1. If you are on the waiting list, and still want to be added to the course, email me mengqian.lu@columbia.edu with subject [W4415 MSI] Waiting list

2. Syllabus available on CourseWorks , after each topic we will have a **practical section, bring your laptop, I will send email reminder.**

3. First assignment release today.

4. Data Visualization Weekend April 5 – 7th
   - 5th – Keynote speaker
   - 6th – students demo day (poster section) with keynote speakers (Visualization tool, software, research), Presentation
   - 7th – Hands on workshops in 3 libraries
   - Proposal in Feb, final list in March

5. R open lab @ Columbia

COLUMBIA UNIVERSITY EVENTS

R Open Lab

Wednesday, January 27, 2016 10:00 AM - 12:00 PM

International Affairs Building 215

Add to Calendar Share

Drop by the Digital Social Science Center to build your skills in R, a free and open source language for statistical computing. Each week we'll have a 10 min tutorial on a dataset or package along with a couple hours of self-guided learning. Email dssc@library.columbia.edu to suggest a topic!

Contact: Julia Marden 212 854 5272

LOCATION:
MORNINGSIDE

TYPE:
OTHER WORKSHOP

CATEGORY:
SOCIAL SCIENCES LIBRARIES

EVENTS OPEN TO:
STUDENT STAFF FACULTY

R Open Labs

Jan 27 - Apr 20

Wednesdays, 10am to 12pm.
215 Lehman
(Bringing personal laptops is highly recommended!)

← BACK TO EVENTS

6. TA Office Hours -- Haolei WENG (wenghl07017@gmail.com)

where: Lounge room, 10th, SSW
when: 3:00-4:00pm, Monday & Wednesday.

7. A Question regarding Estimate of Quantiles from a student

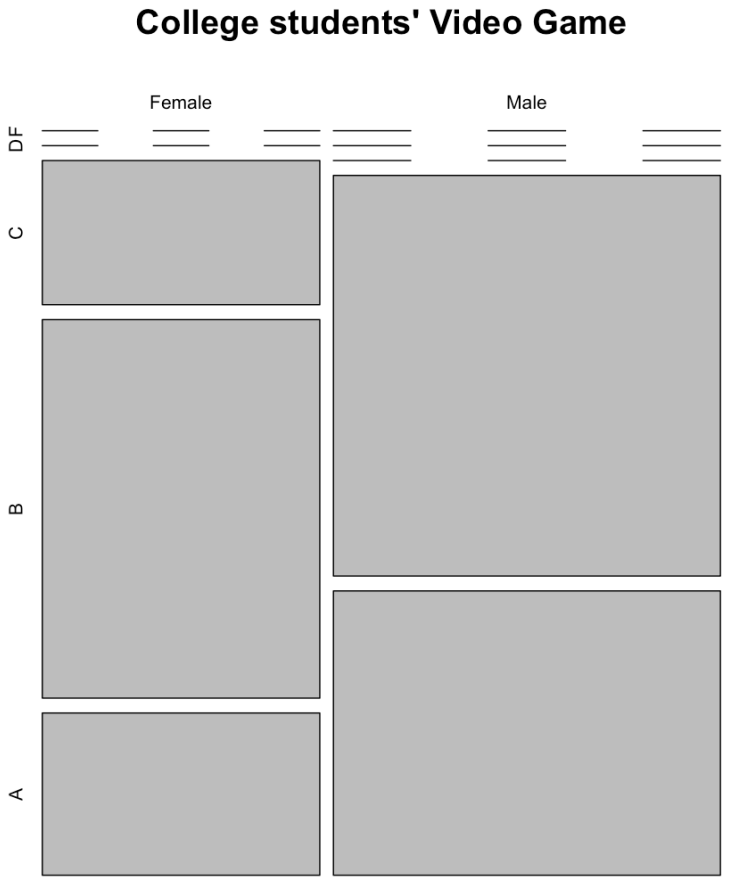
Exploratory Data Analysis and Visualization for Multivariate Data Analysis II

Mengqian LU

Visualizing categorical data: mosaicplot()

```
> table(video$grade, video$sex)
   Female Male
F       0    0
D       0    0
C       8    0
B      21   31
A       9   22
```

Area is proportional to table entry



Chi-Square Test of Independence

	A=1	A=2	Total
B=1	n_{11}	n_{12}	n_{1*}
B=2	n_{21}	n_{22}	n_{2*}
	n_{*1}	n_{*2}	n

H_o: A and B are independent; therefore

$$P(A = i \cap B = j) = P(A = i) \cdot P(B = j)$$

$$\approx \hat{P}(A = i) \cdot \hat{P}(B = j) = \frac{n_{*i}}{n} \cdot \frac{n_{j*}}{n} = \hat{\pi}_{ij}$$

Expected values in entries if H_o is true: $E_{ij} = n \cdot \hat{\pi}_{ij}$

H_a: A and B are not independent.

How different are observed and expected values?

Pearson Chi-Square Statistics:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Contribution of each cell to misfit

If H_o is true, X^2 follows a Chi-Square distribution with degrees of freedom:

$$DF = (I-1)(J-1)$$

if n is large and no empty cells

Compute p-value & comparing to the significant level

Example: Haircolor and Eyecolor

Hair Color	Eye Color			
	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

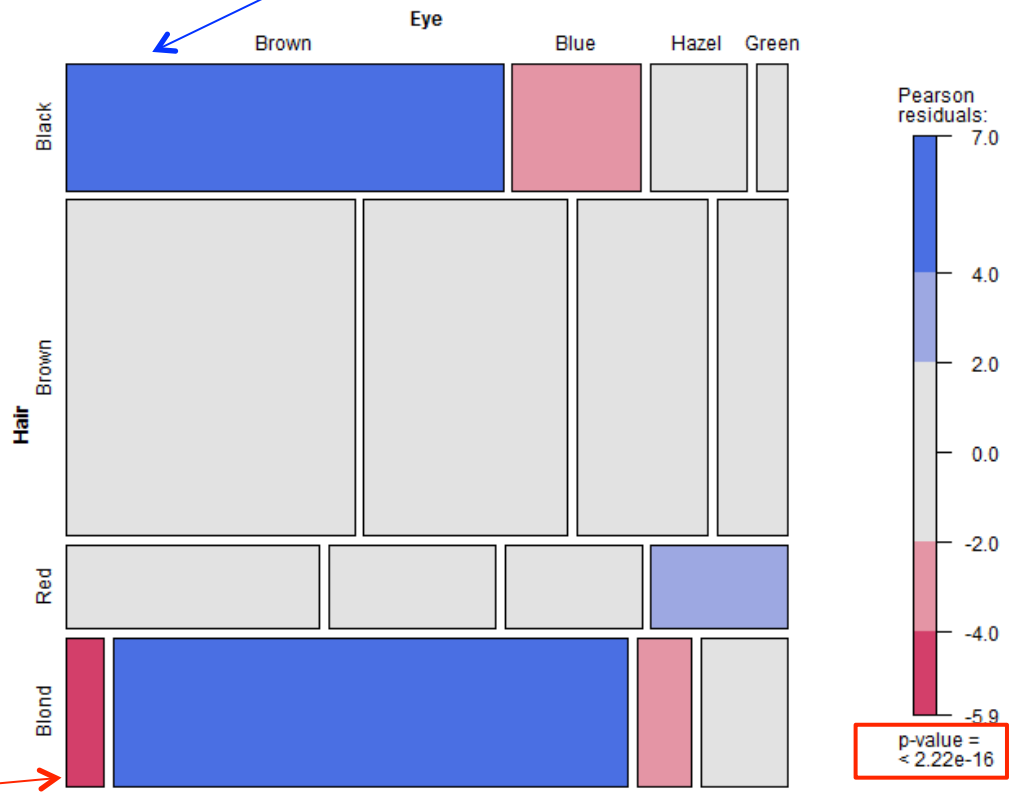
```
Library(vcd) #visualizing categorical data
df = as.data.frame(HairEyeColor)
tab = xtabs(Freq ~ ., data = df)
structable(~ Hair + Eye, data = df)
## Mosaic plot with independence test
mosaic(~ Hair + Eye, data = df, shade = TRUE)
```

