



Subsetting and sorting

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Subsetting - quick review

```
set.seed(13435)
X <- data.frame("var1"=sample(1:5), "var2"=sample(6:10), "var3"=sample(11:15))
X <- X[sample(1:5),]; X$var2[c(1,3)] = NA
X
```

	var1	var2	var3
1	2	NA	15
4	1	10	11
2	3	NA	12
3	5	6	14
5	4	9	13

Subsetting - quick review

```
X[,1]
```

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

```
X["var1"]
```

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

```
X[1:2,"var2"]
```

```
[1] NA 10
```

Logicals ands and ors

```
X[(X$var1 <= 3 & X$var3 > 11),]
```

	var1	var2	var3
1	2	NA	15
2	3	NA	12

```
X[(X$var1 <= 3 | X$var3 > 15),]
```

	var1	var2	var3
1	2	NA	15
4	1	10	11
2	3	NA	12

Dealing with missing values

```
X[which(X$var2 > 8),]
```

	var1	var2	var3
4	1	10	11
5	4	9	13

Sorting

```
sort(X$var1)
```

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

```
sort(X$var1,decreasing=TRUE)
```

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

```
sort(X$var2,na.last=TRUE)
```

```
[1] 6 9 10 NA NA
```

Ordering

```
X[order(X$var1),]
```

	var1	var2	var3
4	1	10	11
1	2	NA	15
2	3	NA	12
5	4	9	13
3	5	6	14

Ordering

```
X[order(X$var1,X$var3),]
```

	var1	var2	var3
4	1	10	11
1	2	NA	15
2	3	NA	12
5	4	9	13
3	5	6	14

Ordering with plyr

```
library(plyr)
arrange(X, var1)
```

	var1	var2	var3
1	1	10	11
2	2	NA	15
3	3	NA	12
4	4	9	13
5	5	6	14

```
arrange(X, desc(var1))
```

	var1	var2	var3
1	5	6	14
2	4	9	13
3	3	NA	12
4	2	NA	15

Adding rows and columns

```
X$var4 <- rnorm(5)  
X
```

	var1	var2	var3	var4
1	2	NA	15	0.18760
4	1	10	11	1.78698
2	3	NA	12	0.49669
3	5	6	14	0.06318
5	4	9	13	-0.53613

Adding rows and columns

```
Y <- cbind(X,rnorm(5))  
Y
```

	var1	var2	var3	var4	rnorm(5)
1	2	NA	15	0.18760	0.62578
4	1	10	11	1.78698	-2.45084
2	3	NA	12	0.49669	0.08909
3	5	6	14	0.06318	0.47839
5	4	9	13	-0.53613	1.00053

Notes and further resources

- R programming in the Data Science Track
- Andrew Jaffe's lecture notes http://www.biostat.jhsph.edu/~ajaffe/lec_winterR/Lecture%202.pdf