Farklı Veri Setlerinin Dağılımlarının Görselleştirilmesi

Koray Demir

```
library(readxl)
library(DALEX)
library(dplyr)
library(gridExtra)
library(ggplot2)
library(ggridges)
library(tidyverse)
```

```
bb <- read_xlsx("bb.xlsx")
ydat <- read_xlsx("ydata.xlsx")
ds <- read_xlsx("Datasci.xlsx")
```

Özet

Bu rapor için kullanılan kütüphaneler yukarıdaki gibidir. Bu ödevde Breaking Bad ve Veri Bilimci maaş veri setleri kullanılmıştır. Breaking Bad veri setinde sezonlara göre bölüm süresi, rating dağılımı ve izlenme sayılarının dağılımları görselleştirilmiştir. Veri bilimci maaş veri setinde ise tecrübe ve firma büyüklüğü ve uzaktan çalışma yüzdesine göre veri bilimcilerin maaşlarının dağılımları görselleştirilmiştir.

1) Breaking Bad

Breaking bad veriseti 5 sezon toplam 62 bölümden oluşmaktadır. Bu veriseti tarih, sezon, bölüm, başlık, yazar, yönetmen , bölüm uzunluğu, özet, imdb puanı, ve toplam izleyici sayısı değişkenlerinden oluşmaktadır.

1.1)Sezonlara göre rating dağılımları

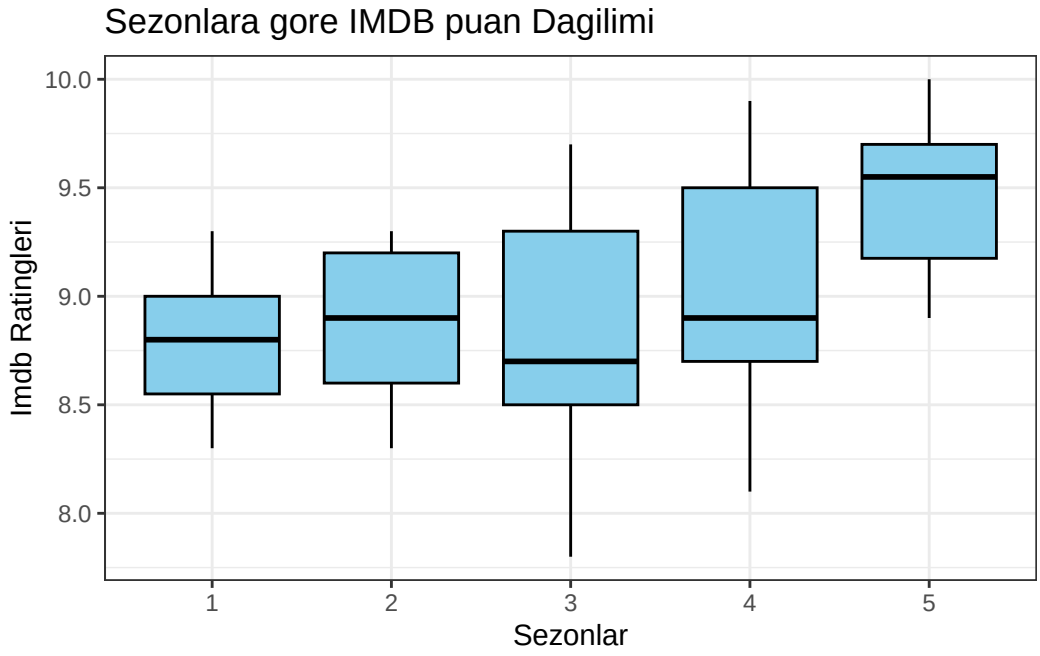
Burada Sezonlar Sıralanarak sezonlara ait imdb puan dağılımları görselleştirilmiştir

```
library(readr)
library(readxl)

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


bb %>%
    ggplot(aes(x = Season, y = Rating_IMDB))  +
  geom_boxplot(color = "black", fill = "skyblue") +

   labs(x="Sezonlar",
         y= "Imdb Ratingleri",
       title = "Sezonlara gore IMDB puan Dagilimi"
       ) + theme_bw()
```