Mosaic plot with independence test, color shading shows Pearson residuals

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$ Pearson Residual

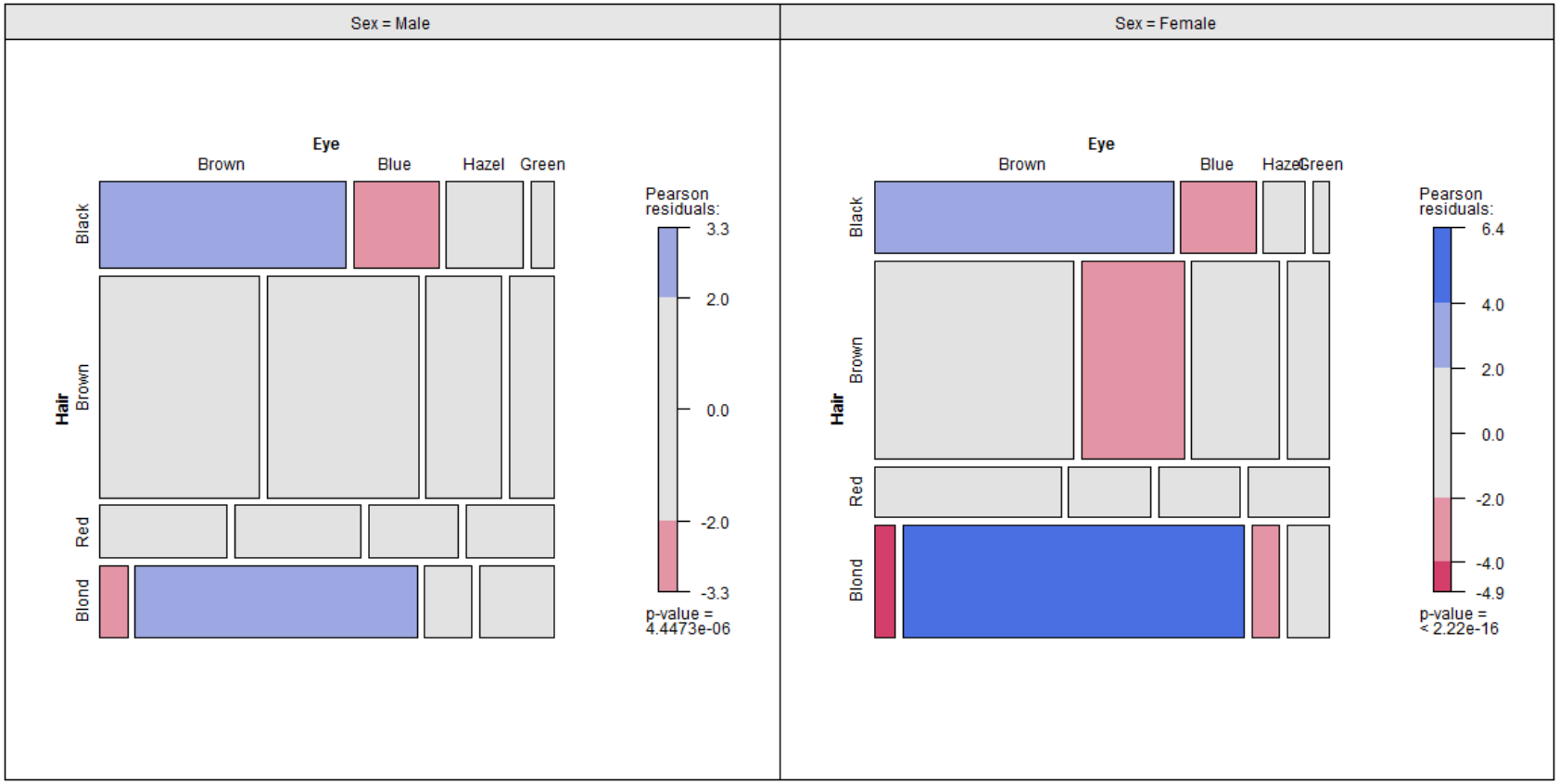
Surprisingly large observed cell count



Surprisingly small observed cell count

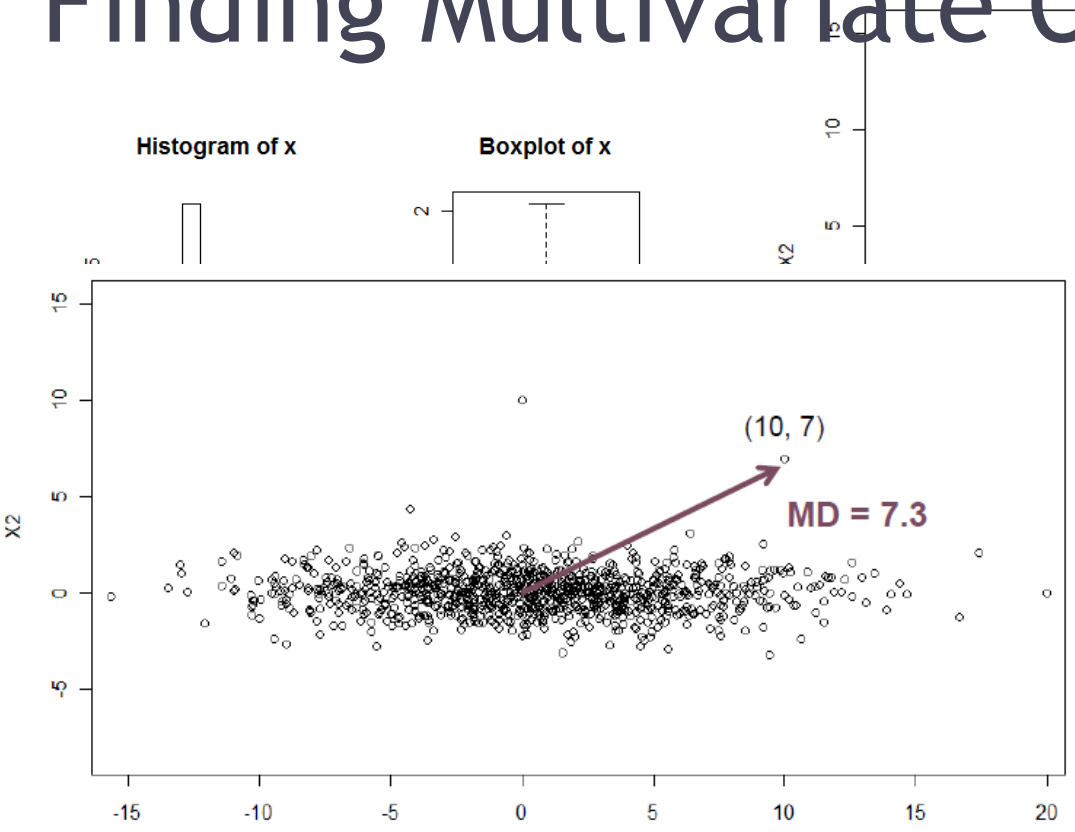
P-value of the test

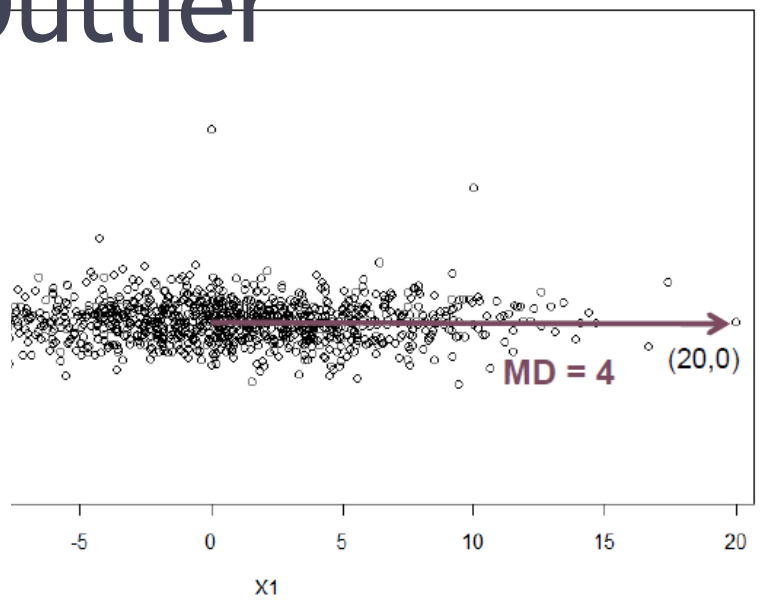
Conditional Mosaic Plot with Shading

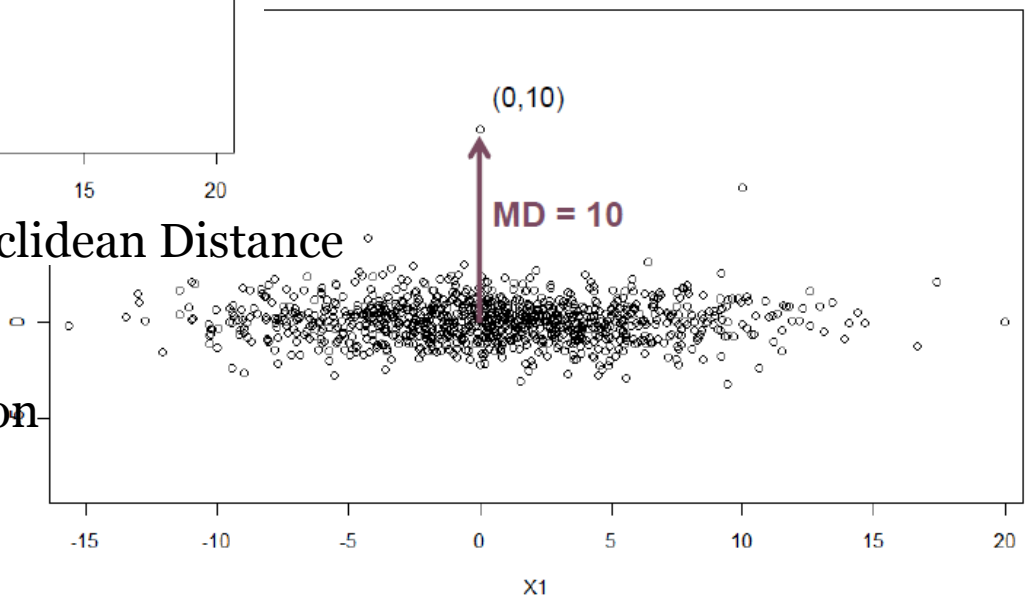



```
cotabplot(~ Hair + Eye | Sex, data = df, shade = TRUE)
```

