Untitled

Zehra Hacioglu, Arzum Gursoy, Omer Sen, Olgun Aslan

23 01 2021

KAGGLE VERİLERİNİN KEŞFİ

Bu proje ile 2017 yılı Kaggle Veri Bilimi Anket sonuçları ile profosyonellerin günlük yaşantılarında kullanmış oldukları dil ve yöntemlerin keşfedilmesi amaçlanmıştır. Paket içeriği 5 ayrı dosyadan oluşmaktadır. Bu projede kullanılan dosya anket sonuçlarının bulunduğu rawMCData veri setidir. Veri seti 16.716 satır ve 228 sütundan oluşmaktadır. Her bir sütun katılımcıların sorulan sorulara verdikleri cevapları içermektedir.

İlk Adım: kullanılacak paketler ve veri setinin yüklenmesi.

```
library(readr)
library(tidyverse)
library(dplyr)
library(ggplot2)
```

```
rawMCData <- read.csv("C:/Users/Zehra/Desktop/multipleChoiceResponses.csv", stringsAsFactors = T, header = T
)
attach(rawMCData)
```

```
head(rawMCData, n = 5)
```

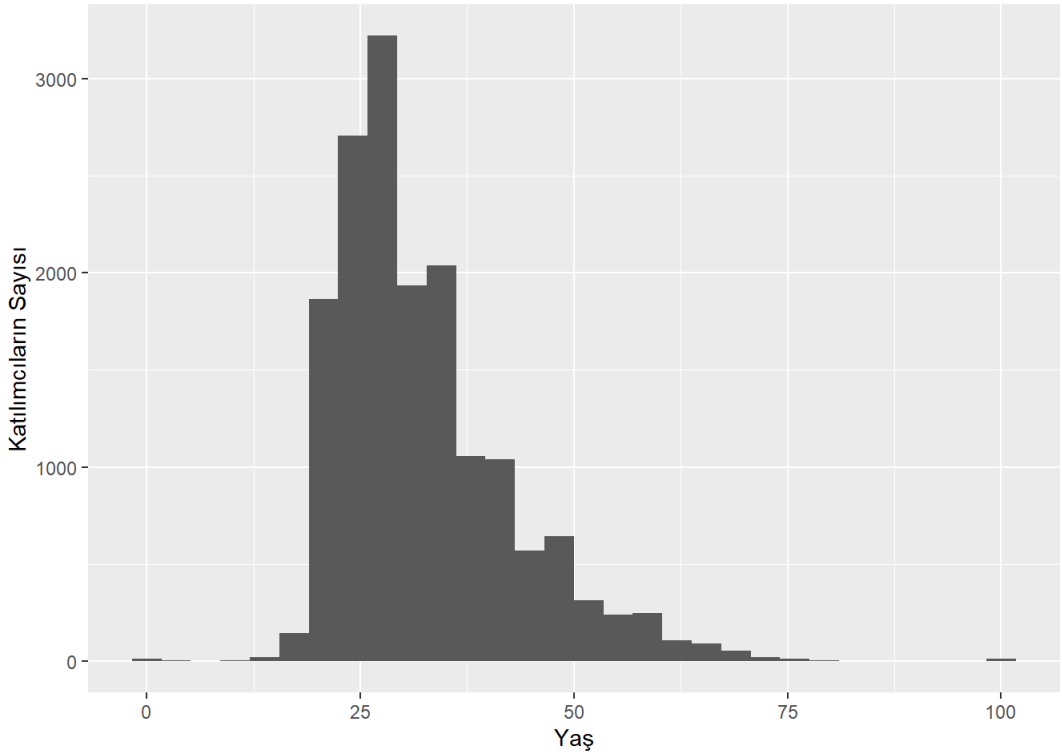
	GenderSelect <fct>	Country <fct>	Age <int>
1	Non-binary, genderqueer, or gender non-conforming		NA
2	Female	United States	30
3	Male	Canada	28
4	Male	United States	56
5	Male	Taiwan	38

5 rows | 1-4 of 229 columns

Kullanılacak olan veri setinin ilk 10 kullanıcıya ait verileri yukarıda gözükmektedir. Katılımcılara ait bazı bilgilerin grafik ve tabloları aşağıda verilmektedir.

İlk olarak katılımcıların yaşlarını gösteren bir histogram çizilsin.

```
ggplot(rawMCData, aes(Age, fill = Age)) +
  geom_histogram() +
  xlab("Yaş") +
  ylab("Katılımcıların Sayısı")
```




```
median(Age, na.rm = TRUE)
```

```
## [1] 30
```

Histogram grifiğine göre katılımcıların yaş aralığı en çok 20-30 aralığındadır.

