

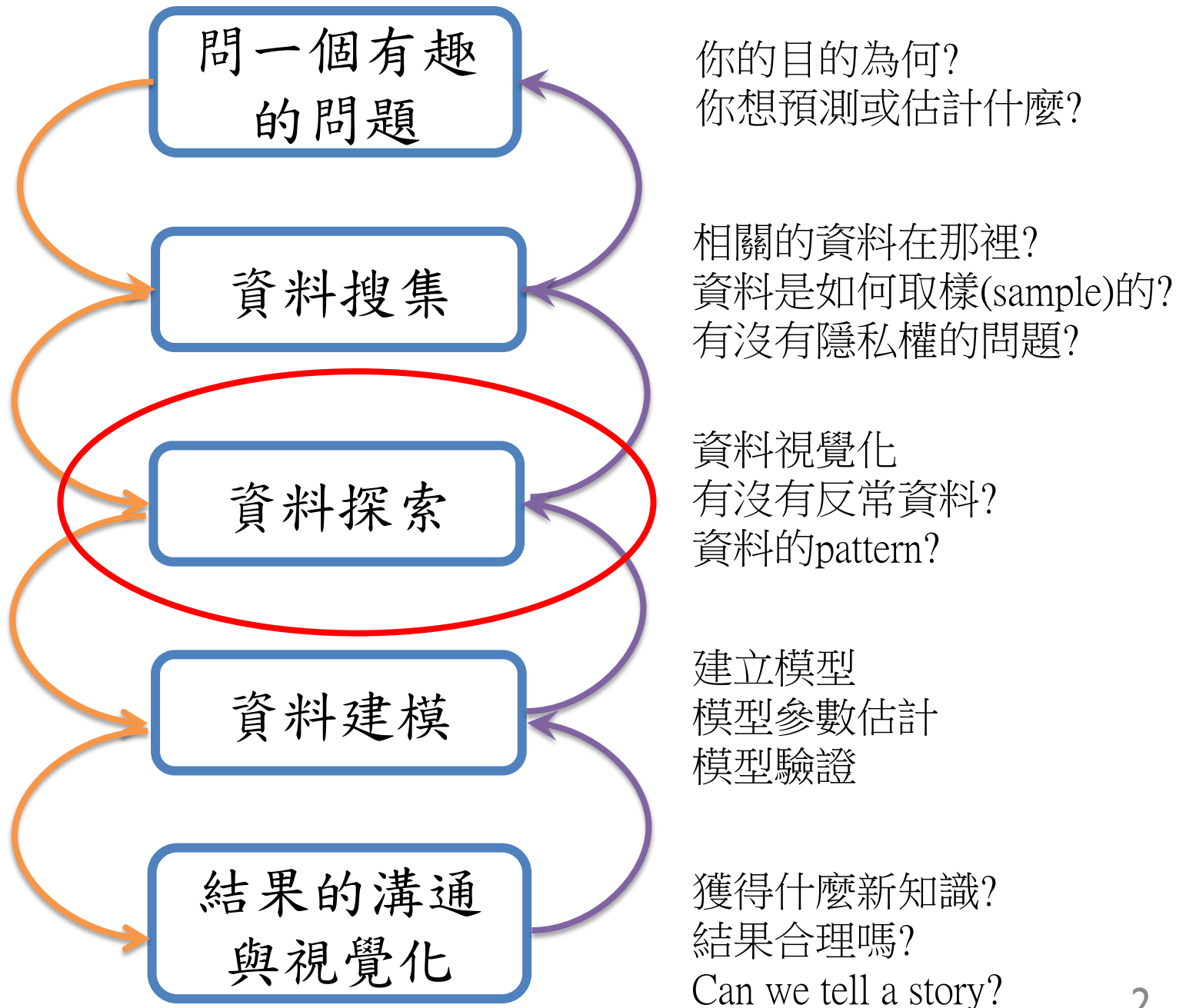
Lecture 7: Exploratory data analysis

IST5573

統計方法 Statistical methods

2016/10/26

分析流程



Exploratory data analysis (EDA)

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

-John Tukey (1915 - 2000)



EDA definition

- An approach to analyzing data sets to summarize their main characteristics, often with **visual methods**, and a statistical model can be used or not. [Wikipedia]
- Primarily EDA is for seeing **what the data can tell us beyond the formal modeling or hypothesis testing task**. [Wikipedia]

Anscombe's quartet

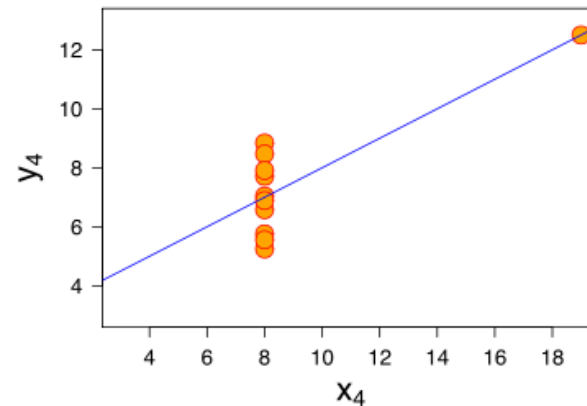
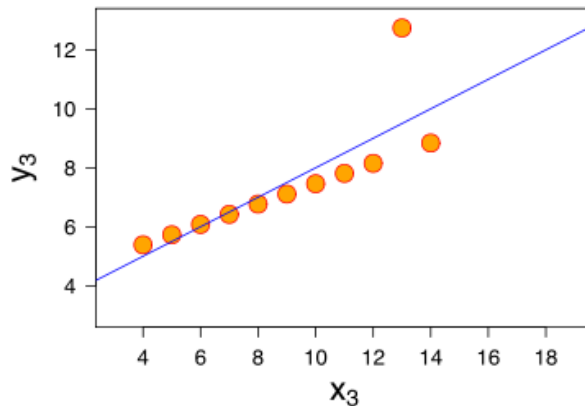
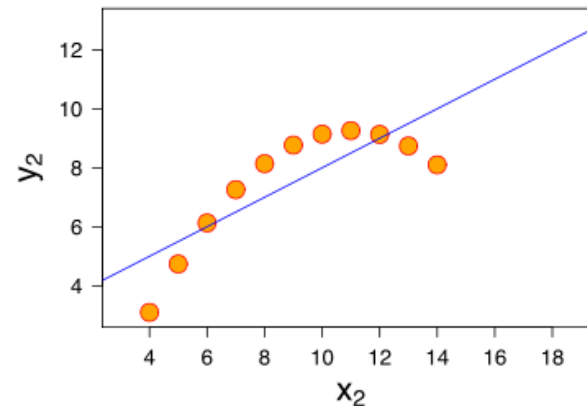
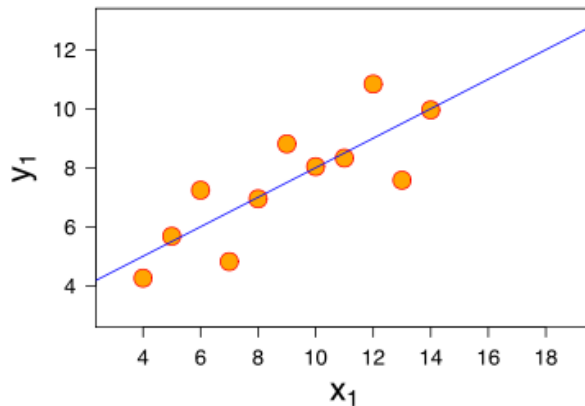
Same mean, variance, correlation, and linear regression line

Anscombe's Quartet: Raw Data								
I		II		III		IV		
x	y	x	y	x	y	x	y	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25	
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50	
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56	
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91	
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89	
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

Anscombe '73

Anscombe's quartet

Same mean, variance, correlation, and linear regression line



http://en.wikipedia.org/wiki/Anscombe's_quartet

Measurement scales

- Nominal (Categorical) (N)
Are = or \neq to other values
Apples, Oranges, Bananas,...
- Ordinal (O)
Obey a $<$ relationship
Small, medium, large
- Quantitative (Q)
Can do arithmetic on them
10 inches, 23 inches, etc.

Measurement scales

- Q - Interval (location of zero arbitrary)
Dates: Jan 19; Location: (Lat, Long)
Like a geometric point. Cannot compare directly.
Only differences (i.e., intervals) can be compared
- Q - Ratio (zero fixed)
Measurements: Length, Mass, Temp, ...
Origin is meaningful, can measure ratios & proportions
Like a geometric vector, origin is meaningful

Data types

Measurement scale	Statistics		Computer programming
interval scale	continuous data		floating-point
ratio scale			
ratio scale	discrete data	count	integer
nominal/ordinal scale		binary	boolean
nominal scale		nominal	integer/character
ordinal scale		ordinal	

Example: Gene expression microarray data

- Data from a study using gene expression profiling to predict breast cancer outcomes (<http://www.nature.com/nature/journal/v415/n6871/full/415530a.html>)
- 78 breast cancer: 44 remained disease-free for an interval of at least five years after their initial diagnosis (good prognosis group), while 34 patients had developed distant metastases within five years (poor prognosis group).

samplexprs.csv

Variable	Description
id	An unique identification number
age	Age at diagnosis of breast cancer (year)
metastases	Developing distant metastases: 0=no (good prognosis group), 1=yes (poor prognosis group)
followup	Follow-up time (year)
ERp	ER- α expression level
J00129	\log_{10} gene expression intensity ratios
Contig29982_RC	\log_{10} gene expression intensity ratios

- RMD_example 7.1

id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100				-0.153
DE65	40	0	6.25	0				0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	-0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Variable names

Example: Gene expression microarray data (samplexprs.csv)

id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.153
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.003	-0.003	0.108
LD44	33	1	1.4	0	0.786	-0.003	-0.003	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Item

Example: Gene expression microarray data
(samplexprs.csv)

id	age	metastases	followup	ERp	100129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961			-0.153
DE65	40	0	6.25	0	-0.749			0.032
HG87	53	0	5.18	0	-0.426			0.429
HG88	37	1	1.09	100	-0.566			-0.336
AB22	37	0	5.8	90	-0.42			-0.048
HG91	30	1	1.03	0	-0.499			0.143
HG92	39	1	3.36	80	-0.465			0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	-0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Attribute
Feature
Variable

Example: Gene expression microarray data
(samplexprs.csv)

id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.153
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0			
SF67	48	1	1.98	0	-0			
LD44	33	1	1.4	0	0			
AA04	41	0	13	50	-0			
AA01	43	0	12.53	80	-0			
GL73	52	1	2.13	0	1			
AA10	49	0	11.16	80	-0			
HG86	54	0	5.89	50	-0			
DE62	40	0	6.97	50	-0			
AB26	41	0	8.17	10	-0			
SF57	41	1	2	0	-0			
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

1 = Continuous (interval)
2 = Continuous (ratio)
3 = Binary
4 = Ordinal
5 = Nominal

Example: Gene expression microarray data
(samplexprs.csv)

Resources for R Graphics

- CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
<http://cran.r-project.org/web/views/Graphics.html>
- R Graphics 2nd Edition
<https://www.stat.auckland.ac.nz/~paul/RG2e/>
- ggplot2
<http://ggplot2.org/>
- Cookbook for R graphs
<http://www.cookbook-r.com/Graphs/>
- GGobi Data Visualization System
<http://www.ggobi.org/>

R graphic package: ggplot2

- `ggplot2` = `g`rammar of `g`raphics
- A replacement for the base graphics in R
- `qplot()` :
 - Short for quick plot
 - Wraps up all the details of `ggplot`
 - Simple to use, but gives the least control
- `ggplot() + ...` : Add, remove or alter components in a plot at a high level of abstraction

Univariate data (1 dimension)

Displaying distributions

- Stem-and-leaf plot
- q-q plot
- Histogram
- Box plot
- Bar chart
- Pie chart

Stem-and-leaf plot

For data x_1, x_2, \dots, x_n ,

1. Divide each number x_i into two parts: a **stem**, consisting of one or more of the leading digits and a **leaf**, consisting of the remaining digit.
 2. List the stem values in a vertical column.
 3. Record the leaf for each observation beside its stem.
 4. Write the units for stems and leaves on the display.
- Displaying the relative density and shape of the data, giving the reader a quick overview of distribution.
 - Retain (most of) the raw numerical data. Also useful for highlighting outliers and finding the mode.

age	J00129
52	-0.795
50	-0.509
54	-0.961
40	-0.749
53	-0.426
37	-0.566
37	-0.42
30	-0.499
39	-0.465
45	-0.189
30	-0.739
48	-0.601
33	0.786
41	-0.819
43	-0.448
52	1.206
49	-0.391
54	-0.234
40	-0.75
41	-0.299
41	-0.455
45	-1.173
48	-0.721
48	-0.416
44	-0.688
38	-0.352
51	-0.734
48	-0.112
36	-0.919

Stem-and-leaf plot

The decimal point is 1 digit(s) to the right of the |

```

2 | 88
3 | 00234
3 | 677788889999
4 | 0011111112333444
4 | 5555566667888888999
5 | 0012222222333444444444

```

age

The decimal point is at the |

```

-2 | 0
-1 |
-1 | 321000
-0 | 99999988888888887777777766666666666666555555555555
-0 | 4444444433221
0 | 01
0 | 689
1 | 2

```

J00129

- RMD_example 7.2

Quantile

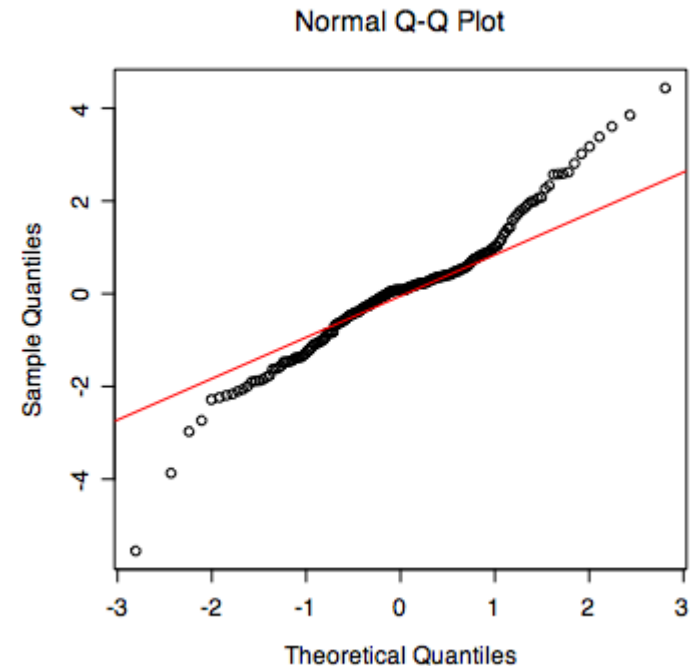
- The $(\frac{p}{100})$ -th quantile (or the p -th percentile) of a list of a distribution x is defined as the number q that is bigger than $p\%$ of numbers

$$\Pr(x \leq q) = \frac{p}{100}$$

- For example, the 50-th percentile is the median.

