Olasılık ve İstatistik

Fırat İsmailoğlu, PhD

Korelasyon – R'a Giriş

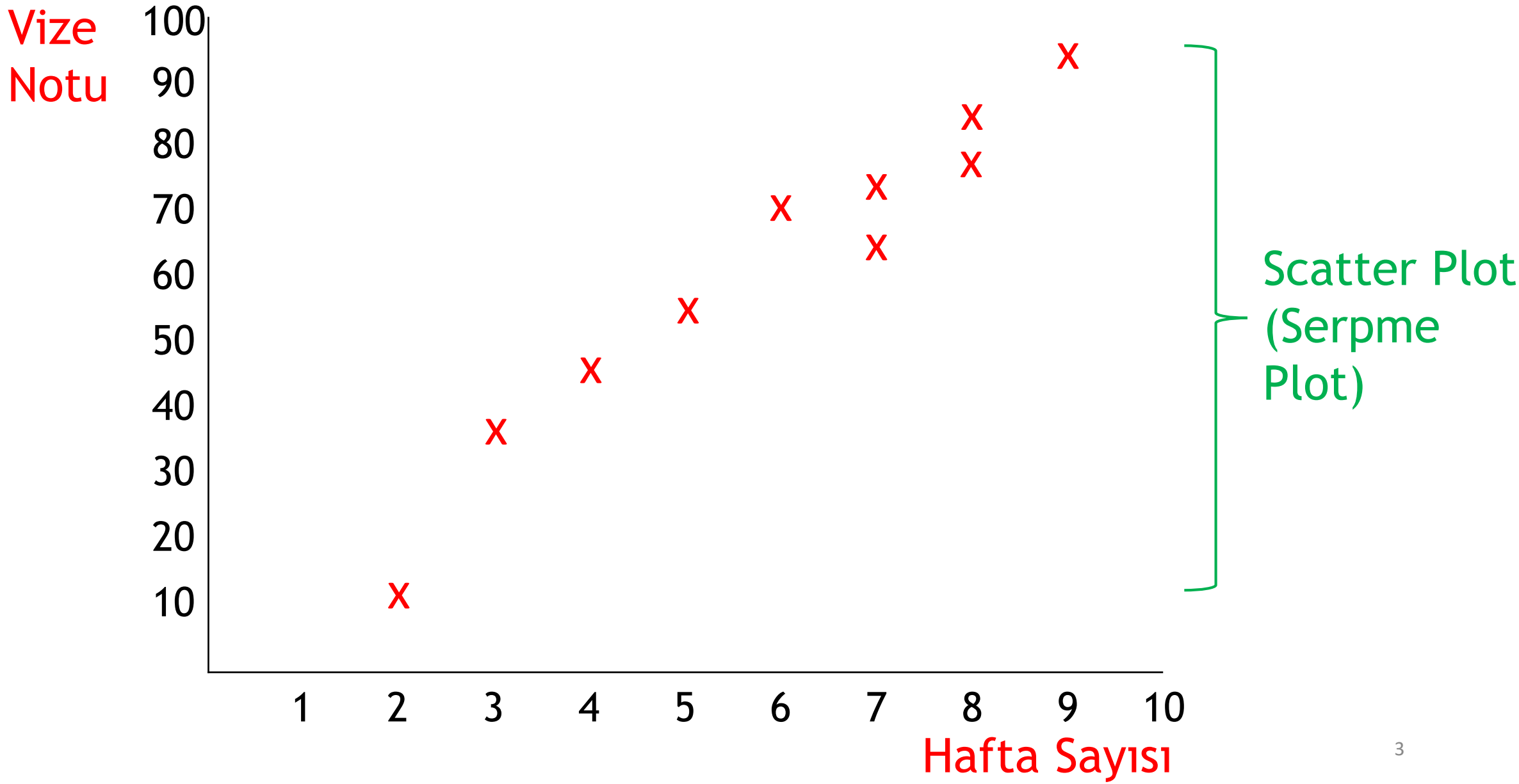
Korelasyon (Correlation)

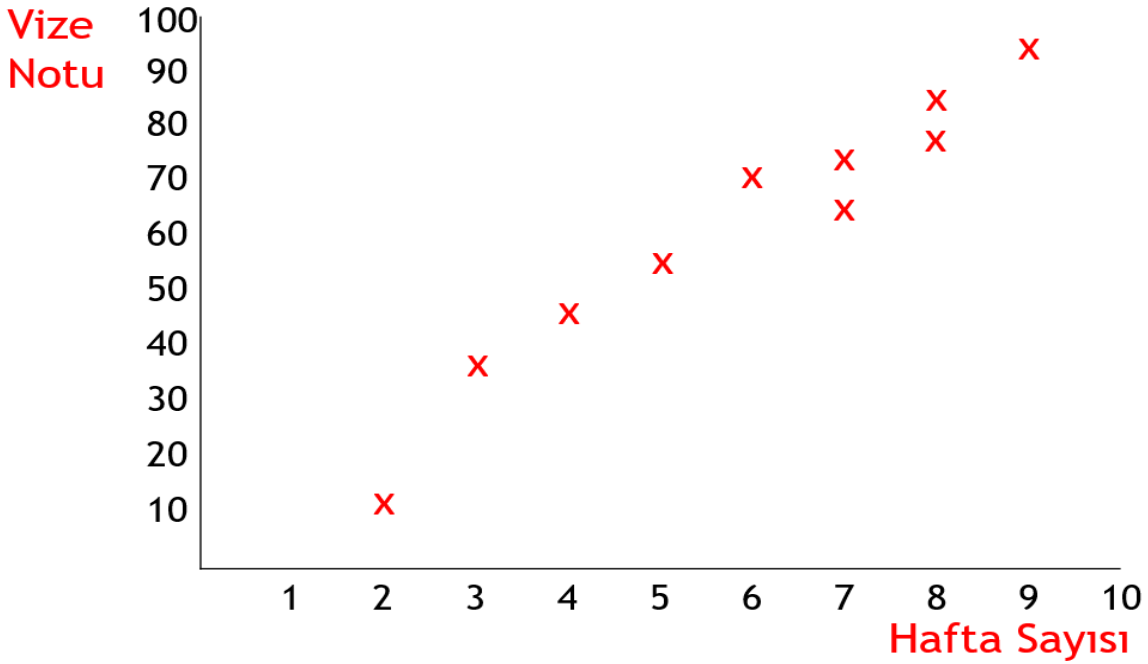
Şimdiye kadar hep tek bir değişken ile ilgilendik (bir değişkenin ortalaması, varyansı…). Şimdi ise diyelimki verimizde birden çok değişken var ve bu değişkenler arasında bir *ilişki* var mı yok mu diye merak ediyoruz.

