Oranları Görselleştirelim

ÖZET:

Amazon'un en çok satan 50 kitabının türlerine göre oranlarının görselleştirilmesi ve en çok satan kitaplarının kullanıcı puanı ve türlerine göre oranları görselleştirilip, yorumlanmıştır. Ayrıca Marvel filmlerinin türlerine göre oranlanıp görselleştirilip beraberinde filmlerin uzunluklarıyla oranlaştırılıp görselleştirilmiştir. Son olarak veri bilimci maaşları veri seti üzerinden tecrübelerine göre veri bilimci sayılarının oranı, veri bilimcilerinin çalışma sistemine göre oranları ve veri bilimcilerinin çalıştığı firma büyüklüğüne göre tecrübe düzeyleri oranlanıp yorumlanmıştır.

Veri Setleri:

1- Amazonda en çok satılan 50 kitabın yanı sıra içerisinde yazar, okuyucu puanı, yorum sayısı, fiyat ve tür bilgisin içeren detaylı bir veri setidir. Buradan veri setine ulaşabilirsiniz.

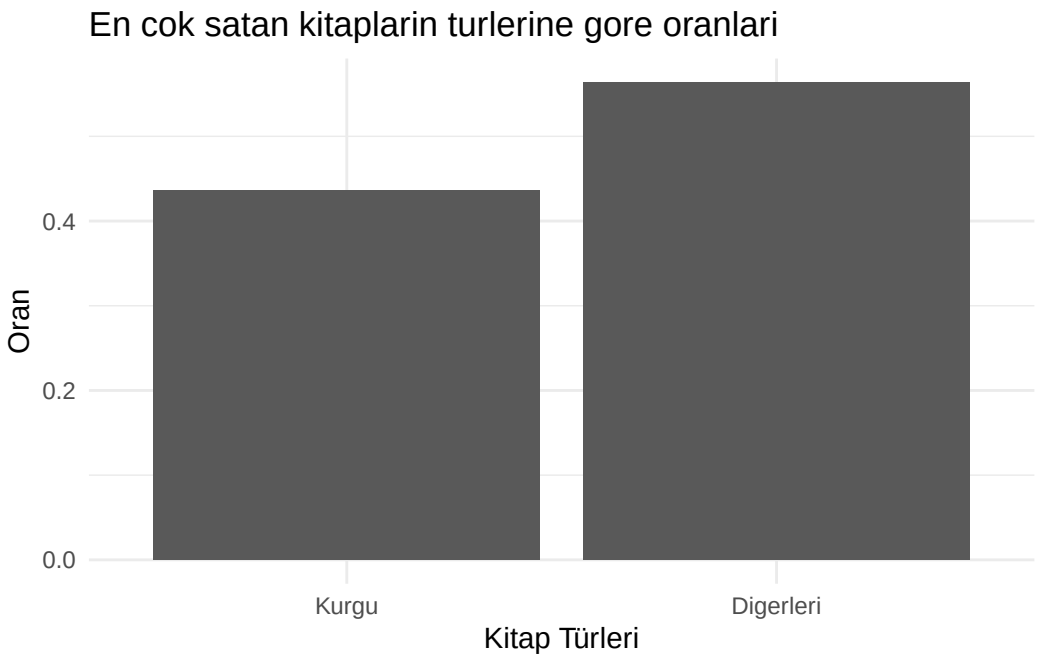
2- Marvel ve DC filmlerinin yer aldığı bu veri setinde film süresi, türü, yılı, imdb puanları ve izleyici yorumları gibi bir çok veriye buradan ulaşabilirsiniz.

3- Veri bilimci maaşları hakkında aşırı detaylı bu veri setinde tecrübe, maaş, çalışma yılı, çalışma düzeni, çalıştığı firma büyüklüğü gibi daha detaylı bu veriye buradan ulaşabilirsiniz.

```
install.packages("rmarkdown")
library(rmarkdown)
install.packages("ggplot2")
library(ggplot2)
install.packages("tidyverse")
library(tidyverse)
kitap = read.csv("bestsellers_with_categories.csv")
install.packages("gridExtra")
library(gridExtra)
```