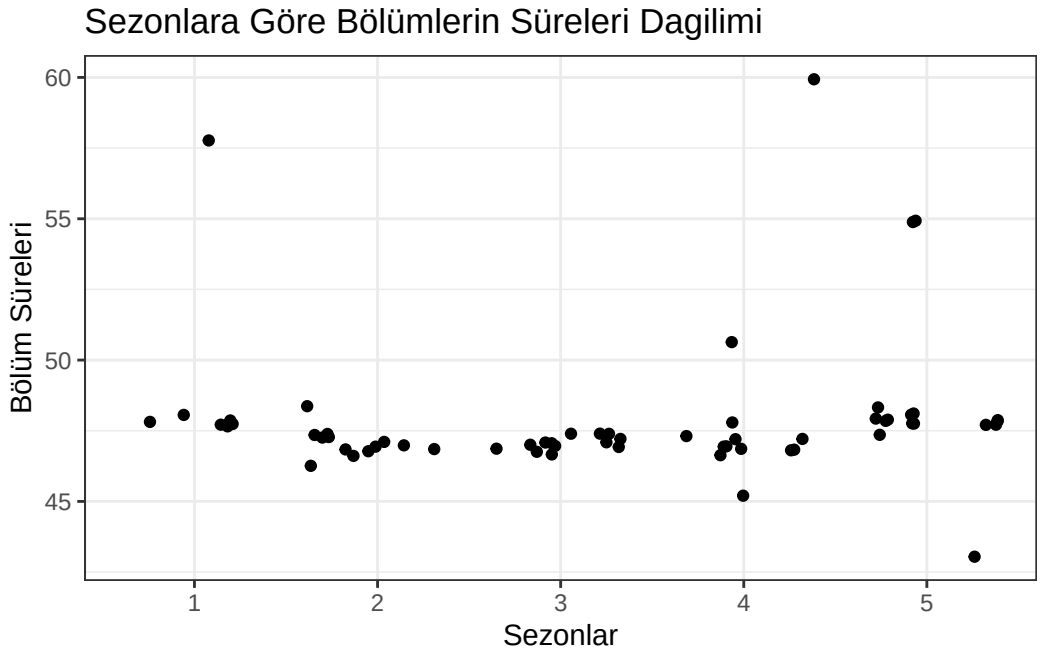
YORUM

Yukarıda ki Grafiğe göre rating değişiminin en düşük olduğu sezon 1. sezon olarak gozlenmiştir ayrıca 1 ve 2. sezonun değişimlerinin de birbirine yakınlığı göze çarpmıştır.

1.2)Sezonlara Göre Bölüm Süreleri Dağılımı

Bu grafikte ise sezonlar bölüm sürelerine göre dağılımı nokta grafik ile görselleştirmek istenmiştir.

```
bb %>%
    ggplot(aes(x = Season, y = Duration_mins))  +
    geom_jitter() +
    labs(x="Sezonlar",
         y= "Bölüm Süreleri",
       title = "Sezonlara Göre Bölümlerin Süreleri Dagilimi")+
  theme_bw()
```



YORUM

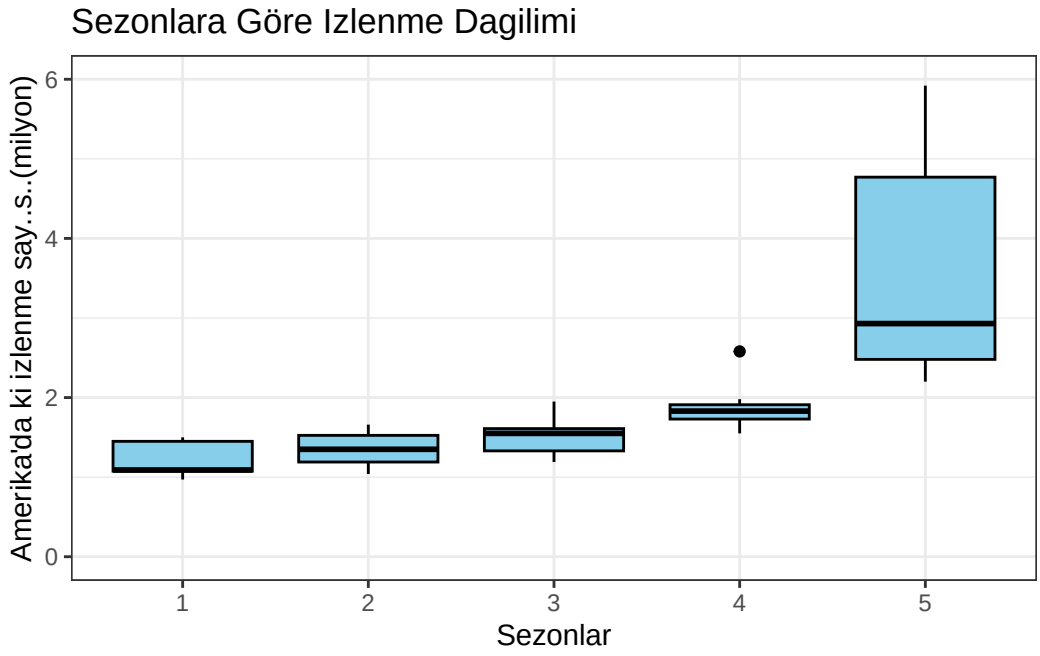
Yukarıda ki bölüm süresi dağılımına göre en yüksek değişimin gözlemlendiği sezon 4. sezon olarak gözlemlenmiştir.