Örneğin derse katılınan hafta sayısı ve vize notları arasında bir ilişki var mi diye merak ediyoruz. Bunun için 10 kişiden aşağıdaki veriyi topluyoruz.

1	Hafta Sayısı	Vize Notu
2	5	55
3	3	37
4	2	11
5	8	84
6	9	94
7	7	65
8	7	74
9	4	46
10	6	71
11	8	78
12		

Bu veriyi yataydaki değişken hafta sayısı, dikeydeki değişken vize notu olacak şekilde görselleştirirsek:



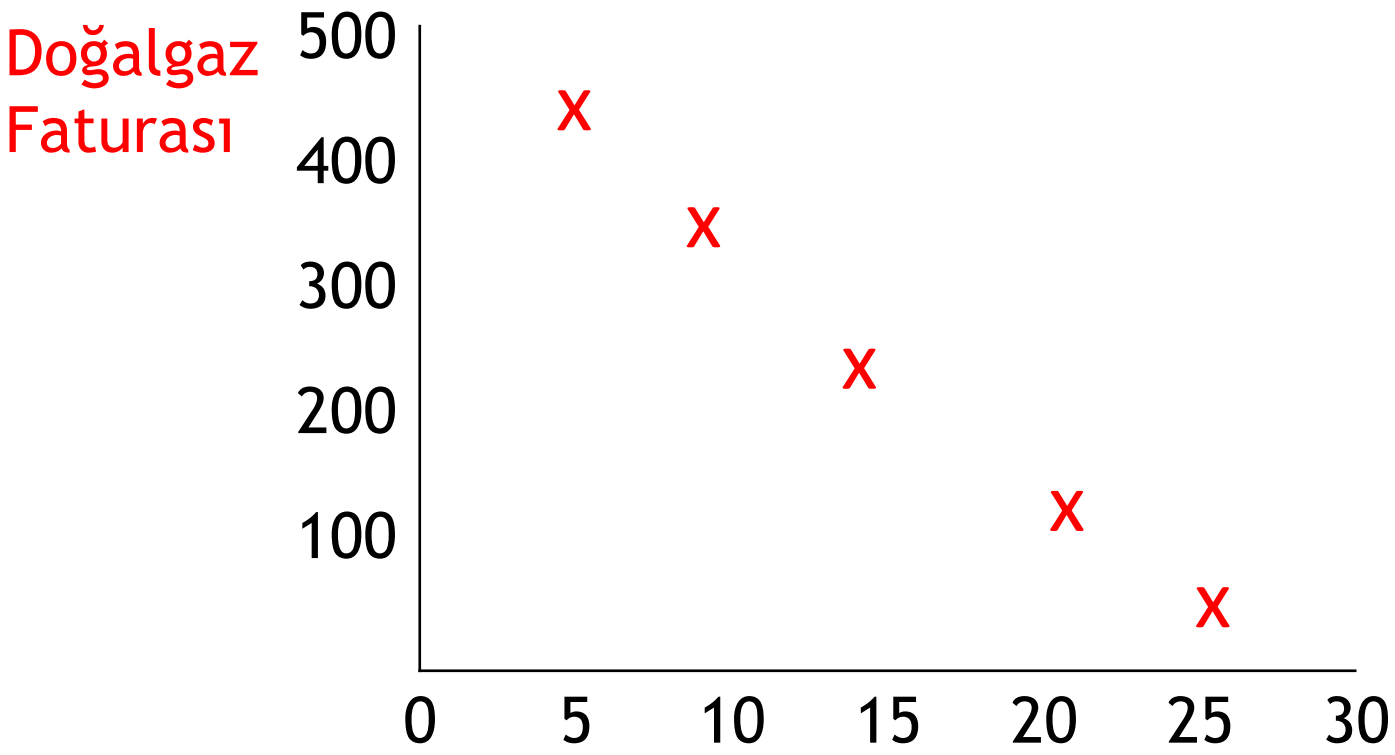


Bu plotta, derse girilien hafta sayısı ile alınan vize notu arasında pozitif bir ilişki görülmektedir. Hafta sayısı arttıkça vize notu artar; hafta sayısı azaldıkça vize notu azalır (iki değişken birlikte hareket ederler). Bu durumda bu iki değişken arasında **pozitif korelasyon** vardır diyeceğiz.

ör.Aşağıdaki tablo aylık ortalama sıcaklık değerlerini ve bu aylardaki doğalgaz faturlarını göstermektedir.

Ortalama Sıcaklık	Doğalgaz Faturası
5	440
9	350
14	220
21	115
26	67

Bu veriyi yataydaki değişken aylık sıcaklık değerleri, dikeydeki değişken doğalgaz faturası olacak şekilde görselleştirirsek:



Bu scatter plotta, aylık ortalama sıcaklık doğalgaz faturasının düştüğü görülür. Burada bir değişken arttıkça diğer değişken azalır (yada bir değişken azalırken diğeri artar). Bu durumda iki değişken arasinda **negatif korelasyon** vardır diyeceğiz.

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Elimizdeki veri $(x_1, y_1),\ \ldots, (x_N, y_N)$ çiftlerinden oluşşsun.

$x_1, \ldots, x_N$ değişkenlerinin ortalaması $\bar{x}$; $y_1, \ldots, y_N$ değişkenlerinin ortalaması $\bar{y}$ olsun.

(x_i, y_i) çiftini ele alalım.

$(x_i - \bar{x})$, x_i değişkeninin kendi merkezi olan $\bar{x}$ 'den sapmasını verir.

$(y_i - \bar{y})$, y_i değişkeninin kendi merkezi olan $\bar{y}$ 'den sapmasını verir.