En Çok Satan Kitapların Türlerine Göre Oranı:

```
kitap2 <- kitap %>%
   group_by(Genre) %>%
     summarise(sayi = n()) %>%
   mutate(oran = sayi / sum(sayi))

 kitap2$Genre=as.factor(kitap2$Genre)
 levels(kitap2$Genre)=c("Kurgu","Digerleri")

 ggplot(kitap2, aes(x = Genre , y = oran)) +
   geom_bar(stat = "identity")+
   scale_fill_viridis_d(option = "inferno", direction = 1) +
   labs(x = "Kitap Türleri", y = "Oran", title = "En cok satan kitaplarin turlerine gore o
   theme_minimal()
```



Grafikte görüldüğü üzere en çok satan kitap türlerinin oranları görselleştirilmiştir. X ekseninde kitap türleri ve Y ekseninde oranları olmak üzere iki değişkene yer verilmiş. Kurgu türüne sahip kitapların oranı kurgu olmayan(diğer) kitaplara nazaran daha az satmıştır. Fakat kurgu olmayan kitap türlerinin sayısının fazla olduğunu düşünecek olursak kurgu türünün çok iyi sattığını söyleyebiliriz ve insanların kurgu türünü sevdiğini söylemek yanlış olmaz.

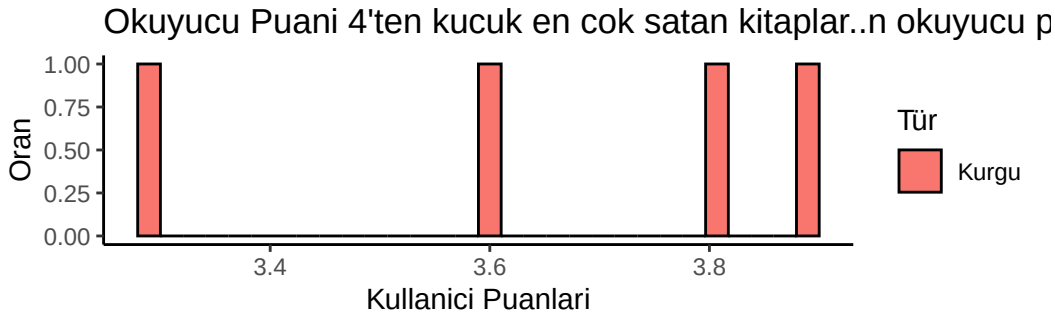
Okuyucu Puanı 4'ten Küçük ve 4'ten Büyük En Çok Satan Kitapların Görselleştirilmesi:

```
kucuk = kitap %>%
   group_by(User.Rating,Genre) %>%
   summarise(sayi = n()) %>%
   mutate(oran = sayi / sum(sayi))


 kucuk= filter(kucuk, User.Rating<4)

  grafik1 =ggplot(kucuk) +
   aes(x = User.Rating, fill = Genre) +
   geom_histogram(bins = 30L,colour="Black") +
   labs(x = "Kullanici Puanlari", y = "Oran", fill = "Tür",title = "Okuyucu Puani 4'ten ku
    scale_fill_discrete(labels=c("Kurgu","Digerleri"))+
   theme_classic()
```

123231

```
buyuk = kitap %>%
    group_by(User.Rating,Genre) %>%
    summarise(sayi = n()) %>%
    mutate(oran = sayi / sum(sayi))

  buyuk= filter(buyuk, User.Rating>4)

 grafik2 = ggplot(buyuk,aes(x=Genre,y = oran))+
    geom_bar(stat="identity",colour="Black",fill="Black")+
    ylim(0, 1)+
    labs(x="Kitap Türü",y="Oran",title="Okuyucu Puani 4'ten buyuk en cok satan kitaplarin
   scale_x_discrete(labels=c('Kurgu',"Digerleri"))+
   theme_classic()
```

```
grid.arrange(
   grafik1,
   grafik2
 )
```