q-q plot

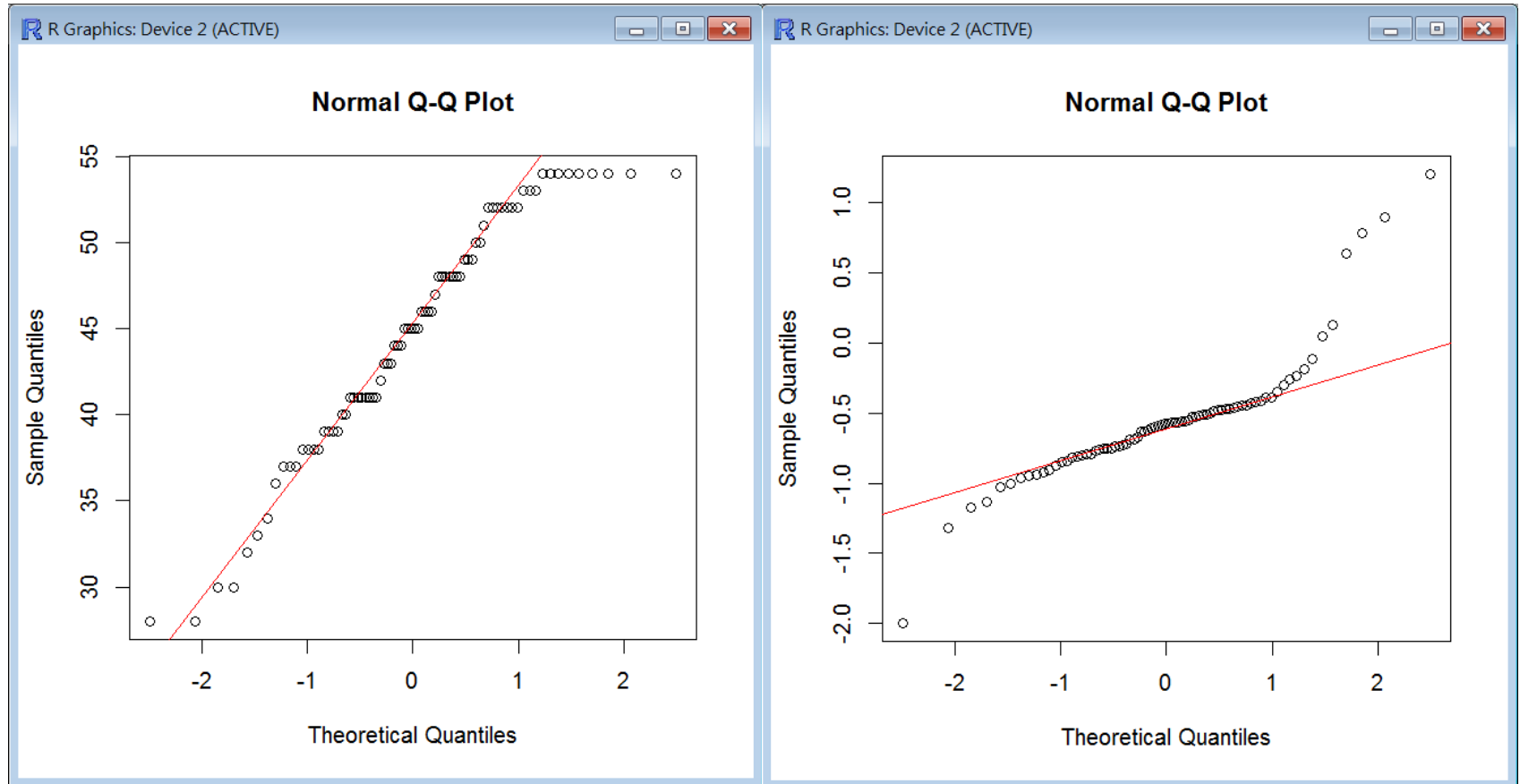
- “q” stands for quantile
- A graphical method for comparing two probability distributions by plotting their quantiles against each other
- If the two distributions being compared are similar, the points in the q–q plot will approximately lie on the line $y = x$



q-q plot

age

J00129

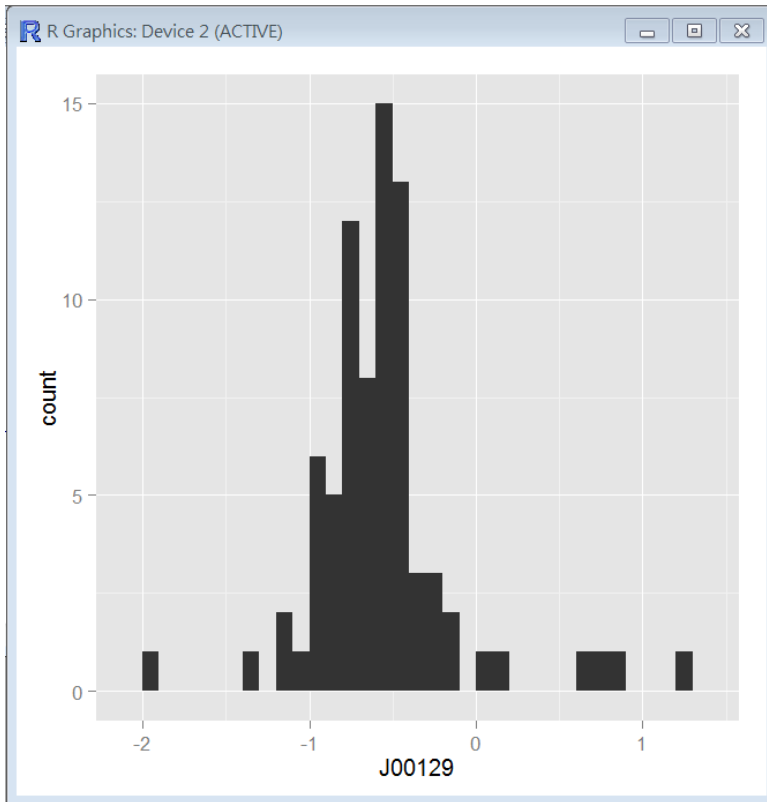


- RMD_example 7.3

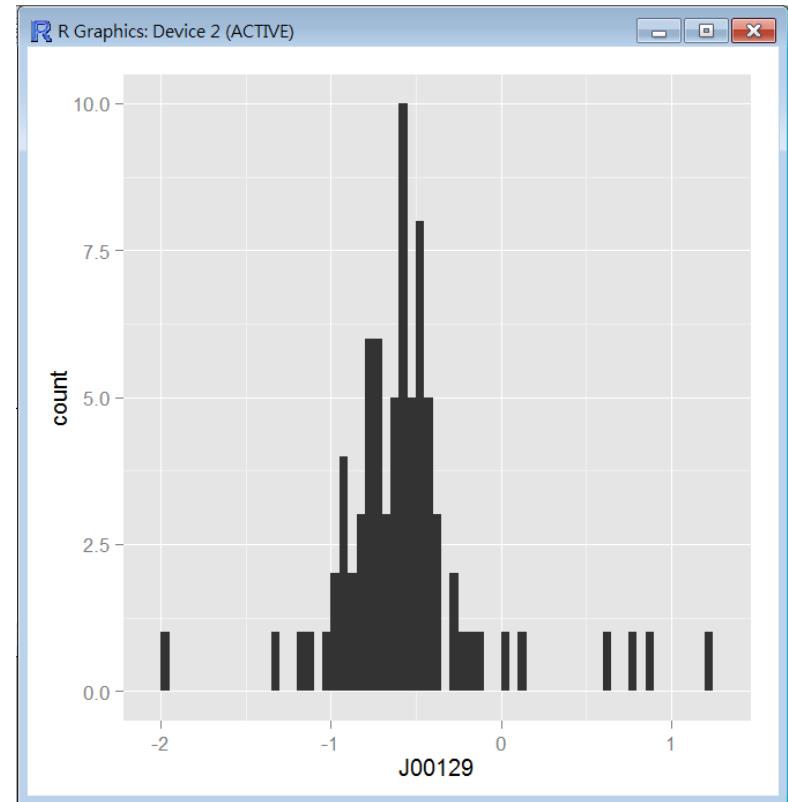
Histogram

To construct a histogram for **continuous data**, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**, or **bins**. If possible, the bins should be of equal width to enhance the visual information in the histogram.