1.3) Sezonlara Göre İzlenme Sayısı Dağılımı

Bu grafikte ise sezonlara göre milyon bazında izlenme sayıları görselleştirilmiştir.Bu görselleştirme için boxplot kullanılmıştır.

```
bb%>%
    ggplot(aes(x = Season, y = as.numeric(bb$`U.S. viewers_million`)))  +
    geom_boxplot(color="black" , fill= "skyblue") +
    labs(x="Sezonlar",
         y= "Amerika'da ki izlenme sayısı(milyon)",
       title = "Sezonlara Göre Izlenme Dagilimi") +

  ylim(0,6)+
    theme_bw()
```



YORUM

Yukarıdaki Görselleştirmeye göre izlenme sayısı değişiminin en düşük olduğu sezon 4. sezon olarak gözlemlenmiştir.

2) Veri Bilimci Maaşları

Veri bilimci maaşları veri seti: çalıştığı yıl, pozisyonu, çalışma durumu, maaş, çalışan lokasyonu, şirket lokasyonu, şirket büyüklüğü ve son olarak uzaktan çalışma yüzdesi değişkenlerinden oluşmaktadır.

2.1) Tecrübe düzeylerine göre veri bilimi pozisyonları aylık maaşları

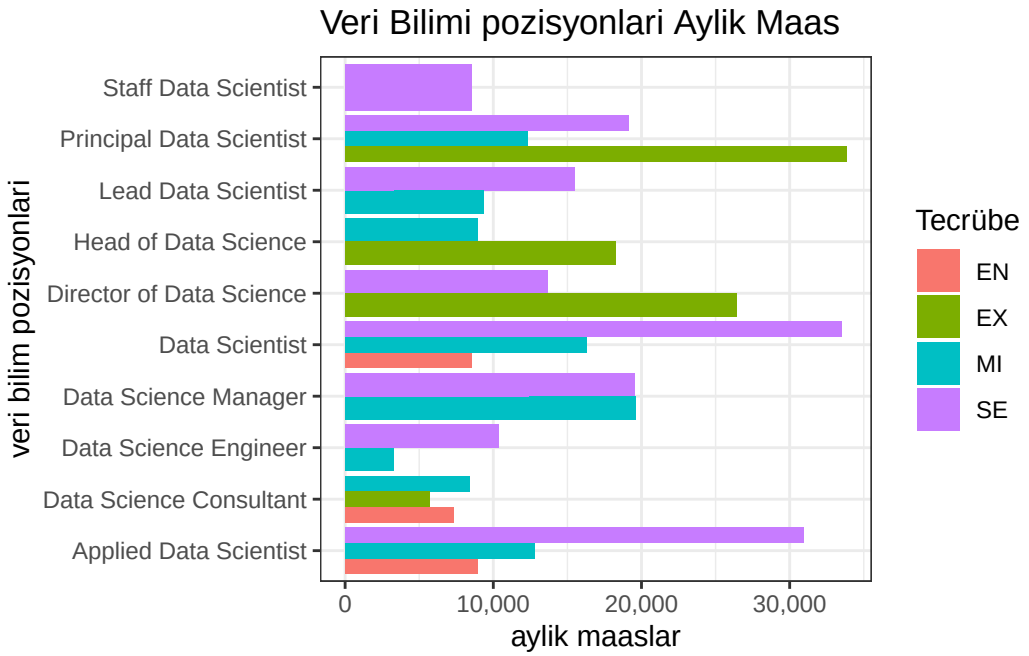
Aşağıdaki Görselde veri bilimiyle ilgili olan meslekler filtrelenmiş ve tecrübe kategorilerine göre aylık maaş miktarları gözlemlenmiştir.

```
ggplot(ydat , aes(x = ydat$Designation ,
                  y = ydat$Salary_In_Dollars/12 ,fill=ydat$Experience)) +
  geom_bar(stat = "identity" , position = "dodge") +
  coord_flip() +
```

```
 scale_y_continuous(labels = scales::comma)  +


labs( x = "veri bilim pozisyonlari" ,
      y = "aylik maaslar",
      title = "Veri Bilimi pozisyonlari Aylik Maas",
      fill= "Tecrübe") +

theme_bw()
```