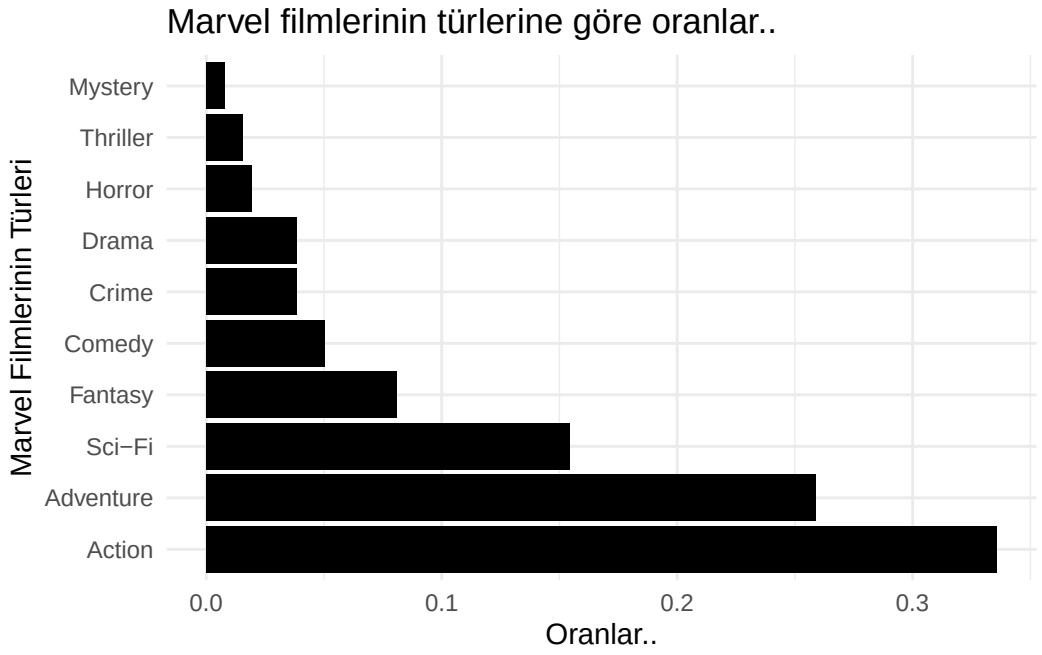
İlk grafikte okuyucu puanının 4'ten küçük olupta türüne göre oranı görselleştirilmiştir. X ekseninde okuyucu puanları ve Y ekseninde oranları verilmiştir. Okuyucu puanının 4'ten küçük olupta kurgu olmayan kitap türünün olmadığını söyleyebiliriz ve en çok satan kurgu olmayan kitap türünün kullanıcı puanlarının 4'ten büyük olduğunu gözlemleyebiliriz.(Kurgu olmayan kitap türünü bir türlü eklemeyi beceremedim)

İkinci grafikte ise okuyucu puanının 4'ten büyük olduğu ve kitap türlerinin oranları görselleştirilmiş. Kitap türleri x ekseninde olup y ekseninde bu kitap türlerinin oranları verilmiş. Okuyucu puanlarının 4'ten büyük olduğu senaryoda okuyucuların kurgu olmayan(diğerleri) türünü daha çok beğendiğini söyleyebiliriz. Bir firma daha çok kitap satmak istiyorsa kurgu olmayan kitap türüne yatırım yapması bu grafiğe göre daha doğru olabilir.

Marvel Filmlerinin Türlerine Göre Oranları

```
marvel = read.csv("mdc.csv")

 marvel2 = marvel %>%
       tidyr::separate_rows(genre, sep = ", ") %>%
       group_by(genre)


     marvel2 = marvel2 %>%
```

```
  group_by(genre) %>%
  summarise(sayi = n()) %>%
  mutate(oran = sayi / sum(sayi))

ggplot(marvel2) +
  aes(x =reorder(genre, -oran), weight = oran) +
  geom_bar(fill = "#000000") +
  labs(
    x = "Marvel Filmlerinin Türleri",
    y = "Oranları",
    title = "Marvel filmlerinin türlerine göre oranları"
  ) +
  coord_flip() +
  theme_minimal()
```