Aşağıda ki fonksiyon istenilen değişken için özet bilgileri veren bir fonksiyondur. Bu fonksiyon ile ilgili değişkenin her bir faktörü için yüzdelik ve miktar hesaplamaları gerçekleştirilecektir.

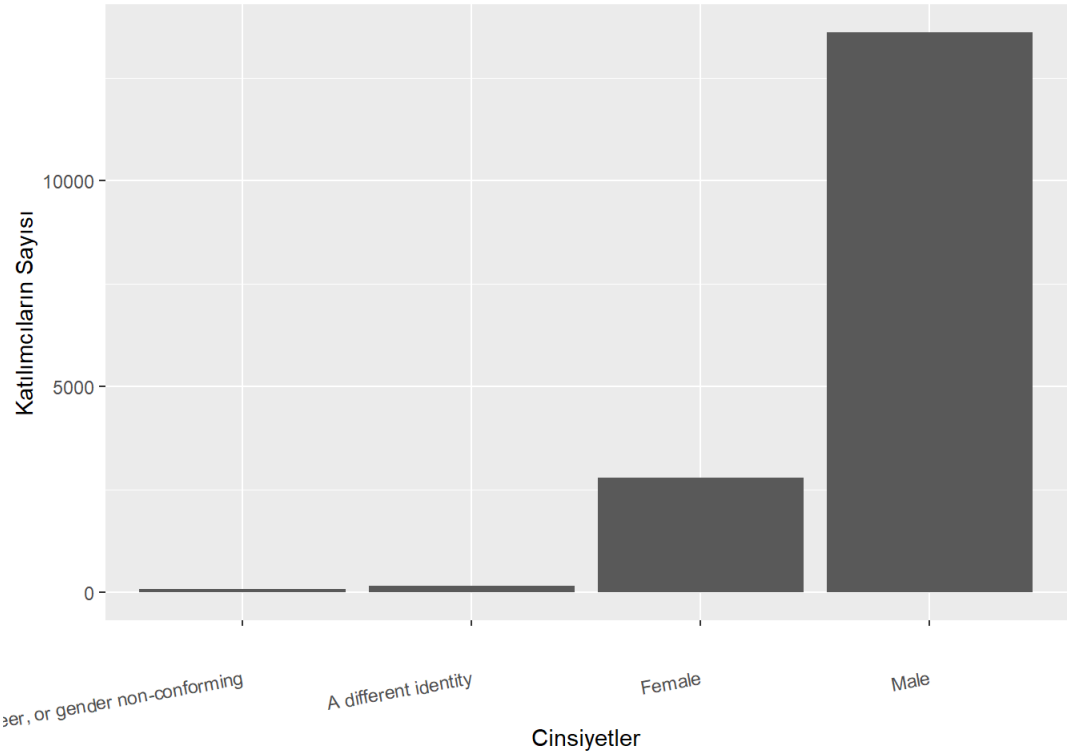
```
chooseOne = function(question, filteredData = rawMCData){

  filteredData %>%
    filter(!UQ(sym(question)) == "") %>%
    group_by_(question) %>%
    summarise(count = n()) %>%
    mutate(percent = (count / sum(count)) * 100) %>%
    arrange(desc(count))

}
```

```
gender <- chooseOne("GenderSelect")
gender
```

GenderSelect <fct>	count <int>	percent <dbl>
Male	13610	81.8843632
Female	2778	16.7137958
A different identity	159	0.9566211
Non-binary, genderqueer, or gender non-conforming	74	0.4452199

