

R ile Programlamaya Giriş ve Uygulamalar



Mustafa Gökçe Baydoğan
Berk Orbay
Endüstri Mühendisliği Bölümü
Boğaziçi Üniversitesi

Yöneylem Araştırması ve Endüstri Mühendisliği Kongresi
12 Eylül 2015
Orta Doğu Teknik Üniversitesi, Ankara

İçerik

- Giriş
- R'ye genel bakış
 - R dili
 - R nedir, ne değildir? Neden R?
 - Arayüz
 - Çalışma alanı
 - Yardım
- R ile çalışmak
 - Paketler
 - Veri okuma/yazma
 - İşleme
 - Grafik oluşturma
 - Uygulamalar
- Sonuç

R'a genel bakış

R dili ve tarihi

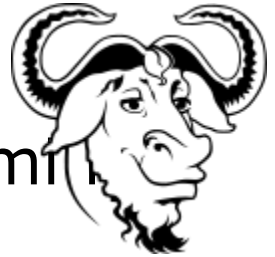
- Temeli 1976 yılından bu yana Bell Laboratuvarları'nda istatistiksel programlama dili olarak geliştirilen **S** diline dayanır.
 - UNIX ile aynı zamanda geliştirilmeye başlandı.
 - Araştırma ve veri analizi için geliştirilmiştir.
 - Sonraları lisanslı olarak S-Plus olarak piyasa sürülmüştür.
- **S** diline benzer ama açık kaynaklı bir platform olarak **R** dili 1990'lı yıllara Yeni Zelanda'daki Auckland Üniversitesi İstatistik Bölümü'nden **Ross Ihaka** ve **Robert Gentleman** tarafından yazılmıştır.
- Daha sonra dünyanın çeşitli yerlerindeki araştırmacılar **R**'yi geliştirmek için bir araya gelmiş ve 1997'de bu gruba "R core team" adı verilmiştir.
- **R** dilinin ilk sürümü "R core team" tarafından 29 Şubat 2000 tarihinde yayınlanmıştır.
- Her iki-üç ayda bir sürümler güncellenmektedir.
 - En son sürümü "R version 3.1.2 (*Pumpkin Helmet*)" 31 Kasım 2014'de yayınlanmıştır.

R'ye genel bakış

R nedir, ne değildir?

- R GNU S'dir.

- Veri işleme, hesaplama ve grafik gösterimi için bir dil ve çevre sağlar.



- Geniş bir yelpazede istatistiksel ve grafiksel teknikleri içerir.

- doğrusal ve doğrusal olmayan modelleme, klasik istatistik testleri, zaman-serileri analizi, sınıflandırma, kümeleme, ...

- Açık kaynak kodlu olması itibariyle geliştirilmeye çok yatkındır.

R'ye genel bakış

R nedir, ne değildir?

- ❑ R dilinin söz dizimi kuralları (syntax) C diline benzerlik gösterir. Fonksiyonel bir programlama dili olan R istatistikçiler ve matematikçiler için kod yazmayı kolaylaştıran fonksiyonlara sahiptir.
- ❑ R, yaygın olarak kullanılan SPSS, SAS gibi istatistik paket programlarının aksine **istatistiksel yazılım geliştirme ortamıdır**.
- ❑ Etkin veri işleme ve saklama özelliğine sahiptir.
- ❑ Dizi ve özellikle matris hesaplamalarında kullanılabilecek **özel operatörler** mevcuttur.
- ❑ Veri analizi için kullanılabilecek **uyumlu ve bir arada kullanılabilen** araçlar içerir.
- ❑ Veri çözümlemede kullanılabilecek grafiksel araçlara sahiptir.

Kaynak: A. F. Özdemir, E. Yıldıztepe ve M. Binar, **Akademik Bilişim 2010**

R'ye genel bakış

R nedir, ne değildir?

■ Özetle R

- Bir programa dilidir.
- İstatiksel bir pakettir.
- Bir yorumlayıcıdır (interpreter).
- Özgür bir yazılımdır.

