Veri Setlerinde Oranlarının Görselleştirilmesi

Koray Demir

11/26/22

ÖZET

Bu rapor için kullanılan kütüphaneler aşağıdaki gibidir. Bu ödevde Fifa 23, Breaking Bad ve Veri bilimci veri setleri kullanılmıştır. Raporun devamında bu 3 veri setiyle ilgili çeşitli oranlar aşağıdaki gibi görselleştirilmiştir.

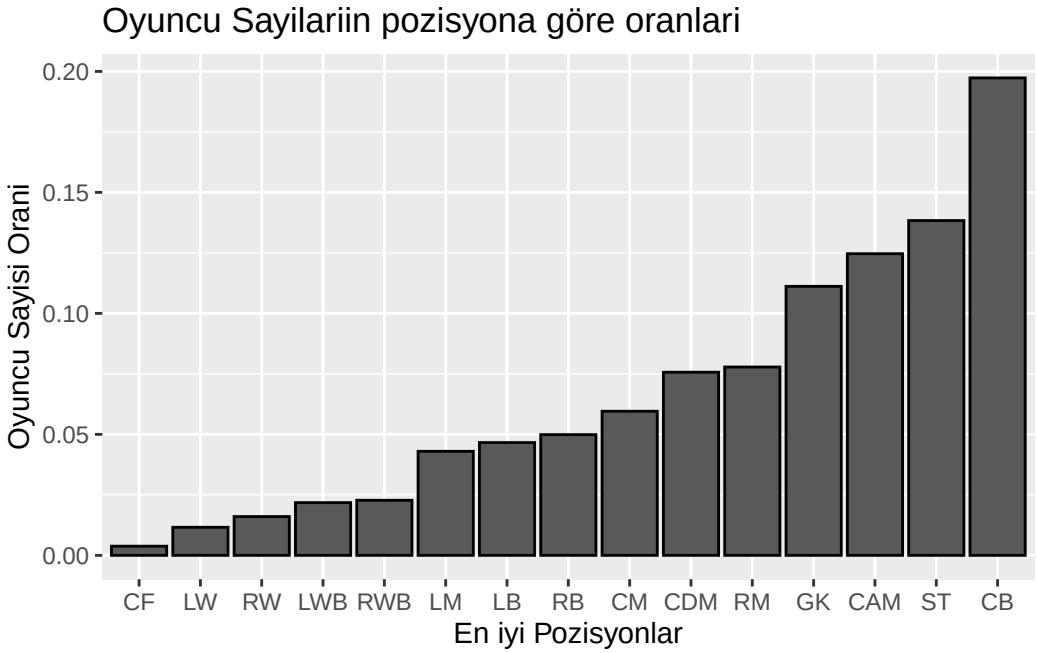
```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(MetBrewer)
library(treemapify)
library(tidyr)

fifa_23 <- read_csv("fifa_23.csv")
breaking_bad <- read_csv("breaking_bad.csv")
Datasci <- read_csv("Datasci.csv")
```

1) FIFA 23 Oyuncuları

Bu veri setinde fifa 23 isimli futbol oynunda bulunan toplam 18539 oyuncunun: oyuncu değerleri, maaşı, en iyi posizyonları, kullandıkları ayakları gibi toplam 89 farklı değişken ile birlikte gözlemlenen verileri bulunmaktadır.

1.1) Oyuncu sayılarının, en iyi oynadıkları pozisyona göre oransal görselleştirmesi

```
ofifa <- fifa_23 %>%
  group_by(poz =`Best Position`) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))
ggplot(ofifa, aes(x = reorder(ofifa$poz, +ofifa$oran), y = ofifa$oran  )) +
   geom_bar(stat = "identity" , position = "dodge" , color="black"  )+
  labs( x= "En iyi Pozisyonlar" ,
        y = "Oyuncu Sayısı Oranı",
        title="Oyuncu Sayılarıın pozisyona göre oranları") +
  theme(legend.position = "none" )
```



GK:Kaleci

LB:Solbek CB:stoper RB:Sağ bek LWB:Sol kanat bek RWB:Sağ kanat bek CDM:Defansif Orta saha