YORUM

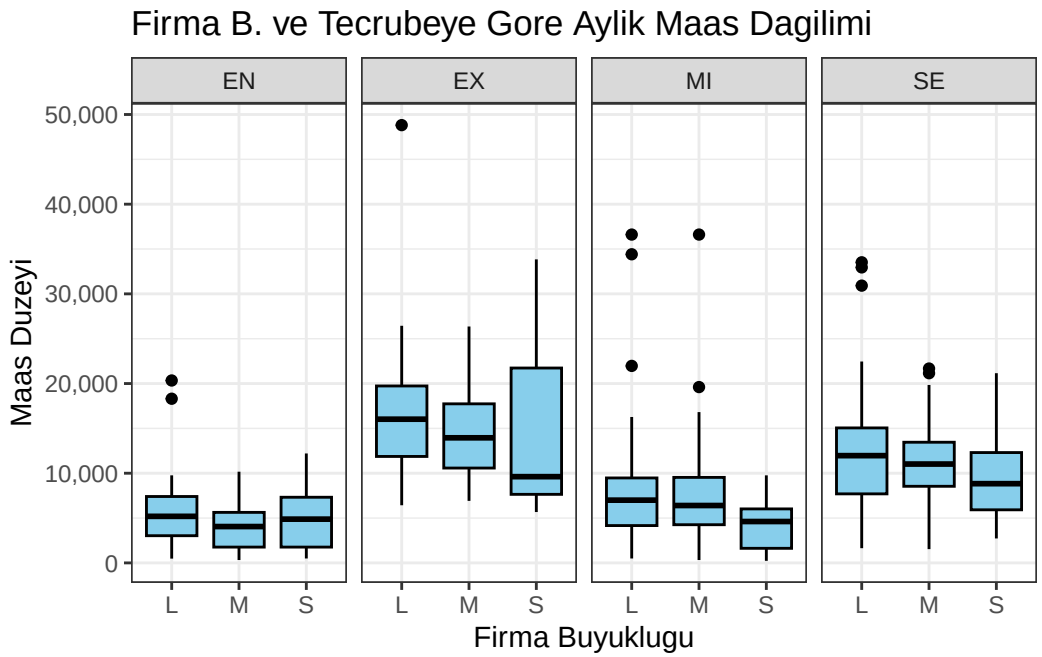
Bu görsele bakılarak yonetici(`EX`) kişiler için en yüksek maaş `principal data scienist` pozisyonun da gözlemlenmiştir. Buna ek olarak ilginç şekilde `Staff Data Scienist` pozisyonunda sadece `SE`ler gözlemlenmiştir. En tecrübesiz adayların(`EN`) 10 pozisyondan sadece 3'ünde yer aldığı gözlemleniştir.Bu da veri bilimi alanında tecrübesiz adaylara pek fazla yer verilmediği anlamına gelmektedir.Son olarak da en düşük maaşı `Data Science Engineer` pozisyonun da orta seviye adayların aldığı gözlemlenmiştir.

2.2) Firma büyüklüğüne ve tecrübe düzeyine göre aylık maaş tutar dağılımları

Bu veri setinde ise firma büyüklüğü ve tecrübeye göre maaş miktarları ayrı ayrı boxplotlar kullanılarak görselleştirilmiştir.