Finding Multivariate Outlier

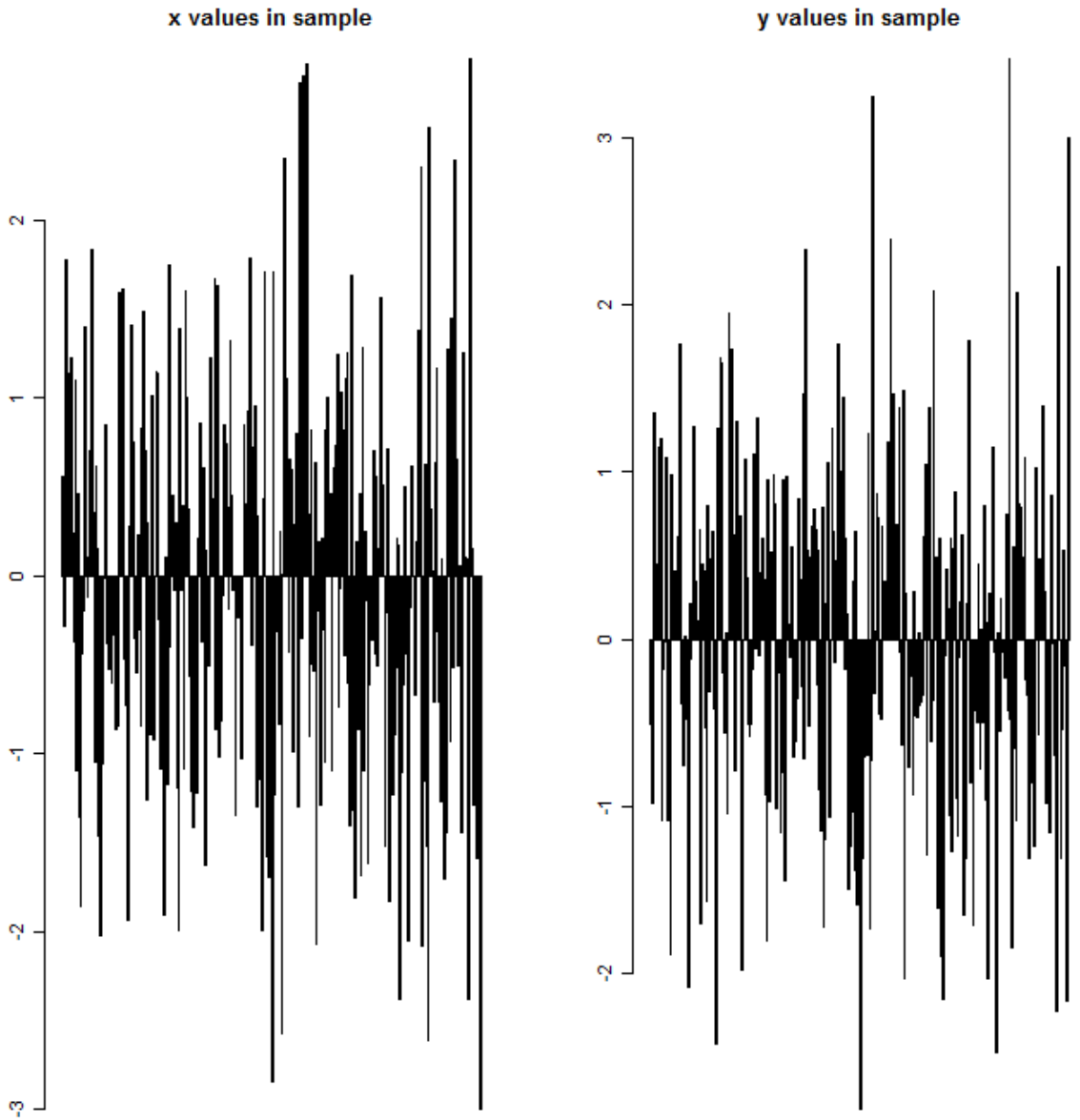






- Mahalanobis Distance (MD) vs. Euclidean Distance
  - MD is unitless
  - scale-invariant
  - take into account the correlation
  - often used to detect outliers
- Covariance & Correlation

Any outlier?



No outlier in x or y

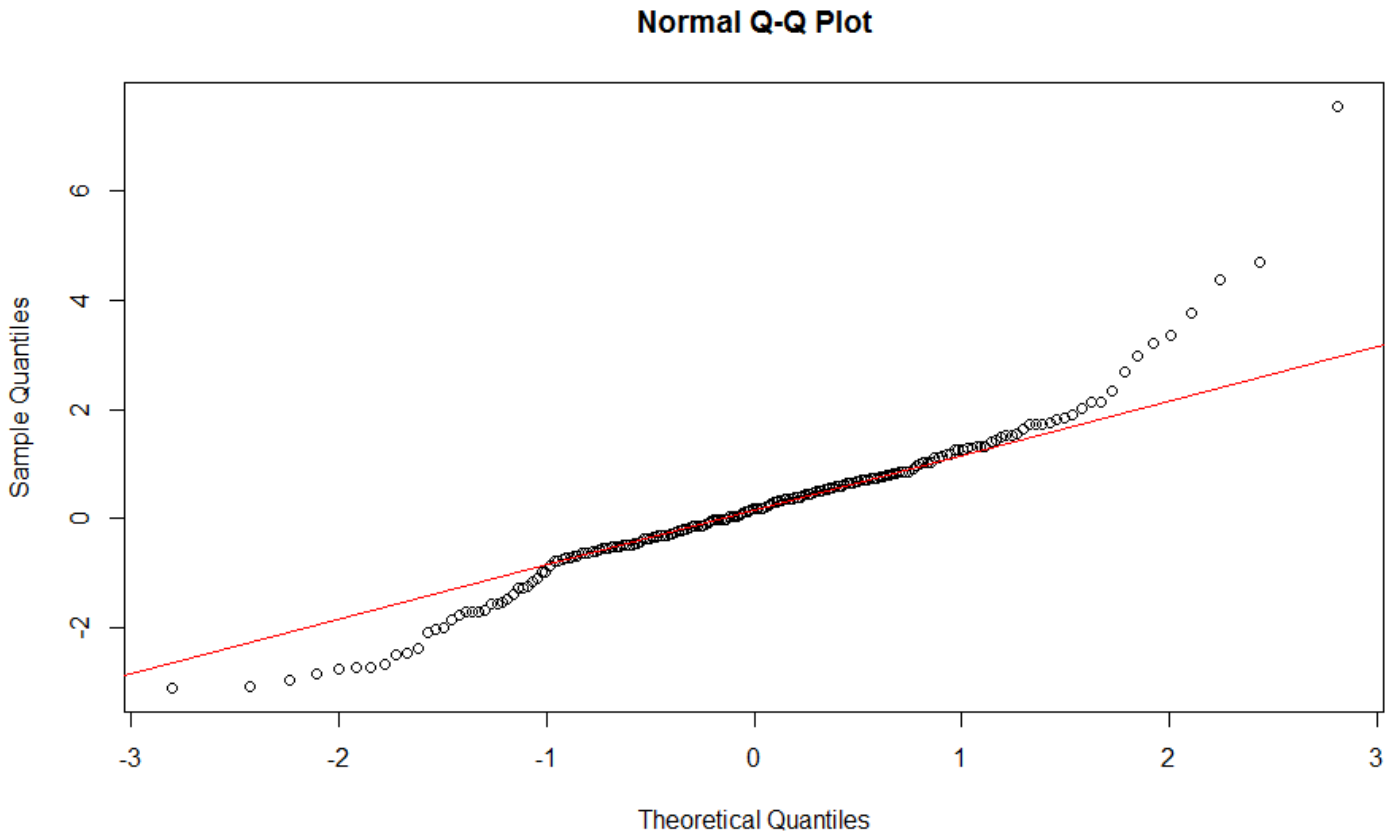
Theory of Mahalanobis Distance

$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left(-\frac{1}{2}\cdot(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu) \sim \chi^2_p$$

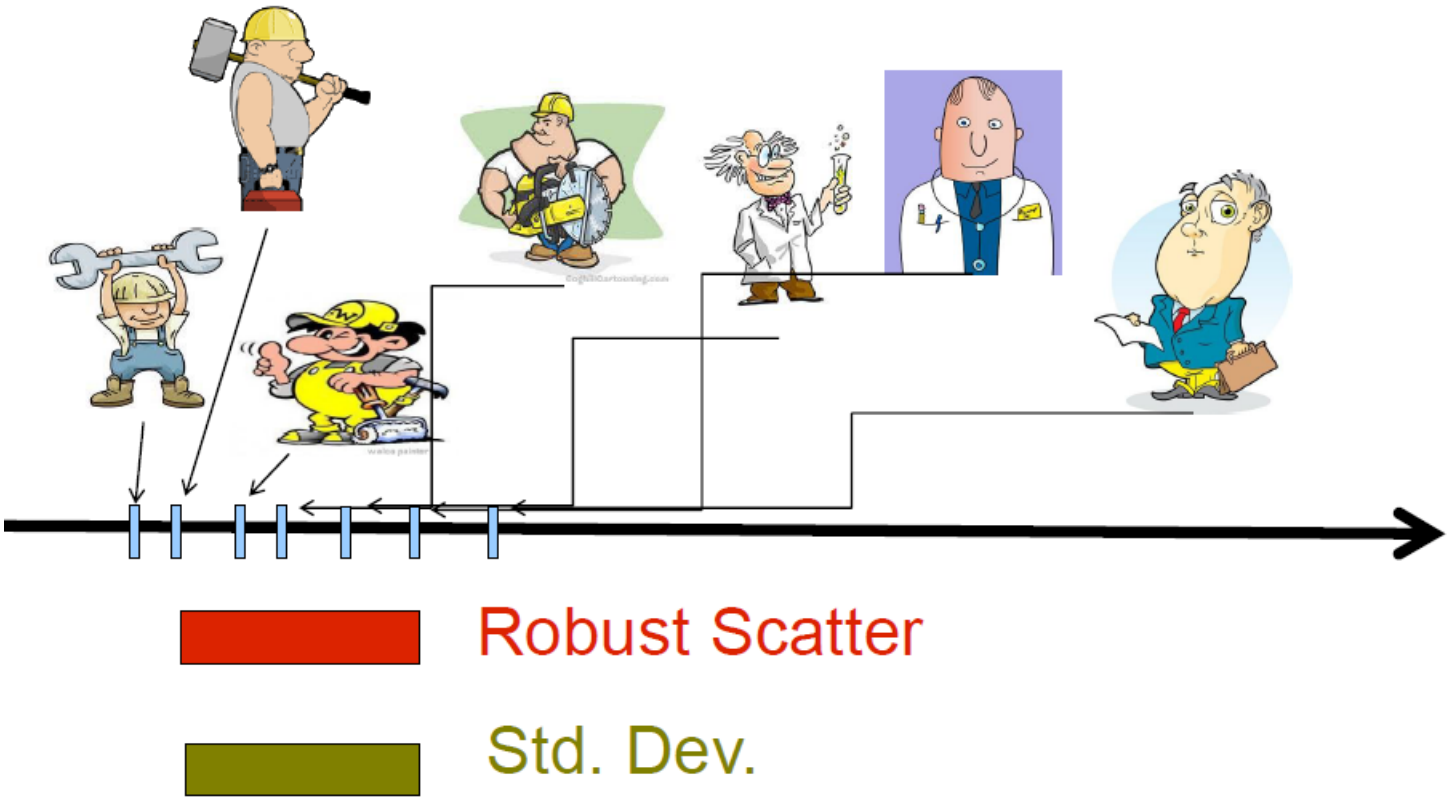
Definition: chi-square distribution with p degrees of freedom is the distribution of a sum of the squares of p independent standard normal random variables