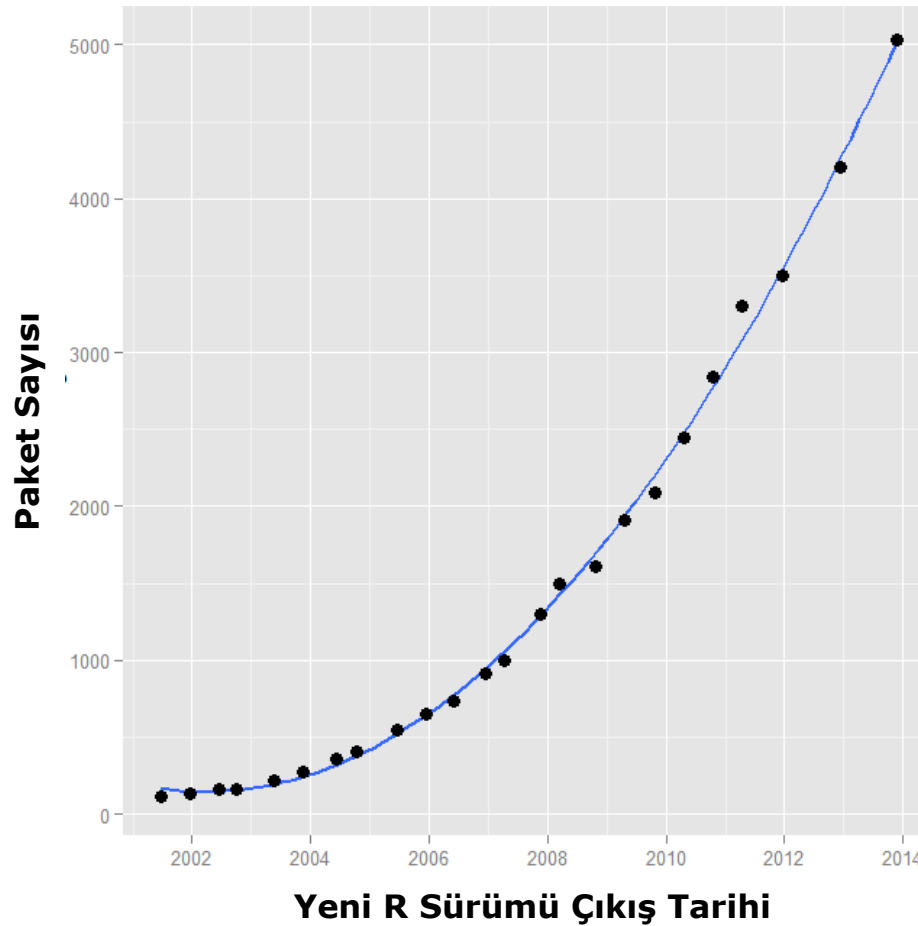


■ Fakat R

- Bir veri tabanı değildir ama veri tabanlarına bağlanabilir.
- Kullanıcı dostu olmasa da java gibi diller aracılığıyla arayüz desteğine sahip bir yazılım geliştirme ortamıdır.
- Tablolardan oluşan yazılım paketi (Excel, Minitab gibi) değildir ama bunlara bağlanabilir.
- Profesyonel veya ticari desteğe tabi bir yazılım değildir.
- Kapalı kutu yazılımlardan oluşan bir yazılım değildir.

R'ye genel bakış

Neden R?

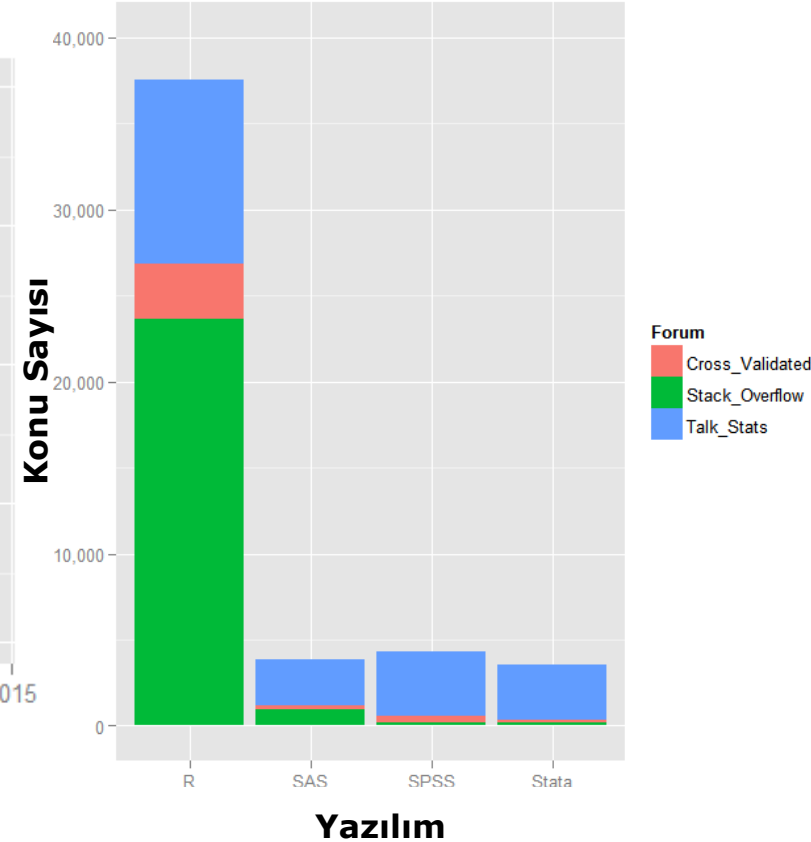
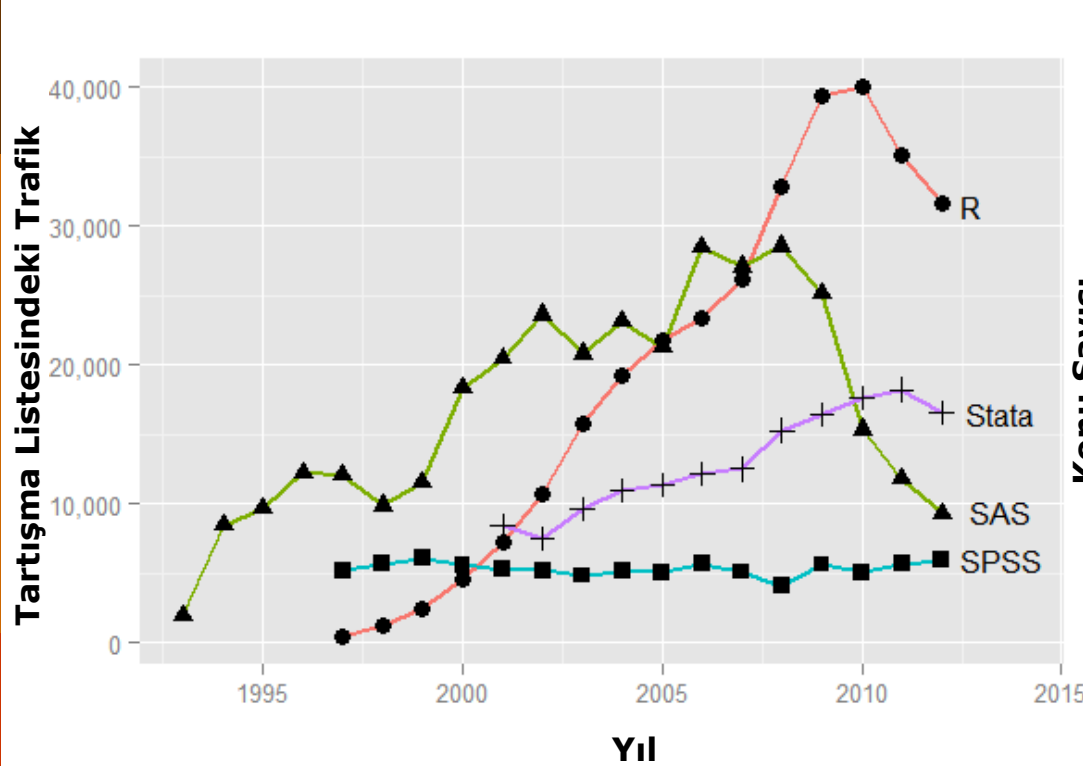


