Data Wrangling - Cheat Sheet

zahidmmmian@gmail.com

10-Dec-2016

Syntax—Creating DataFrames

	EmpID	Age	Name
1st	1	24	Adam
2nd	2	30	Mary
3rd	3	50	John
4th	4	35	David
5th	5	55	Sarah

```
# data frame with values
df <- data.frame(
   "EmpID" = 1:5, "Age" = c(24, 30, 50, 35, 55),
   "Name" = c("Adam", "Mary", "John", "David", "Sarah"),
   stringsAsFactors = F)
# assign row names
rownames(df) <- c("1st", "2nd", "3rd", "4th", "5th")
# new column
df$Title <- c("Mr", "Ms", "Mr", "Mr", "Mrs")
# remove column
df[["Title"]] <- NULL
```

	A	B

```
# empty data frame with only column names
data.frame(A=integer(), B=character(),
   stringsAsFactors = F)
```

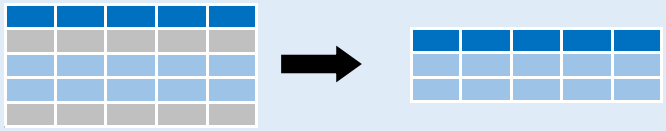
Metadata

```
str(df) # structure
'data.frame':   5 obs. of  3 variables:
 $ EmpID: int  1 2 3 4 5
 $ Age  : num  24 30 50 35 55
 $ Name : chr  "Adam" "Mary" "John" "David" ...
```

```
dim(df) # dimensions
[1]  5 3
ncol(df) # number of cols
5
nrow(df) # number of cols
3
```

```
dimnames(df) # dimensions
[[1]]
[1] "1" "2" "3" "4" "5"
[[2]]
[1] "EmpID" "Age"   "Name"
```

Subset Observations (Rows)



df[df$Age>30,]
Extract rows that meet logical criteria

head(df, n)
Selects first n rows

tail(df, n)
Select last n rows

df[sample(nrow(df), n),]
Randomly selects n rows

unique(df)
df[!duplicated(df),]
Selects unique rows (no dups)

df[rows, cols] # indexed selection syntax

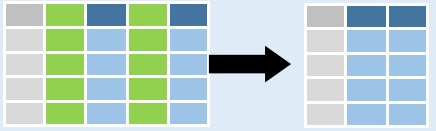
df[1,]
First row, all columns

df[2:4,]
df[c(2,3,4),]
Rows 2-4, all columns

df[-c(1),]
All rows except the first

df["1st",]
Named row (assumes row is named correctly)

Subset Variables (Columns)



df[,1]
First column, all rows

df[,1:3]
First three column, all rows

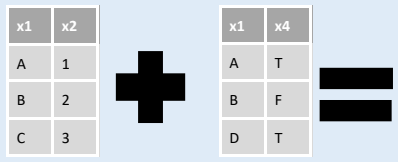
df[,c(1,3)]
First, Second columns, all rows

df[,-c(1)]
All but the first column, all rows

df[,c("Age", "Name")]
Age and Name columns, all rows

df[,(ncol(df) - 1):ncol(df)]
Last two columns, all rows

Combining Data



df1 df2

x1	x2.x	x2.y
A	1	T
B	2	F

```
# natural join
merge(df1, df2,
  by = "x1",
  all = FALSE)
```

x1	x2.x	x2.y
A	1	T
B	2	F
C	3	*NA*
D	*NA*	T

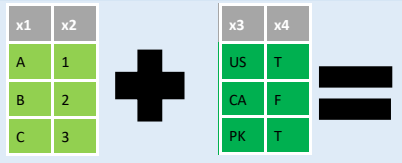
```
# full outer join
merge(df1, df2,
  by = "x1",
  all = TRUE)
```

x1	x2.x	x2.y
A	1	T
B	2	F
C	3	*NA*

```
# left outer join
merge(df1, df2,
  by = "x1",
  all.x = TRUE)
```

x1	x2.x	x2.y
A	1	T
B	2	F
D	*NA*	T

```
# right outer join
merge(df1, df2,
  by = "x1",
  all.x = TRUE)
```



x1	x2	x3	x4
A	1	US	T
B	2	CA	F
C	3	PK	T

```
# append columns from df2 to df1
# no need for join
cbind(df1, df2)
```