm 4. sezonda iken en kısa bölüm 5.sezondadır. Bölümler çoğunlukla yaklaşık 45 dakikadır.

2.3 Sezonlara göre izlenme sayısı dağılımları

```
breaking_bad %>%
    ggplot(aes(x = Season, y = as.numeric(`U.S. viewers_million`)))  +
    geom_boxplot() +
    labs(x="Sezonlar",
         y= "Amerika'da ki izlenme sayisi(milyon)",
       caption = "Veri Kaynagi: https://www.kaggle.com")+
  ylim(0,6)+
    theme_drwhy()
```



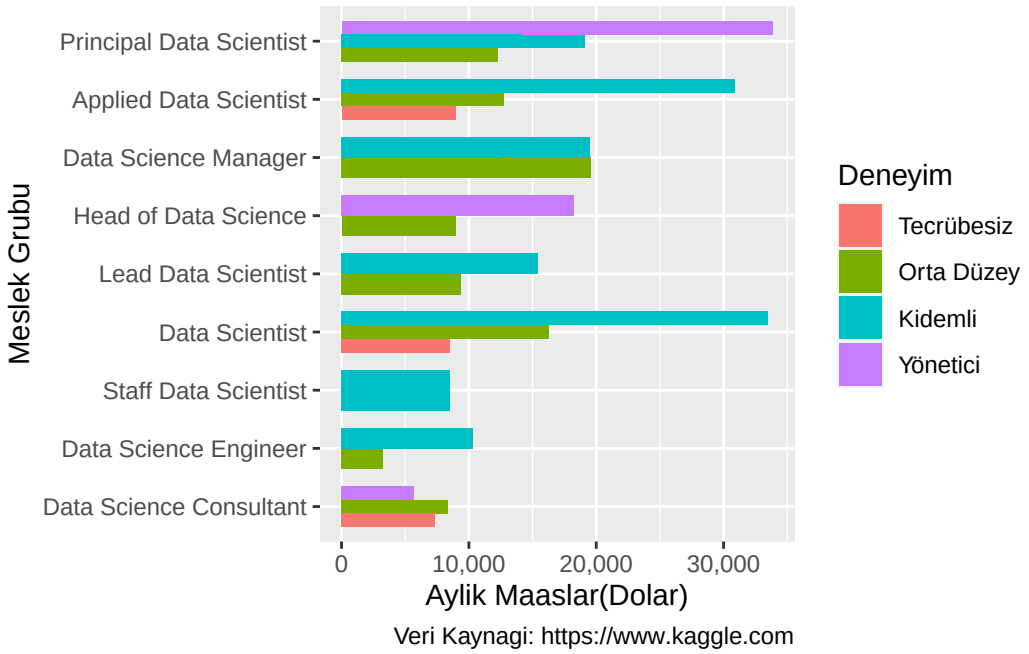
Yukarıdaki grafikte sezonlara göre izlenme sayısı dağılımları incelenmiştir. En az izlenme ilk sezonda iken en yüksek izlenme 5. sezondadır. En düşük değişim 4. sezonda ve en yüksek değişim 5. sezonda olmuştur. Bölümlerin izlenmesi sezonlar ilerledikçe artmıştır.

3. Veri bilimci maaşları

Bu veri seti pozisyon, deneyim, maaş, şirket büyüklüğü vb. başlıklar altında Veri bilimci maaşları için veriler içermektedir.