□ Kaynak

- <http://r4stats.com/>
- <http://blog.revolutionanalytics.com/r-is-hot/>

R'ye genel bakış

Neden R?



■ Kaynak

- <http://r4stats.com/>
- <http://blog.revolutionanalytics.com/r-is-hot/>

R'ye genel bakış

Neden R?

■ Artıları

- Hızlı ve ücretsiz
 - Hesaplama yoğun işlemlerde başarı
- Güncel
 - İstatistik alanında çalışan araştırmacılar algoritmalarını R ortamında paylaşmaktalar.
 - Yaygın kullanım ve kullanıcı desteği
- Analizin nasıl yapılması gerektiği hakkında düşündürür.
- Diğer diller ve programlar ile bağlantı desteği
- İşletim sisteminden bağımsız olarak çalışır.

■ Eksileri

- Öğrenme süreci uzundur.
 - Profesyonel destek eksikliği problemlerin kullanıcı tarafından çözülmesini gerektirir.
- Kullanıcı dostu değildir.
 - Basit seviyede bir kullanıcı arayüzüne sahiptir.
- Hata yapmak kolaydır ve tespit edebilmesi zor olabilir.
- Veriyi işlenecek hale getirmek zaman alıcı ve hataya açık bir süreçtir.
- Tüm işlemler hafızada gerçekleştirilir.
 - Çok büyük veriler fazla RAM gerektirir.

R'ye genel bakış

R'ye giriş

□ Yükleme

■ R-Project web sayfası

□ <http://www.r-project.org/>

Windows, Linux, Mac OS X, source

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2014-10-31, Pumpkin Helmet) [R-3.1.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

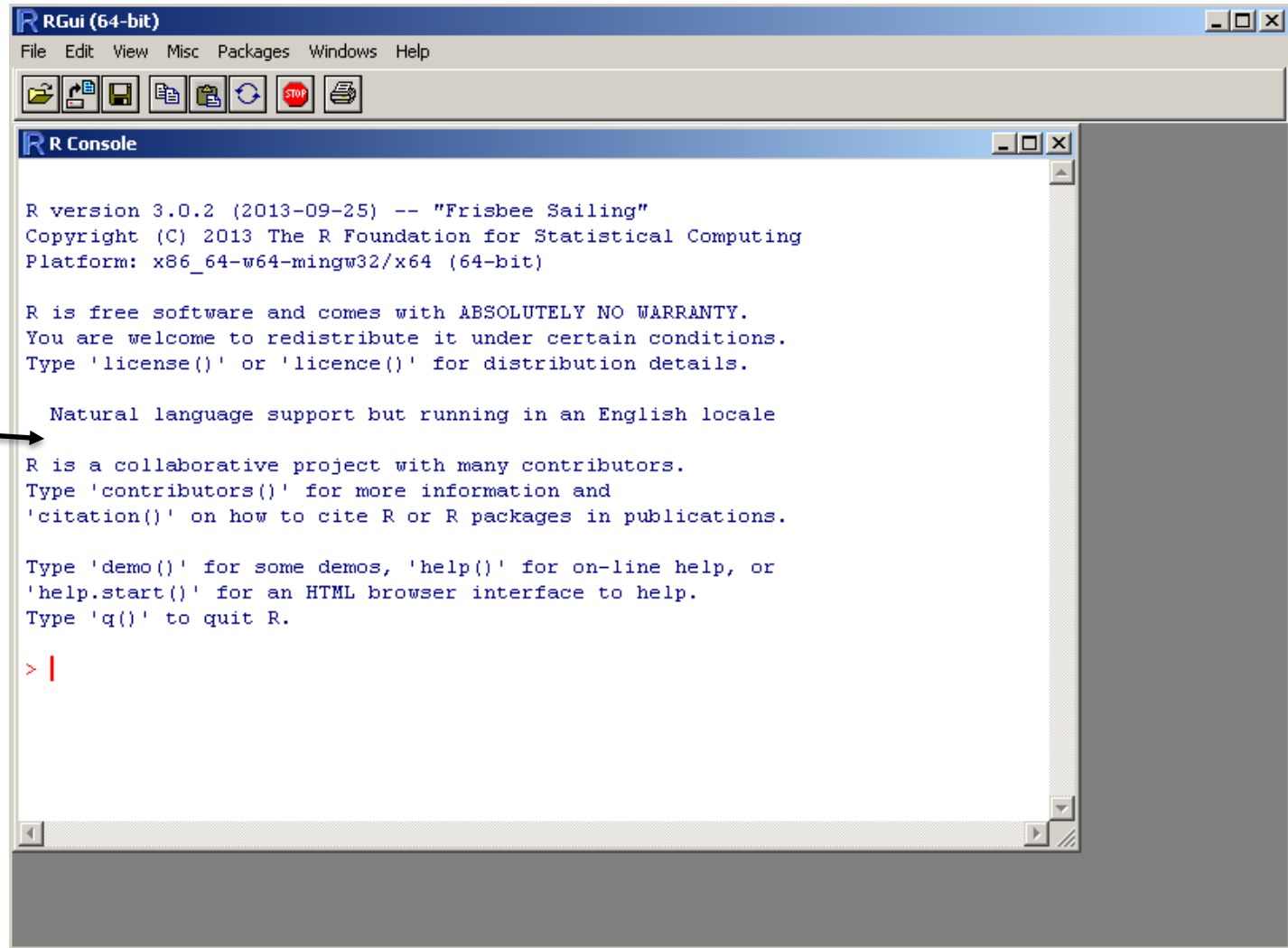
Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