LM:Sol Orta saha CM:Orta saha RM:Sağ OrtaSaha CDM:Ofansif Orta saha

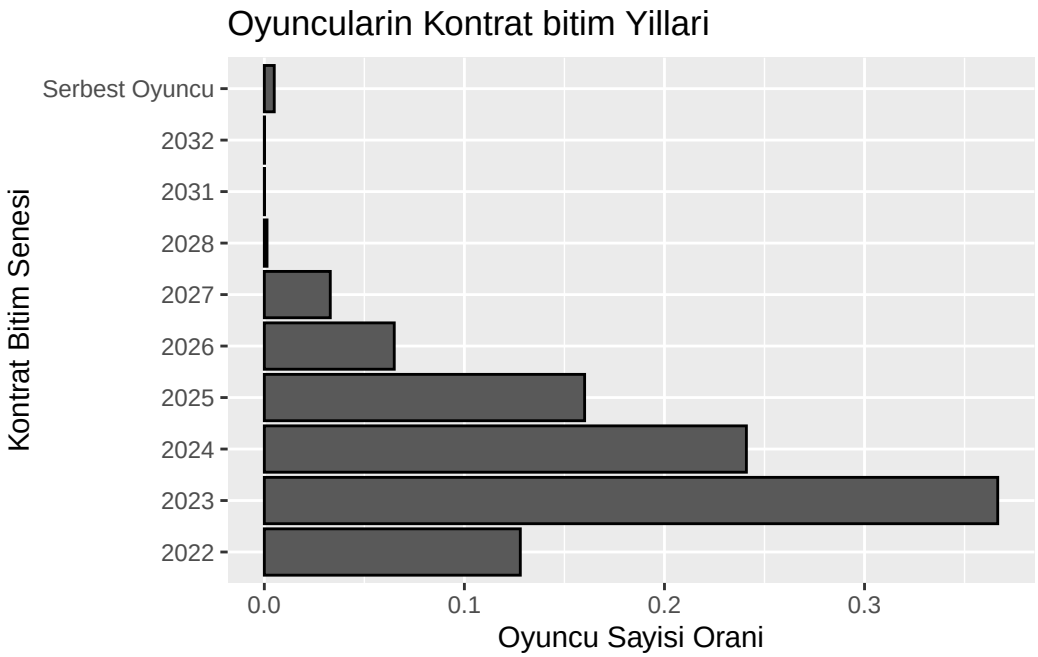
LW:Sol Kanat RW:Sağ Kanat CF:Santrfor ST:Pivot Santrfor

YORUM

Yukarıdaki grafikte görüldüğü gibi en çok oyuncu stoper mevkisinde bulunmaktadır.Bu oran tüm oyuncular içinde neredeyse $\frac{1}{5}$ oranındadır. En az oyuncu oranı ise görüldüğü gibi santrfor mevkisindedir.

1.2) Oyuncu sayılarının, kontrat sürelerine göre oranlarının görselleştirmesi

```
fifa_23 <- fifa_23 %>%
  add_column(kontrat=if_else(fifa_23$`Contract Until` =="-", "Serbest Oyuncu",
                             fifa_23$`Contract Until`))


fifa1 <- fifa_23 %>%

  group_by(kontrat) %>%
  dplyr::summarize(sayi = n())%>%
mutate(oran = sayi / sum(sayi))



ggplot(fifa1, aes(x = kontrat, y = fifa1$oran )) +

  geom_bar(stat = "identity" , position = "dodge" , color="black" ) +

  labs( x= "Kontrat Bitim Senesi" ,
        y = "Oyuncu Sayisi Oranı",
        title ="Oyuncuların Kontrat bitim Yılları" ) +
 theme(legend.position = "none") +
   coord_flip()
```



YORUM

Yukarıda ki grafiğe bakılarak en çok 2023 yılında futbolcuların sözleşmesinin biteceği gözükmektedir.Bunun Sebebi ise genellikle kulüplerin daha az risk almak için oyuncular ile 1 senelik sözleşme imzalaması olabilir. Buna ek olarak 2031 yılında ise çok az oranlı futbolcunun sözleşmesinin olduğu gözükmektedir. En üstte ise kontratı bitmiş Serbest durumdaki oyuncular yer almaktadır.

