INTRODUCTION TO R

Explore the Data Frame

Datasets

- Observations
- Variables
- Example: people
  - each person = observation
  - properties (name, age ...) = variables
- Matrix? Need different types
- List? Not very practical

name	age	child
Anne	28	FALSE
Pete	30	TRUE
Frank	21	TRUE
Julia	39	FALSE
Cath	35	TRUE

Data Frame

- Specifically for datasets
- Rows = observations (persons)
- Columns = variables (age, name, ...)
- Contain elements of different types
- Elements in same column: same type

name	age	child
Anne	28	FALSE
Pete	30	TRUE
Frank	21	TRUE
Julia	39	FALSE
Cath	35	TRUE

Create Data Frame

- Import from data source
- CSV file
- Relational Database (e.g. SQL)
- Software packages (Excel, SPSS ...)

Create Data Frame `data.frame()`

```
> name <- c("Anne", "Pete", "Frank", "Julia", "Cath")
> age <- c(28, 30, 21, 39, 35)
> child <- c(FALSE, TRUE, TRUE, FALSE, TRUE)

> df <- data.frame(name, age, child)
```

column names match variable names

```
> df
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Name Data Frame

```
> names(df) <- c("Name", "Age", "Child")
> df
   Name Age Child
1  Anne  28 FALSE
2  Pete  30  TRUE
   ...
5  Cath  35  TRUE

> df <- data.frame(Name = name, Age = age, Child = child)
> df
   Name Age Child
1  Anne  28 FALSE
2  Pete  30  TRUE
   ...
5  Cath  35  TRUE
```

Data Frame Structure

Factor instead of character

```
> str(df)
'data.frame':5 obs. of  3 variables:
 $ Name : Factor w/ 5 levels "Anne","Cath",..: 1 5 3 4 2
 $ Age  : num  28 30 21 39 35
 $ Child: logi  FALSE TRUE TRUE FALSE TRUE

> data.frame(name[-1], age, child)
Error : arguments imply differing number of rows: 4, 5

> df <- data.frame(name, age, child,
                   stringsAsFactors = FALSE)
> str(df)
'data.frame':5 obs. of  3 variables:
 $ name : chr  "Anne" "Pete" "Frank" "Julia" ...
 $ age  : num  28 30 21 39 35
 $ child: logi  FALSE TRUE TRUE FALSE TRUE
```

INTRODUCTION TO R

Subset - Extend - Sort Data Frames

Subset Data Frame

- Subsetting syntax from matrices and lists
- `[` from matrices
- `[[` and `$` from lists

people

```
> name <- c("Anne", "Pete", "Frank", "Julia", "Cath")
> age <- c(28, 30, 21, 39, 35)
> child <- c(FALSE, TRUE, TRUE, FALSE, TRUE)
> people <- data.frame(name, age, child,
                       stringsAsFactors = FALSE)

> people
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Subset Data Frame

```
> people[3,2]
[1] 21

> people[3,"age"]
[1] 21

> people[3,]
   name age child
3 Frank  21  TRUE

> people[,"age"]
[1] 28 30 21 39 35
```

```
> people
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Subset Data Frame

```
> people[c(3, 5), c("age", "child")]
  age child
3  21  TRUE
5  35  TRUE

> people[2]
  age
1  28
2  30
3  21
4  39
5  35
```

```
> people
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Data Frame ~ List

```
> people$age
[1] 28 30 21 39 35

> people[["age"]]
[1] 28 30 21 39 35

> people[[2]]
[1] 28 30 21 39 35
```

```
> people
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Data Frame ~ List

```
> people["age"]
  age
1  28
2  30
3  21
4  39
5  35

> people[2]
  age
1  28
2  30
3  21
4  39
5  35
```

```
> people
   name age child
1  Anne  28 FALSE
2  Pete  30  TRUE
3 Frank  21  TRUE
4 Julia  39 FALSE
5  Cath  35  TRUE
```

Extend Data Frame

- Add columns = add variables
- Add rows = add observations

Add column

```
> height <- c(163, 177, 163, 162, 157)

> people$height <- height

> people[["height"]] <- height

> people
   name age child height
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
3 Frank  21  TRUE    163
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
```

Add column

```
> weight <- c(74, 63, 68, 55, 56)

> cbind(people, weight)
   name age child height weight
1  Anne  28 FALSE    163     74
2  Pete  30  TRUE    177     63
3 Frank  21  TRUE    163     68
4 Julia  39 FALSE    162     55
5  Cath  35  TRUE    157     56
```

Add row

```
> tom <- data.frame("Tom", 37, FALSE, 183)

> rbind(people, tom)
Error : names do not match previous names

> tom <- data.frame(name = "Tom", age = 37,
                    child = FALSE, height = 183)
> rbind(people, tom)
   name age child height
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
3 Frank  21  TRUE    163
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
6   Tom  37 FALSE    183
```

Sorting

```
> sort(people$age)
[1] 21 28 30 35 39

> ranks <- order(people$age)
> ranks
[1] 3 1 2 5 4

> people$age
[1] 28 30 21 39 35
```

```
> people
   name age child height
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
3 Frank  21  TRUE    163
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
```

21 is lowest: its index, 3, comes first in ranks

28 is second lowest: its index, 1, comes second in ranks

39 is highest: its index, 4, comes last in ranks

Sorting

```
> sort(people$age)
[1] 21 28 30 35 39

> ranks <- order(people$age)
> ranks
[1] 3 1 2 5 4

> people[ranks, ]
   name age child height
3 Frank  21  TRUE    163
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
5  Cath  35  TRUE    157
4 Julia  39 FALSE    162
```

```
> people
   name age child height
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
3 Frank  21  TRUE    163
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
```

Sorting

```
> sort(people$age)
[1] 21 28 30 35 39

> ranks <- order(people$age)
> ranks
[1] 3 1 2 5 4

> people[order(people$age, decreasing = TRUE), ]
   name age child height
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
2  Pete  30  TRUE    177
1  Anne  28 FALSE    163
3 Frank  21  TRUE    163
```

```
> people
   name age child height
1  Anne  28 FALSE    163
2  Pete  30  TRUE    177
3 Frank  21  TRUE    163
4 Julia  39 FALSE    162
5  Cath  35  TRUE    157
```