R'ye genel bakış

R Ara yüzü

R terminali



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

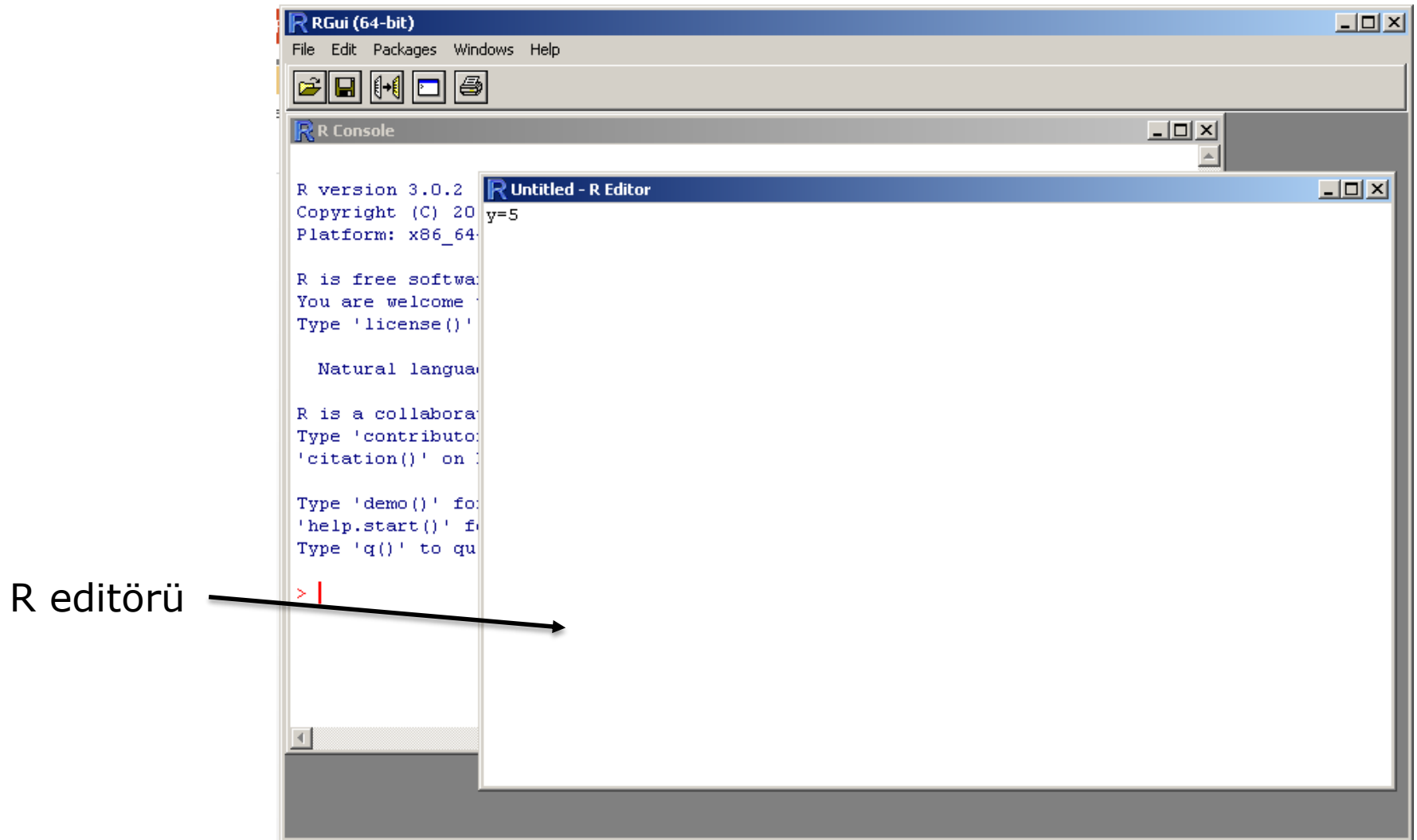
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

R'ye genel bakış

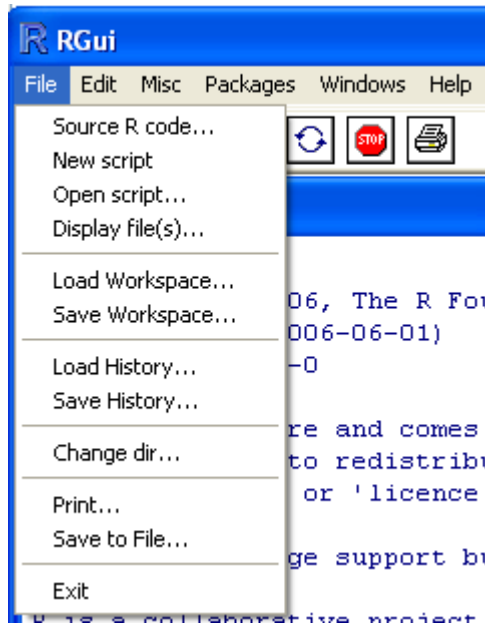
R Ara yüzü



R'ye genel bakış

R Ara yüzü

- R dilinde komut satırına girilen söz dizim kuralları (syntax) aynı zamanda metin dosyalarına da yazılabilir.
- Bu durumda metin dosyası uzantısı "*.R" olarak kaydedilir. Bu şekilde kaydedilmiş bir dosya artık R script dosyasıdır.