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımında eğer:

1. $x_i > \bar{x}$ ve $y_i > \bar{y}$ ise bu çarpım pozitif olur (bu durumda her iki değer de merkezinden daha büyük)

Yada

2. $x_i < \bar{x}$ ve $y_i < \bar{y}$ ise bu çarpım yine pozitif olur (bu durumda her iki değer de merkezinden daha küçük).

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Sonuç olarak iki durumda da; iki değişken benzer hareket ediyorlar (ya ikisi birden merkezlerinden daha büyük, ya ikisi birden merkezlerinden daha küçük). O halde (x_i, y_i) çifti benzerdir; ve x ve y değişkenlerinin korelasyonuna pozitif katkıda bulunur.

Tersi olarak

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımında eğer:

1. $x_i > \bar{x}$ ve $y_i < \bar{y}$ ise bu çarpım negatif olur olur (bu durumda birinci değişken merkezinden büyük iken ikinci değişken merkezinden küçüktür; bu iki değişken arasında uyumsuzluk vardır)
2. $x_i < \bar{x}$ ve $y_i > \bar{y}$ ise bu çarpım yine negatif olur olur (bu durumda birinci değişken merkezinden küçük iken ikinci değişken merkezinden büyüktür; bu iki değişken arasında yine uyumsuzluk vardır)

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

O halde

$$(x_i - \bar{x})(y_i - \bar{y})$$

çarpımı negatif iken (x_i, y_i) çifti benzer değildir; ve x ve y değişkenlerinin korelasyonuna negatif katkıda bulunur.

$(x_1, y_1), \ \ldots, (x_N, y_N)$ çiftlerinin x ve y değişkenlerinin korelasyonuna olan katkılarının ortalamasını alırsak:

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Bu değere **kovaryans** denir σ_{xy} ile gösterilir. Fakat kovaryans değeri $(-\infty, +\infty)$ arasında bir değerdir. Yani herhangi bir sınırlanması yoktur. Veri setinden veri setine çok farklılık gösterebilir.

Elde ettigimiz benzerlik değerinin herkesçe anlaşılır olması için, bir anlam ifade etmesi için standart bir aralık olan $[-1\,, 1]$ aralığında yer alması gerekir. Bunun için kovaryansı x'in ve y'nin standart sapmalarına böleceğiz.

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

x'in standard sapması σ_x :

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

y'nin standard sapması σ_y :

$$\sigma_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Kovaryans σ_{xy}, yukarıdaki standart sapmaların çarpımına bölünürse:

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Korelasyonun Hesaplanması (Korelasyonun Sayısal Olarak Ölçülmesi)

Bu ifade sadeleştirilerek korelasyon hesağlanır:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Bu değere *Pearson Korelasyon Katsayısı* diyeceğiz. Bu değer her zaman $[-1, 1]$ arasındadır.

Eğer x ve y değişkenleri arasında negatif korelasyon varsa korelasyon negatiftir $[-1, 0)$ arası değer alır;

eğer x ve y değişkenleri arasında pozitif korelasyon varsa korelasyon pozitiftir $(0, 1]$ arası bir değer alır;

eğer x ve y değişkenleri arasında herhangi bir korelasyon yoksa, korelasyon 0'dır.

ör. Daha önce gördüğümüz hafta – sayısı vize notu arasındaki korelasyonu hesaplayalım.

Hafta Sayısı	Ortalamdan Sapma	Sapmanın Karesi	Vize Notu	Ortalamdan Sapma	Sapmanın Karesi	Sapmaların Çarpımı
5	(5-5.9)=-0.9	0.81	55	(55-61.5)=-6.5	42.25	-0.9 ·-6.5=5.85
3	(3-5.9)=-2.9	8.41	37	(37-61.5)=-24.5	600.25	-2.9·-24.5=71.05
2	(2-5.9)=-3.9	15.21	11	(11-61.5)=-50.5	2550.25	-3.9·-50.5=196.95
8	(8-5.9)=2.1	4.41	84	(84-61.5)=22.5	506.25	2.1·22.5=47.25
9	(9-5.9)=3.1	15.61	94	(94-61.5)=32.5	1056.25	3.1·32.5=100.75
7	(7-5.9)=1.1	1.21	65	(65-61.5)=3.5	12.25	1.1·3.5=3.85
7	(7-5.9)=1.1	1.21	74	(74-61.5)=12.5	156.25	1.1·12.5=13.75
4	(4-5.9)=-1.9	3.61	46	(46-61.5)=-15.5	240.25	-1.9·-15.5=29.45
6	(6-5.9)=0.1	0.01	71	(71-61.5)=9.5	90.25	0.1·9.5=0.95
8	(8-5.9)=2.1	4.41	78	(78-61.5)=16.5	272.25	2.1·16.5=34.65
Ortalama: 5.9		Toplam: 48.9	Ortalama: 61.5		Toplam: 5526.5	Toplam: 504.5

Şu halde korelasyon $\rho = \frac{504.5}{\sqrt{48.9 \cdot 5526.5}} = 0.97$ (çok yüksek pozitif korelasyon)

ör. Daha önce gördüğümüz ortalama sıcaklık - doğalgaz faturası değişkenlerinin korelasyonuna bakalım.

x_i $(i \in \{1, \dots, 5\})$ değişkenleri sıcaklıkları göstersin. Bu değişkenlerin ortalaması: $\bar{x} = 15$

Bu değişkenlerin ortalamadan farklarının karelerinin toplamı:

$$(5-15)^2 + (9-15)^2 + (14-15)^2 + (21-15)^2 + (26-15)^2 = 294$$

y_i $(i \in \{1, \dots, 5\})$ değişkenleri doğalgaz faturalarını göstersin. Bu değişkenlerin ortalaması: $\bar{y} = 238.4$