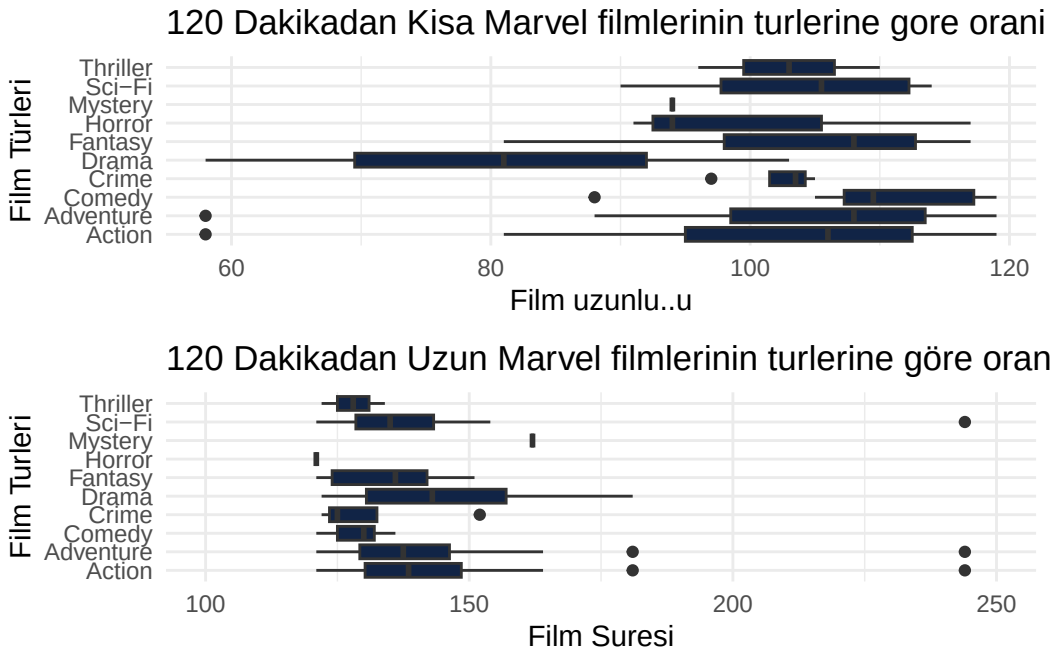
Grafikte Marvel filmlerinin türlerine göre oranlarının görselleştirilmesi verilmiş. Y ekseninde film türleri bulunuyorken x ekseninde bu film türlerinin oranları bulunuyor. Açık bir şekilde Marvel en çok aksiyon türünde film üretirken en az gizem türünde film çekmiş. Bu veriye bakarak en fazla hasılat yapan yada izleyicilerin sevdiği tür aksiyon yorumu yapılabilir. İzleyecek film arayıp aynı zamanda aksiyon seven birisine Marvel önerisi yapılabilir. Aynı zamanda gizem, drama ve korku türünü seven birisi ise Marvel filmlerlerinden uzak durmalı yorumu yapılabilir. Bu veriler aynı zamanda reklam politikası için altın değerindedir, genel olarak

aksiyon filmi izleyen birisine Marvel filmlerinin reklamını algoritmalarla sunmak yanlış bir aksiyom olmayacaktır.

###Marvel Filmlerinin Süresi

```
 marvelkucuk = marvel %>%
       tidyr::separate_rows(genre, sep = ", ") %>%
       group_by(genre,runtime) %>%
       summarise(n())

marvelkucuk =filter(marvelkucuk, runtime <120)




     m1 = ggplot(marvelkucuk) +
       aes(x = runtime, y = genre) +
       geom_boxplot(fill = "#112446") +
       labs(
         x = "Film uzunluğu",
         y = "Film Türleri",
         title = "120 Dakikadan Kisa Marvel filmlerinin turlerine gore orani"
       ) +
       theme_minimal()



     marvelbuyuk = marvel %>%
       tidyr::separate_rows(genre, sep = ", ") %>%
       group_by(genre,runtime) %>%
       summarise(n())

 marvelbuyuk2 = marvelbuyuk %>%
       group_by(runtime,genre) %>%
       summarise(sayi = n()) %>%
       mutate(oran = sayi / sum(sayi))

 marvelbuyuk2 =filter(marvelbuyuk2, runtime >120)

 m2 =  ggplot(marvelbuyuk2) +
       aes(x = runtime, y = genre) +
       geom_boxplot(fill = "#112446") +
```

```
      labs(
        x = "Film Suresi",
        y = "Film Turleri",
        title = "120 Dakikadan Uzun Marvel filmlerinin turlerine göre orani"
      ) +
      theme_minimal() +
      xlim(100, 250)

grid.arrange(
  m1,
  m2
)
```