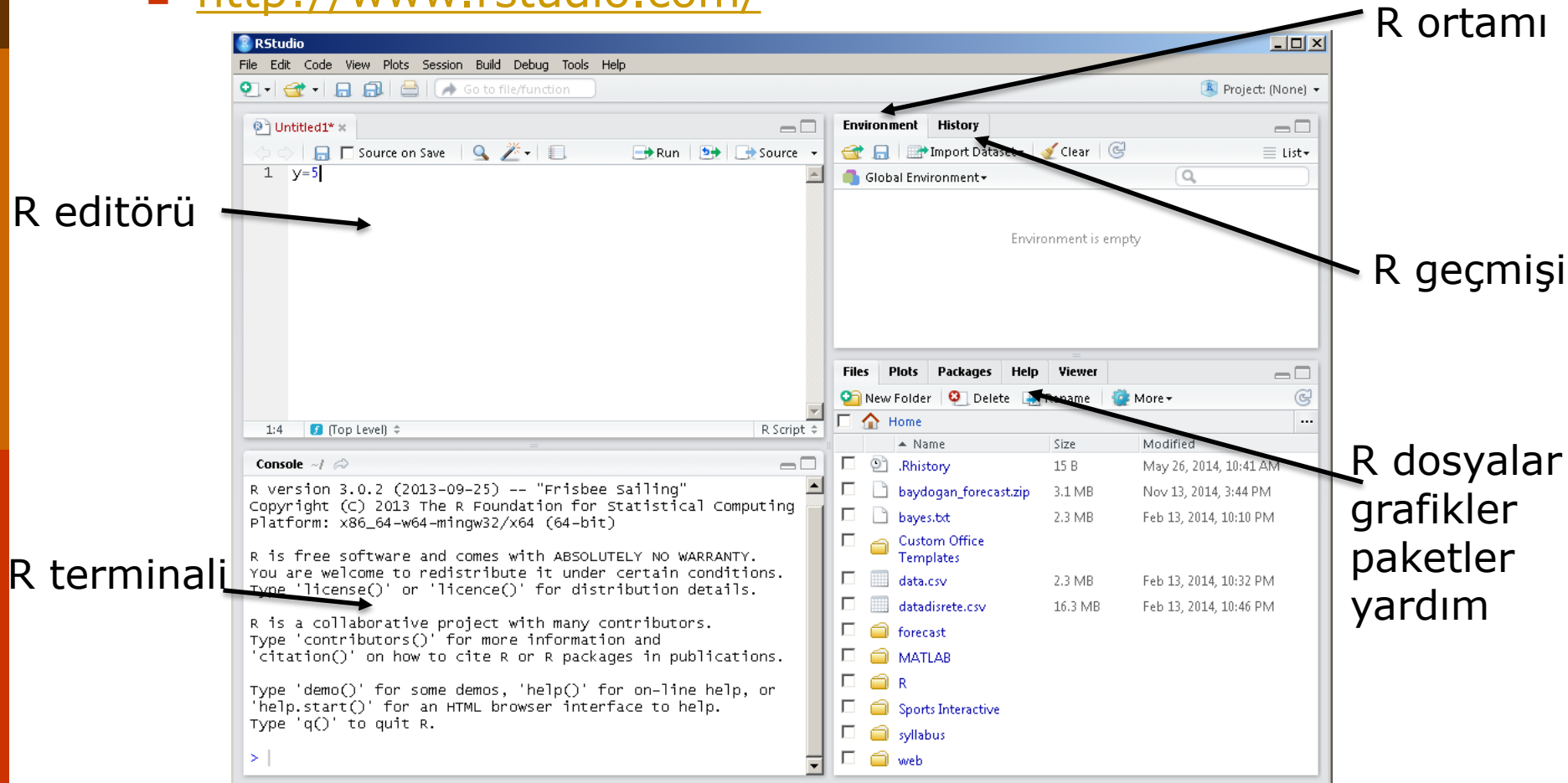
Kaynak: A. F. Özdemir, E. Yıldıztepe ve M. Binar, Akademik Bilişim 2010

R'ye genel bakış

Alternatif editörler ve ara yüzler

■ En yaygın kullanılan editör + ara yüz RStudio'dur.

■ <http://www.rstudio.com/>



R'ye genel bakış

Alternatif ücretsiz editörler ve ara yüzler

- ❑ Geany
 - <http://www.geany.org/>
- ❑ Notepad++
 - <http://notepad-plus-plus.org/>
- ❑ RWinEdt
 - <http://cran.r-project.org/web/packages/RWinEdt/index.html>
- ❑ Tinn-R
 - <http://sourceforge.net/projects/tinn-r/>
- ❑ JGR (R için Java ara yüzü)
 - <http://www.rforge.net/JGR/>
- ❑ Emacs + ESS
 - <http://www.gnu.org/software/emacs/>
 - <http://ess.r-project.org/>
- ❑ Rattle
 - <http://rattle.togaware.com/>
- ❑ Playwith (grafikler için)
 - <https://code.google.com/p/playwith/>