QQ-plot

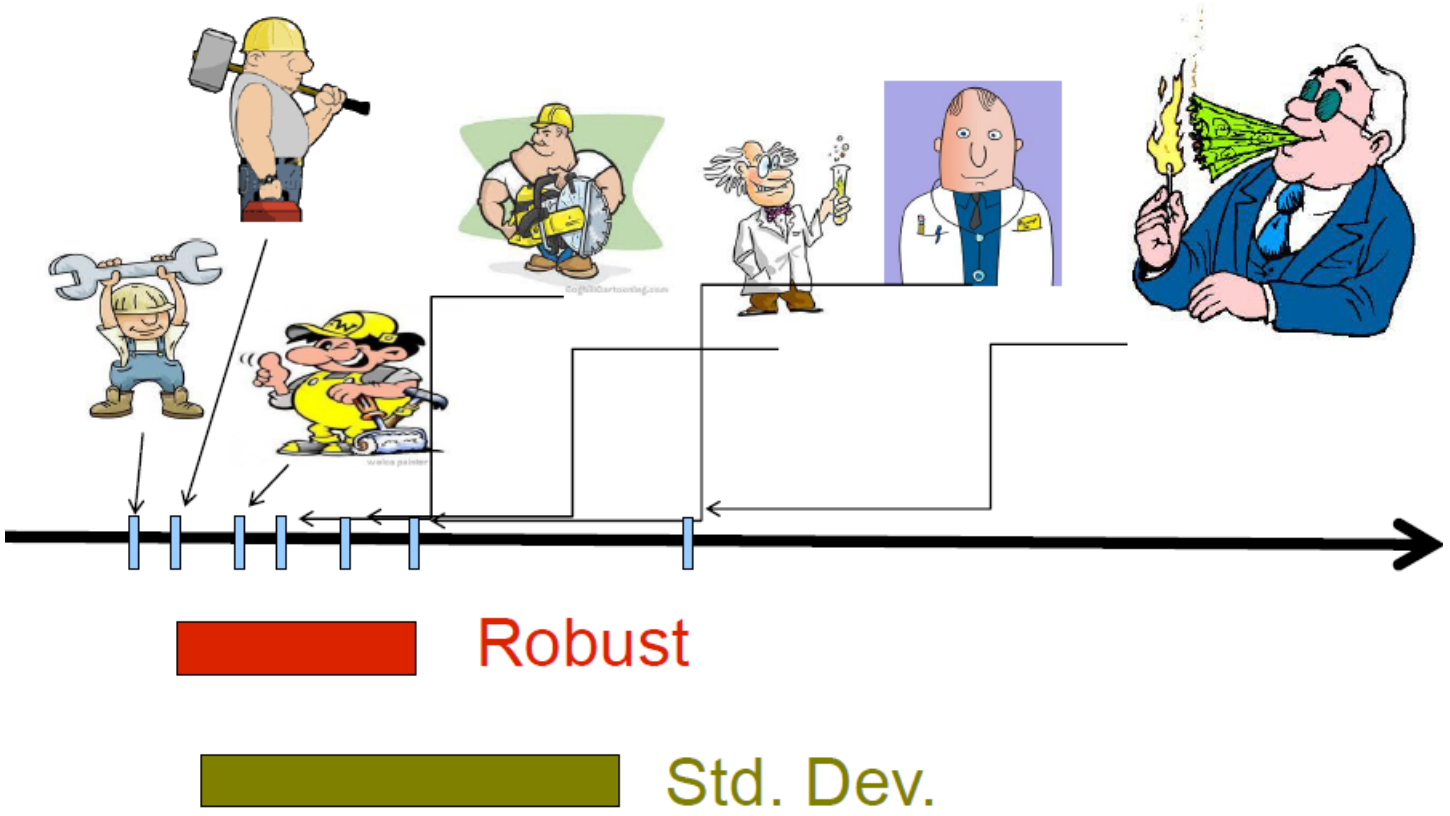


```
qqplot(qchisq(ppoints(500), df = 3), rchisq(500, df = 3))
qqline(y, distribution = function(p) qchisq(p, df = 3), col=2)
```

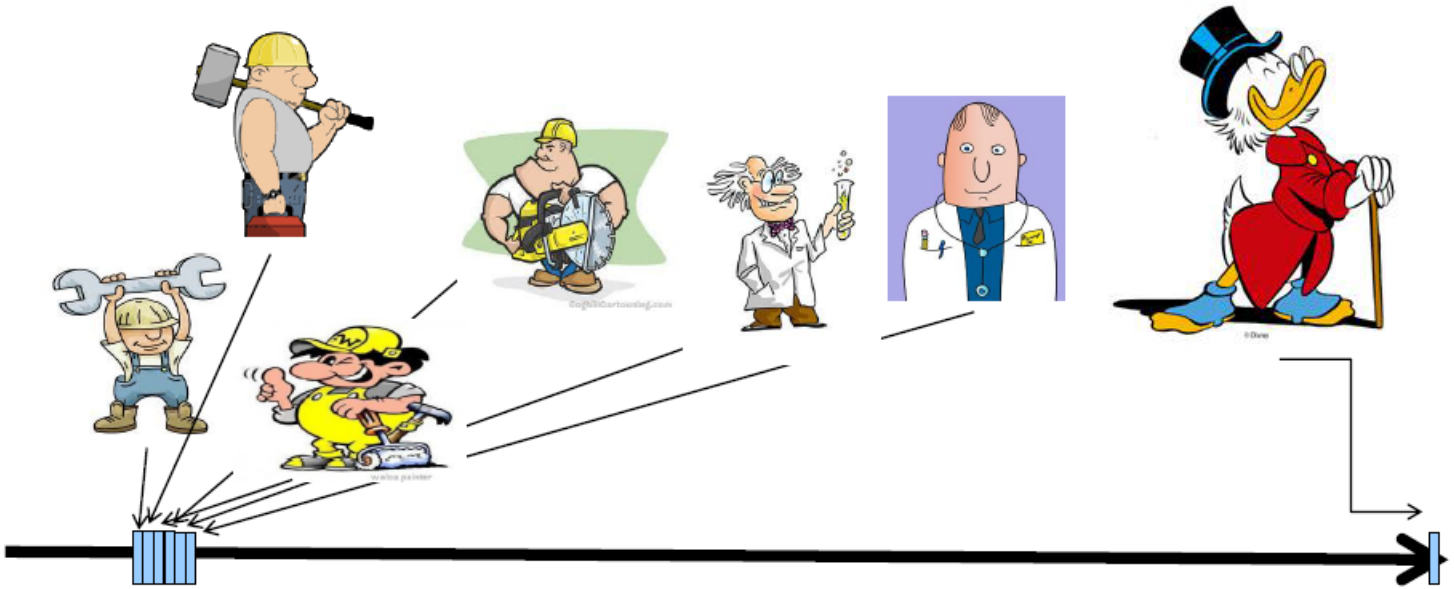
Robust Estimates (Suggested Reading 1)