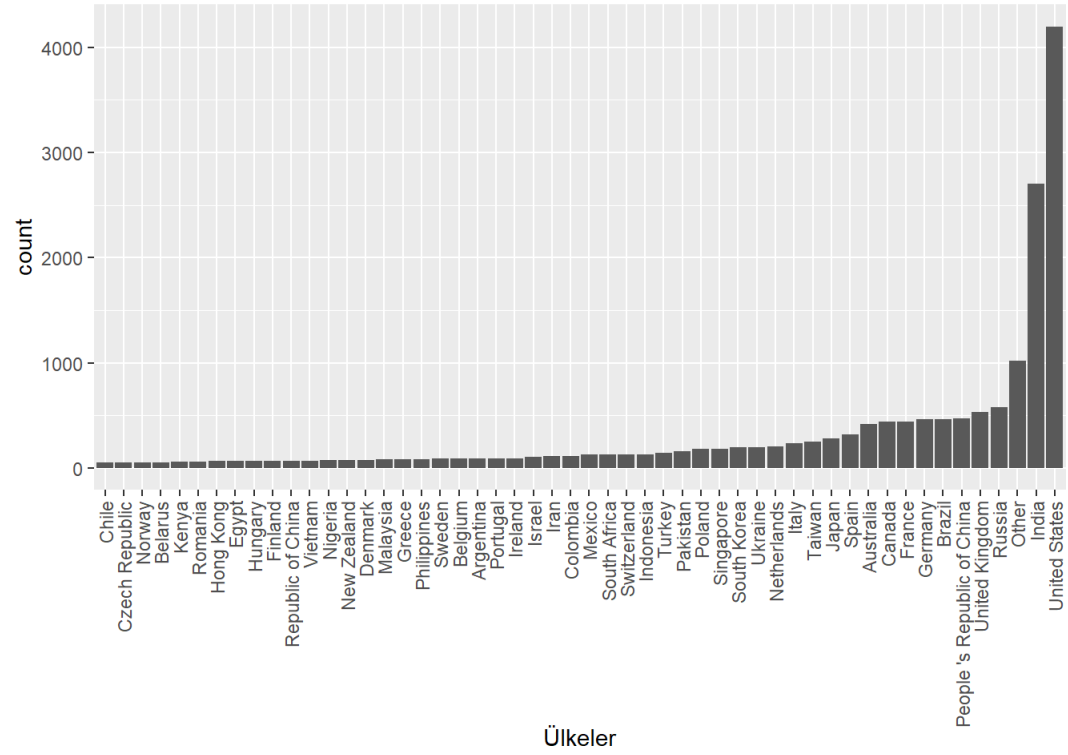
4 rows

```
ggplot(chooseOne("GenderSelect"),aes(x = reorder(GenderSelect, count), y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 10,
                                   vjust = 0.5,
                                   hjust = 1))+
  xlab("Cinsiyetler")+
  ylab("Katılımcıların Sayısı")
```



İstatistikler ve grafik sonuçlarına göre rawMCData veri setinin %81,8 kadar büyük kısmını erkek katılımcılar oluşturmaktadır.

```
ggplot(chooseOne("Country"),aes(x = reorder(Country, count), y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 1))+
  xlab("Ülkeler")
```




```
head(chooseOne("Country") , n = 5)
```

Country <fct>	count <int>	percent <dbl>
United States	4197	25.290750
India	2704	16.294064
Other	1023	6.164507
Russia	578	3.482977
United Kingdom	535	3.223863