Histogram



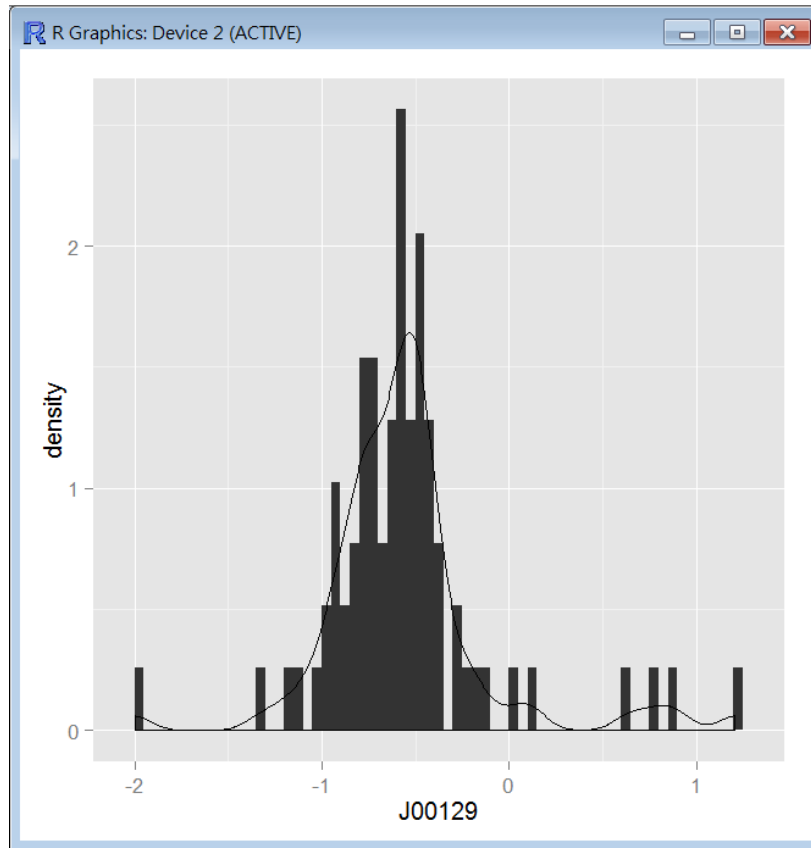
binwidth=0.1



binwidth=0.05

- RMD_example 7.4

Histogram



Display a smooth
density estimate

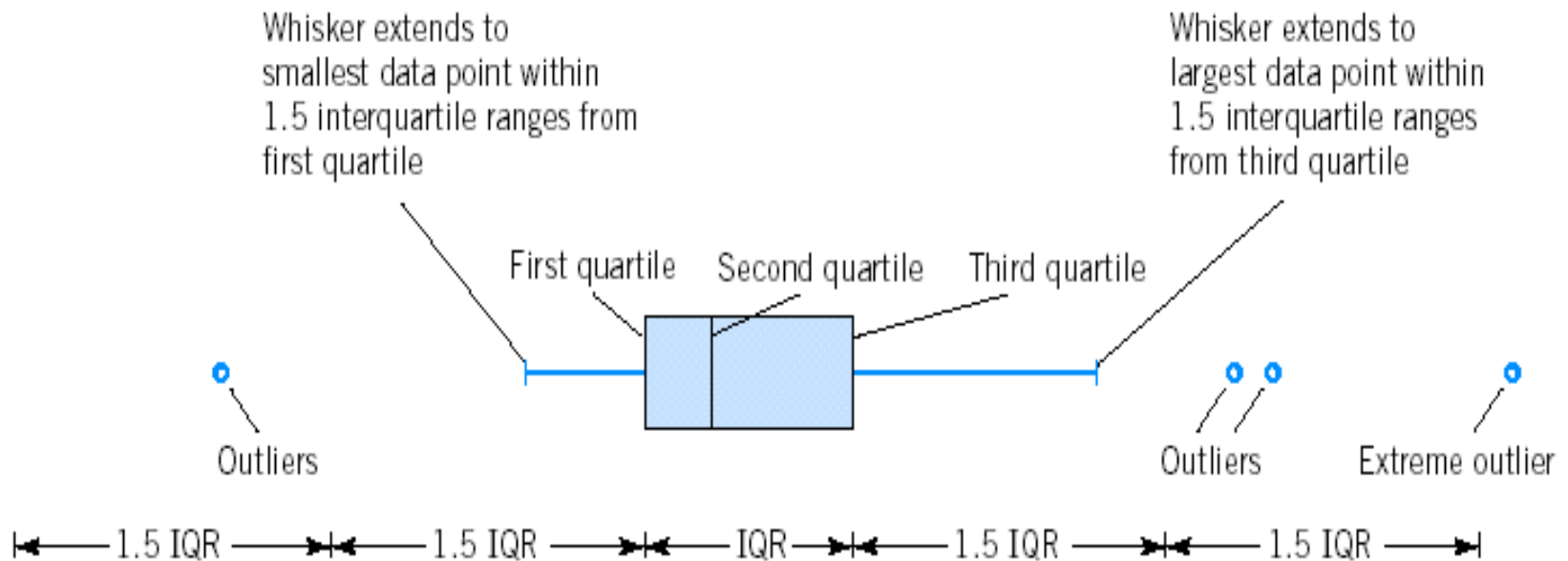
- RMD_example 7.4

Quartile

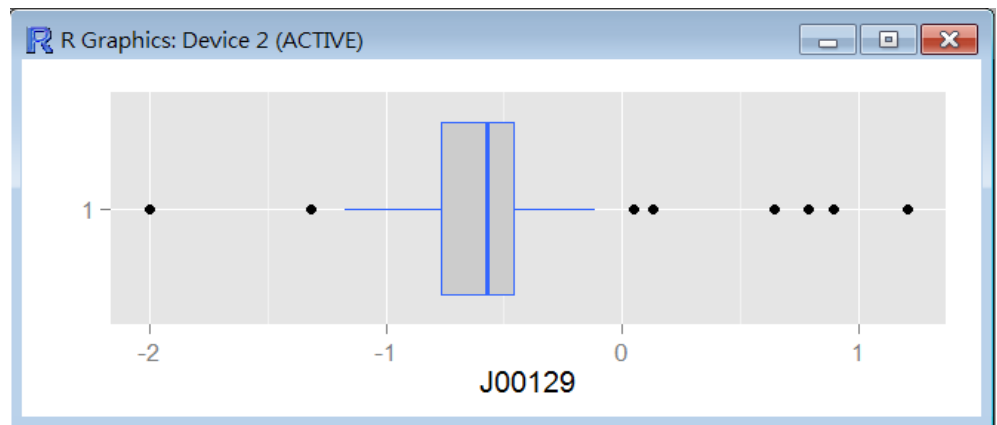
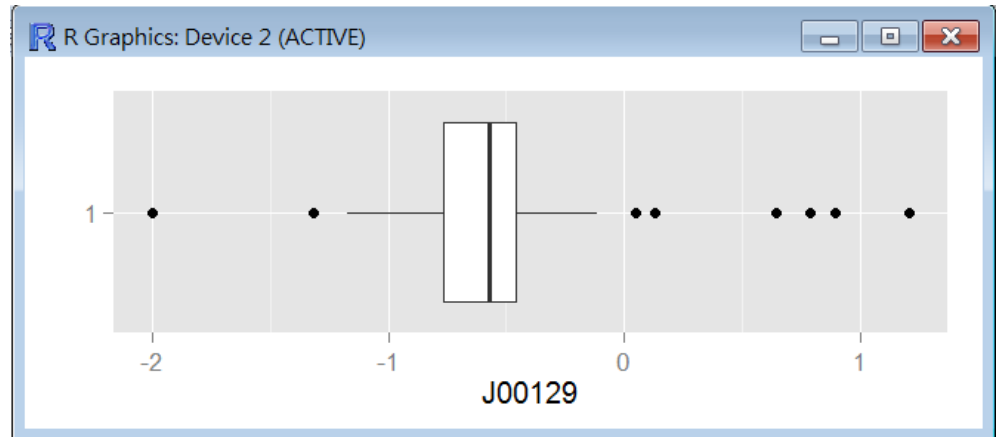
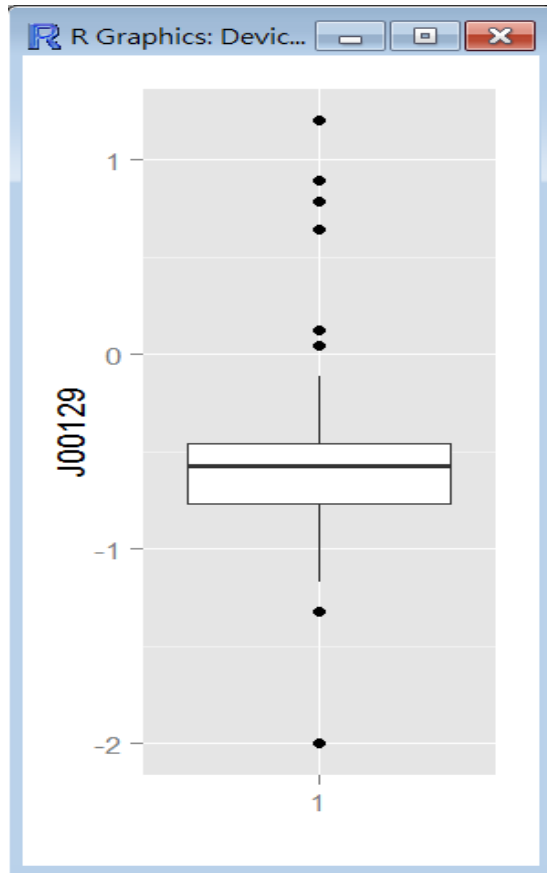
- First quartile ($Q1$) = 25-th percentile
- Second quartile ($Q2$) = 50-th percentile
- Third quartile ($Q3$) = 75-th percentile
- Interquartile range (IQR) = $|Q3 - Q1|$

Boxplot

The boxplot describes center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data.



Boxplot

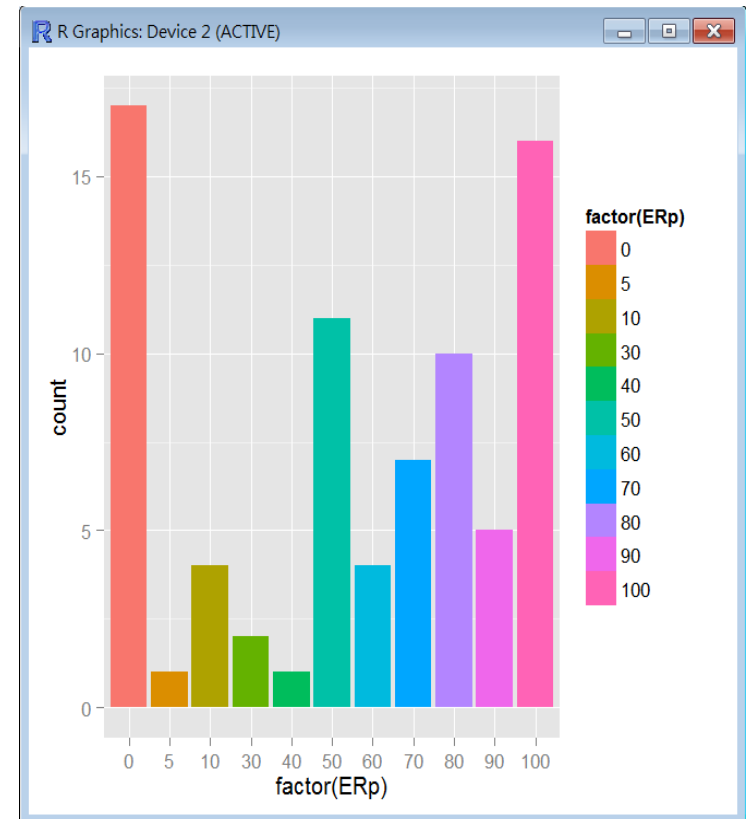
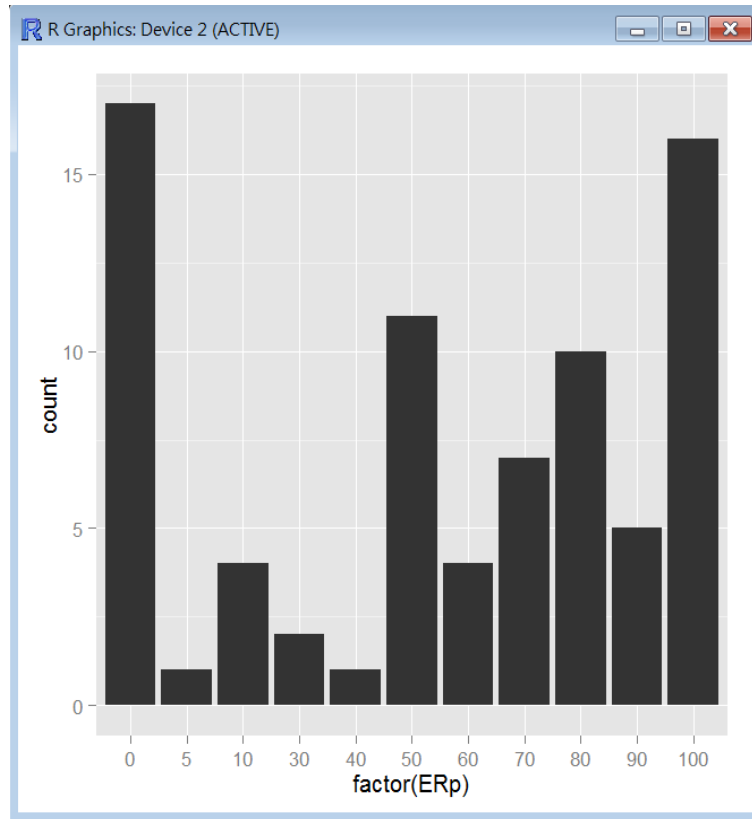


- RMD_example 7.5

ERp
100
0
100
0
0
100
90
0
80
50
70
0
0
50
80
0
80
50
50
10
0
100
90
100
50
80
5
100
30

Bar chart

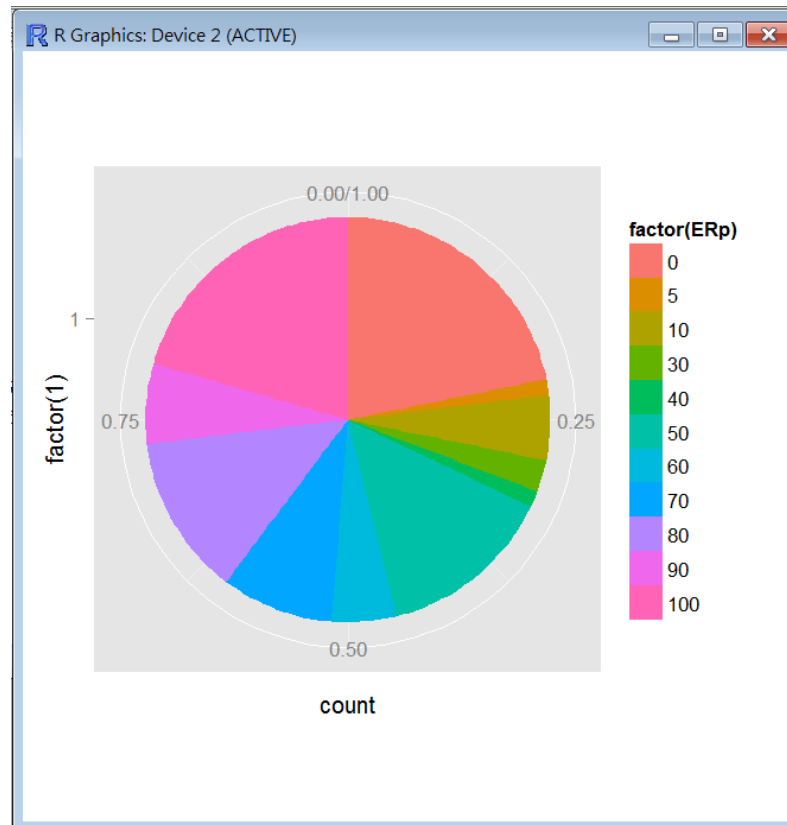
For categorical data



- RMD_example 7.6

Pie chart

For categorical data with proportions



- RMD_example 7.7

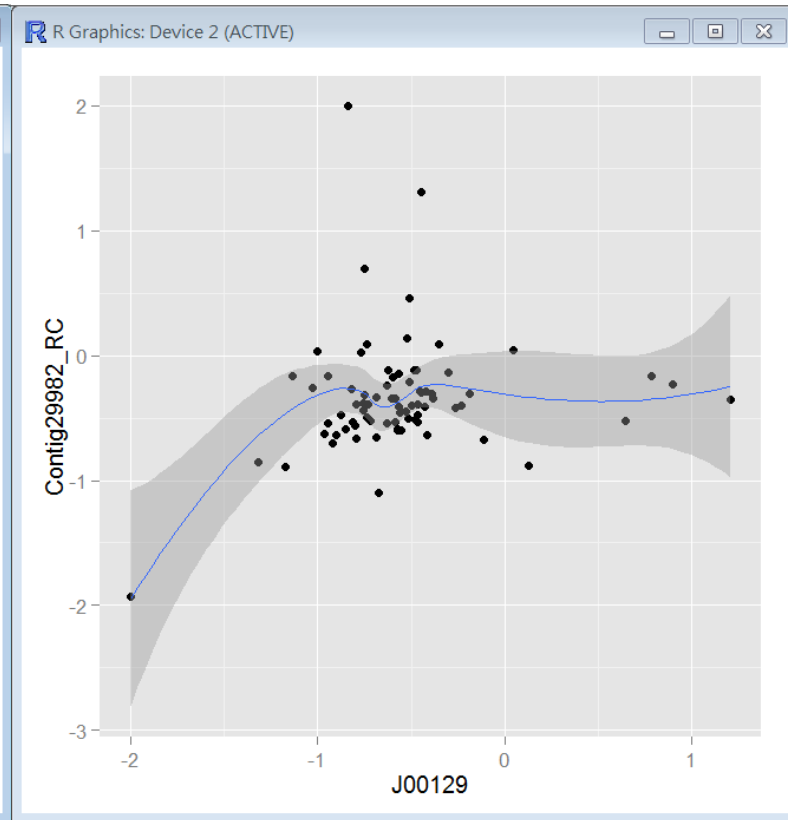
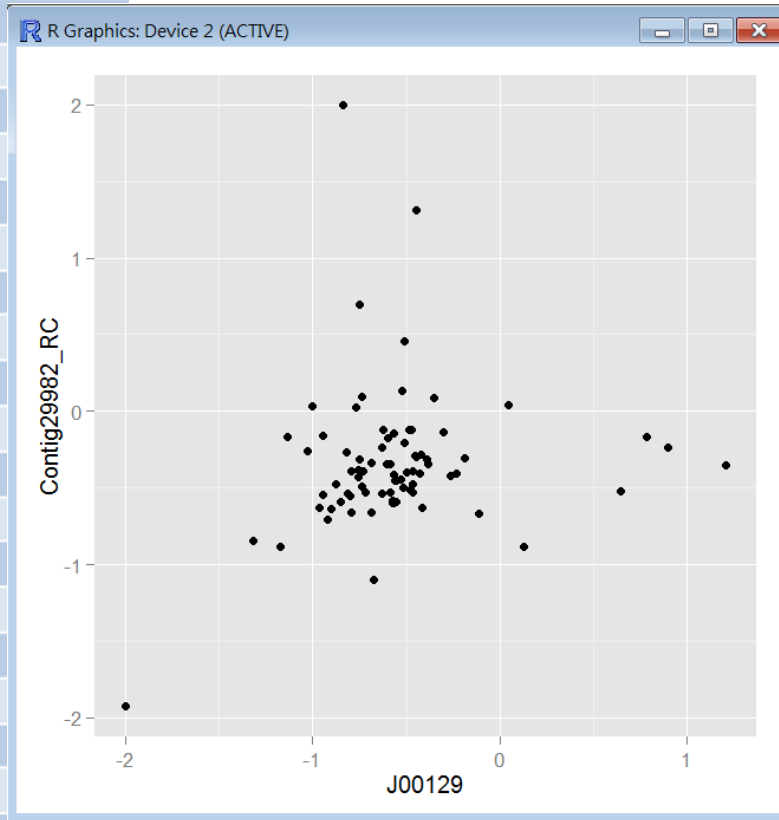
Bivariate data (2 dimensions)

Displaying correlations

- Scatterplot
- Box plots
- Stacked bar chart
- Faceting bar charts
- Stacked area chart
- Time series plot

J00129	Contig29982_RC
-0.795	-0.387
-0.509	0.459
-0.961	-0.631
-0.749	0.699
-0.426	-0.406
-0.566	-0.596
-0.42	-0.286
-0.499	-0.402
-0.465	-0.533
-0.189	-0.309
-0.739	0.093
-0.601	-0.177
0.786	-0.164
-0.819	-0.267
-0.448	-0.296
1.206	-0.353
-0.391	-0.31
-0.234	-0.404
-0.75	-0.316
-0.299	-0.137
-0.455	-0.288
-1.173	-0.887
-0.721	-0.527
-0.416	-0.633
-0.688	-0.659
-0.352	0.088
-0.734	-0.493
-0.112	-0.67
-0.919	-0.704

