R - Import Data

Abhishek Kumar
ItsAbhishekKumar.com
@MeAbhishekKumar

pluralsight
hardcore dev and IT training

Outline

Local & remote files

Built-in datasets

Database

Working Directory

- Default location for any file to read or write
- Until explicit path is specified
- Make sure you are working in a desired directory

```
> getwd() #get working directory
> setwd(path) #set working directory
```

Import CSV Files

- CSV (Comma Separated Value) files
- Very common file format for storing data

CSVTtest.csv

```
1  student.physics.marks,student.chemistry.marks
2  70,60
3  75,70
4  80,85
5  85,70
6
```

Notepad Application

	A	B
1	student.physics.marks	student.chemistry.marks
2	70	60
3	75	70
4	80	85
5	85	70
6		

Spreadsheet Application

Import CSV Files

- Use read.csv() function

```
> read.csv(file)
> read.csv(file,header = TRUE)
> read.csv(file,header = TRUE,sep =“,”)
```

- Store values in a data frame

```
> my.data <- read.csv(file)
```

Import Table

- **Tabular data**

File type	Separator	For decimal points
csv	, (comma)	. (dot)
csv2	; (semicolon)	, (comma)
tab delimited	\t (tab)	. (dot)
tab delimited 2	\t (tab)	, (comma)

Import Table

- Use read.table() function

```
> read.table(file)
> read.table(file, header = FALSE, sep ="", dec=".")
```

- Other important attributes
  - stringsAsFactors : whether to treat strings as factors or not
  - colclasses : vector of classes for columns
  - skip : number of lines to skip before reading data
  - nrows : number of rows to read

Import From URL

- File not available on local disk
- Download from a URL (Universal Resource Locator)
- In-built web server in R

```
> my.data <- read.table(url)
> download.file(url,localFile)
> my.data <- read.table(localFile)
```

Import XML Files

- **XML (Extensible Markup Language)**


```
<?xml version="1.0"?>
<classroom>
  <student>
    <name>Raj</name>
    <physics>70</physics>
    <chemistry>60</chemistry>
  </student>
  <student>
    <name>Rahul</name>
    <physics>75</physics>
    <chemistry>70</chemistry>
  </student>
</classroom>
```

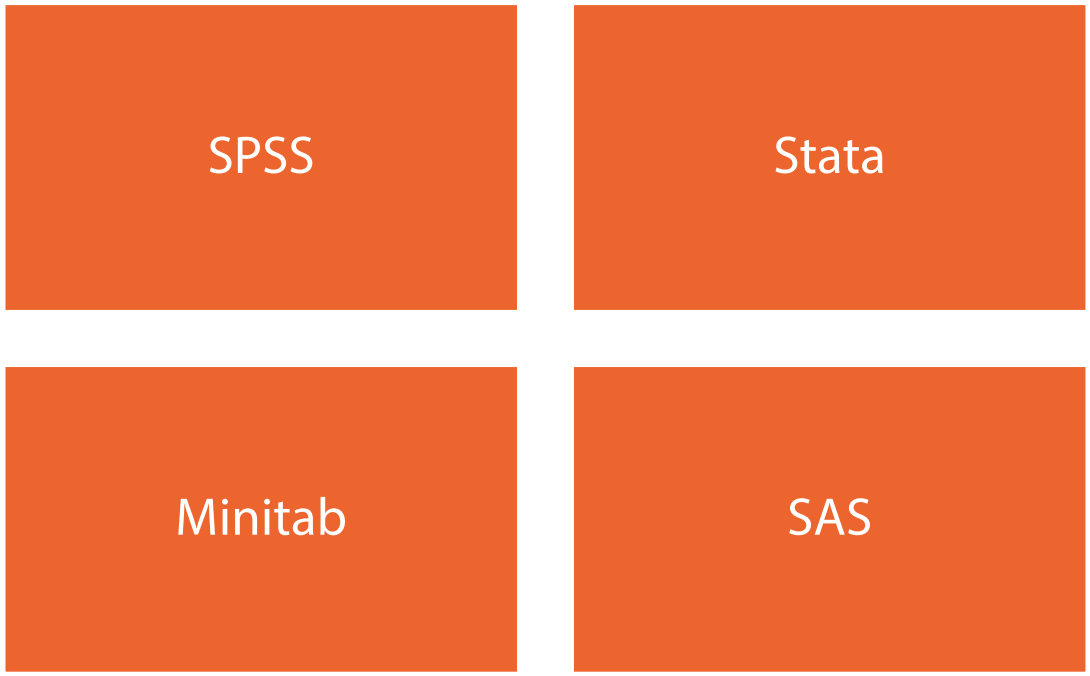
Import Excel Files

- Use XLConnect package
- Only requires Java installed on your machine
- Work with both .xls and .xlsx files

```
> install.packages("XLConnect")
> library(XLConnect)
> my.data <- readWorksheetFromFile(file,sheet=1)
> my.data <- readWorksheetFromFile(file,sheet="name")
```

Import Other File Types

- Use foreign package



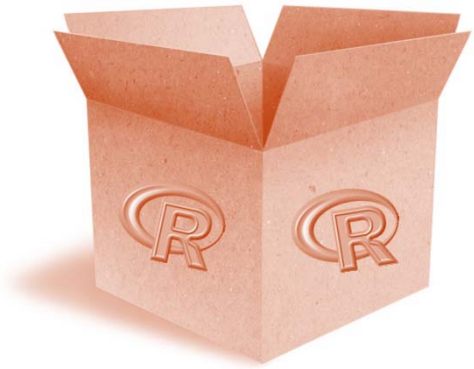
Import Other File Types

```
> library(foreign)
> my.data <- read.spps(file) #SPSS data file
> my.data <- read.dta(file) #Stata binary file
> my.data <- read.mtp(file) #Minitab portable worksheet
> my.data <- read.xport(file) #SAS XPORT format library
```

```
> help(package="foreign")
```

Built-in Datasets

- **Many packages contain datasets**
- **Can be used in**
  - Learning
  - Testing



```
> library(package)
> data(dataset)
```

Import From Database

Relational databases

NoSQL databases

Import Data From Database

- **General packages for relational databases**
  - RODBC
  - DBI
- **Specific packages for relational databases**

Package	Database
MySQL	RMySQL
Oracle	ROracle
PostgreSQL	RPostgreSQL
SQLite	RSQLite

Import Data From Database

- **NoSQL databases**

Package	Database
MongoDB	RMongo, rmongodb
Cassandra	RCassandra
CouchDB	R4CouchDB
HBase (Hadoop ecosystem)	rhbase (RHadoop)

Using RODBC Package

- **Configure a ODBC DSN (Data Source Name)**
  - Windows : http://bit.ly/odbcdsnwindows
  - Mac : http://bit.ly/odbcdsnmac
  - Unix : http://bit.ly/odbcdsnunix

Summary

Local & remote files

read.csv()

read.table()

XML

XLConnect

Foreign

Built-in datasets

data()

Database

Relational

NoSQL

RODBC