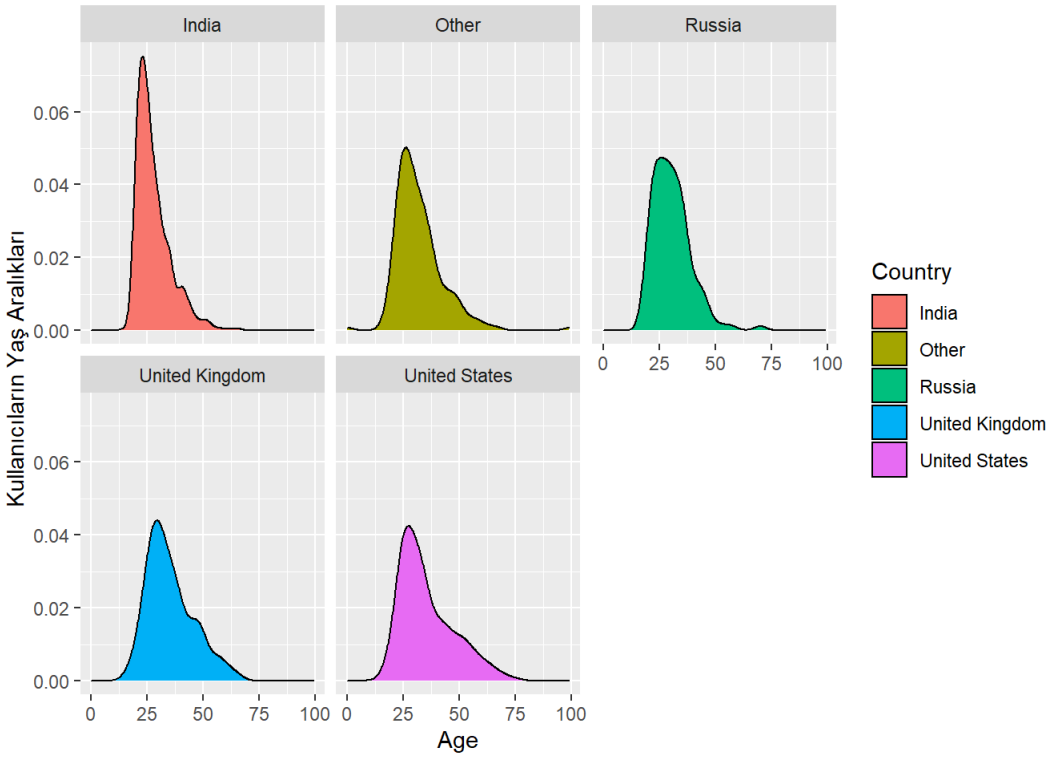
5 rows

Yukarıdaki grafik katılımcıların yaşadıkları ülkelerin artan sıraya göre grafiğini göstermektedir. Katılımcıların en çok katılım gösterdikleri ilk 5 ülke %25.2 oranında Birleşik Devletler, %16.2 oranında Hindistan, %6.1 oranında diğer ülkeler, %3.4 oranında Rusya ve son olarak %3.2 oranında Birleşik Krallık olacak şekilde bir sıralamaya sahiptir.

En çok katılımın gerçekleştiği 5 ülke filtrelenerek ve yeni bir veri setine atansın ve bu yeni veri seti ile en çok katılımın sağlandığı 5 ülkenin yaş aralıkları grafik ile gösterilsin.

```
newData <- rawMCData%>%
  filter(Country == c("United States","India","Other","Russia","United Kingdom"))
```

```
ggplot(newData, aes(x = Age, fill = Country)) +
  geom_density() +
  facet_wrap(~Country) +
  ylab("Kullanıcıların Yaş Aralıkları")
```



Yaş aralığı en geniş olan grafik Birşekik Devletler'e aitken Hindistan diğer ülkelere kıyasla daha dar ve genç bir aralığa sahiptir. Bunların yanı sıra yaş aralığının en çok ABD ve Rusya'da olduğu görülmektedir.

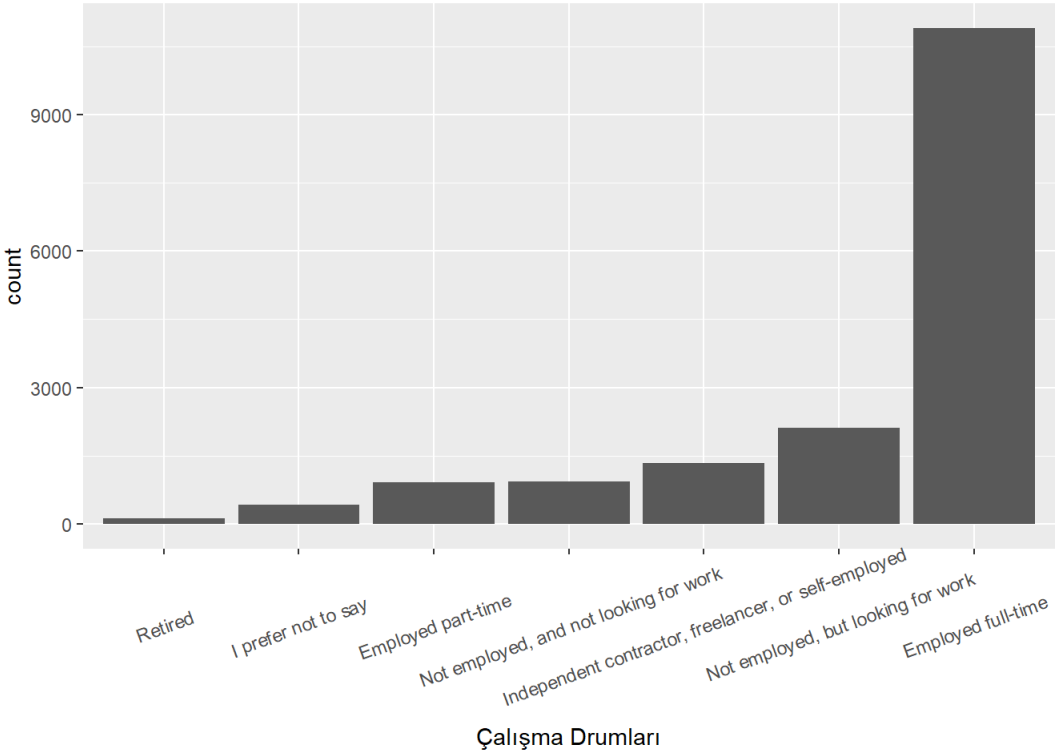
Katılımcıların çalışma durumları nedir?

```
chooseOne("EmploymentStatus")
```