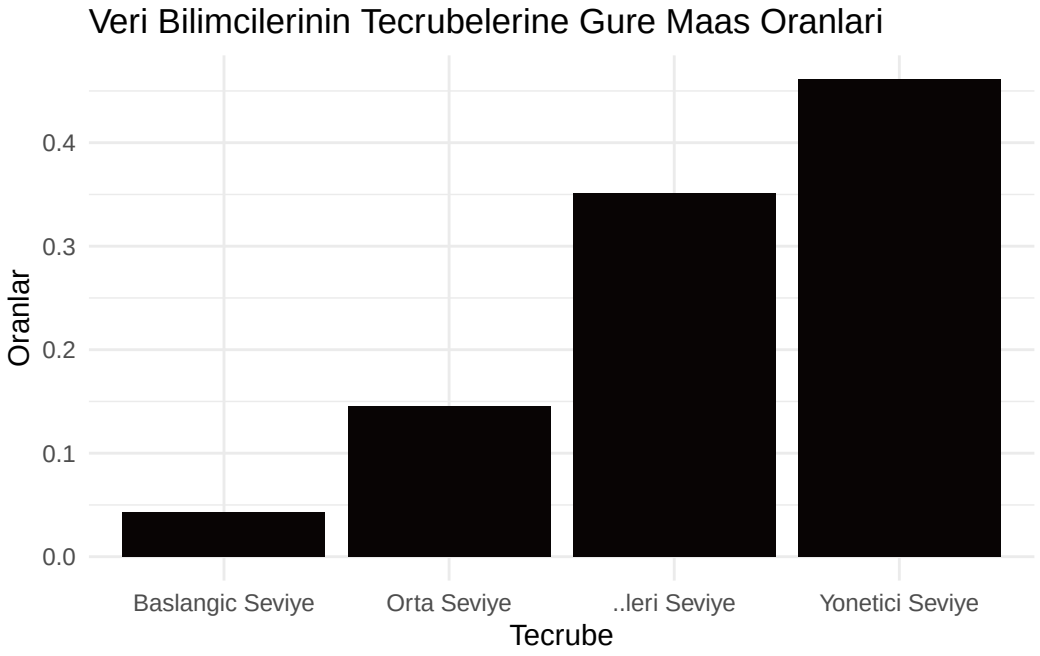
Birinci grafikte 120 dakikadan daha kısa marvel filmlerinin türlerine göre görselleştirilmesi verilmiştir. Film türleri Y ekseninde alırken, film süresi x ekseninde bulunuyor. Bariz bir şekilde en kısa film türü dramadır. Yeteri kadar hasılat yapmadığı için Marvel da daha az bütçe ayırarak daha kısa filmler çekiyor olma ihtimali yüksektir. Öte yandan diğer film türlerinin süreleri birbirlerine oldukça yakındır. Marvel spesifik olarak drama türünü sevmediği söylenebilir.

İkinci grafikte ise 120 dakikadan daha uzun Marvel filmlerinin türlerine göre görselleştirilmesi bulunıyor. Aykırı değerlerin fazla olması nedeniyle Marvel burada hasılattan çok film konusunun uzunluğu doğrultusunda filmin süresine karar verildiği yorumu yapılabilir. Gizem ve korku

türünde çok az film çekildiği için de bu türler belli bir ortalama süreye karar vermek zordur. Bir cüretkarlık yapıp bu filmlerin iyi hasılat yapmadığı için çok az sayıda çekildiği yorumu bu grafiğe bakarak yapmak zordur ama yine de bu ihtimal yüksektir.

Veri Bilimci Maaşları

```
veriB = read.csv("Data_Science_Fields_Salary_Categorization.csv")


   veriB$Experience=as.factor(veriB$Experience)

   veriB2 = veriB %>%
     group_by(Experience) %>%
     summarise(sayi = n()) %>%
     mutate(oran = sayi / sum(sayi))

   ggplot(veriB2) +
     aes(x =reorder(Experience,+oran), y = oran) +
     geom_col(fill = "#080404") +
     labs(
       x = "Tecrube",
       y = "Oranlar",
       title = "Veri Bilimcilerinin Tecrubelerine Gure Maas Oranlari"
     )+
     scale_x_discrete(labels=c("Baslangic Seviye","Orta Seviye","İleri Seviye","Yonetici S
     theme_minimal()
```