3.1 Tecrübe düzeylerine göre veri bilimi pozisyonları aylık maaşlarının dolar karşılığı bazında dağılımı

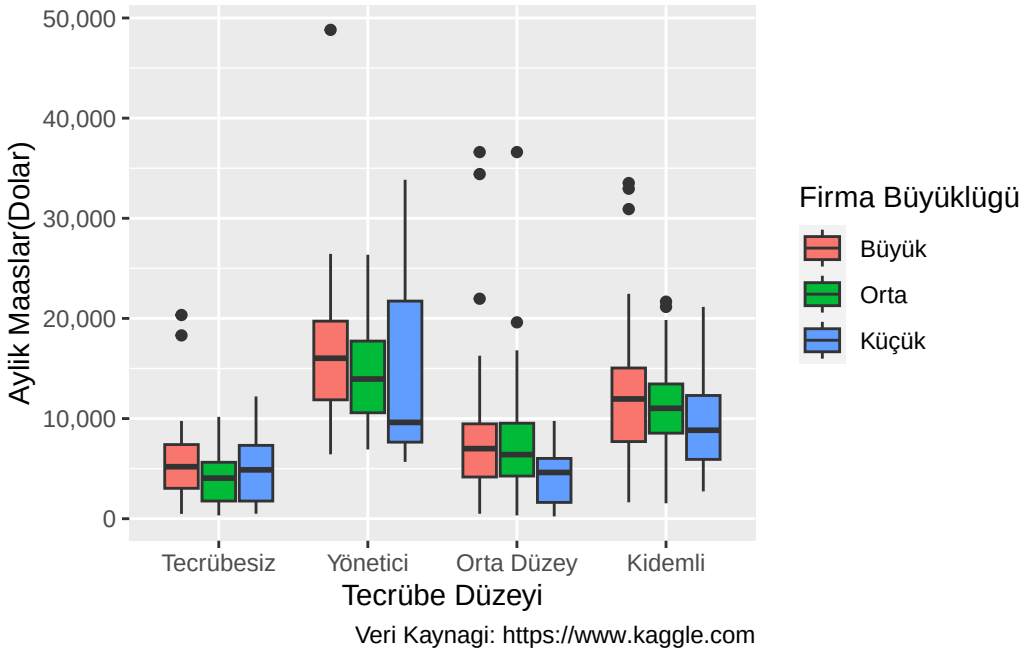
```
Data_Science_Fields_Salary_Categorization <- read_excel("Data_Science_Fields_Salary_Catego
data3 <- read_excel("Data_Science_Fields_Salary_Categorization.xlsx")
Data_Science_Fields_Salary_Categorization$Experience <- factor(Data_Science_Fields_Salary_
data2 <- filter(Data_Science_Fields_Salary_Categorization,Designation=='Data Scientist'| D
Director of Data Science'| Designation=='Head of Data Science'| Designation=='Principal Da
salary5 <- gsub(","," " , data2$Salary_In_Rupees)
salary6 <- gsub(","," " , data3$Salary_In_Rupees)
data2 <- data2%>%
  group_by(Experience)
ggplot(data2, aes(x = reorder(Designation, +as.numeric(salary5) ) , y = as.numeric(salary5
    geom_bar(stat = "identity", position = "dodge", width = 0.7)+
   coord_flip()+
   scale_y_continuous(labels = scales::comma)+
    labs(x="Meslek Grubu",
         y= "Aylik Maaslar(Dolar)",
         fill="Deneyim",
       caption = "Veri Kaynagi: https://www.kaggle.com")+
   scale_fill_discrete(labels= c("Tecrübesiz","Orta Düzey", "Kidemli","Yönetici" ))
```



Yukarıdaki grafikte tecrübe düzeylerine göre veri bilimi pozisyonları aylık maaşlarının dolar karşılığı bazında dağılımı incelenmiştir. En yüksek maaş "Yönetici" deneyimine sahip "Principal Data Scientist" pozisyonundadır ve en düşük maaş "Orta Düzey" deneyimine sahip "Data Science Engineer" pozisyonundadır. Aynı pozisyon içerisindeki deneyime göre alınan maaşta en düşük değişim "Data Sciennce Manager" pozisyonundadır ve en yükseği "Data Scientist" pozisyonundadır.