EmploymentStatus <fct>	count <int>	percent <dbl>
Employed full-time	10897	65.1890404
Not employed, but looking for work	2110	12.6226370
Independent contractor, freelancer, or self-employed	1330	7.9564489
Not employed, and not looking for work	924	5.5276382
Employed part-time	917	5.4857621
I prefer not to say	420	2.5125628
Retired	118	0.7059105

7 rows

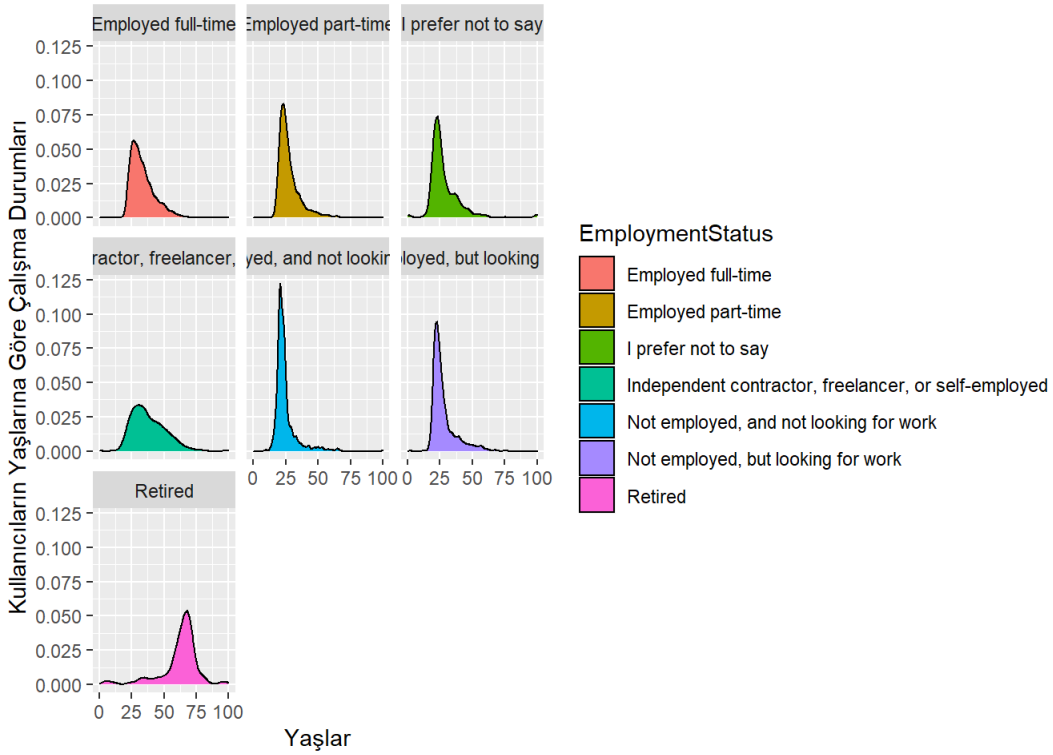
```
ggplot(chooseOne("EmploymentStatus"), aes(x = reorder(EmploymentStatus,count), y = count))+
  geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 20,
                                   vjust = 0.6)) +
  xlab("Çalışma Drumları")
```



Ankete katılanların %65'i tam zamanlı bir işte çalışmaktadır, %12.6'sı çalışmıyor fakat iş arıyor. Genel olarak katılımcıların %78.5'si çalışıyor, %18.3'i çalışmıyor, %2.5'i çalışma durumunu belirtmek istemiyor ve %0.7'si ise emeklidir.

Çalışma durumu ile yaş değişkenleri arasındaki ilişki kontrolünün yapılması.Katılımcıların çalışma durumları "EmploymentStatus" kolonunda yer almaktadır.

```
ggplot(rawMCData, aes(x = Age, fill = EmploymentStatus)) +
  geom_density() +
  facet_wrap(~EmploymentStatus)+
  xlab("Yaşlar") +
  ylab("Kullanıcıların Yaşlarına Göre Çalışma Durumları")
```



Anketörler, çalışma durumlarını “çalışmıyorum fakat iş arıyorum”, “çalışmıyorum ve iş aramıyorum” ve “çalışma durumumu belirtmek istemiyorum” olarak cevaplayan katılımcılara yüksek dereceli bir okulda okuyup okumadıklarını sordu.Katılımcıların bu soruya verdikleri cevaplar “StudentStatus” kolonunda yer almaktadır. Verilen cevaplara göre katılımcıların eğitim seviyeleri belirlenecektir.

```
chooseOne("StudentStatus")
```

StudentStatus <fct>	count <int>	percent <dbl>
Yes	981	76.64062
No	299	23.35938

2 rows

İstatistikler ankete katılanların %76.64’ünün yüksek dereceli bir okulda eğitim aldıklarını söylemektedir.