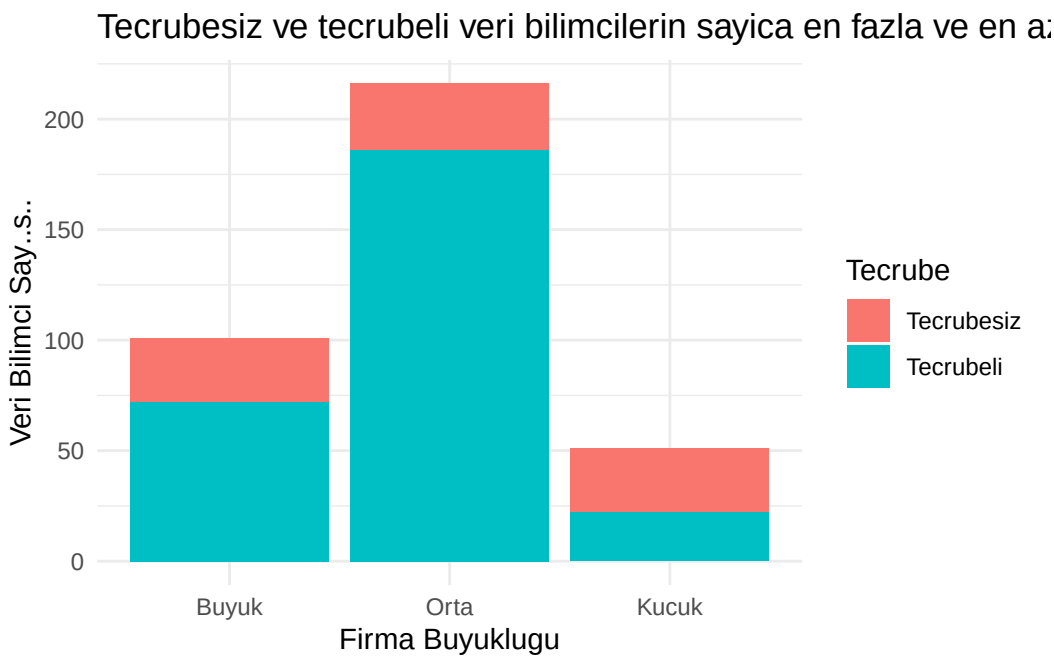


Grafikte veri bilimcilerinin tecrübesi doğrultusunda aldığı maaş oranları verilmiştir. Şaşırtmayacak şekilde Yönetici seviye(en tecrübeli) en yüksek maaşı alırken en az maaşı ise işe yeni başlamış tecrübesiz bir veri bilimci bulunuyor. Oranların farkına bakarsak veri bilimciler için tecrübe altın değerindedir çünkü tecrübesiz ve tecrübeli veri bilimci maaşı arasında dağlar kadar fark vardır. İşe yeni başlamış bir veri bilimci çok para kazanmayı beklemek yerine sabredip tecrübelenmeye odaklanması yorumu yapılabilir.

Tecrübeli ve Tecrübesiz Veri Bilimcilerinin Çalıştığı Firma Büyüklükleri:

```
veriBtecrubesiz = filter(veriB, Experience %in% c("EN","SE"))

    veriBtecrubesiz = veriBtecrubesiz %>%
      group_by(Experience,Company_Size) %>%
      summarise(n())

    ggplot(veriBtecrubesiz) +
      aes(x = Company_Size, y = `n()`, fill = Experience) +
      geom_col() +
      scale_fill_hue(direction = 1) +
      labs(
        x = "Firma Buyuklugu",
        y = "Veri Bilimci Sayısı",
```

```
    title = "Tecrubesiz ve tecrubeli veri bilimcilerin sayica en fazla ve en az calis
    fill = "Tecrube"
  ) +
  scale_x_discrete(labels=c("Buyuk","Orta","Kucuk"))+
  scale_fill_discrete(labels=c("Tecrubesiz","Tecrubeli"))+
  theme_minimal()
```