R'ye genel bakış

R dili

■ Örnekler (Hesap makinesi olarak R)

```
> log2(32)
```

```
[1] 5
```

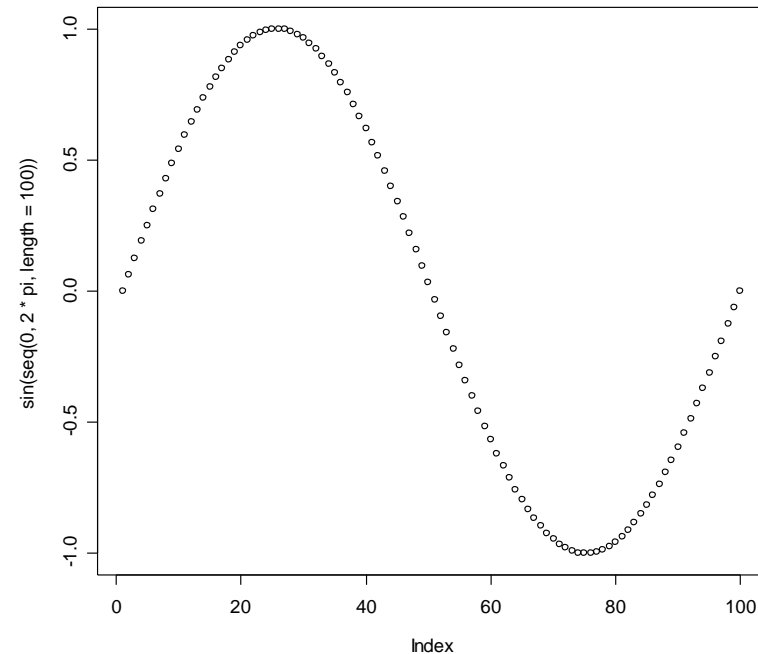
```
> sqrt(2)
```

```
[1] 1.414214
```

```
> seq(0, 5, length=6)
```

```
[1] 0 1 2 3 4 5
```

```
> plot(sin(seq(0, 2*pi, length=100)))
```



R'ye genel bakış

R dili

- R, belleğe direkt erişim yerine özel veri yapılarını kullanır.
- R'deki temel nesne türleri:
 - numeric
 - integer, double, complex
 - character
 - logical
 - function
- Bu nesneler kullanılarak aşağıdaki objeler oluşturulabilir
 - **Vektörler:** aynı tipte nesneleri barındıran dizilerdir.
 - **Listeler:** Listeler de vektördür ancak listedeki elemanlar farklı tiplerde olabilir.

R'ye genel bakış

R dili

- Değişkenleri çalışma sırasında tanımlanır.
 - Önceden tanımlamaya gerek duyulmaz.

```
> a = 49
> sqrt(a)
[1] 7

> a = "Kedi ödevimi yedi"
> sub("Köpek", "Kedi", a)
[1] "Köpek ödevimi yedi"

> a = (1+1==3)
> a
[1] FALSE
```

numeric

function

character string

logical

R'ye genel bakış

R dili

▣ Vektörler, matrisler ve diziler

```
> a = c(1,2,3)
> a
[1] 1 2 3
> a[1]
[1] 1
> a[-1]
[1] 2 3
> a[2]
[1] 2
> a[4]
[1] NA
> a[5]="c"
> a
[1] "1" "2" "3" NA  "c"
> a[10]="deneme"
> a
[1] "1" "2" "3" NA  "c" NA NA  NA
[9] NA  "deneme"
> length(a)
[1] 10
```

NA (not available)



R'ye genel bakış

R dili

▣ Vektörler, matrisler ve diziler

■ Operatörler

▣ <-

▣ =

```
> x <- c(0,1,2,3,4)
```

```
> x
```

```
[1] 0 1 2 3 4
```

```
> y = 1:5
```

```
> y
```

```
[1] 1 2 3 4 5
```

```
> median(x = 1:10)
```

```
> x
```

```
Error: object 'x' not found
```

```
> median(x <- 1:10)
```

```
> x
```

```
[1] 1 2 3 4 5 6 7 8 9 10
```

```
> a = c(1,2,3)
```

```
> a
```

```
[1] 1 2 3
```

```
> a[1]
```

```
[1] 1
```

```
> a[-1]
```

```
[1] 2 3
```

```
> a[0]
```

```
numeric(0)
```

```
> a[2]
```

```
[1] 2
```

```
> a[4]
```

```
[1] NA
```

```
> str(a)
```

```
num [1:3] 1 2 3
```

R'ye genel bakış

R dili

▣ Vektörler ile matematiksel işlemler

```
> x <- c(0,1,2,3,4)
> y <- 1:5
> z <- 1:50
> x + y
[1] 1 3 5 7 9
> x * y
[1] 0 2 6 12 20
> x * z
[1] 0 2 6 12 20 0 7 16 27 40 0
[12] 12 26 42 60 0 17 36 57 80 0 22
[23] 46 72 100 0 27 56 87 120 0 32 66
[34] 102 140 0 37 76 117 160 0 42 86 132
[45] 180 0 47 96 147 200
```

R'ye genel bakış

R dili

- Vektör: aynı tipe sahip veriler topluluğu
 - `a = c(1,2,3)`
- Matris: aynı tipe sahip iki boyutlu veri
 - `a = matrix(0,5,10)`
 - Örnek: 5 öğrencinin 10 günlük yoklama bilgisi
- Dizi: ikiden daha fazla boyutlu matris
 - `a = array(1:60, dim=c(3,4,5))`
 - Örnek: Renkli resim
 - R, G, B (Kırmızı, Yeşil ve Mavi) kanallarındaki piksel yoğunlukları

R'ye genel bakış

R dili

- ❑ Liste: farklı tipte sıralı veriler topluluğu
- ❑ Genel olarak vektörler indeks (sayı) ile listeler ise elemanlarının isimleriyle erişilir.
 - Listeler indeksi de destekler.

```
> denemeList=list(isim="mustafa",yas=31,evliMi=F)
> str(denemeList)
List of 3
 $ isim   : chr "mustafa"
 $ yas    : num 31
 $ evliMi : logi FALSE
> denemeList[1]
$isim
[1] "mustafa"
> denemeList[[1]]
[1] "mustafa"
> denemeList$yas
[1] 31
```