Katılımcılara veri bilimi becerilerini geliştirmeye odaklı olup olmadıkları sorulmuştur.

```
chooseOne("LearningDataScience")
```

LearningDataScience <fct>	count <int>	percent <dbl>
Yes, I'm focused on learning mostly data science skills	800	62.305296
Yes, but data science is a small part of what I'm focused on learning	429	33.411215
No, I am not focused on learning data science skills	55	4.283489

3 rows

Katılımcıların %62’si veri bilimi becerilerini geliştirmeye odaklandıklarını söylemektedir.

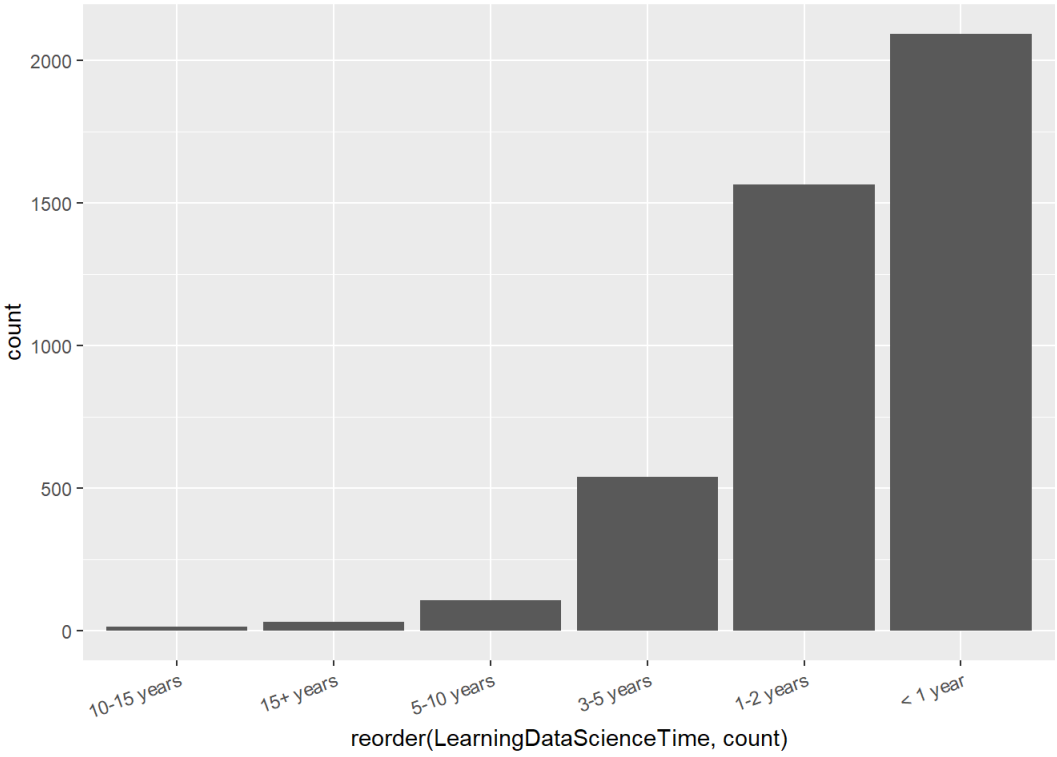
Anketörlerin, ne kadar süredir veri bilimi ile ilgilendikleri sorusuna katılımcıların vermiş oldukları cevaplar aşağıdaki gibir.

```
lds_time <- chooseOne("LearningDataScienceTime")
lds_time
```

LearningDataScienceTime <fct>	count <int>	percent <dbl>
< 1 year	2093	48.1260060
1-2 years	1566	36.0082778
3-5 years	540	12.4166475
5-10 years	106	2.4373419
15+ years	30	0.6898138
10-15 years	14	0.3219131

6 rows

```
ggplot(chooseOne("LearningDataScienceTime"),aes(x = reorder(LearningDataScienceTime, count), y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 20,
                                   vjust = 1,
                                   hjust = 1))
```



İstatistikler, katılımcıların %84 gibi büyük bir çoğunluğunun 2 yıldan kısa bir süredir veri bilimi ile ilgilendiklerini göstermektedir.