3.2 Firma büyüklüğüne ve tecrübe düzeyine göre aylık maaş tutarı dağılımları

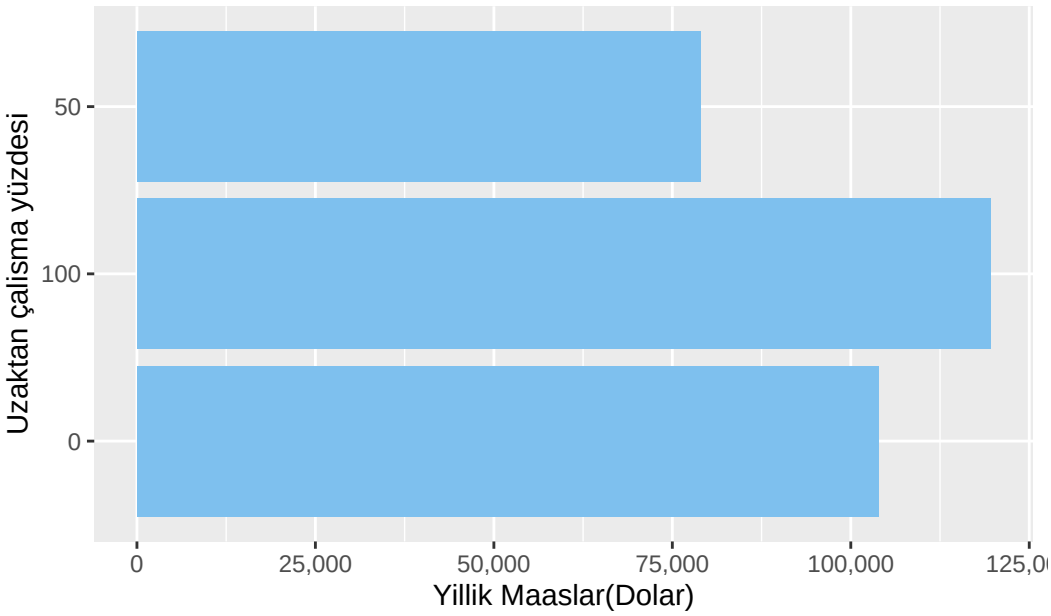
```
 data3 <- data3%>%
  group_by(Experience)
ggplot(data3, aes(x = Experience , y = as.numeric(salary6)/978, fill = Company_Size ))  +
    geom_boxplot()+
   scale_y_continuous(labels = scales::comma) +
    labs(x="Tecrübe Düzeyi",
         y= "Aylik Maaslar(Dolar)",
         fill = "Firma Büyüklügü",
       caption = "Veri Kaynagi: https://www.kaggle.com")+
   scale_fill_discrete(labels= c("Büyük","Orta", "Küçük"))+
  scale_x_discrete(labels = c("Tecrübesiz","Yönetici", "Orta Düzey", "Kidemli" ))
```



Yukarıdaki grafikte firma büyüklüğüne ve tecrübe düzeyine göre aylık maaş tutarı dağılımları incelenmiştir. En yüksek maaşlar genelde "Yönetici" deneyime sahip büyük şirketlerde çalışan kişilerdedir. "Orta Düzey" deneyime sahip kişiler en düşük maaşları küçük firmalarda almaktadır. En yüksek değişim "Yönetici" deneyime sahip küçük firma çalışanlarındadır ve en düşüğü ise "Tecrübesiz" deneyime sahip olup büyük firmalarda çalışan kişilerdedir. Bu grafiğe göre tecrübesiz kişilerin küçük şirketleri tercih etmesi maaş bakımından mantıklı bir tercih olur.

3.3 Uzaktan çalışma yüzdesi sistemine göre ortalama yıllık maaş ücreti dağılımları

```
data10 <- filter(Data_Science_Fields_Salary_Categorization,Remote_Working_Ratio=='0')
salary11 <- gsub("," ,"" , data10$Salary_In_Rupees)
salary12 <- as.double(salary11)
data11 <- filter(Data_Science_Fields_Salary_Categorization,Remote_Working_Ratio=='50')
salary15 <- gsub("," ,"" , data11$Salary_In_Rupees)
salary16 <- as.double(salary15)
data21 <- filter(Data_Science_Fields_Salary_Categorization,Remote_Working_Ratio=='100')
salary25 <- gsub("," ,"" , data21$Salary_In_Rupees)
salary26 <- as.double(salary25)
emp.data <- data.frame(
   emp_id = c(mean(salary12), mean(salary16), mean(salary26)),
   Remote_Working_Ratio3 = c("0","50","100"),
   stringsAsFactors = FALSE
)
ggplot(emp.data, aes(y=emp_id/81.5, x=as.factor(Remote_Working_Ratio3)))+
  geom_col(fill = "skyblue2")+
  coord_flip()+
  scale_y_continuous(labels = scales::comma)+
  labs(x="Uzaktan çalisma yüzdesi",
         y= "Yillik Maaslar(Dolar)",
       caption = "Veri Kaynagi: https://www.kaggle.com")
```



Yukarıdaki grafikte uzaktan çalışma yüzdesi sistemine göre yıllık maaş ücreti dağılımları incelenmiştir. Tamamen uzaktan çalışan kişiler diğerine ortalamaya göre daha yüksek maaş almaktadır ve en düşük maaş ortalaması %50 uzaktan çalışan kişilerindir.