Scatterplot



Add trend line

- RMD_example 7.8

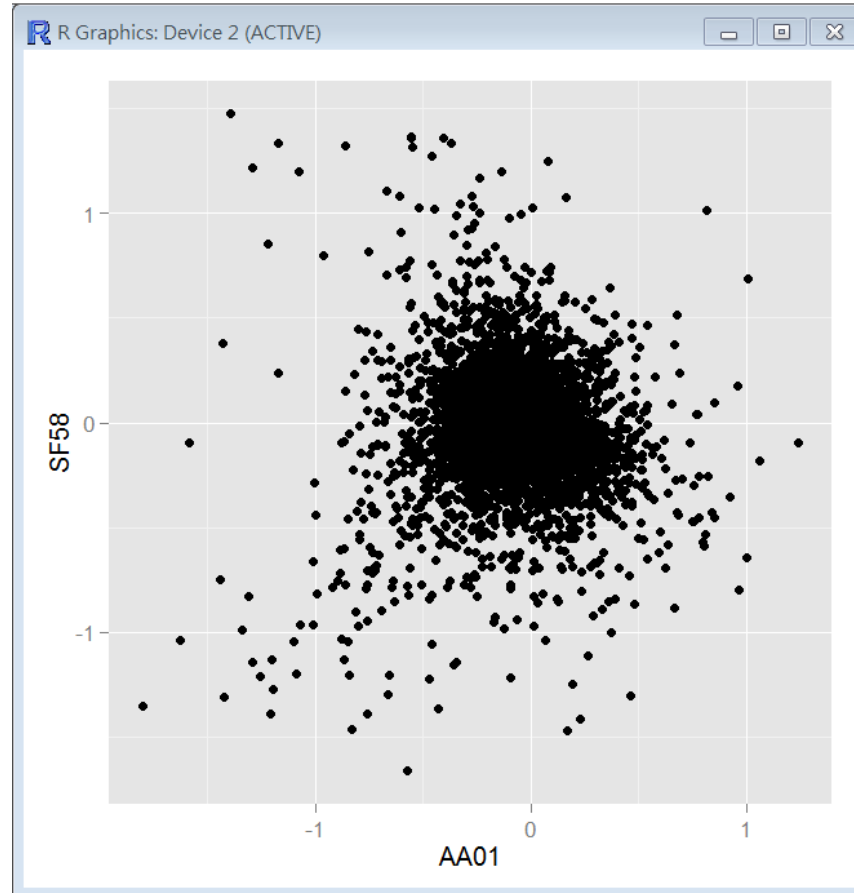
exprs_sig.csv

	FG80	SF58	DE72	DE65
J00129	-0.795	-0.509	-0.961	-0.749
Contig29982_RC	-0.387	0.459	-0.631	0.699
Contig42854	0.199	-0.257	0.037	-0.346
Contig42014_RC	-0.247	-0.065	-0.153	0.032
Contig27915_RC	0.176	0.129	0.144	0.3
Contig20156_RC	-0.129	0.009	-0.202	-0.025
Contig50634_RC	-0.111	0.021	0.192	-0.067
Contig42615_RC	0.119	0	-0.19	-0.226
Contig56678_RC	0.231	-0.649	-0.086	-0.018
Contig48659_RC	0.118	0.058	-0.052	-0.278
Contig49388_RC	0.035	-0.038	0.055	0.13
Contig1970_RC	-0.482	-0.105	0.013	-0.338
Contig26343_RC	0.015	0.053	-0.123	0.038
Contig53047_RC	-1.389	-0.601	-1.378	-0.007
Contig43945_RC	-0.011	0.005	0.113	-0.277
Contig19551	-0.092	0.295	-0.806	-1.106

- RMD_example 7.9

	AA01	SF58
J00129	-0.448	-0.509
Contig299	-0.296	0.459
Contig428	-0.1	-0.257
Contig420	-0.177	-0.065
Contig279	-0.107	0.129
Contig201	-0.11	0.009
Contig506	-0.095	0.021
Contig426	-0.076	0
Contig566	-0.134	-0.649
Contig486	-0.14	0.058
Contig493	0.006	-0.038
Contig197	0.111	-0.105
Contig263	-0.236	0.053
Contig530	-0.866	-0.601
Contig439	0.126	0.005
Contig195	-0.692	0.295
Contig104	0.132	0.006
Contig472	0.095	-0.25
Contig207	0.252	-0.384
AL157502	0.139	-0.185
Contig366	-0.097	-0.775
D31887	0.113	-0.04
AB033006	-0.209	0.608
AB033007	0.107	-0.13
M83822	0.098	0.046
AB033025	0.11	-0.127
AF114264	0.096	-0.108
Contig406	0.305	-0.008
Contig173	0.055	-0.142

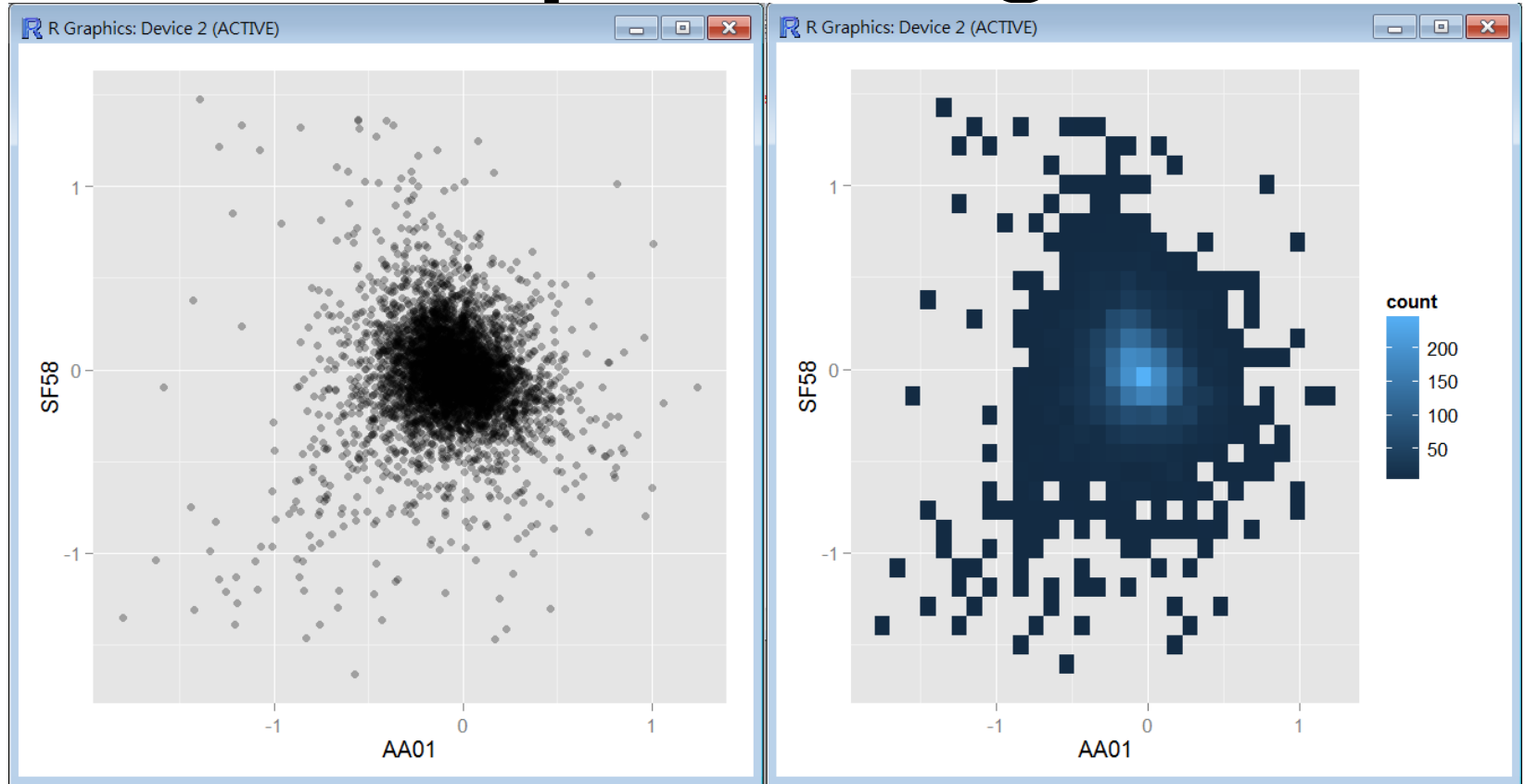
Scatterplot for big data



For large datasets with overplotting

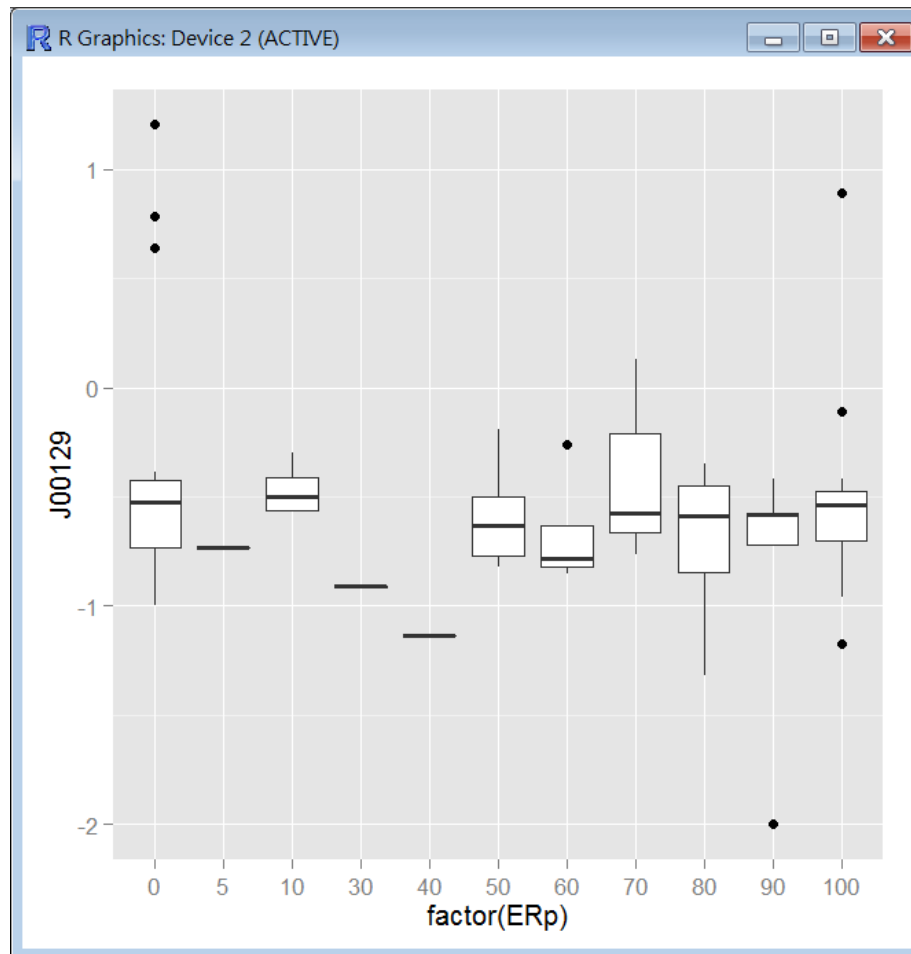
- RMD_example 7.10

Scatterplot for big data



- Alpha aesthetic makes the points more transparent
- Heatmap shows the density
 - **RMD_example 7.10**