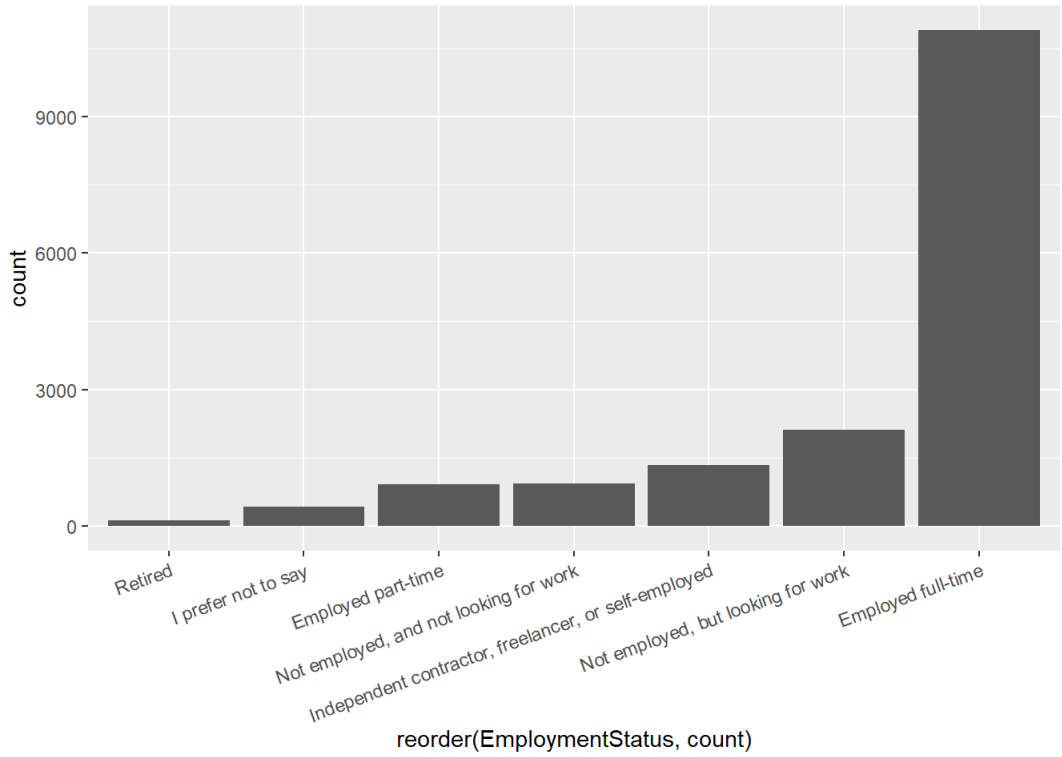
Katılımcılara mesleklerinin ne olduğu sorulmaktadır

```
chooseOne("EmploymentStatus")
```

EmploymentStatus <fct>	count <int>	percent <dbl>
Employed full-time	10897	65.1890404
Not employed, but looking for work	2110	12.6226370
Independent contractor, freelancer, or self-employed	1330	7.9564489
Not employed, and not looking for work	924	5.5276382
Employed part-time	917	5.4857621
I prefer not to say	420	2.5125628
Retired	118	0.7059105

7 rows

```
ggplot(chooseOne("EmploymentStatus"),aes(x = reorder(EmploymentStatus, count), y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 20,
                                   vjust = 1,
                                   hjust = 1))
```



Katılımcıların %20.5' i Veri Bilimci, %14.8'i Yazılım Mühendisi, %10.4'ü farklı meslek sahibi, %10.2'si Veri Analisti olduklarını söylemektedirler. Kaggle kullanıcılarının yaklaşık% 45'i Veri Bilimciler, Yazılım Geliştiriciler / Mühendisler veya Veri Analistleridir.