Robust Estimates



Robust Estimates

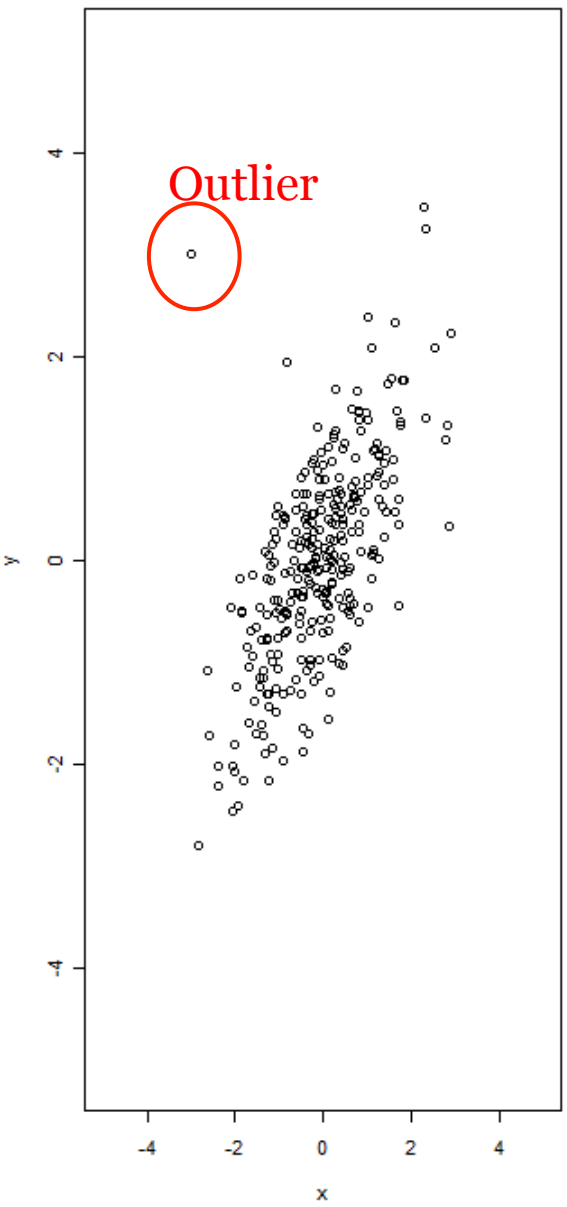


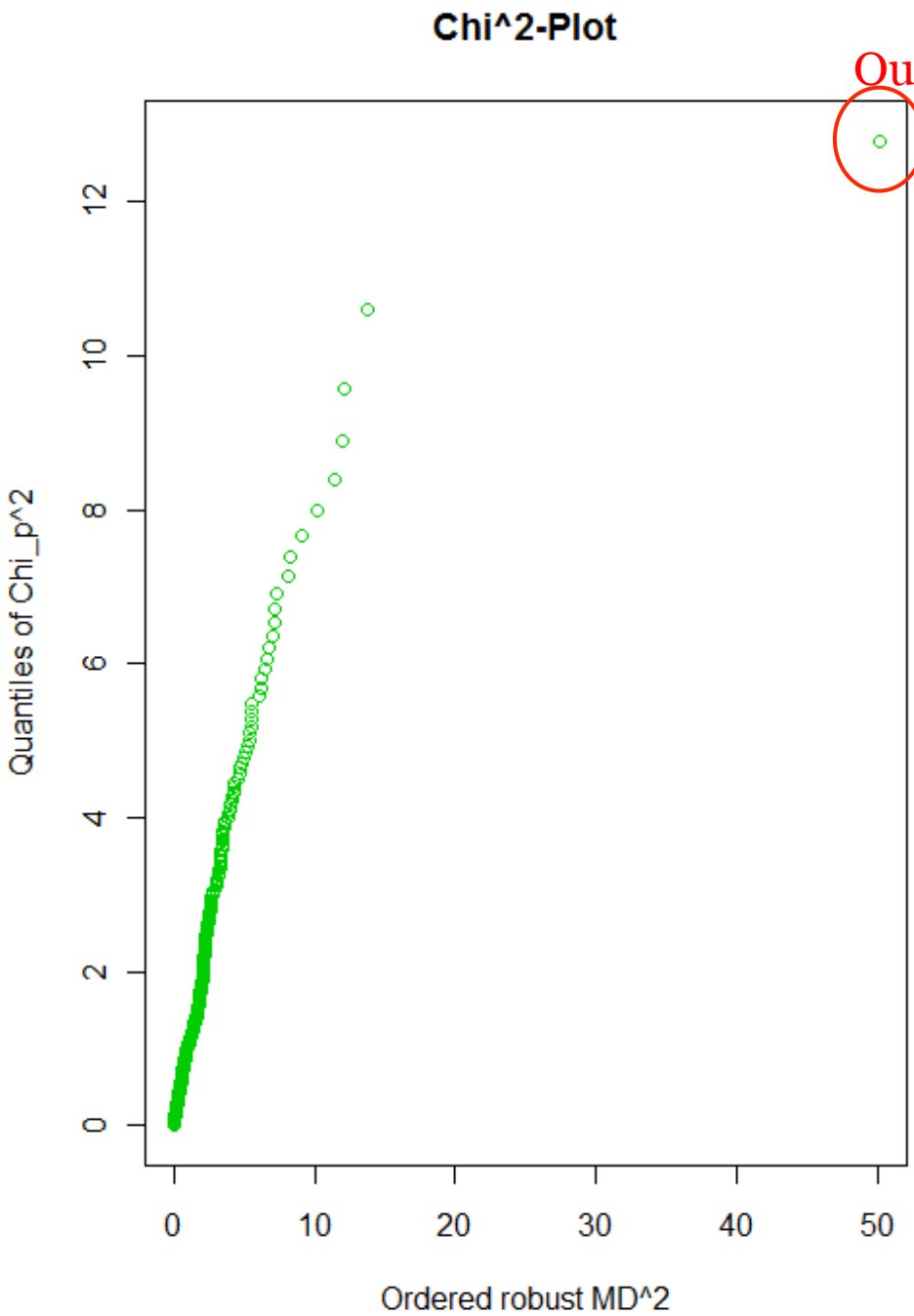
Robust

Std. Dev.

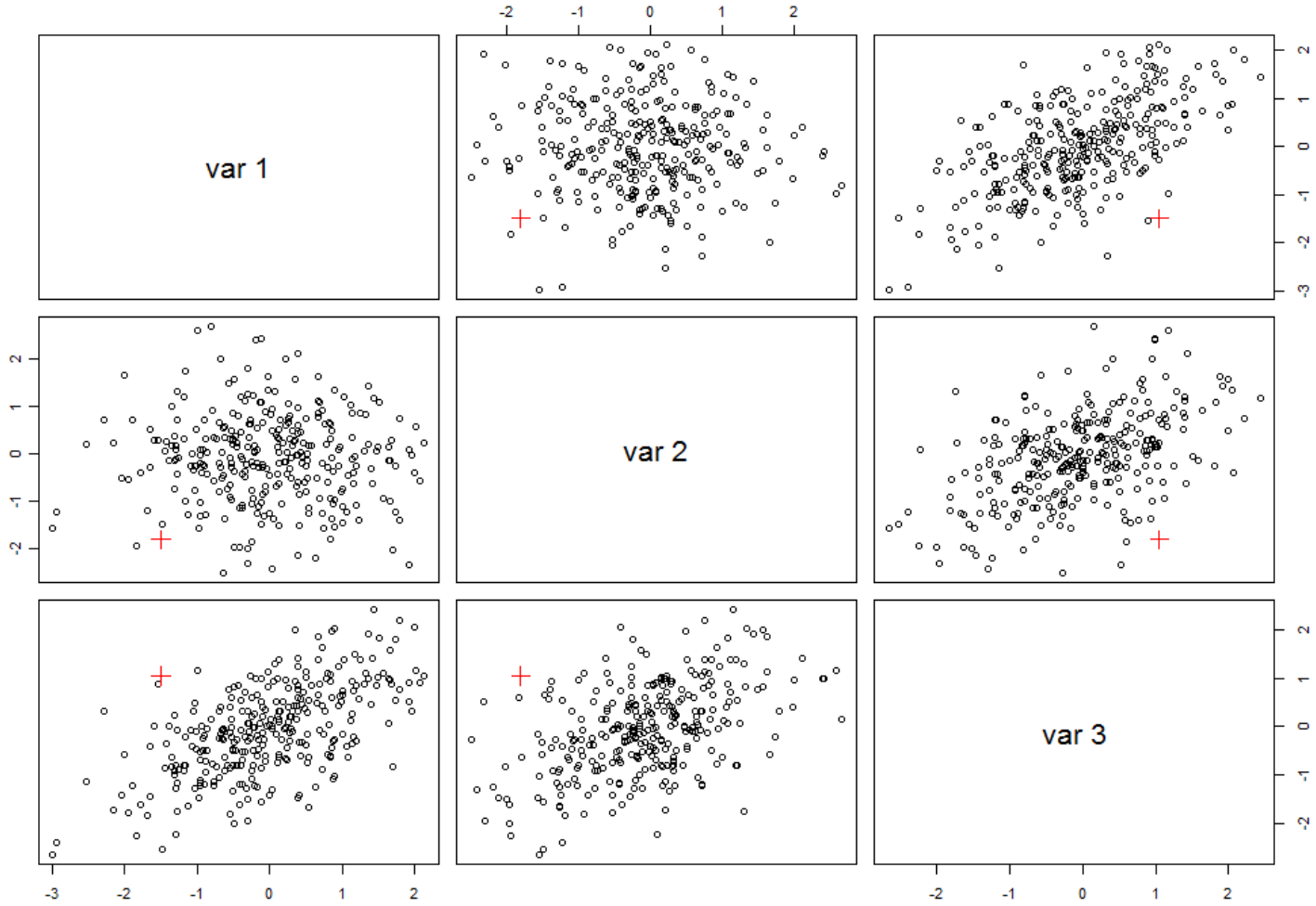
Robust Estimates

- Suggested reading 1 by *Peter Filzmoser & Karel Hron*
- R package: *mvoutlier* – an interactive play for your to identify outliers


Outlier
y
x
4
2
0
-2
-4
-4
-2
0
2
4


Chi^2-Plot
Outlier
Quantiles of Chi_p^2
Ordered robust MD^2
0
2
4
6
8
10
12
0
10
20
30
40
50

Outliers > 2d?