EN: entry , EX:executive , MI: mıddle, SE: Senior

```
ggplot(ds , aes(x = as.factor(ds$Company_Size) , y=ds$Salary_In_Rupees/978)) +
  scale_y_continuous(labels = scales::comma) +
  labs(
    title = "Firma B. ve Tecrubeye Gore Aylik Maas Dagilimi",
    x ="Firma Buyuklugu" ,
    y ="Maas Duzeyi"
  ) +
  geom_boxplot(color="black" , fill= "skyblue") +
  facet_grid(~ds$Experience) + theme_bw()
```



YORUM

Bu grafiğe bakılarak tecrübesiz adaylar için (`EN`) işe başlamak adına en uygun olan firma ölçeği `L` olarak gözlemlenmiştir.Bunun sebebi ise tecrübesiz aday için değişimin en az olduğu şirket ölçeği `L` olarak gözlemlenmesidir.

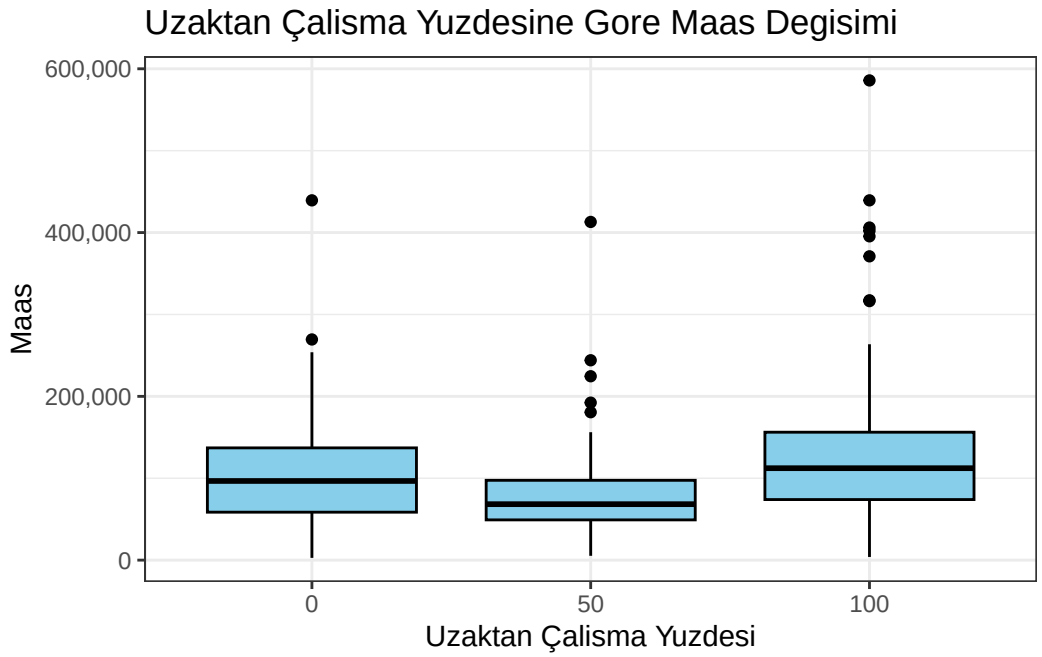
2.3) Uzaktan çalışma yüzdesi sistemine göre yıllık maaş ücreti dağılımları

Bu görselde Calısma yuzdeleri 3 ayrı kategoriye ayrılıp, bu kategorilere göre olan maaş dağılımları anlatılmak istenmiştir.

```
ds$Remote_Working_Ratio <- factor(ds$Remote_Working_Ratio, ordered = TRUE,
                                          levels = c("0","50","100"))

ggplot(ds, aes(x=ds$Remote_Working_Ratio , y =ds$Salary_In_Rupees/81.5)) +
  labs(title="Uzaktan Çalisma Yuzdesine Gore Maas Degisimi",
       x ="Uzaktan Çalisma Yuzdesi" ,
       y= " Maas ") +
  scale_y_continuous(labels = scales::comma) +

geom_boxplot(color = "black", fill = "skyBlue" ) +
  theme_bw()
```



YORUM

Yukarıdaki Grafikte uzakdan çalışma yüzdesine göre değişimin en az oldugu yuzdelik %50 olarak gözlemlenmiştir.Buna karşıt olarak %100 uzaktan çalışan bireyinde en fazla maaş alıdıgını, bunun yanında ise değişimin en yüksek %100 çalışma oranı olduğu gösterilmiştir.