Bu ankete katılım sağlayanlara günlük hayatta kullanmış oldukları araçların neler olduğu sorulmuştur.

```
head(chooseOne("HardwarePersonalProjectsSelect"), n = 5)
```

HardwarePersonalProjectsSelect <fct>	**count** <int>	**percent** <dbl>
Basic laptop (Macbook)	1713	40.72753
Gaming Laptop (Laptop + CUDA capable GPU)	427	10.15216
Laptop + Cloud service (AWS, Azure, GCE ...)	313	7.44175
Traditional Workstation	304	7.22777
Laptop or Workstation and local IT supported servers	268	6.37185

5 rows

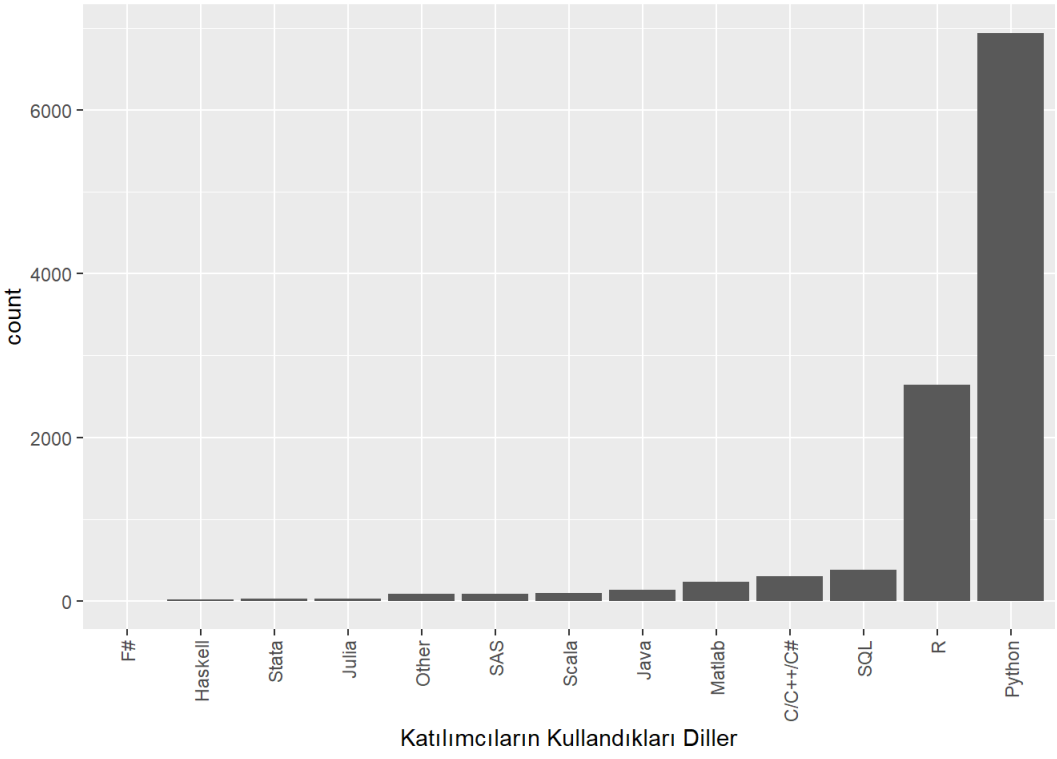
Yukarıdaki tablo katılımcıların günlük hayatlarında kullandıkları araçları göstermektedir. Kullanıcıların %40.7'si dizüstü bilgisayar(Macbook), %10.1'i oyun bilgisayarı, %7.4'ü dizüstü bilgisayar ve bulut hizmeti, %7.2'si iş istasyonu (teknik) bilgisayarı, %6.3'ü dizüstü veya iş istasyonu ve yerel BT destekli sunucuları kullandıklarını söylediler. Genel olarak çoğunluğun %40.7 oranında dizüstü bilgisayar kullandıkları saptanmıştır.

```
chooseOne("LanguageRecommendationSelect")
```

LanguageRecommendationSelect <fct>	count <int>	percent <dbl>
Python	6941	63.11147481
R	2643	24.03164212
SQL	385	3.50063648
C/C++/C#	307	2.79141662
Matlab	238	2.16402982
Java	138	1.25477360
Scala	94	0.85470085
SAS	88	0.80014548
Other	85	0.77286779
Julia	30	0.27277687

1-10 of 13 rows Previous **1** 2 Next

```
ggplot(chooseOne("LanguageRecommendationSelect"),aes(x = reorder(LanguageRecommendationSelect, count), y = c
ount)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 1))+
  xlab("Katılımcıların Kullandıkları Diller")
```



Görüldüğü üzere katılımcıların %63 gibi büyük bir kısmı veri analizleri sırasında Python'ı tercih ediyor. R ise %24 kullanım oranıyla Python'dan sonra en çok tercih edilen dil oluyor.

Sonuç olarak istatistiklere göre kullanıcıların en çok kullandıkları araç %40.7 oranıyla dizüstü bilgisayar(Macbook) ve en çok kullandıkları dil %63 oranıyla Python oluyor.