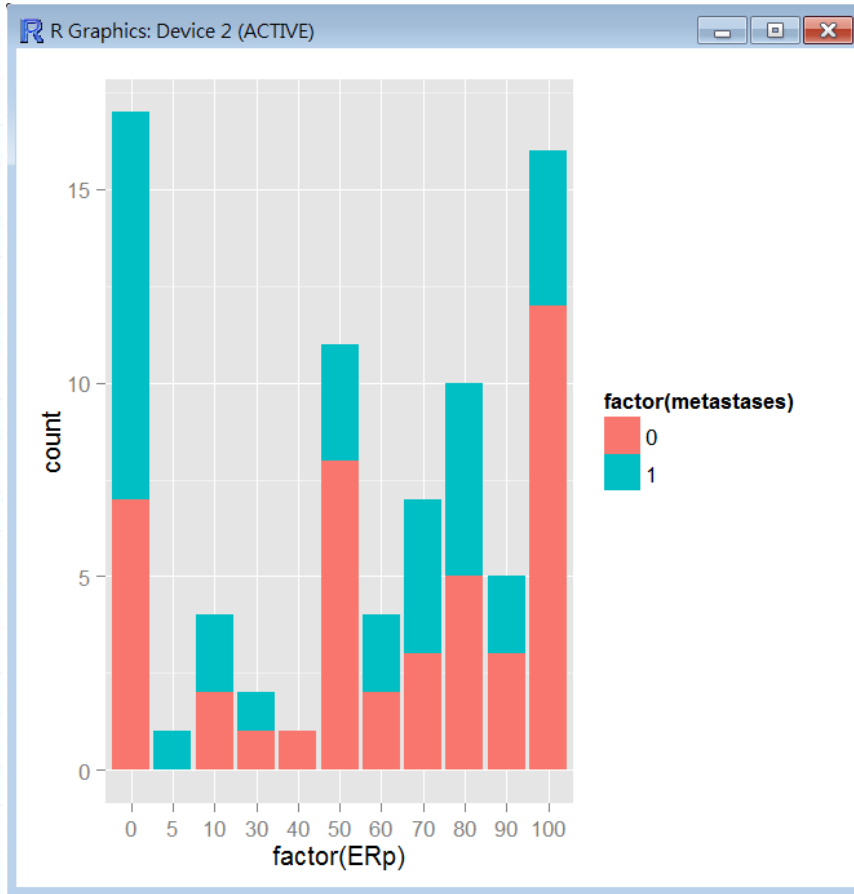
Box plots for different ERp



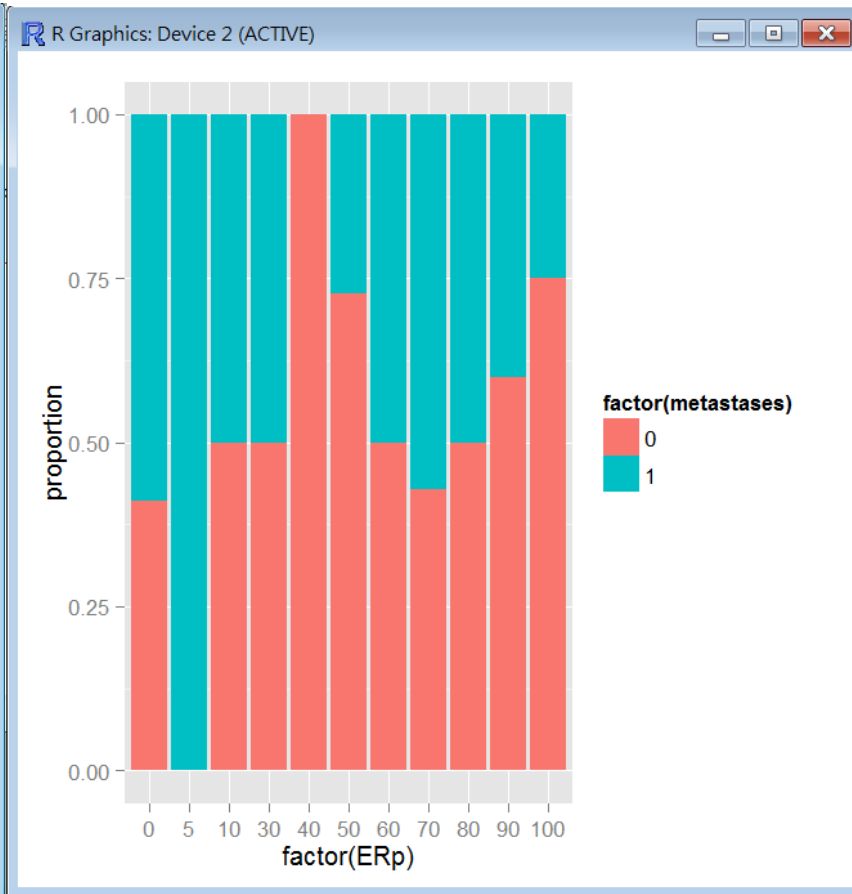
- RMD_example 7.11

Stacked bar chart

For counts

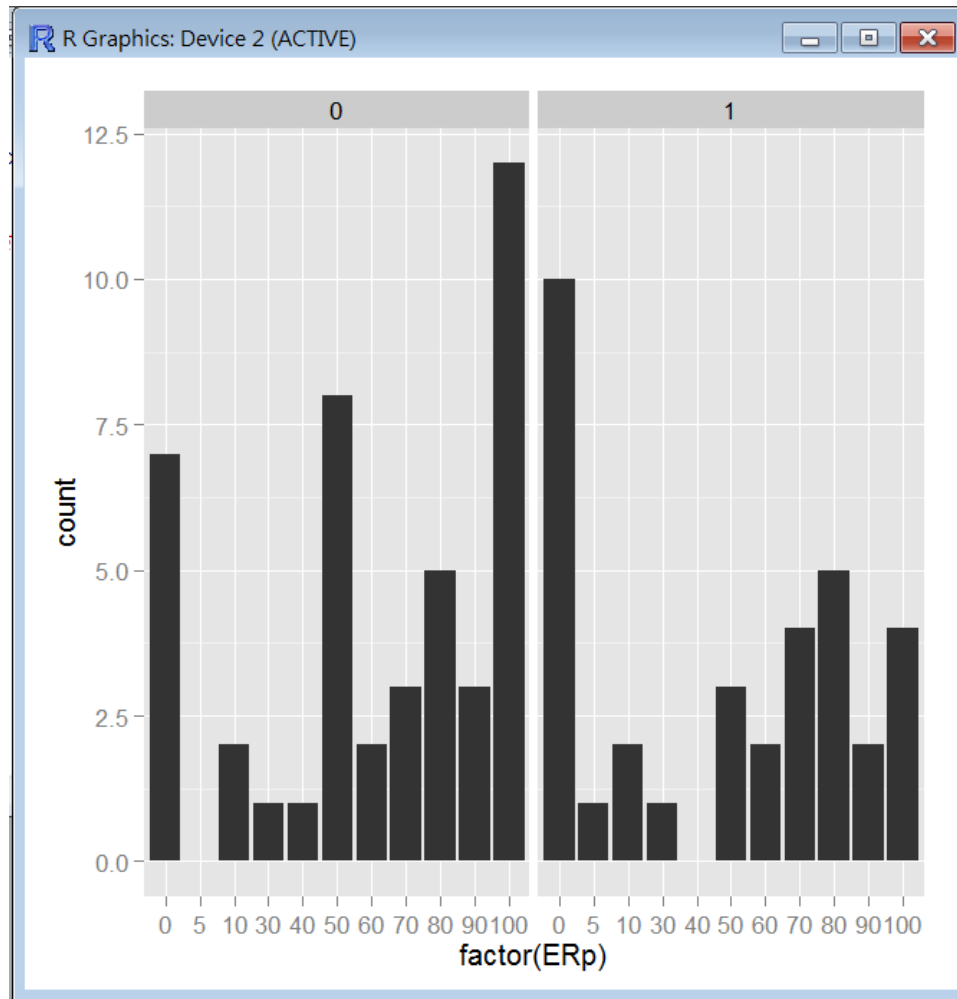


For proportions



- RMD_example 7.12

Faceting bar charts

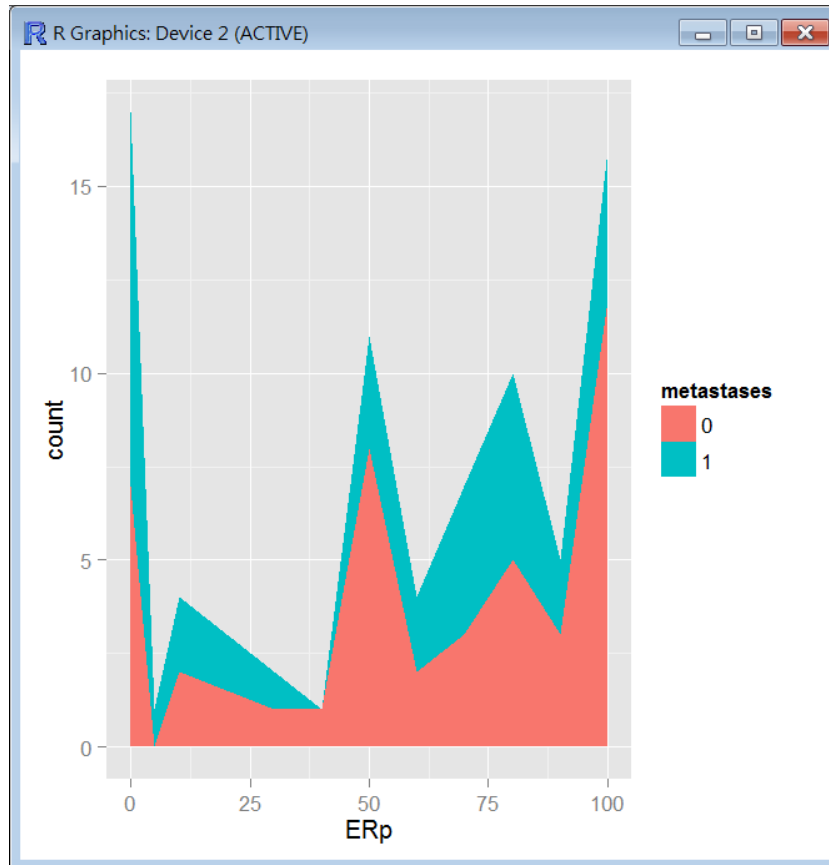


Bar charts of
ERp for
different
metastases

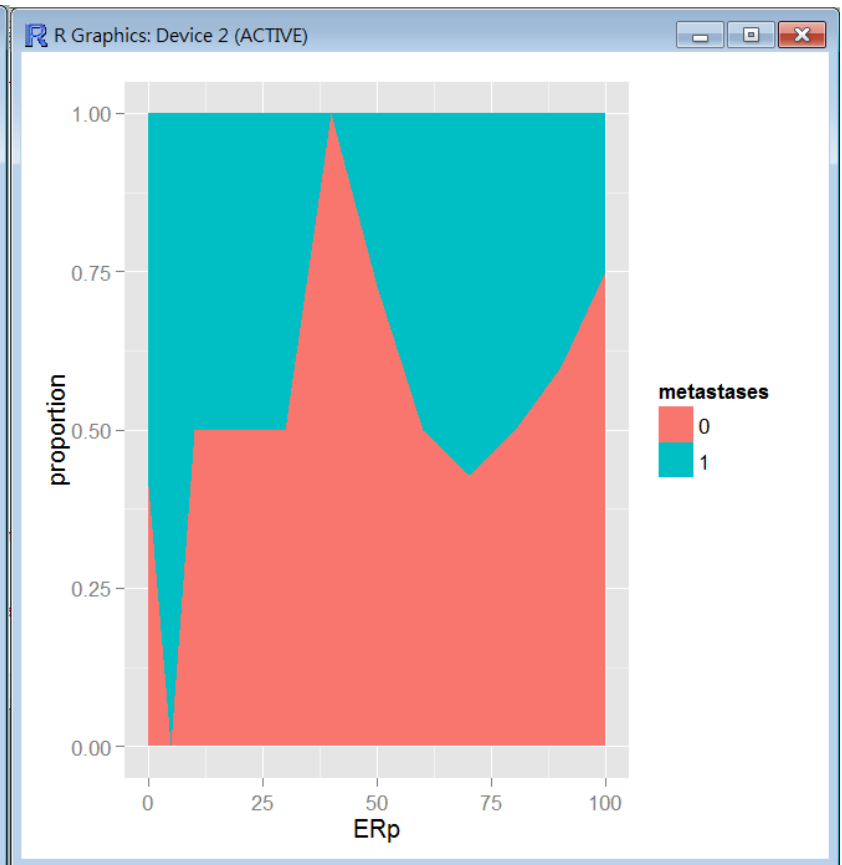
- RMD_example 7.13

Stacked area chart

For counts



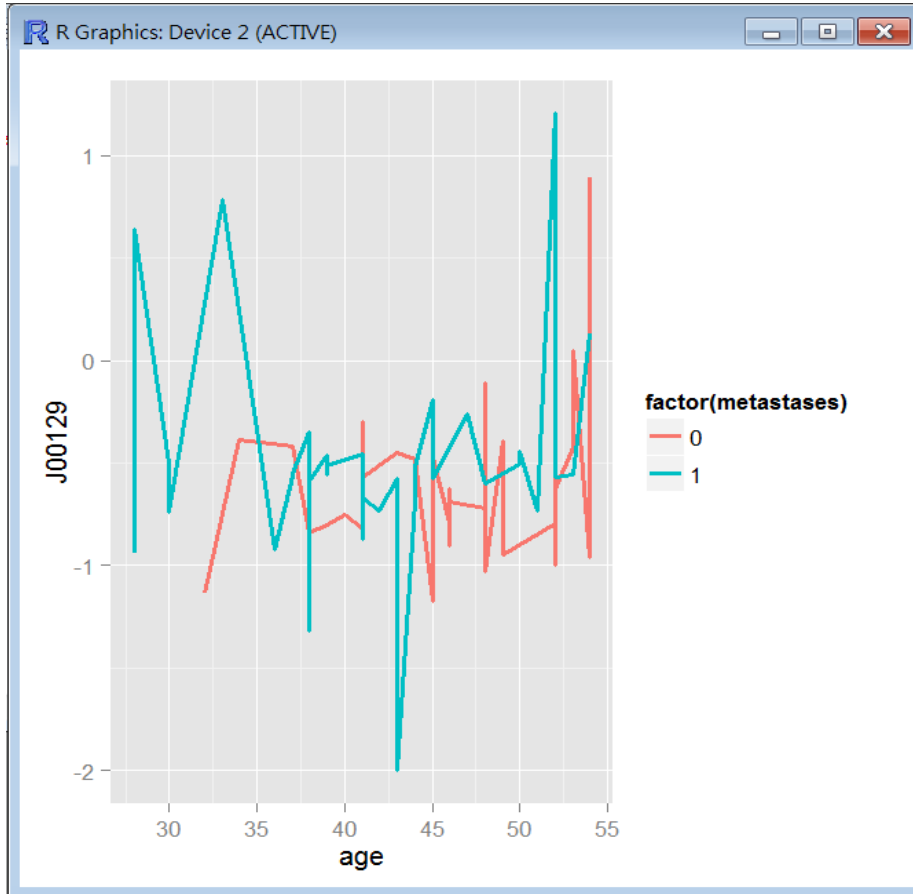
For proportions



Treat ERp as continuous!

- RMD_example 7.14

Time series plot



- Connect observations, ordered by x value
- Measurements are plotted as a time series
- See trends, cycles over time

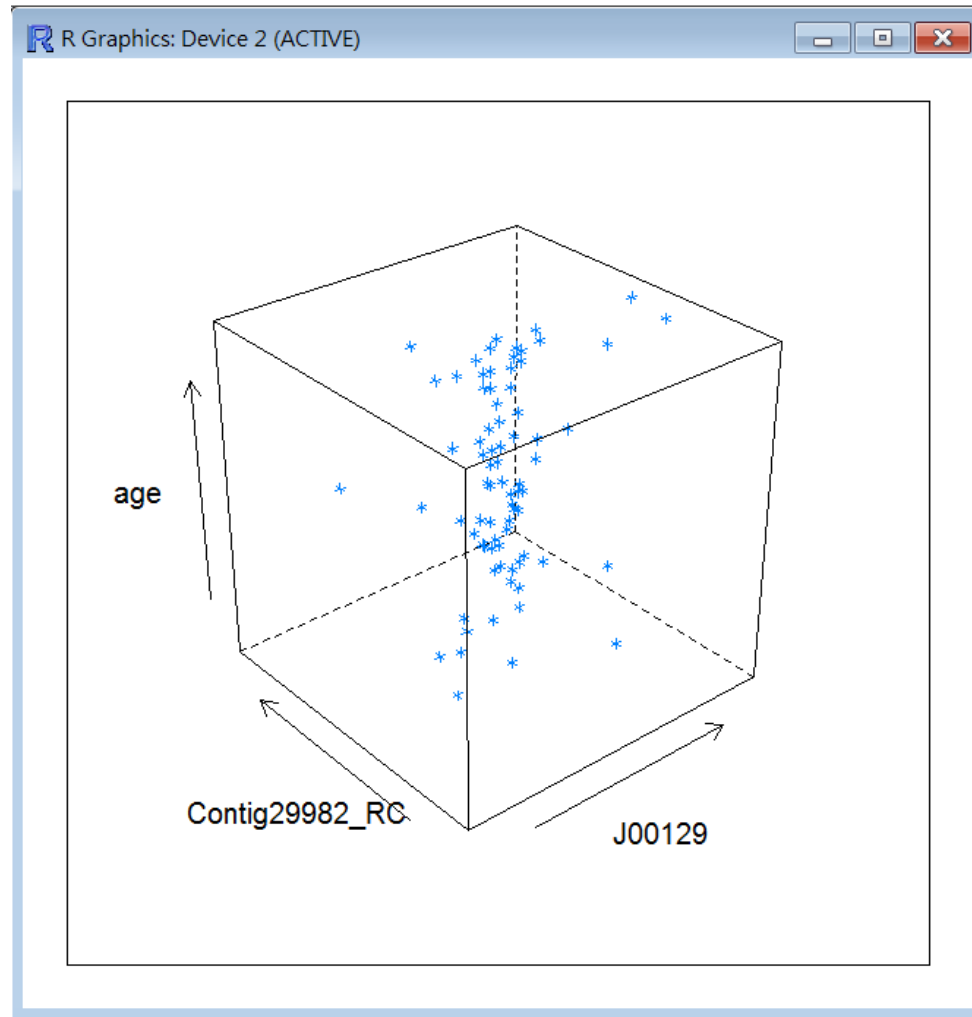
● RMD_example 7.15

Multivariate data (≥ 3 dimensions)