R'ye genel bakış

R dili

- Data frame: Özelleşmiş bir liste türüdür.
 - R'nin veri okuma fonksiyonlarının çoğu varsayılan tip olarak data frame tipinde bir nesne oluşturur.
 - read.table, read.csv

```
> path='C:/Mustafa/Research/Presentations/inet/ornek.csv'
> ornekdata=read.csv(path)
> ornekdata
  Col1 Col2 Col3
1  100   a1  b1
2  200   a2  b2
3  300   a3  b3
> str(ornekdata)
'data.frame':   3 obs. of  3 variables:
 $ Col1: int   100  200  300
 $ Col2: Factor w/ 3 levels "a1","a2","a3": 1 2 3
 $ Col3: Factor w/ 3 levels "b1 ", "b2 ", "b3 ": 1 2 3
```


R'ye genel bakış

R dili

▣ Alt kümeleme

```
> ornekdata[1,]  
  Col1 Col2 Col3  
1  100   a1   b1  
> ornekdata[,2]  
[1] a1 a2 a3  
Levels: a1 a2 a3  
> ornekdata[,2:3]  
  Col2 Col3  
1   a1   b1  
2   a2   b2  
3  a3   b3  
> ornekdata$Col1  
[1] 100 200 300
```

Faktörler:

Karakterden farklı olarak
belirli sayıda seviyeye sahip
olan veri tipi
Örnek: günler

R'ye genel bakış

R dili

□ Fonksiyonlar

- Diğer dillerdeki gibi tanımlanır.
- Argüman listesi vardır.
- Herhangi bir veri tipinde değer dönebilir.

```
> ornekFonk <- function(x){  
    2*sqrt(x)  
}  
  
> ornekFonk(4)  
[1] 4  
> x <- c(0,1,9,25)  
> ornekFonk(x)  
[1] 0 2 6 10
```

R'ye genel bakış

R dili

▣ Yardım almak (Tarayıcıda açılır)

■ help()

```
> help(read.table)
starting httpd help server ... done
```

■ help.search()

```
> help.search('median')
```

■ Arama motorları

R'ye genel bakış

R dili

`read.table` {utils}

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"",  
  dec = ".", row.names, col.names,  
  as.is = !stringsAsFactors,  
  na.strings = "NA", colClasses = NA, nrows = -1,  
  skip = 0, check.names = TRUE, fill = !blank.lines.skip,  
  strip.white = FALSE, blank.lines.skip = TRUE,  
  comment.char = "#",  
  allowEscapes = FALSE, flush = FALSE,  
  stringsAsFactors = default.stringsAsFactors(),  
  fileEncoding = "", encoding = "unknown", text)
```

```
read.csv(file, header = TRUE, sep = ",", quote = "\"",  
  dec = ".", fill = TRUE, comment.char = "", ...)
```

```
read.csv2(file, header = TRUE, sep = ";", quote = "\"",  
  dec = ",", fill = TRUE, comment.char = "", ...)
```

```
read.delim(file, header = TRUE, sep = "\t", quote = "\"",  
  dec = ".", fill = TRUE, comment.char = "", ...)
```

```
read.delim2(file, header = TRUE, sep = "\t", quote = "\"",  
  dec = ",", fill = TRUE, comment.char = "", ...)
```

Arguments

file the name of the file which the data are to be read from. Each row of the table appears as one line of the file. Tilde-expansion is performed where supported. This can be a compressed file (see [file](#)).

Alternatively, `file` can be a readable text mode [connection](#) (which will be opened for reading if necessary).

■ `help("read.table")`

■ `?read.table`

■ <http://127.0.0.1:25645/library/utils/html/read.table.html>

R'ye genel bakış

R dili

□ `help.search("median")`

- <http://127.0.0.1:25645/doc/html/Search?pattern=median>

The search string was **"median"**

Help pages:

stats::mad	Median Absolute Deviation
stats::median	Median Value
stats::medpolish	Median Polish of a Matrix
stats::runmed	Running Medians - Robust Scatter Plot Smoothing
stats::smooth	Tukey's (Running Median) Smoothing
stats::smoothEnds	End Points Smoothing (for Running Medians)

R'ye genel bakış

R Oturumu (Session) ve Yönetimi

- ❑ Çalışma klasörü (working directory)
 - Kaydedilen (diske) her türlü bilgi bu klasöre yazılır (eğer uygun bir yol belirtilmemişse).

- ❑ `getwd()`

```
> getwd()  
[1] "C:/Users/baydogan/Documents"
```

- ❑ `setwd(path)` komutu ile yeni klasör belirlenebilir.

- ❑ Çalışma alanında tanımlı nesneler

- `ls()`

```
> ls()  
[1] "a"          "denemeList" "m"          "ornekdata"  "path"
```

- Takibi hafıza kullanımı açısından önemlidir.

R'ye genel bakış

R Oturumu (Session) ve Yönetimi

▣ Objeleri silme

- Hafıza yönetimi oldukça önemlidir.
- `rm()`

```
> ls()  
[1] "a" "denemeList" "m" "ornekdata" "ornekFonk" "path" "x"  
> rm("denemeList", "m")  
> ls()  
[1] "a" "ornekdata" "ornekFonk" "path" "x"  
> gc()
```

	used (Mb)	gc trigger (Mb)	max used (Mb)
Ncells	266520 14.3	531268 28.4	350000 18.7
Vcells	502038 3.9	1031040 7.9	1007484 7.7

- `gc()`
 - ▣ Çöp toplayıcısı

R ile çalışmak

Koşullar

- ▣ Söz dizim kuralları dışında döngü mantığı diğer diller ile aynıdır.