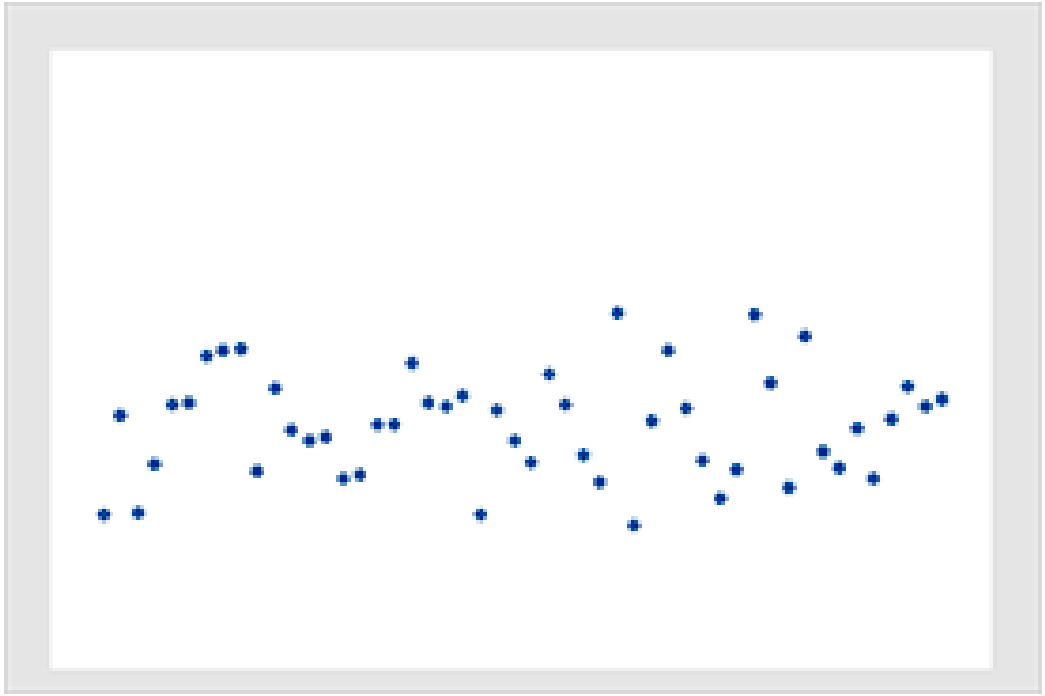
Bu değişkenlerin ortalamadan farklarının karelerinin toplamı:

$$(440-238.4)^2 + (350-238.4)^2 + (220-238.4)^2 + (115-238.4)^2 + (67-238.4)^2 = 98041.2$$

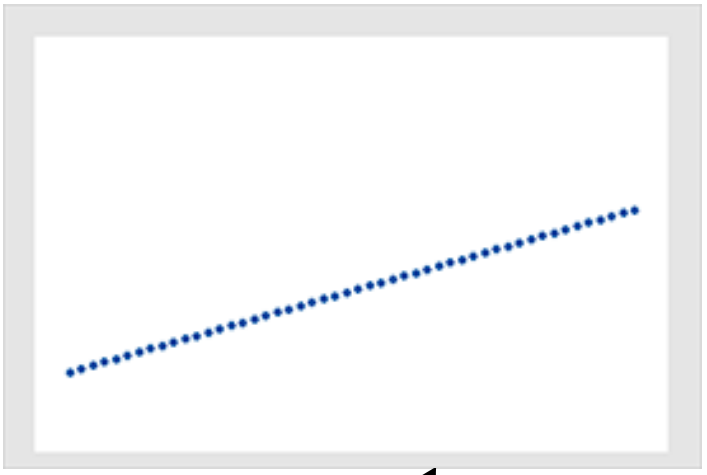
$$(5-15)(440-238.4)+(9-15)(350-238.4)+(14-15)(220-238.4)$$
$$+(21-15)(115-238.4)+(26-15)(67-238.4)=-5293$$

Korelasyon $\rho=\frac{-5293}{\sqrt{294\cdot 98041.2}}=-0.98$ (çok yüksek negatif korelasyon)

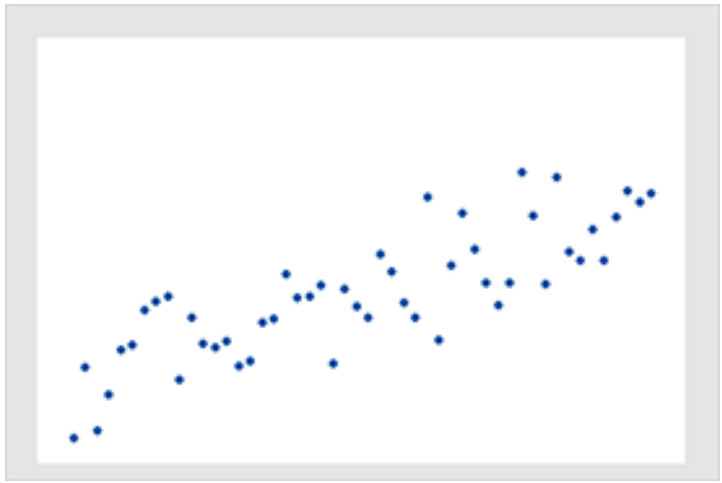
ör. Aşağıda gösterilen scatter plotta iki değişken arasında herhangi bir korelasyon yoktur ($\rho=0$).