Displaying association

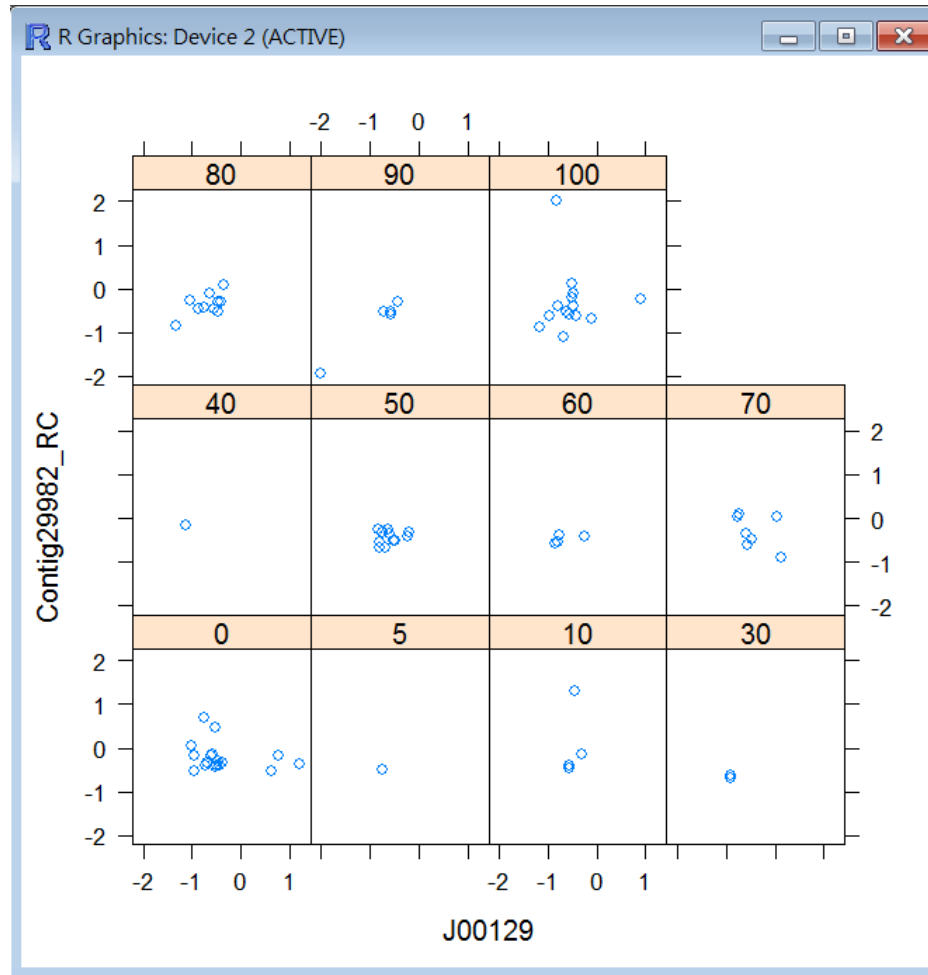
- 3d scatterplot
- Lattice in the 3rd dim
- Map the 3rd dim to colors
- Lay out panels in the 3rd dim
- Scatterplot matrices
- Heatmap

3d scatterplot



- RMD_example 7.16

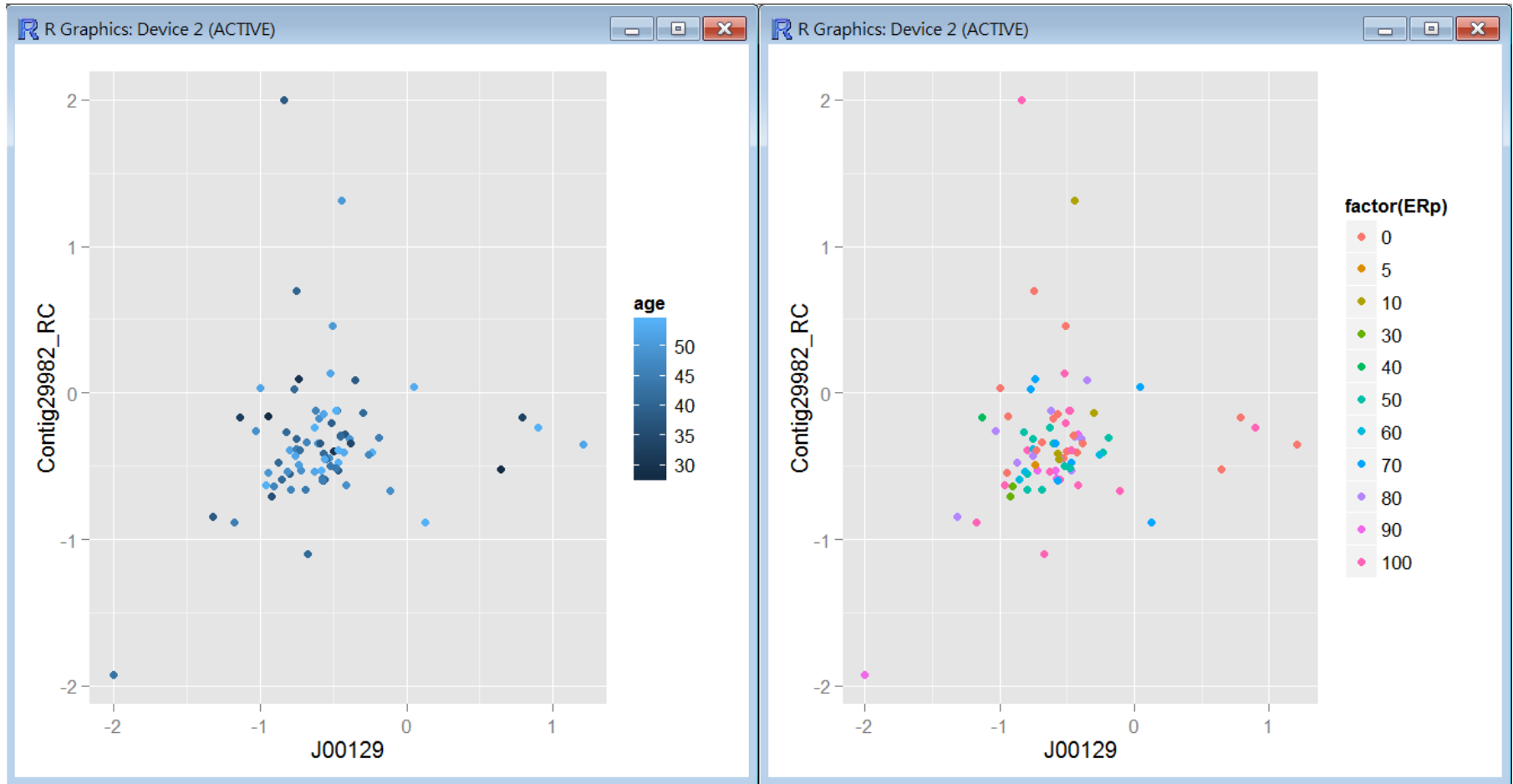
Lattice in the 3rd dim



Lattices
in ERp

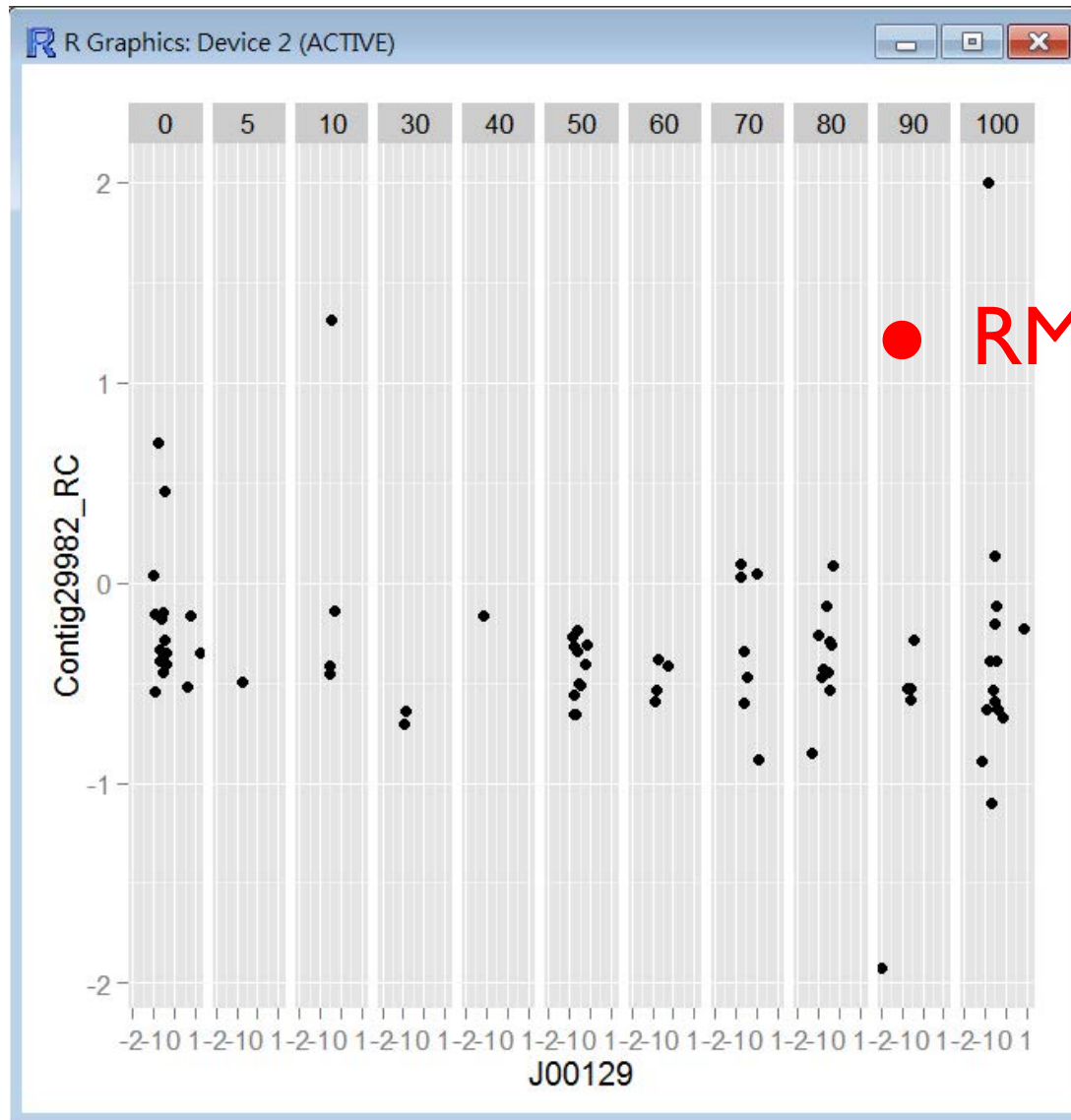
- RMD_example 7.16

Map the 3rd dim to colors

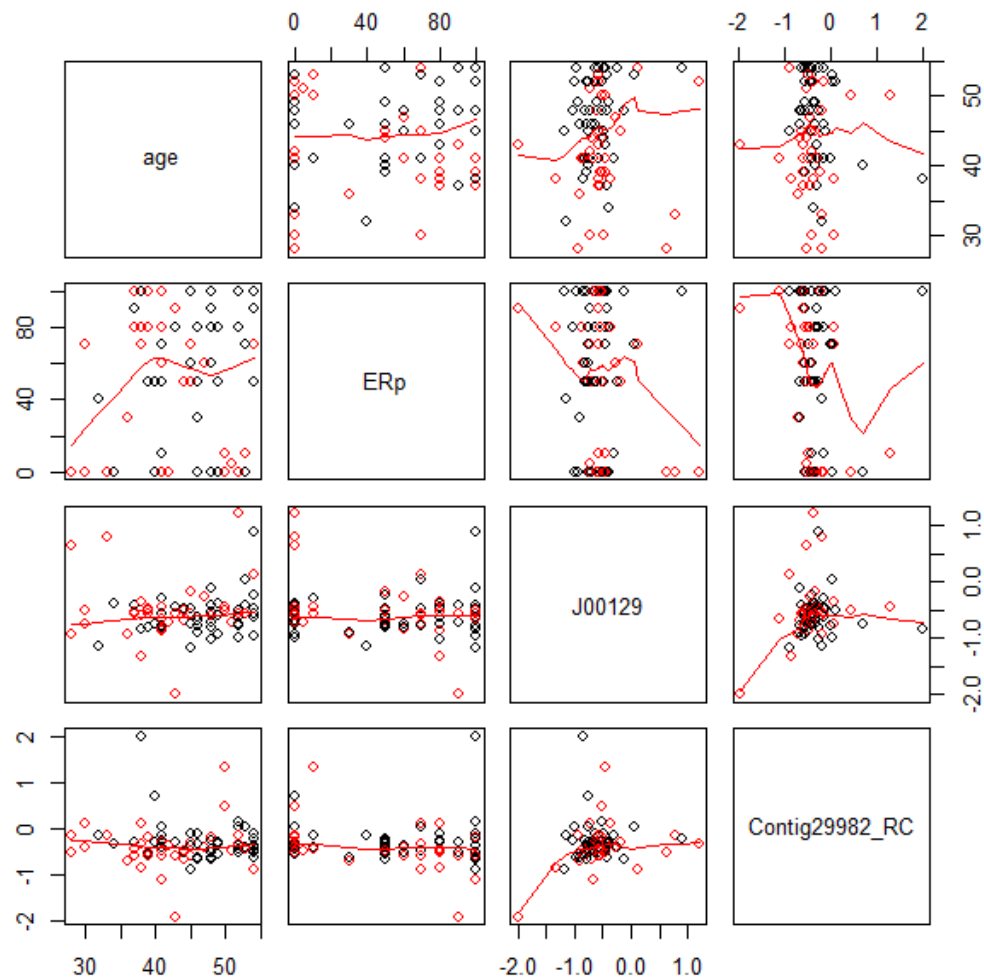


- RMD_example 7.16

Lay out panels in the 3rd dim



● RMD_example 7.16



Scatterplot matrices

Color in metastases
Add smooth lines

- RMD_example 7.17

Heatmap

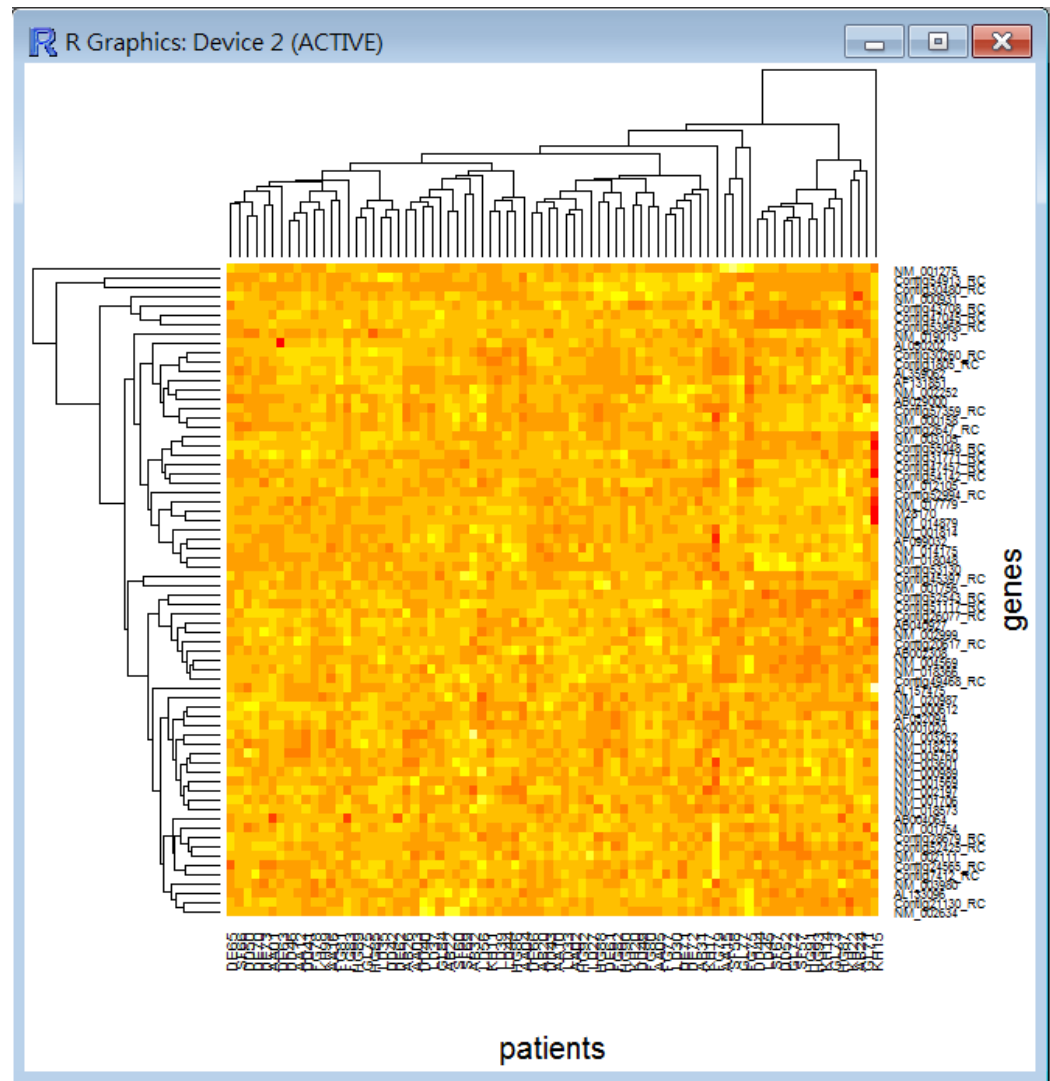
A graphical representation of data where the individual values contained in a **matrix** are represented as **colors**. [wikipedia]