```
> x = 1:9

if (length(x) <= 10)
{
  x <- c(x, 10:20);
  print(x)
}
else
{
  print(x[1])
}
```


R ile çalışmak

Döngüler

- Söz dizim kuralları dışında döngü mantığı diğer diller ile aynıdır.

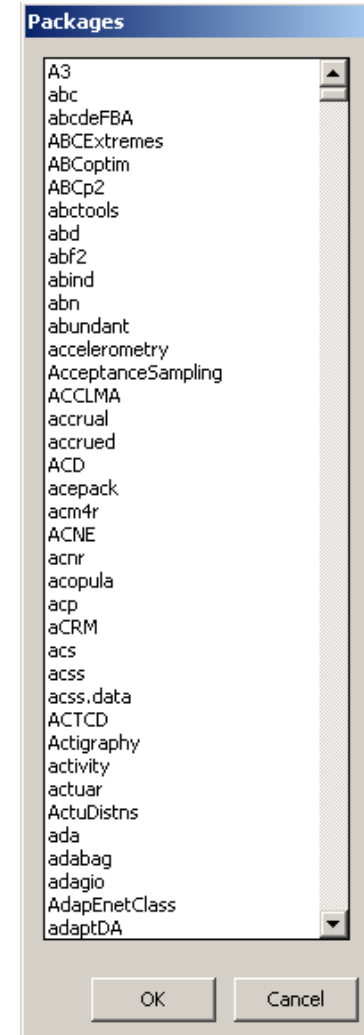
```
> for(i in 1:10) {  
  x[i] <- rnorm(1)  
}  
j = 1  
while( j < 10) {  
  print(j)  
  j <- j + 2  
}
```

- C gibi temel dillere kıyasla döngüler yavaş çalışır.
 - Vektörler üzerindeki işlemleri vektörel olarak kodlamak önemlidir.
 - Örneğin bir vektörün (a olsun) her elamanını 5 ile çarpak için bir döngü yazmak yerine `sonuc=5*a` kullanılabilir.
 - `lapply`, `sapply` ve `apply` fonksiyonları önemlidir.

R ile çalışmak

Paket yapısı

- ❑ R fonksiyonları ayrı paketler halinde düzenlenmişlerdir.*
 - Böylece gerekli paketlerle çalışarak daha az bellek kullanımı ve hızlı işlem gücü sağlanır.
 - Bu paketlerin bir başka avantajı da yazılan fonksiyonlardan oluşan paketlerin R web sitesinden temin edilerek yüklenebilmesidir.
- ❑ Her paketin bir yaratıcısı ve kendisine ait bir yardım dosyası bulunur.
 - <http://cran.r-project.org/web/packages/LPStimeSeries/index.html>



*Kaynak: A. F. Özdemir, E. Yıldıztepe ve M. Binar, Akademik Bilişim 2010

R ile çalışmak

Paket yapısı

- Paketler ara yüz aracılığıyla yüklenebilir.
 - Terminalden `install.packages(paketismi)` komutu kullanarak da yüklenebilir.
 - Paketin indirileceği bir sunucu seçilmesi gereklidir.
- Paketlere ait fonksiyonlar kullanılacağı zaman paket çağrılmalıdır.

```
> require(LPSTimeSeries)
Loading required package: LPSTimeSeries
LPSTimeSeries 1.0
> library(LPSTimeSeries)
```

R ile çalışmak

Paket yapısı

- ▣ Klasik veri yapıları (örneğin vektör) haricinde tanımlanmış farklı veri yapıları mevcuttur
 - Data Table (data.table paketi)
 - Sparse Matrix (Matrix paketi)
 - Time Series (zoo veya ts paketi)
 - Adjacency matrix (igraph paketi)
 - Big matrix (bigmemory paketi)
 - ...

R ile çalışmak

Veri alışverişi

- ❑ Kullanılacak olan veri dosyalarının R ortamına alınabilmesi için farklı seçenekler vardır:
 - metin dosyalarından (txt, csv),
 - gerekli paketleri yükleyerek
 - ❑ binary ve dbase (dbf) dosyalarından,
 - ❑ hesap tablosu dosyalarından (xls, sav),
 - ❑ farklı veri tabanlarından (MySQL, MS Access, Microsoft SQL Server, Postgre SQL, Oracle, IBM DB2)
 - ❑ diğer programların çıktılarından (SPSS, SAS, WEKA)
 - ❑ web tabanlı json, xml dosyalarından

Daha fazla bilgi için:

<http://www.r-tutor.com/r-introduction/data-frame/data-import>

R ile çalışmak

Parallelleştirme ve Büyük Veri

- Unix ortamında birden çok çekirdekli işlemcilere sahip bilgisayarda işler farklı işlemcilere dağıtılabilir.
 - doMC paketi bunu sağlayan örnek paketlerdendir.
- Bilgisayar hafızasına sığmayacak büyük veriler ile çalışıldığında çeşitli indeksleme seçenekleri sağlayan paketler kullanılabilir.
 - bigmemory paketi bunu sağlayan örnek paketlerden biridir.
- Zaman alan ve hafıza tutan işlemlerin bir kısmını daha temel dillerde (C gibi) yapıp R'a entegre edilebilir.
 - R C, Fortran vb. gibi dillere bir ara yüz sağlamaktadır.

Detaylı bilgi:

<http://cran.r-project.org/web/views/HighPerformanceComputing.html>

Sonuç

- Bu çalışmada, son yıllarda yaygın olarak kullanılan R programlama dilinin tanıtılması hedeflenmiştir.
- R, ücretsiz olarak temin edilmesi ve birçok araştırmacının bu dilin gelişimine destek vermesi sonucunda, özellikle veri madenciliği alanlarında çalışan uygulamacıların dikkatini çekmiştir.
- SAS, SPSS ve STATA gibi programlar ile R arasındaki en önemli fark R'nin istatistiksel yazılım geliştirme ortamı ve programlama dili olmasıdır.
- Kişisel olarak hem danışmanlık faaliyetlerinde hem de akademik çalışmalarda oldukça başarılı sonuçlar elde edilmiştir.

Teşekkürler

Sorular