1.3) Oyuncu sayılarının en iyi oynadıkları pozisyona ve kullandıkları

ayaklarına göre oranlarının görselleştirilmesi

```
fifaoran <- fifa_23 %>%
  group_by(`Best Position` , `Preferred Foot`) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / nrow(fifa_23))



ggplot(fifaoran, aes(area = fifaoran$oran,
```

```
                    label = fifaoran$`Preferred Foot`,
                    subgroup = fifaoran$`Best Position`)) +

  geom_treemap(fill="forestgreen") +
  labs(title="Oyuncu sayılarının en iyi oynadıkları pozisyona ve kullandıkları
ayaklarına göre oranı") +

  geom_treemap_text(colour = "black",
                    place = "centre",
                    size = 15) +


  geom_treemap_subgroup_border(colour = "white",
                               size = 10) +
  geom_treemap_subgroup_text(place = "centre",
                             grow = TRUE,

                             alpha = 0.23,
                             colour = "black",
                             fontface = "plain") +
  theme(legend.position = "none")
```

YORUM

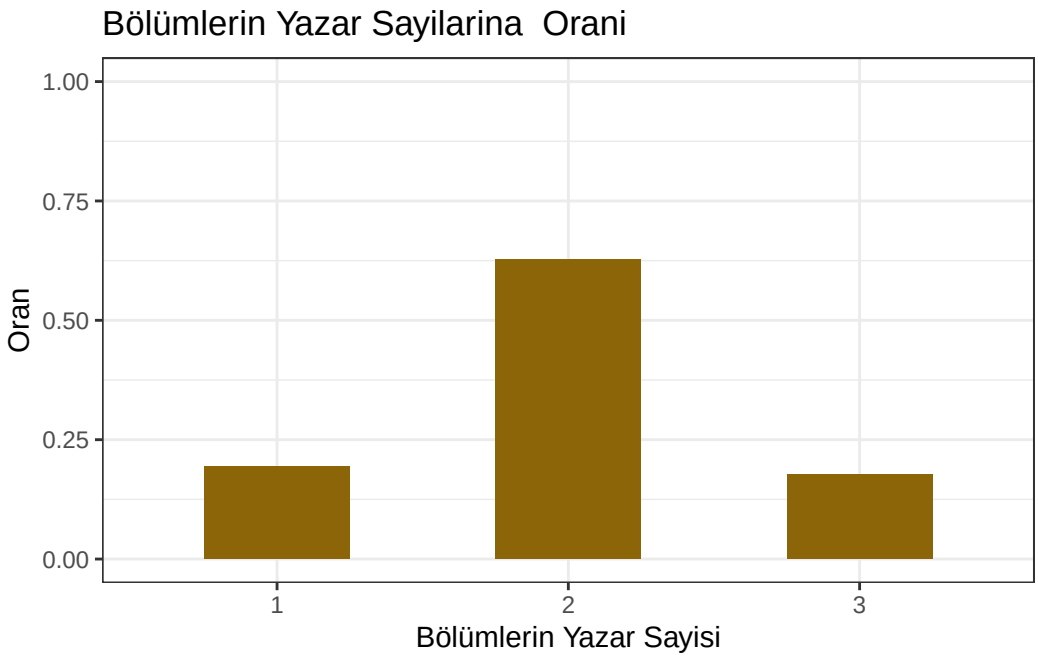
Yukarıda ki grafiğe bakarak ilk olarak `CB`,`ST`,`CAM` mevkilerinin diğerine göre oranının daha fazla olduğu görülmektedir. `RWB`: sağ kanat bek mevkiisindeki oyuncularının hepsinin sağ ayağını kullandığı görülmektedir. Buna karşılık olarak ise `LWB`: Sol kanatbek , `LB`:Solbek ve `LM`:SolOrtasaha mevkilerdeki oyuncuların büyük bir çoğunluğuunun sol ayak kullanma oranın yüksek olduğu görülmektedir.

2) Breaking Bad dizisi

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.

2.1) Bölümlerin, yazar sayısına göre oranlarının görselleştirmesi

```
bb1 <- breaking_bad %>%
  tidyr::separate_rows(`Written by`,
                       sep = ", ") %>%
  group_by(Season,Episode,Rating_IMDB) %>%
  summarise(yazar = n())%>%
  add_column(imdb =
               if_else(breaking_bad$Rating_IMDB < 9, "1", "2"))

ggplot(bb1, aes(x = as.factor(yazar) ,
               y=frequency(yazar)/nrow(bb1)))+

  geom_bar(stat = "identity", fill = "darkgoldenrod4", width = 0.5)+

  labs(x = "Bölümlerin Yazar Sayisi",
       y = "Oran",
       title = "Bölümlerin Yazar Sayilarına  Oranı")+
  ylim(0,1) + theme_bw()
```