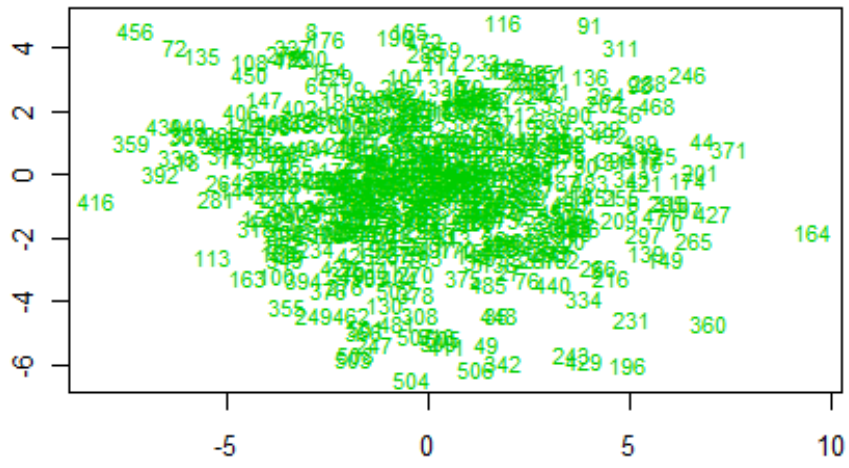


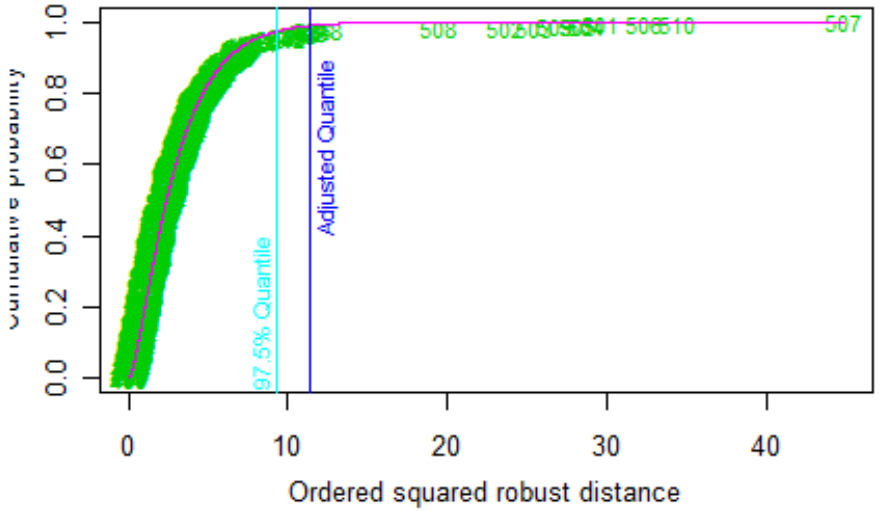
I: Quantile of χ^2

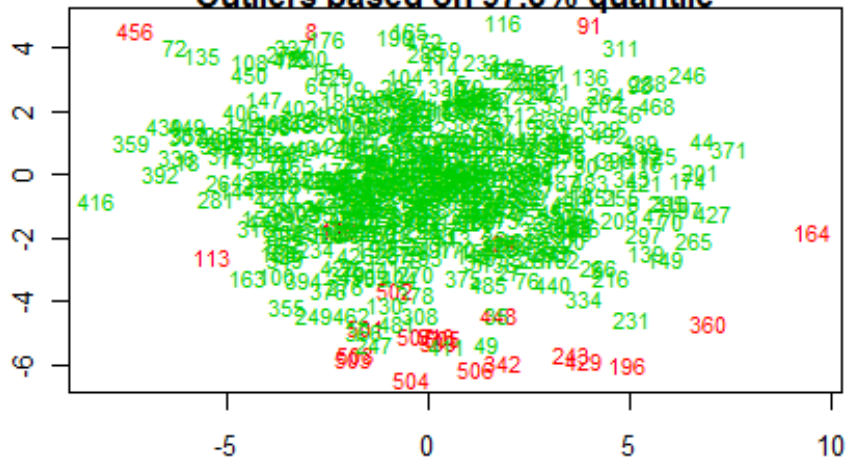
1. Compute robustly estimated MD for each point;
2. Compute the 97.5% Quantile, Q, of the χ^2_p distribution
3. Samples with MD > Q are declared outliers

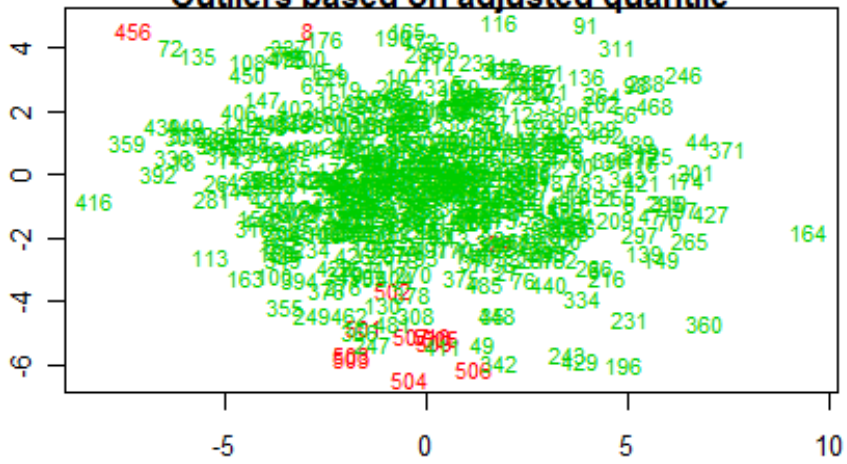
II: Adjusted Quantile for Outliers

1. Compare cdf of χ^2_p with the ecdf of samples at tails
2. Outliers have “abnormally large” deviations at tails
3. aq.plot() in R package *mvoutlier*




Adjusted Quantile
97.5% Quantile
Ordered squared robust distance


Outliers based on 97.5% quantile


Outliers based on adjusted quantile

III: *pcout()* – trailer for PCA

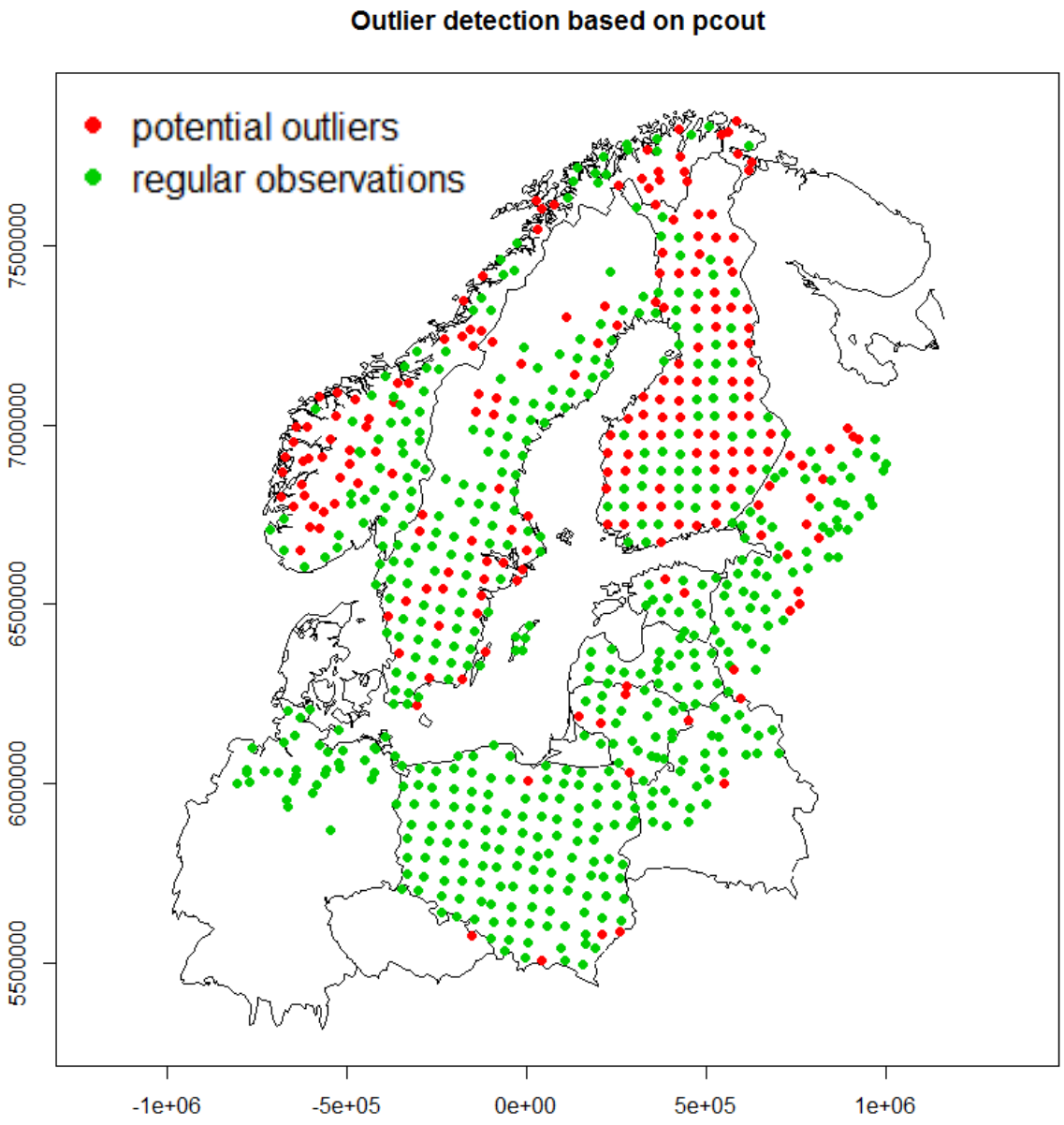
Topic for next week

- Robust principal components
- Complex, very much involved, very fast, especially good for high dimensions