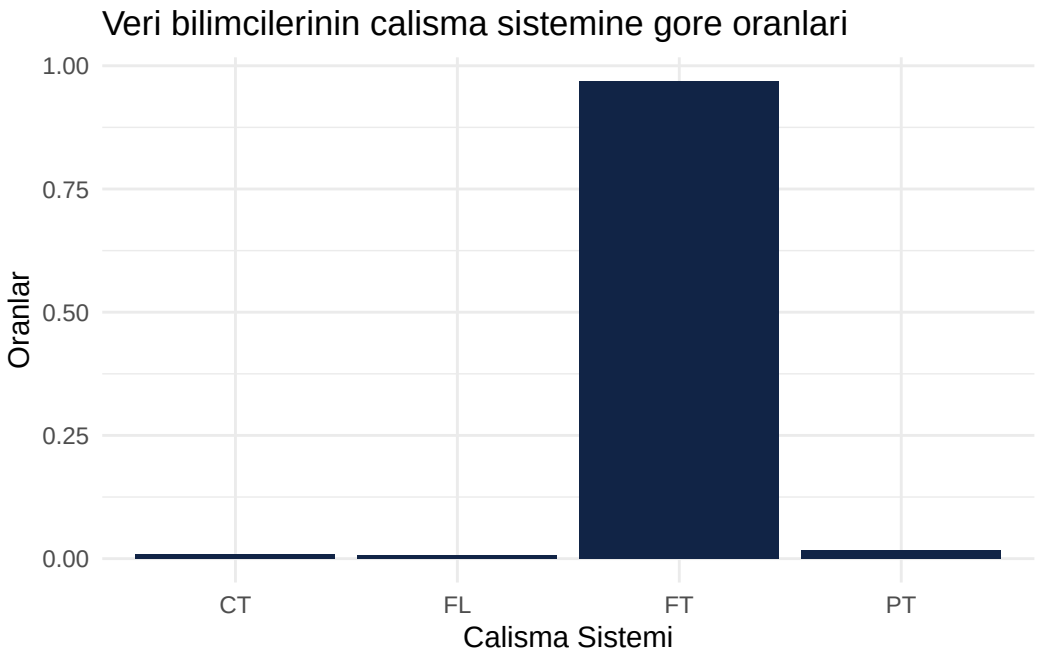


Grafikten görüleceği üzere tecrübeli ve tecrübesiz veri bilimcilerinin hangi büyüklükteki firmalarda daha çok çalıştığı görülmektedir. Firma büyüklüğü ile tecrübe doğru orantılıdır yorumu yapılabilir. İşe yeni başlayan tecrübesiz veri bilimciler oransal olarak daha çok küçük firmalarda çalışıyor ve tecrübe kazandıkça daha büyük firmalara geçiyorlar. Orta büyüklükteki firma için şu yorum yapılabilir; Orta büyüklükteki firmalar tecrübeye çok önem verip daha tecrübeli insanları işe aldıkları bariz şekilde görülmektedir. Veri bilimci olarak hayata atılacak biri oransal olarak orta büyüklükteki firma yerine küçük veya büyük firmaları tercih edebilir.

###Veri bilimcilerinin çalışma sistemine göre oranları

```
veriBB = veriB %>%
      group_by(Employment_Status) %>%
      summarise(sayi = n()) %>%
      mutate(oran = sayi / sum(sayi))
```

```r
ggplot(veriBB) +
  aes(x = Employment_Status, weight = oran) +
  geom_bar(fill = "#112446") +
  labs(
    x = "Calisma Sistemi",
    y = "Oranlar",
    title = "Veri bilimcilerinin calisma sistemine gore oranlari"
  ) +
  theme_minimal()
```



(Çalışma sisteminin ne olduğunu anlayamadım, umarım doğrudur)

Grafikte görüldüğü üzere x ekseninde veri bilimcilerin çalışma sistemi ve y ekseninde bu çalışma sisteminin oranları verilmiştir. Çok açık bir şekilde veri bilimciler FT çalışma sisteminde çalışmaktadır. Diğer çalışma sistemleri 0'a çok yakın ve çok az tercih edilmektedir. Bu yüzden hayata yeni atılan bir veri bilimci hangi çalışma sisteminde çalışacağı hakkında soru işaretleri varsa grafiğe bakarak net bir şekilde FT çalışma sistemini tercih etmesi söylenebilir.