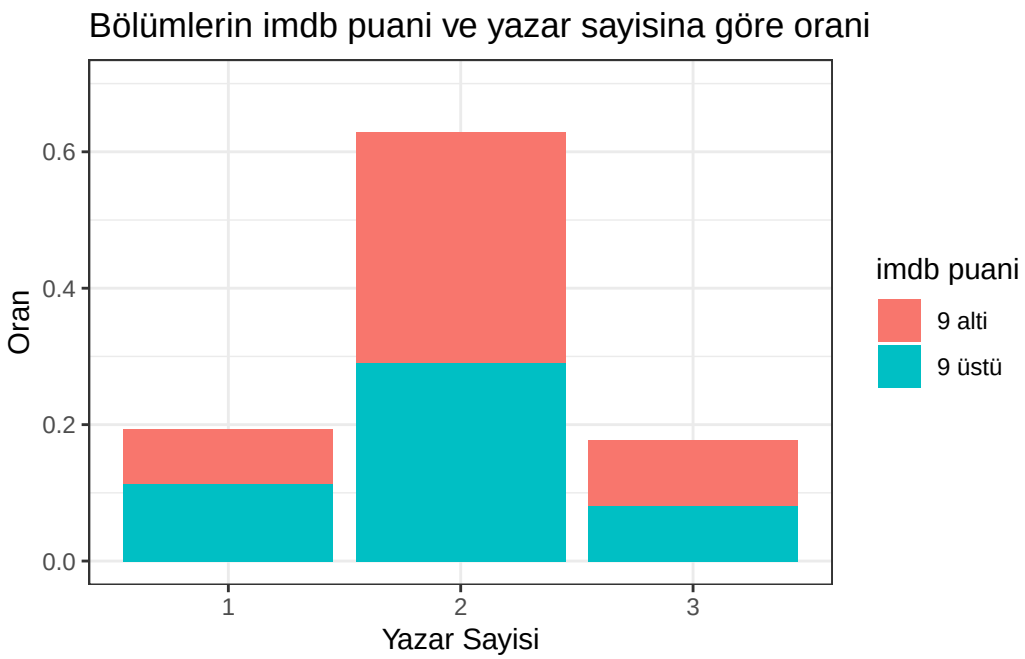


YORUM

Burada 2 yazarlı bölümlerin oranının diğerlerinin 2 katından fazla olduğunu, 1 ve 3 yazarlı bölümlerin oranının ise birbirine yakın olduğunu görürüz.

2.2) Bölümlerin, yazar sayısına ve reyting puanlarına göre oranlarının veri görselleştirmesi

```
ggplot(bb1, aes(x = as.character(yazar), fill = as.factor(imdb),
                y=(frequency(yazar)/nrow(bb1))))+
  geom_bar(stat = "identity"  )+


  labs(x = "Yazar Sayisi",
       y = "Oran",
       title = "Bölümlerin imdb puanı ve yazar sayısına göre oranı",
       fill="imdb puanı")+
  ylim(0,0.7)+

  scale_fill_discrete(labels = c("9 altı", "9 üstü")) + theme_bw()
```



YORUM

Yukarıda ki görselde 1 yazarlı bölümler arasında 9 puan üstü bölümlerin oranının daha fazla olduğunu, 2 yazarlı bölümler arasında 9 puan altı bölüm oranının fazla olduğunu ve 3 yazarlı bölümler arasında ise 9 puan altı bölüm oranının daha çok olduğunu görürüz.