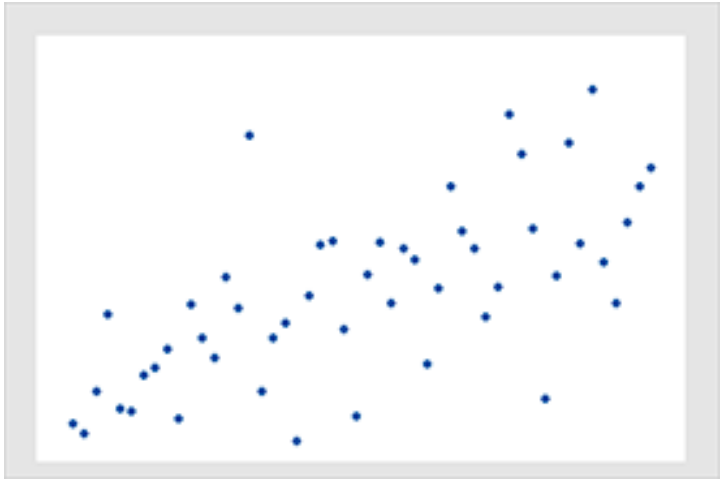
Farklı Büyüklükte Korelasyonlar



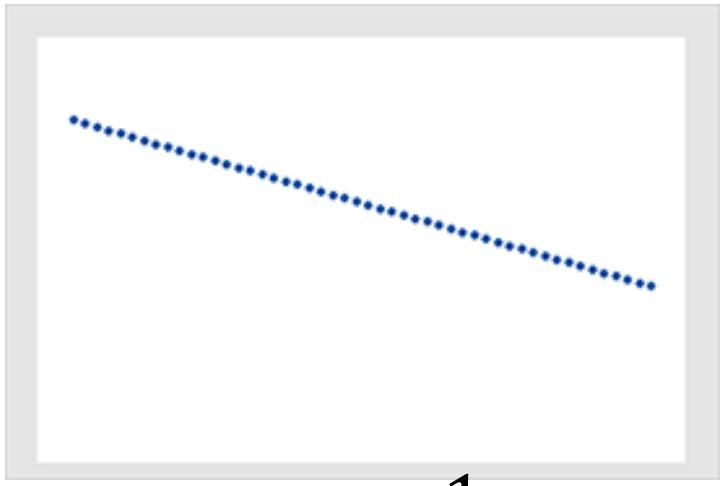
$\rho = 1$



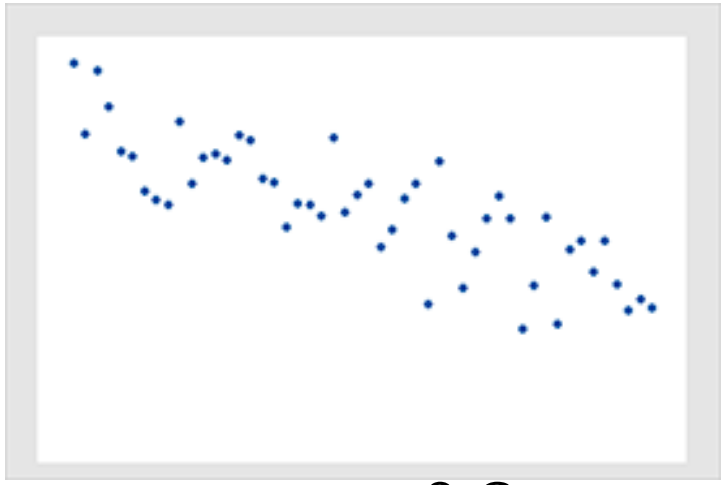
$\rho = 0.8$



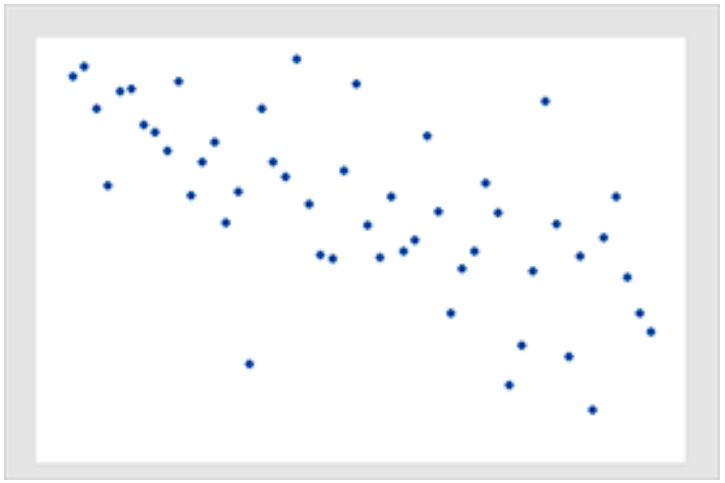
$\rho = 0.6$



$\rho = -1$



$\rho = -0.8$



R

R basit arayüzlü, tamamen açık kaynak kodlu bir istatistik programıdır.

R ile ileri derece hesaplama yapabilir, elinizdeki veriyi görselleştirebilir, manipüle edebilir, değiştirebilirsiniz.

R'da birçok hazır fonksiyon mevcuttur. Bu fonksiyonlar genelde package (yani bir bakima library) icinde bulunur. Bunun haricinde bildiğimiz anlamda kendi fonksiyonlarinizi da yazabilirsiniz.

R'da komut girecegimiz konsol `>` yada `>>` satırı ile komut alır.

Bir de fraklı olarak R'da atama (assign etme) operatörü `<-` 'dır. Bunu sağdan sola bir ok gibi düşünün. Zaten herhangi bir programlama dilinde atama her zaman soldan sağa olur

```
> x<-3
```

Not:Isterseniz atamayı bildiğimiz = ile de yapabilirsiniz, program hata vermez

```
>x=3
```