欄1	FG80	SF58	DE72	DE65	HG87	HG88	AB22	HG91
J00129	-0.795	-0.509	-0.961	-0.749	-0.426	-0.566	-0.42	-0.499
Contig299	-0.387	0.459	-0.631	0.699	-0.406	-0.596	-0.286	-0.402
Contig428	0.199	-0.257	0.037	-0.346	-0.355	-0.352	-0.09	0.181
Contig420	-0.247	-0.065	-0.153	0.032	0.429	-0.336	-0.048	0.143
Contig279	0.176	0.129	0.144	0.3	-0.036	0.037	0.291	-0.268
Contig201	-0.129	0.009	-0.202	-0.025	0.191	-0.147	-0.166	0.849
Contig506	-0.111	0.021	0.192	-0.067	0.091	-0.081	0.264	0
Contig426	0.119	0	-0.19	-0.226	0.2	0.037	0.026	0.268
Contig566	0.231	-0.649	-0.086	-0.018	-1.23	0.383	0.253	-1.198
Contig486	0.118	0.058	-0.052	-0.278	-0.058	0.049	0.127	-0.188
Contig493	0.035	-0.038	0.055	0.13	-0.303	0.383	-0.352	-0.31
Contig197	-0.482	-0.105	0.013	-0.338	-0.465	-0.161	0.52	-0.387
Contig263	0.015	0.053	-0.123	0.038	-0.175	-0.042	-0.012	0.226
Contig530	-1.389	-0.601	-1.378	-0.007	0.63	-1.082	-1.264	0.346
Contig439	-0.011	0.005	0.113	-0.277	-0.258	-0.024	-0.333	0.331
Contig195	-0.092	0.295	-0.806	-1.106	-0.201	0.071	0.272	-0.57
Contig104	-0.058	0.006	0.132	-0.216	-0.169	-0.188	0.176	0.374
Contig472	-0.548	-0.25	0.456	-0.967	-0.544	-0.447	-0.628	-0.367
Contig207	-0.106	-0.384	0.296	-1.087	-0.054	0.093	-0.111	0.628
AL157502	0.363	-0.185	-0.179	0.33	-0.355	-0.12	-0.115	-0.61
Contig366	-0.139	-0.775	0.244	-1.806	-1.207	-0.252	-0.635	-0.958
D31887	-0.061	-0.04	0.067	-0.008	0.13	-0.069	-0.049	0.315
AB033006	0.2	0.608	-0.298	-0.118	0.09	0.27	-0.156	0.191
AB033007	0.041	-0.13	0.178	0.139	0.154	0.363	0.227	-0.037
M83822	0.037	0.046	0.023	-0.154	-0.398	-0.137	0.152	-0.351
AB033025	-0.48	-0.127	0.156	-0.567	0.294	-0.263	-0.007	-0.256
AF114264	0.091	-0.108	0.221	0.337	0.088	-0.238	0.109	-0.212
Contig406	-0.159	-0.008	0.415	-0.373	-0.47	0.525	-0.129	-0.177
Contig173	0.084	-0.142	-0.012	-0.254	0.04	-0.149	-0.124	0.194
AB033034	-0.08	0.076	0.059	0.117	0.329	-0.112	-0.006	0.089
AB033035	0.022	-0.019	-0.044	0.029	0.583	-0.076	-0.072	0.187

Heatmap

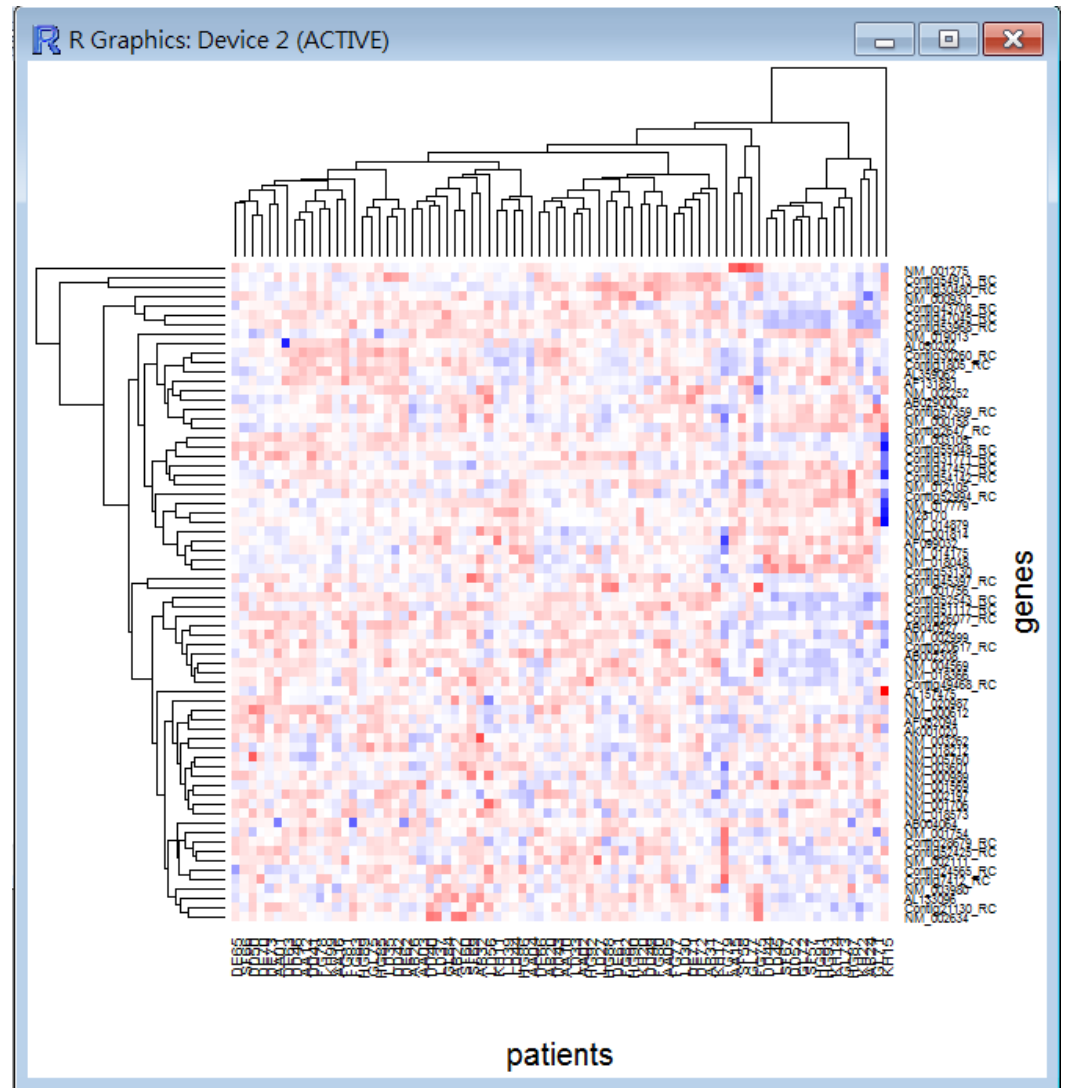
A graphical representation of data where the individual values contained in a **matrix** are represented as **colors**. [wikipedia]

- RMD_example 7.18



Heatmap
















With different
color schemes



- RMD_example 7.18

Colors

Color model- HSL/HSB

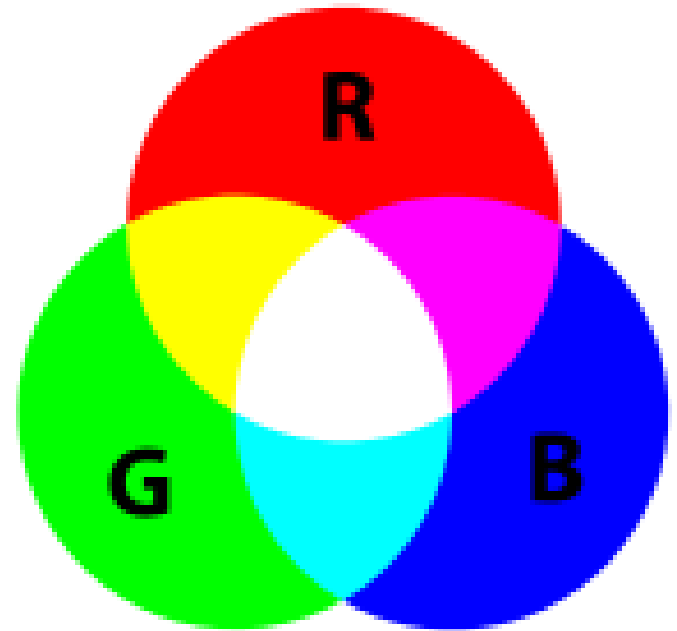
Luminance (明度)		Brightness	Saturation(飽和度)	Hue(色相)
				
				
				
				
				

Perceived as ordered

Not as much

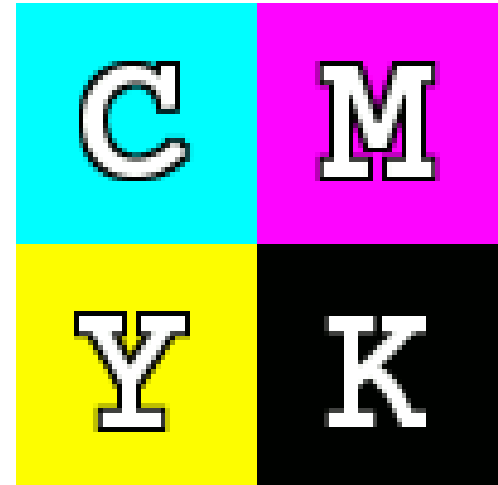
Color model- RGB

- R=red, G=green, B=blue
- An additive color model in which red, green and blue light are added together in various ways to reproduce a broad array of colors.



Color model- CMYK

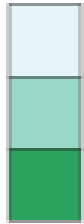
- C : Cyan = 青色，又稱為『天藍色』或是『湛藍』
- M : Magenta = 品紅色，又稱為『洋紅色』
- Y : Yellow = 黃色
- K : Key (black) = 定位套版色（黑色）
- A subtractive color model, used in color printing, and is also used to describe the printing process itself.



RGB ▼

CMYK ▼

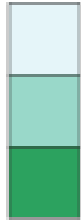
HEX ▼



229,245,249

153,216,201

44,162,95



10,0,0,0

40,0,15,0

83,0,70,0



#e5f5f9

#99d8c9

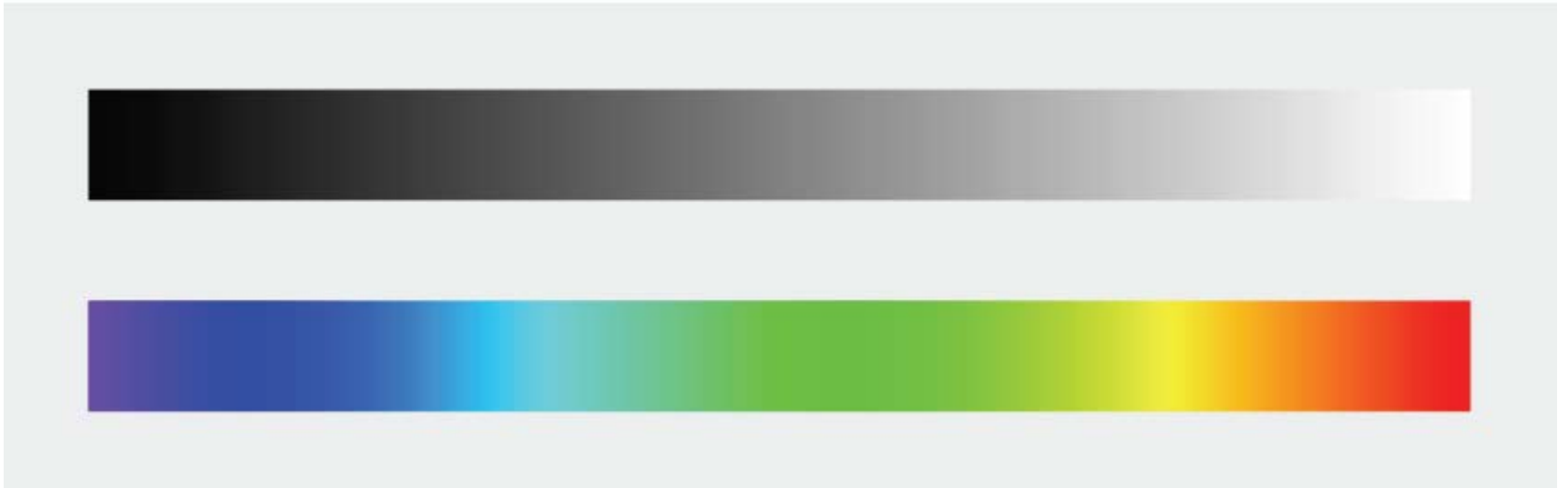
#2ca25f

Color codes

- An abstract mathematical model describing the way colors can be represented as tuples of numbers, typically as three or four values or color components
 - RGB: the 3 component values (for R, G and B) are often stored as integer numbers in the range 0 to 255 (decimal)
 - CMYK: representation of color of the form (C%, M%, Y%, K%), where C, M, Y, and K are the percent values for the cyan, magenta, yellow, and black values of the color.
 - HEX: representing each RGB component number as 2 hexadecimal
 - HSL/HSB: <http://colorizer.org/>

Rainbow Colormap

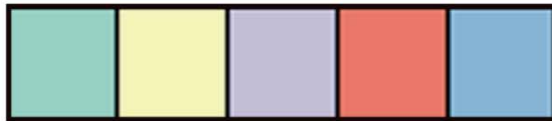
Rainbow colormap is perceptually nonlinear