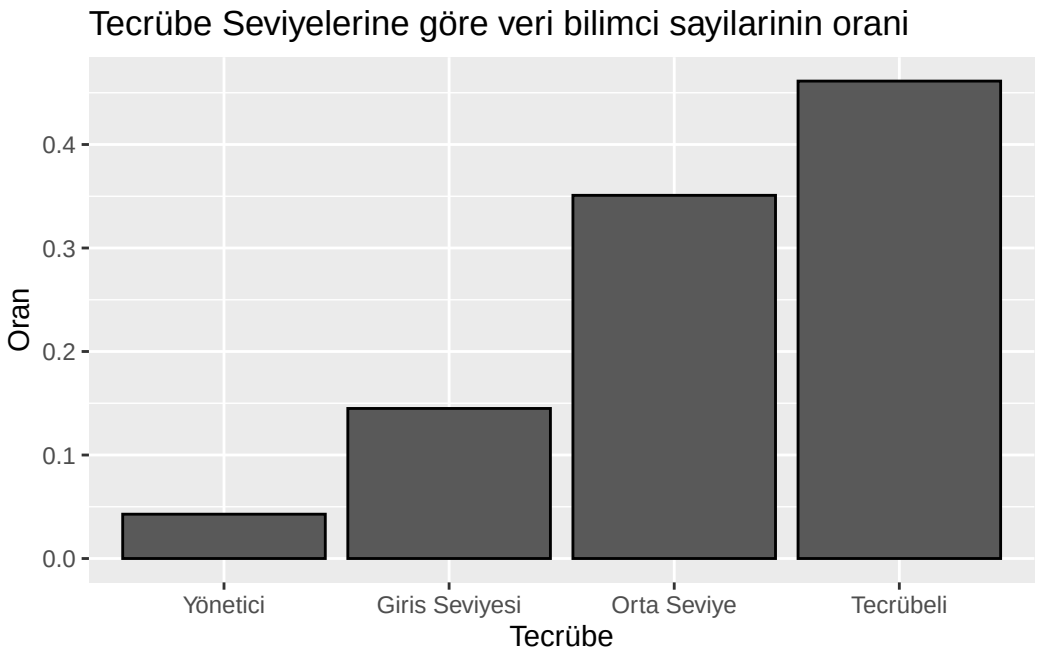
3) Veri bilimci maaşları

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.

3.1) Tecrübelerine göre veri bilimci sayılarının oranlarının veri görselleştirmesi

```
d1 <- Datasci %>%
  group_by(tec = Experience) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))
ggplot(d1, aes(x = reorder(d1$tec, +d1$oran), y = d1$oran  )) +
```

```
  geom_bar(stat = "identity" , position = "dodge" , color="black" )+
  labs( x= "Tecrübe" ,
        y = "Oran",
        title="Tecrübe Seviyelerine göre veri bilimci sayılarının oranı" ,
        fill= "Tecrübe Seviyesi") +
scale_x_discrete(labels=c("Yönetici","Giriş Seviyesi","Orta Seviye","Tecrübeli")) +
  theme(legend.position = "none")
```



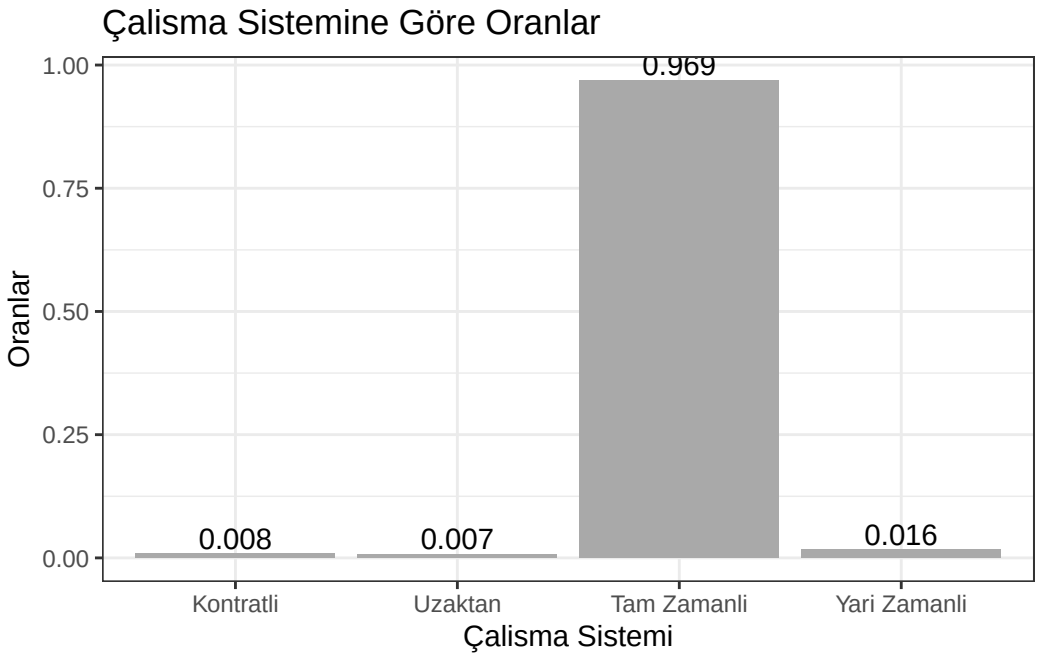
YORUM

Bu grafikte ise veri bilimciler arasında en düşük oranın yönetici olduğunu, en yüksek oranın ise Yönetici seviyesinde olduğunu görürüz