- Yet, it is ready for you in R: *pcout()* in *mvoutlier*

Example:

data(*bsstop*)$_{768X46}$



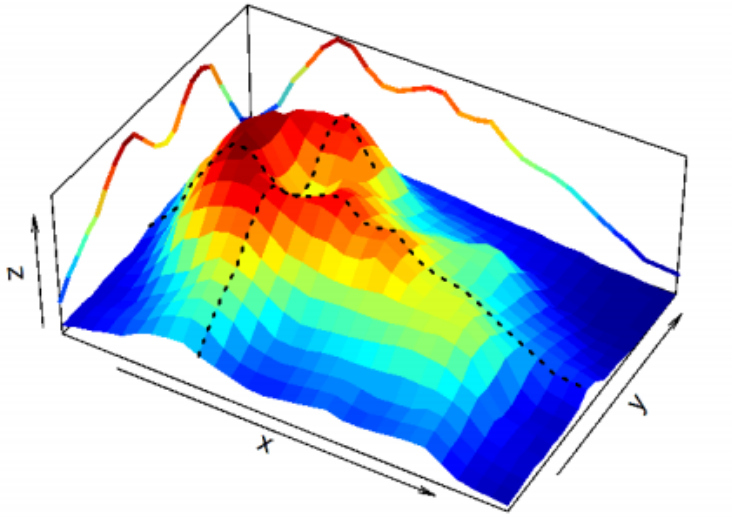
A Quick Summary

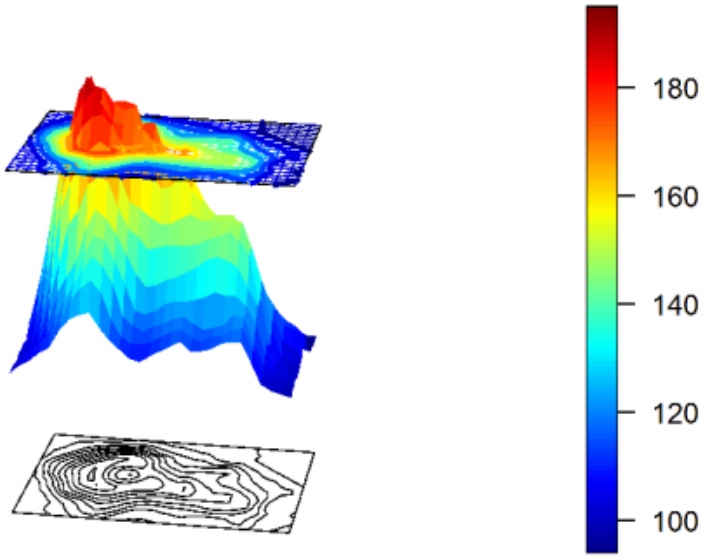
- Detecting multivariate outliers with
  - estimated Mahalanobis distance
  - QQ-plot
  - *Chisq.plot*, *pcout* (package('mvoutlier'))
- Technical details:
  - Estimated Mahalanobis Distance from sample
  - Use robust estimates for μ, Σ

More Examples of EDA-V

The Key to EDA is VISUALIZATION

SOME EXAMPLES



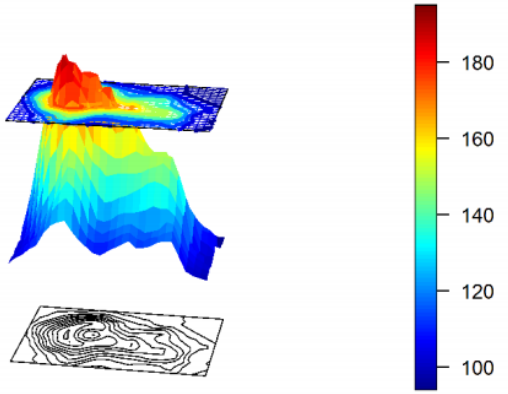


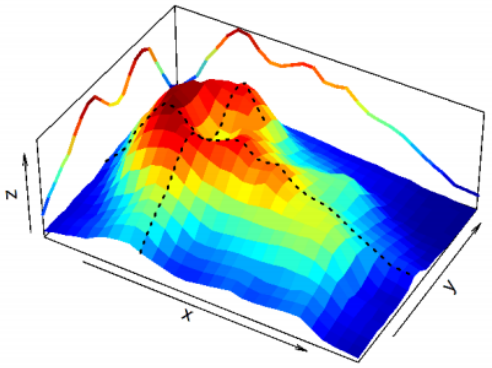
Package('plot3D')

Gridded data

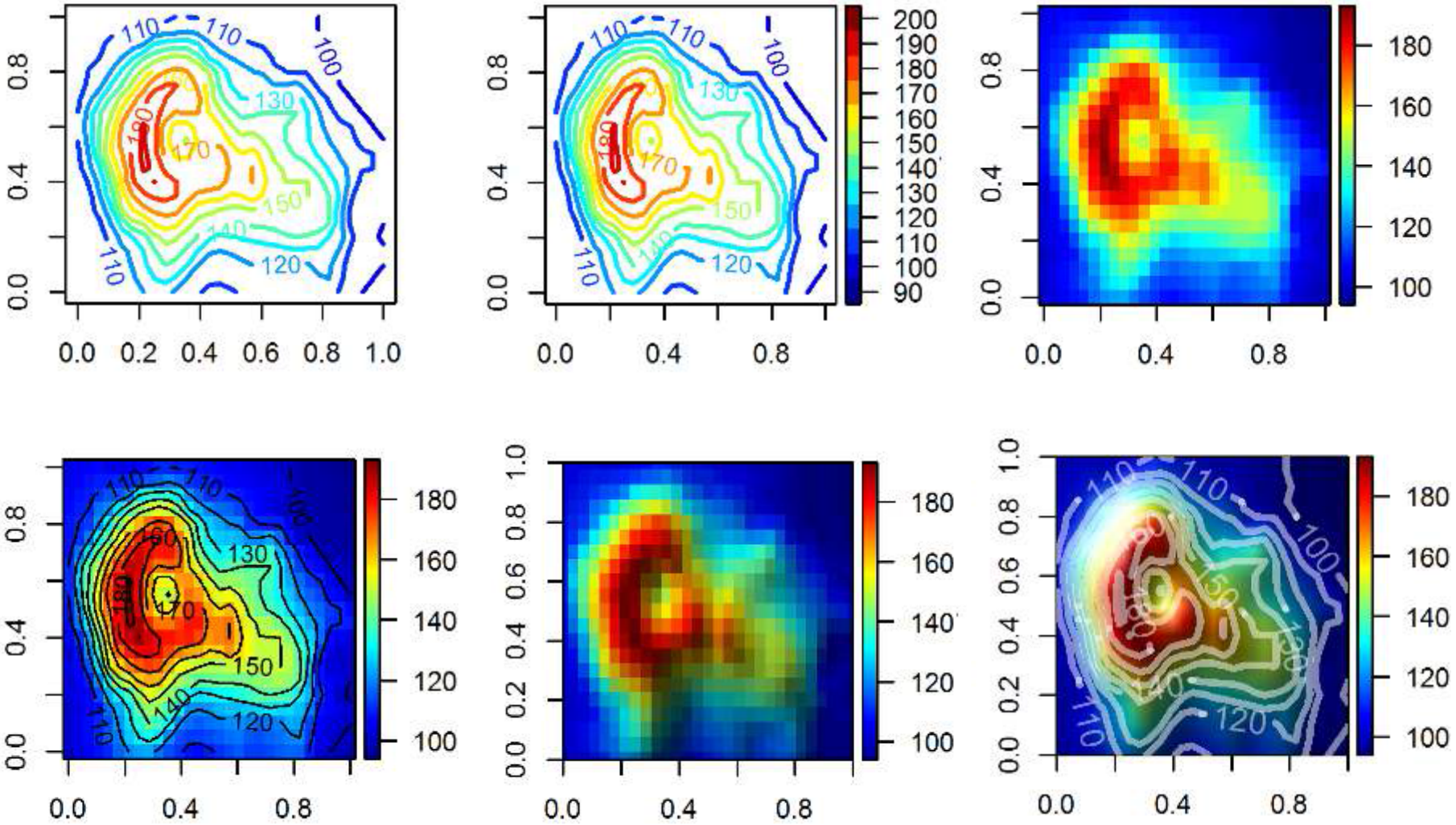
Volcano

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
1	100	101	101	100	102	102	103	101	104	107	108	110
2	103	104	104	103	104	105	106	107	111	118	120	124
3	105	107	107	105	106	109	114	120	123	129	140	142
4	108	110	110	108	113	120	127	136	141	150	158	157
5	110	115	114	117	124	133	150	155	161	165	169	174
6	116	118	121	123	130	147	160	170	179	181	183	187
7	120	126	128	130	136	152	167	178	186	191	193	191
8	122	130	135	139	147	161	172	182	190	189	184	182
9	123	133	140	146	154	164	175	183	185	177	167	164
10	118	129	137	145	151	163	173	180	180	169	158	153
11	114	120	131	138	146	154	164	174	179	169	157	149
12	111	114	120	130	139	147	155	168	177	174	166	161
13	108	112	117	121	132	144	153	164	178	179	176	170
14	107	112	115	120	128	140	150	164	174	179	176	166
15	109	113	117	121	129	141	148	159	166	168	164	159
16	111	115	118	124	131	142	148	160	168	168	160	153
17	113	117	120	125	132	142	150	166	170	170	163	155
18	115	118	121	125	134	142	152	159	162	160	157	150
19	112	115	119	126	136	143	150	155	155	152	148	145
20	112	114	117	127	139	145	150	150	150	149	142	140
21	113	116	118	129	140	146	150	150	150	147	139	136