Brewer scales

Nominal

Qualitative Scale



Ordinal

Sequential Scale




0 → Max

Diverging Scale




Max ← 0 → Max

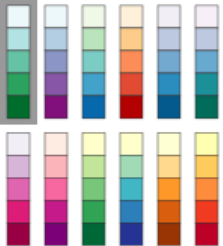
[Cynthia Brewer, Color Use Guidelines for Mapping and Visualization](#)

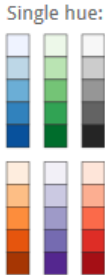
Number of data classes: 3  how to use | updates | downloads | credits


COLORBREWER 2.0
color advice for cartography


Nature of your data: 
☒ sequential ☐ diverging ☐ qualitative



Pick a color scheme:

Multi-hue: 


Single hue: 




Only show: 
☐ colorblind safe
☐ print friendly
☐ photocopy safe





Context: 
☐ roads
☐ cities
☒ borders

Background:
☒ solid color 
☐ terrain

color transparency


3-class BuGn

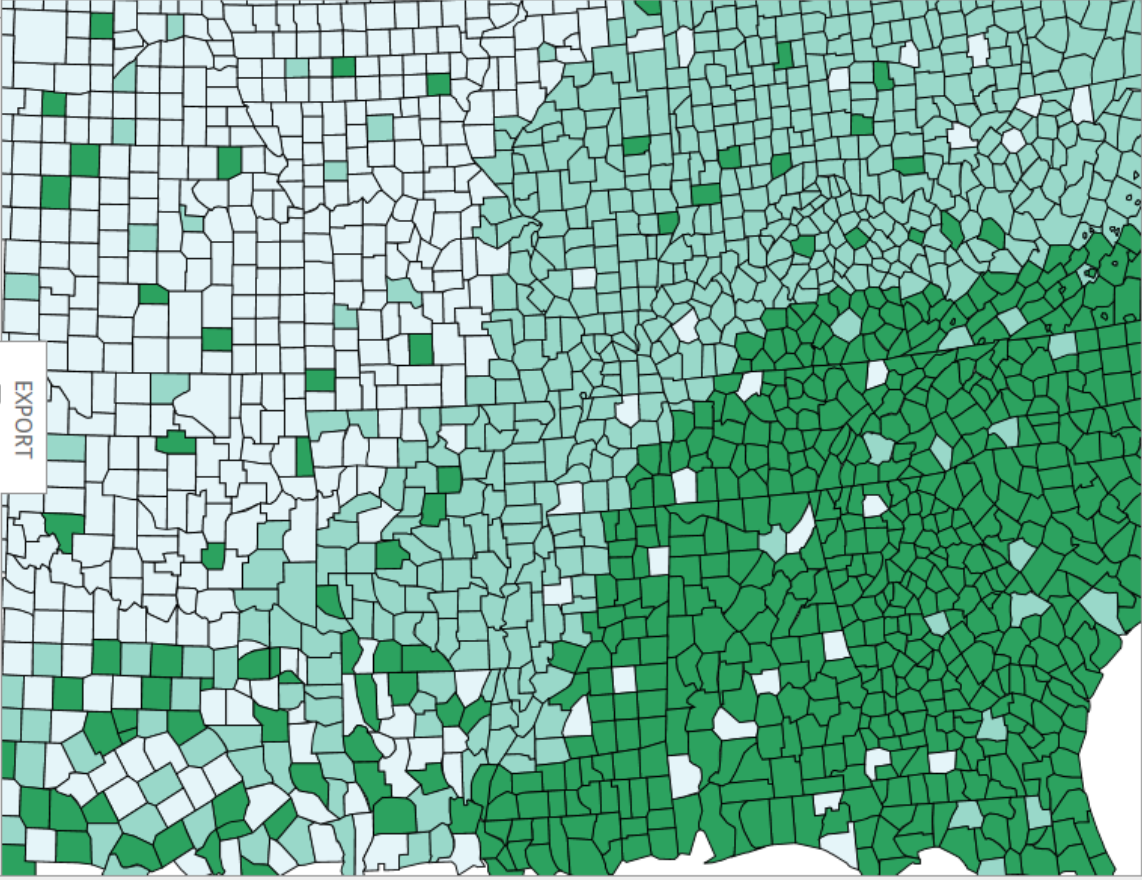
HEX 

 #e5f5f9
 #99d8c9
 #2ca25f

EXPORT    

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University
Support
[Back to Flash version](#)
[Back to ColorBrewer 1.0](#)





<http://colorbrewer2.org/>

Pick up color scheme and get color codes

Use colors in a graph with ggplot2

Cookbook for R

[index](#) » [Graphs](#) » [Colors_\(ggplot2\)](#)

Colors (ggplot2)

This page was recently updated to reflect changes in the new version of ggplot2, 0.9.3. See [Installing and using packages](#) to make sure you have the latest version of ggplot2.

Problem

You want to use colors in a graph with ggplot2.

Solution

The default colors in ggplot2 can be difficult to distinguish from one another because they have equal luminance. They are also not friendly for colorblind viewers.

A good general-purpose solution is to just use the [colorblind-friendly palette](#) below.

Table of contents

- [Colors \(ggplot2\)](#)
 - [Problem](#)
 - [Solution](#)
 - [Sample data](#)
 - [Simple color assignment](#)
 - [Mapping variable values to colors](#)
 - [A colorblind-friendly palette](#)
 - [Color selection](#)
 - [Setting luminance and saturation \(chromaticity\)](#)
 - [Palettes: Color Brewer](#)
 - [Palettes: manually-defined](#)
 - [Continuous colors](#)
- [Color charts](#)
 - [Hexadecimal color code chart](#)
 - [RColorBrewer palette chart](#)

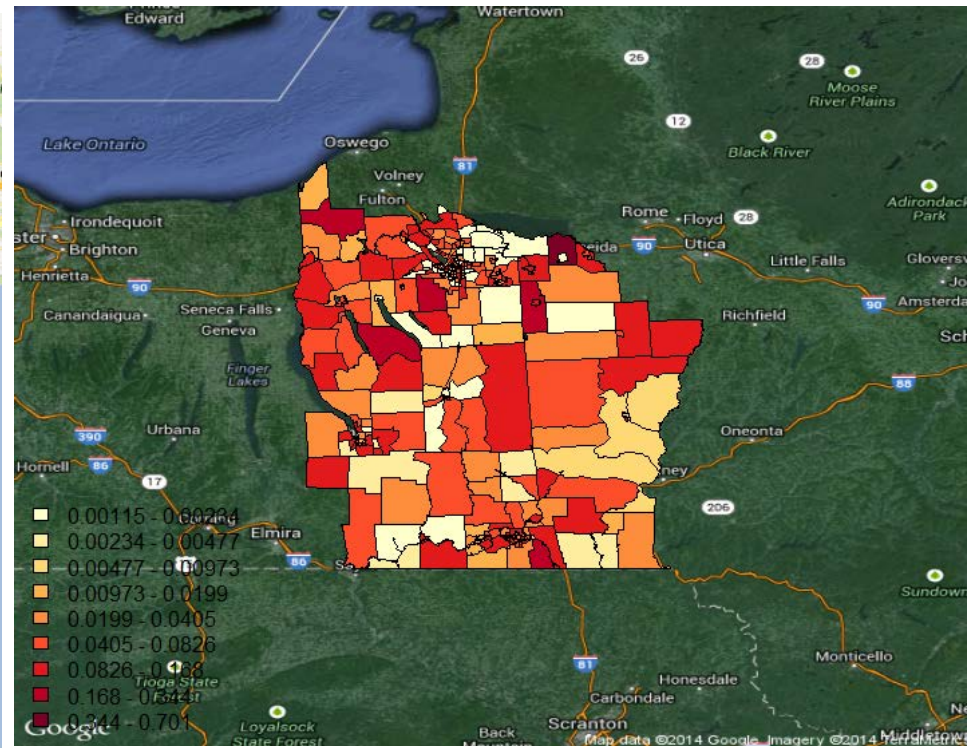
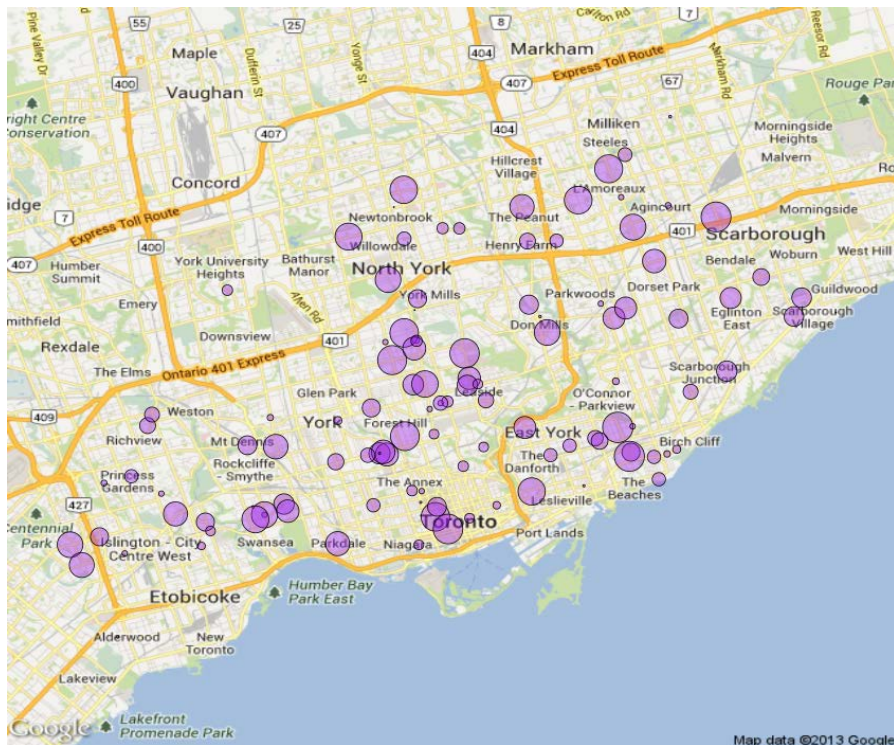
[http://www.cookbook-r.com/Graphs/Colors_\(ggplot2\)/](http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/)

Other useful graphic tools:
google ↔ R

RgoogleMaps

R package for plotting on Google maps in R

<http://www.nr.no/~thordis/files/5-RgoogleMaps-WithSolution.pptx>

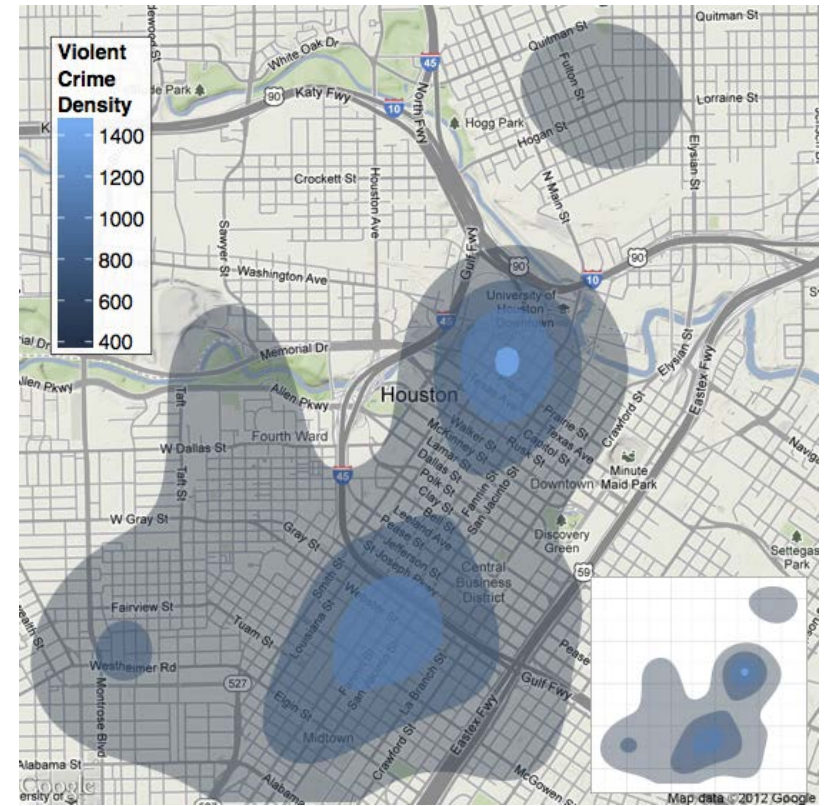
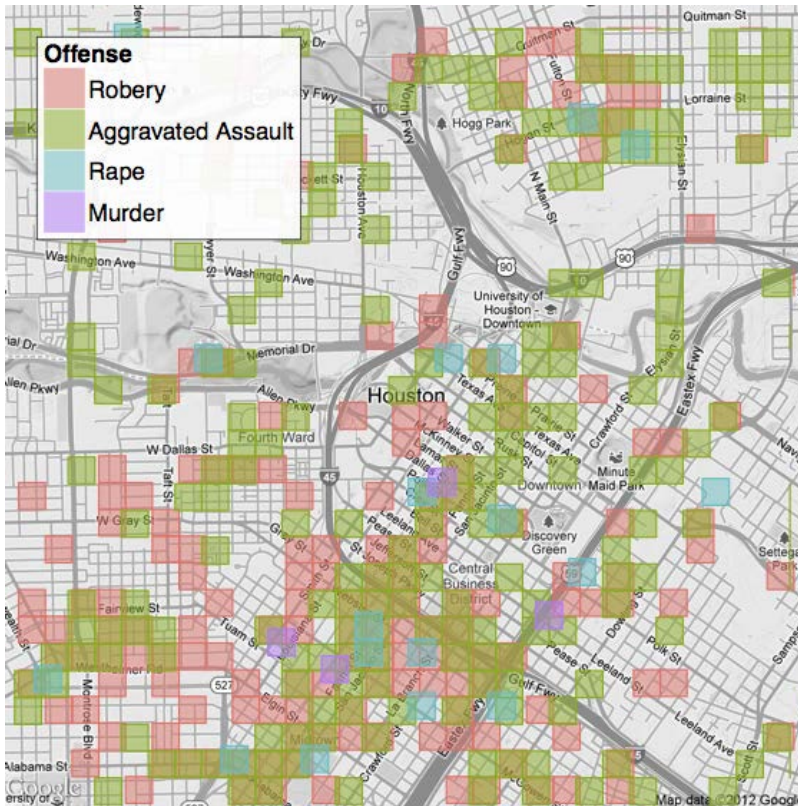


<http://rforwork.info/tag/rgooglemaps/>

<http://somethinglikescience.wordpress.com/2014/03/13/r365-day-31-rgooglemaps/>

ggmap

R package for plotting on Google maps using ggplot2



<http://journal.r-project.org/archive/2003-1/kahle-wickham.pdf>

googleVis

<http://github.com/mages/googleVis>

- R package provides an interface between R and the Google Charts

Google Charts

<https://developers.google.com/chart/>

- Provide a way to visualize data on web sites
- The API makes it easy to create interactive charts
- Use JavaScript and DataTable / JSON as input
- Output is either HTML5/SVG or Flash



googleVis

- The functions of the package allow users to visualize data with the Google Charts without uploading their data to Google
- The output functions is html code that contains the data and references to JavaScript functions hosted by Google