R Studio

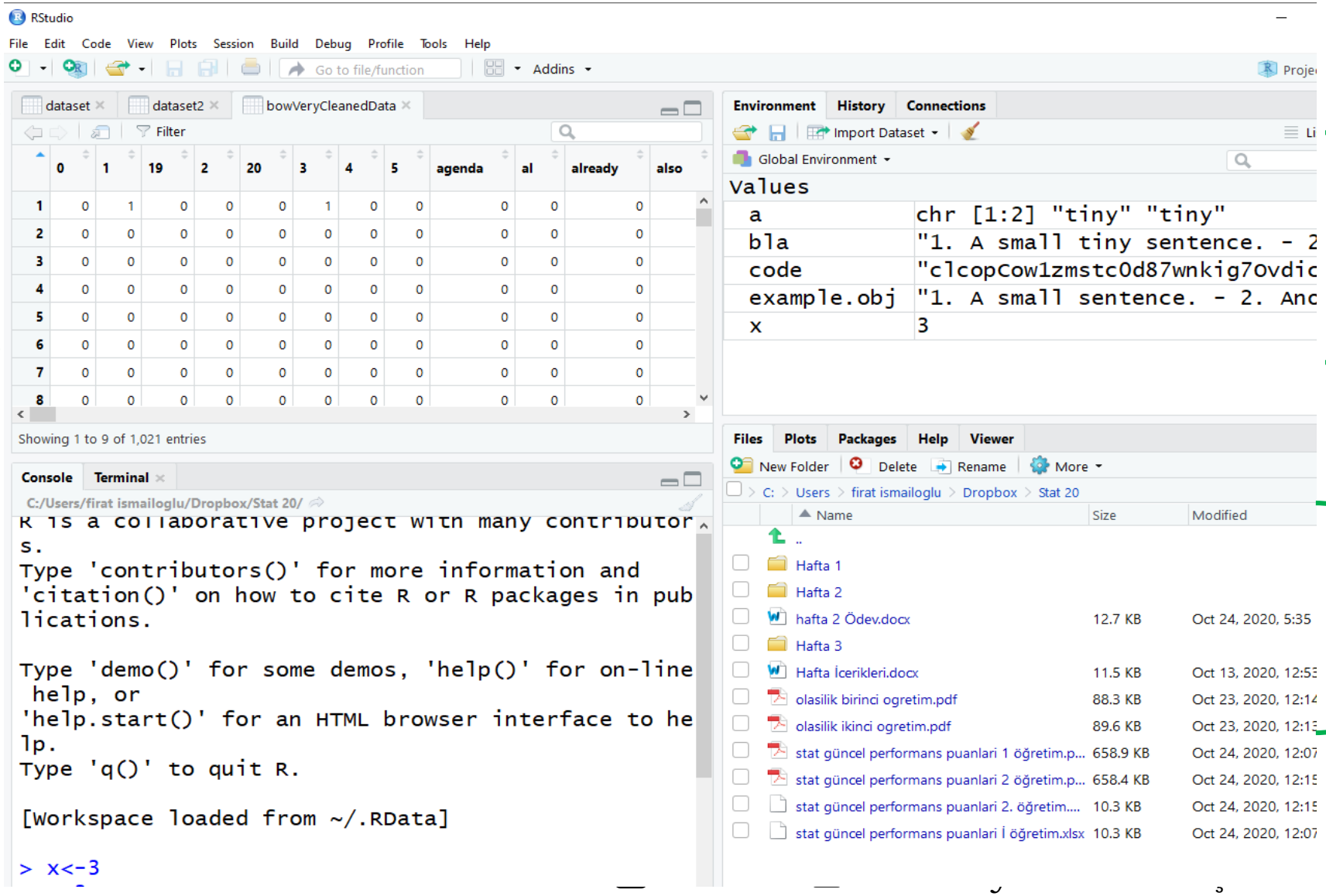
R'ı indiridiğinizde çok basit bir arayüzle gelir. Bu arayüz her türlü komutu çalıştırmak için yeterli olsada, biz kullanım kolaylığı açısından R'ın en poüler ide'lerinden biri olan R Studio'yu kullanacağız.

Veri setlerinin ön gösterimi

R Konsolu

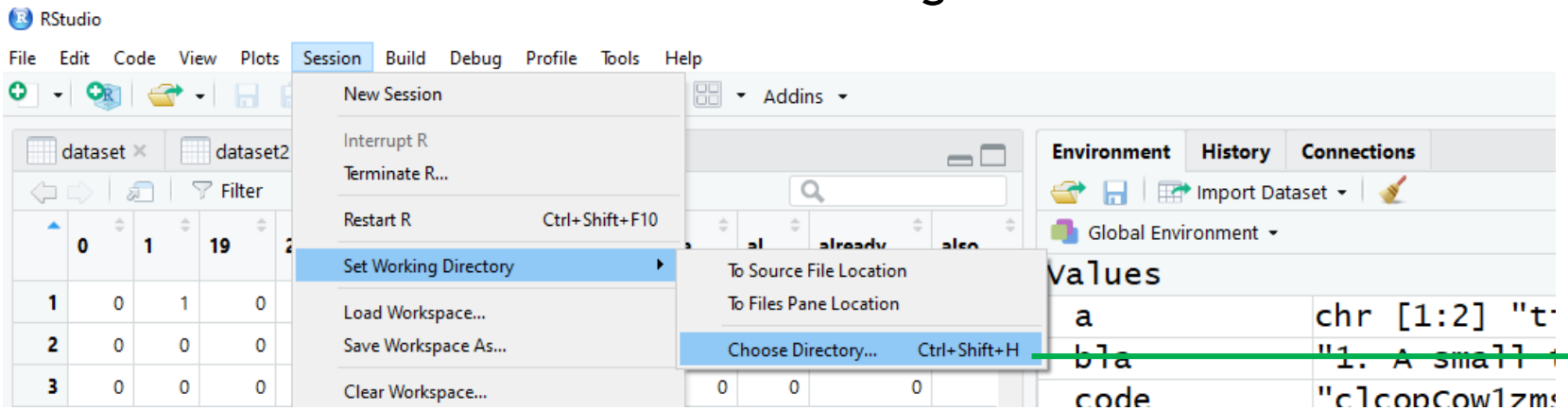
yüklü değişkenler

working directory'deki dosyalar



R Studio'da Working Directory'i Ayarlamak

R'da normalde `getwd()` komutu ile o anki working directory (wd) görülebilir; `setwd()` komutuyla, örneğin `setwd("C:/Users/firat/Dropbox/Stat 20")`, komutuyla wd değiştirilebilir. Fakat R studio'da bu komutu hatırlamak zorunda değilsiniz.



Working directory'i buradan değiştirebilirsiniz.

R'da Package Yüklemek

R'da `install.packages("<paket adı>")` komutu ile package (yani library) yukleyebiliriz. Ornegin en meshur paketlerden biri data gorsellestirmede kullancagimiz ggplot2 iki paketidir:

```
>install.packages("ggplot2")
```

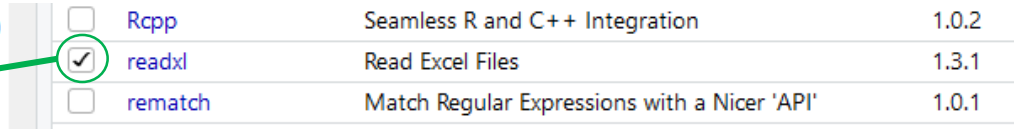
Bir paketin yüklenmesi (install) edilmesi demek o paketi hemen kullanabilecegimiz anlamina gelmez. Bunun için o paketi library komutuyla çağırmalıyız:

```
>library("ggplot2")
```