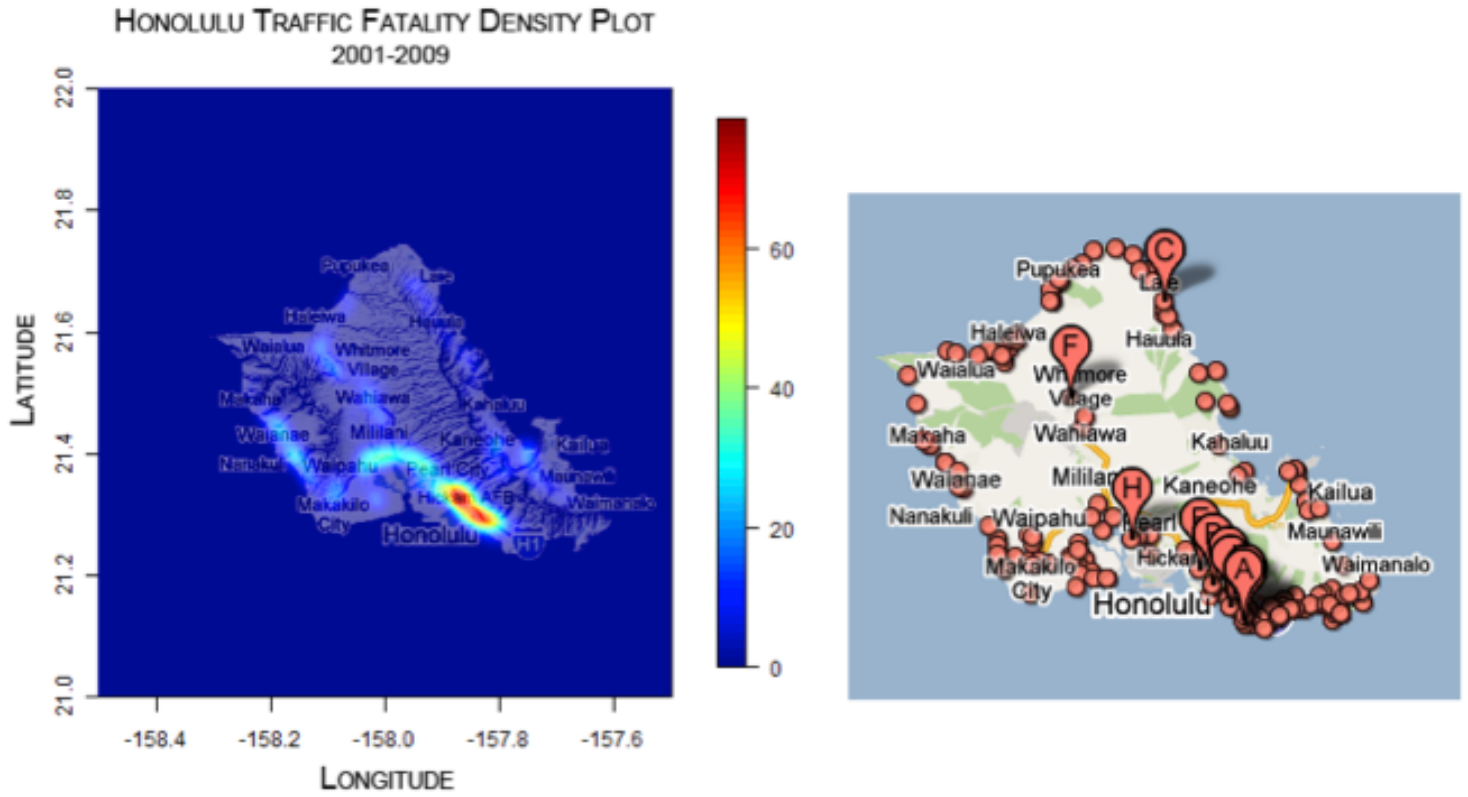


Numerous ways…



Location, date and time

Example: Traffic fatalities in Honolulu 2001 – 2009



Locations (start, finish), date and time

Volcano | NYCflights.R | MUCflights.R | airports | flight.info

6344 observations of 11 variables

	AirportID	Name	City	Country	IATA	ICAO	Latitude	Longitude	A
1	1	Goroka	Goroka	Papua New Guinea	GKA	AYGA	-6.081689	145.391881	5
2	2	Madang	Madang	Papua New Guinea	MAG	AYMD	-5.207083	145.788700	2
3	3	Mount Hagen	Mount Hagen	Papua New Guinea	HGU	AYMH	-5.826789	144.295861	5
4	4	Nadzab	Nadzab	Papua New Guinea	LAE	AYNZ	-6.569828	146.726242	2
5	5	Port Moresby Jacksons Intl	Port Moresby	Papua New Guinea	POM	AYPY	-9.443383	147.220050	1
6	6	Wewak Intl	Wewak	Papua New Guinea	WWK	AYWK	-3.583828	143.669186	1
7	7	Narsarsuaq	Narssarssuaq	Greenland	UAK	BGBW	61.160517	-45.425978	1
8	8	Nuuk	Godthaab	Greenland	GOH	BGGH	64.190922	-51.678064	2
9	9	Sondre Stromfjord	Sondrestrom	Greenland	SFJ	BGSF	67.016969	-50.689325	1
10	10	Thule Air Base	Thule	Greenland	THU	BGTL	76.531203	-68.703161	2
11	11	Akureyri	Akureyri	Iceland	AEY	BIAR	65.659994	-18.072703	6
12	12	Egilsstadir	Egilsstadir	Iceland	EGS	BIEG	65.283333	-14.401389	7
13	13	Hornafjordur	Hofn	Iceland	HFN	BIHN	64.295556	-15.227222	2
14	14	Husavik	Husavik	Iceland	HZK	BIHU	65.952328	-17.425978	4
15	15	Isafjordur	Isafjordur	Iceland	IFJ	BIIS	66.058056	-23.135278	8
16	16	Keflavik Nas	Keflavik	Iceland	KEF	BIKF	63.985000	-22.605556	1
17	17	Patreksfjordur	Patreksfjordur	Iceland	PFJ	BIPA	65.555833	-23.965000	1

Displayed 1000 rows of 6344 (5344 omitted)

Console ~/

```
      lsk     fnr                          lvg hal ha2 ha3            haf          hafen
52126   L SQ  328 Singapore Airlines          SIN                Singapur      Singapore
52127   L LH 2557 Lufthansa                    TBS                 Tiflis         Tbilisi
52128   L LH 1789 Lufthansa                    YEI                 Bursa          Bursa
52129   L LH  765 Lufthansa                    BOM        Mumbai (Bombay) Mumbai (Bombay)
52130   L QR  009 Qatar Airways                DOH                   Doha           Doha
52131   L LH 2541 Lufthansa                    LED           St.Petersburg   St Petersburg
                      stt                 ett                   lde                len ter ber
52126 2011-01-07 05:10:00 2011-01-07 05:10:00             Singapur           Singapur   2   H
52127 2011-01-07 05:55:00 2011-01-07 05:40:00             Georgien           Georgien   2   H
52128 2011-01-07 05:55:00 2011-01-07 05:35:00               T}rkei             T}rkei   2   H
52129 2011-01-07 06:00:00 2011-01-07 07:50:00               Indien             Indien   2   H
52130 2011-01-07 06:45:00 2011-01-07 06:35:00                Katar              Katar   2   H
52131 2011-01-07 06:50:00 2011-01-07 08:20:00 Russische F|deration Russische F|deration   2   H
       typ ver saa gat
52126 B77W      49 I32
52127 A319 100  49 I30
52128 CRJ9      49
52129 A343      49 I08
52130 A332      49 I38
52131 A320 200  49 I20
> View(airports)
>
```

Environment | History

Import Dataset | Clear | List

Global Environment

Data

airports	6344 obs. of 11 variables
flight.info	774 obs. of 18 variables
flights	0 obs. of 0 variables
myflights	9 obs. of 18 variables
myroutes	1553 obs. of 35 variables

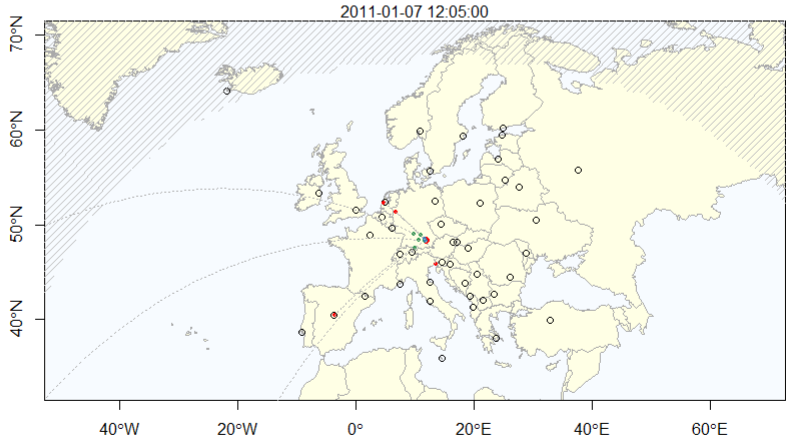
Values

all	List of 0
missing	int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
months	int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
needed	chr [1:12] "2013-1.csv" "2013-2.csv" "2013-3.csv" "2013-4.cs...

Functions

Files | Plots | Packages | Help | Viewer

Zoom | Export | Clear All



Example: Flights information from Munich Airport