3.2) Veri bilimcilerinin çalışma sistemine göre oranlarının görselleştirmesi

```
dsis <- Datasci %>%
  group_by(sis = Employment_Status) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))
```

```
ggplot(dsis , aes( x = sis , y = oran , fill= sis)) +
  geom_bar(stat="identity" , fill="darkgray") +
  geom_text(aes(label=format(round(dsis$oran, 3),nsmall = 3)), vjust=-0.25) +
  labs( x = "Çalışma Sistemi" ,
        y = "Oranlar" ,
    title = "Çalışma Sistemine Göre Oranlar") +
   scale_x_discrete(labels=c("Kontratlı" , "Uzaktan","Tam Zamanlı" , "Yarı Zamanlı")) +
  theme_bw()
```



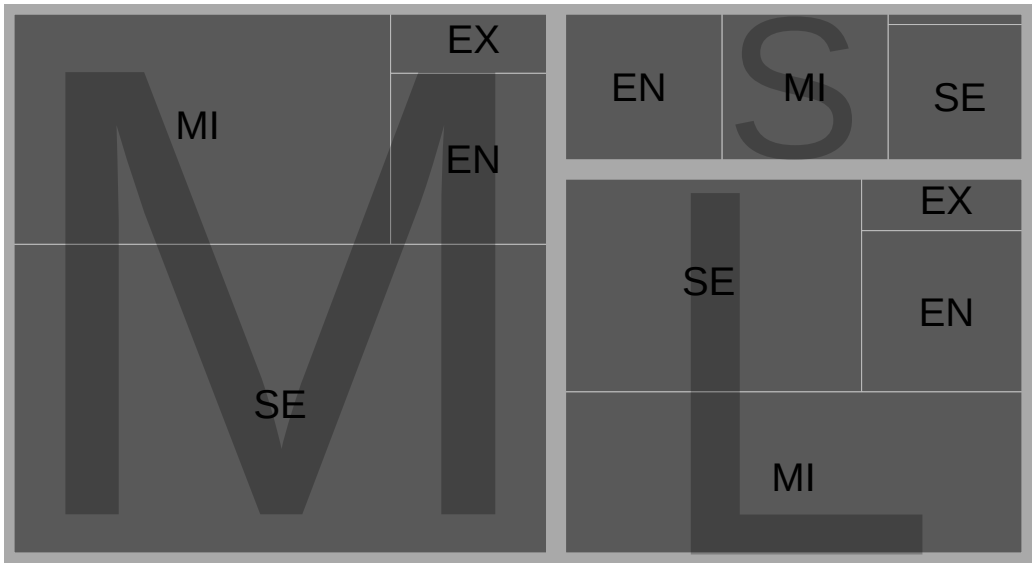
YORUM

Veri bilimcilerin çalışma sistemine göre oranı grafiğine bakıldığında çok büyük bir oranın Tam zamanlı çalıştığını görmekteyiz. Veri bilimci olmak isteyen birinin çok büyük bir ihtimalle tam zamanlı çalışması gerektiği söylenebilir.

3.3) Firma büyüklüğü ve tecrübe düzeyine göre veri bilimci sayılarının oranı Görseli

```
options(dplyr.summarise.inform = FALSE)
d3 <- Datasci %>%
  group_by(Company_Size , Experience) %>%
  dplyr::summarise(sayi = n()) %>%
  mutate(oran = sayi / nrow(Datasci))



ggplot(d3, aes(area = d3$oran,

                   label = d3$Experience,
                   subgroup = d3$Company_Size)) +
  geom_treemap() +
  labs(title="Firma büyüklüğüne ve tecrübe düzeyine göre veri bilimci
sayılarının oranı") +

  geom_treemap_text(colour = "black",
                    place = "centre",
                    size = 15) +

  geom_treemap_subgroup_border(colour = "darkgray",
                               size = 10) +
  geom_treemap_subgroup_text(place = "centre",
                             grow = TRUE,
                             alpha = 0.25,
                             colour = "black",
                             fontface = "plain") +
  theme(legend.position = "none")
```

Firma büyüklügüne ve tecrübe düzeyine göre veri bilimci sayilarinin orani



YORUM

Şirketler büyüklükleri arasında orta firma büyüklüğünün diğerlerine göre daha fazla olduğu görülmektedir. Tecrübesiz(`EN`) veri bilimcilerin sayıca en fazla küçük firmalarda çalıştığını, Tecrübeli(`SE`) veribilimcilerin ise en çok orta büyüklükteki firmalarda çalıştığını görmekteyiz. Bu grafiğe bakılarak eğer veri bilimine giriş yapmak isterseniz en uygun tercihin küçük firmalarda başlamak olduğu söyleyebilriz çünkü giriş seviye veri bilimciler küçük firmalarda daha çok yer oluşturmaktadır.