R Studio'da Package Yüklemek

Daha önce `install.packages()` komutuyla yüklenmiş paketler R Studio'ile şu şekilde de çağrılabilir:

```
> library("readxl", lib.loc="~/R/win-library/3.5")
>
>
```



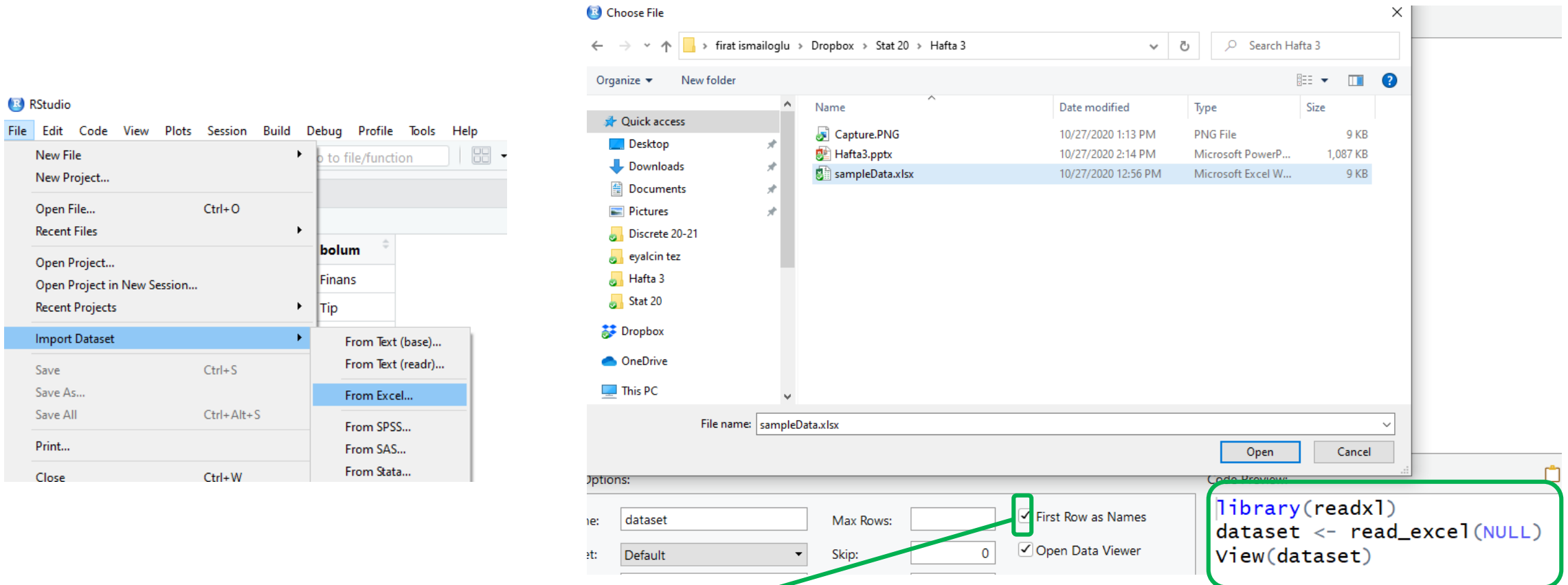
R Studio'da sağ altta yer alan package tab'ında yüklü paketler gösterilir. Burdan çağırmak istediğiniz paket'i tiklarsaniz, konsolda otomatik olarak library komutuyla bu paketi cagirmis olursunuz.

R Studio'da Data Yüklemek (import/read etmek)

R'da txt, .csv, .xlsx ve SPSS dataları kolayca import edilebilir. Zaten normalde veriyi elimizle tek tek oluşturmayız. Onun yerine bize verilen bir data ile uğraşır, bu datayı R'da çağırırız ederiz.

Örneğin xlsx uzantılı bir excel dosyasını çağrıken, önce library(readxl) komutuyla ilgili paketi yuklemek daha sonra `read_excel` komutuyla working directory'de olan datayi yukleriz. R Studio'da bu arayüzü sayesinde daha kolay yapılır:

R Studio'da Data Yüklemek (import/read etmek)



Eğer excel'deki ilk satırınız da kolon adları yoksa bunu seçmeyiniz. Bunu seçerseniz col_names=TRUE olur secmezseniz col_names=FALSE olur)

Burda kodu kendi oluşturdu.

R’da Merkezi Eğilim ve Korelasyon Hesabı

R’da bir yukledigimiz Excel dosyasi "data table" olarak tutulur, yani degisken turu data table’dır. Şu şekilde bir sampleData.xlsx adında Excel dosyamız olsun:

isim	yas	boy	agirlik	bolum
Yamac Kocovali	30	180	70	Finans
Ates Hekimoglu	44	189	76	Tip
Miran Aslanbey	24	186	74	Lojistik
Asya Arslan	39	176	50	Tip
Ali Vefa	34	179	76	Tip

Simdi bu datayi oyuncular adı ile R’da okuyalım (isterseniz direkt R studio ile okuyabilirsiniz, boylece sintaksi hatirlamaniz gerek kalmaz)

```
>oyuncular<-read_excel("sampleData.xlsx",col_names=TRUE)
```

names komutu bir data table’daki colon adlarini getirir.

```
>names(oyuncular)
 "isim" "yas" "boy" "agirlik "bolum"
```

Tek bir kolona erismek istedigimizde $ isaretini kullaniriz. Örnegin boy kolonunu cağıralım:

```
>oyuncular$boy
 "isim" "yas" "boy" "agirlik "bolum"
```

R'da Merkezi Eğilim ve Korelasyon Hesabı

Boyun ortalamasini alalım:

```
>mean(oyuncular$boy)
```

```
182
```

Benzer olarak var ile varyasyon, sd ile standart sapma hesaplanabilir.

Boy ve Agirlik kolonlarinin kovaryansini hesaplayalim:

```
>cov(oyuncular$boy,oyuncular$agirlik)
```

```
40
```

Simdi ise boy ve agirlik arasindaki Pearson korelasyon katsayisini hesaplayalim. Hatirlarsak bunun icin kovaryans, standart sapmalarin carpimlarina bolunurdu:

```
>cov(oyuncular$boy,oyuncular$agirlik)/(sd(oyuncular$boy)*sd(oyuncular$agirlik))
```

```
0.68
```

Yada direkt cor komutuyla Pearson korealasyonu hesaplnabilir:

```
>cor(oyuncular$boy,oyuncular$agirlik)
```

```
 0.68
```

R’da Merkezi Eğilim ve Korelasyon Hesabı

Not: Bir data table’da kolonları (yada satırları) indekslerle, yani kolon (satir) numaralariylada cagirabiliriz. Burada önemli nokta numarlandirmanin 1’den basladigidir

Birinci kolonu cagiralim:

```
>oyuncular[,1]
# A tibble: 5 x 1
  isim <chr>
1 Yamac Kocovali
2 Ates Hekimoglu
3 Miran Aslanbey
4 Asya Arslan
5 Ali Vefa
```

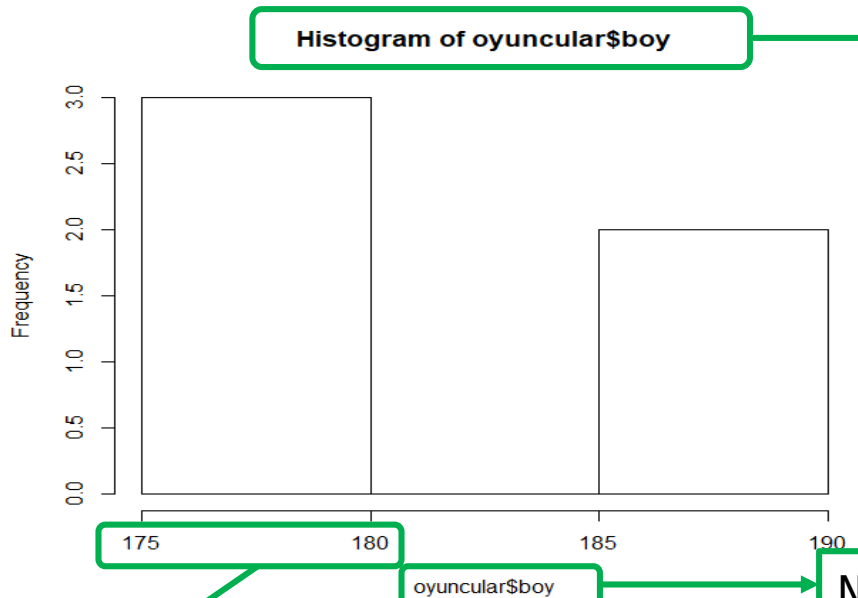
Ikinci satiri cagiralim:

```
>oyuncular[2,]
isim  yas   boy agirlik bolum
  <chr>      <dbl> <dbl>   <dbl> <chr>
Ates Hekimoglu 44   189      76    Tip
```

R’da Histogram

R’da histogram hist komutuyla yapılır.

```
>hist(oyuncular$boy)
```



Histogramin basligi (main parametresiyle degistirilebilir)

Neyin histogrami oldugu xlab parametresiyle degistirilebilir

Bin genisligini R otomatik olarak 5 secmis, başka bir seyde secebilirdi, data’dan dataya degisir. Biz istersek bunu break parametresiyle degistirebiliriz

R'da Histogram

hist komutunun parametreleri:

* `col` (renk icin) örnegin: `col="green"`

* `xlab` (x ekseninie isim (label) vermek icin or. `xlab="Oyuncularin Boylari"`

* `main` : Histograma başlık eklemek icin: ör. `main= "BOYLAR"`

* `xlim`: x eksenin nerden baslayip nerde bitecegini belirlerken ör. `xlim=c(160,180)`

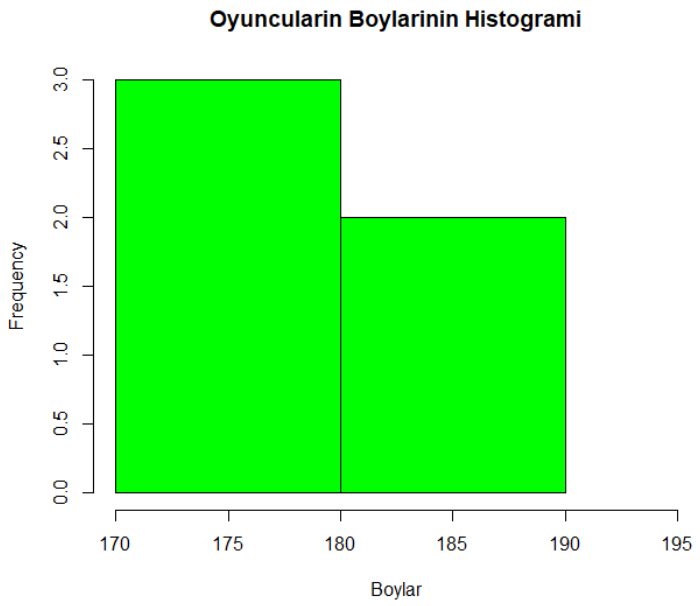
* `breaks`: bin (kutu) sayisini ve istenirse kutu genisligini ayarlar. Örnegin `breaks=3` olarak belirtilirse 4 adet kutu oluşturulabilir. Break sayisindan bir fazla kutu oluşturulur. Çunku aslında breaks kırılma (bir kutunun bitip diger kutunun basladigi nokta) demektir. 3 adet kırılma olması için 4 adet kutu olmasi gerekir.

Breaks ile bin (kutu) genisliği de ayarlanabilir. Örnegin 170-190 arasini 10'ar 10'a kutulamak istiyoruz. Bu durumda `breaks=seq(170,190,10)` parametresini eklememiz gerekir. (seq burada sequence: sıra/dizi 'nin kısaltması)

Önemli Not: hist komutunun icinde bu parametreler herhangi bir sırayla yer alabilir, hangisinin once geldigi onemli degildir!

R'da Histogram

```
>hist(oyuncular$boy, col="green", xlab="Boylar", main="Oyuncularin Boylarinin Histogrami", xlim=c(170,195),breaks=1)
```




```
>hist(oyuncular$boy, col="green", xlab="Boylar", main="Oyuncularin Boylarinin Histogrami", xlim=c(170,195),breaks=seq(170,190, 2))
```