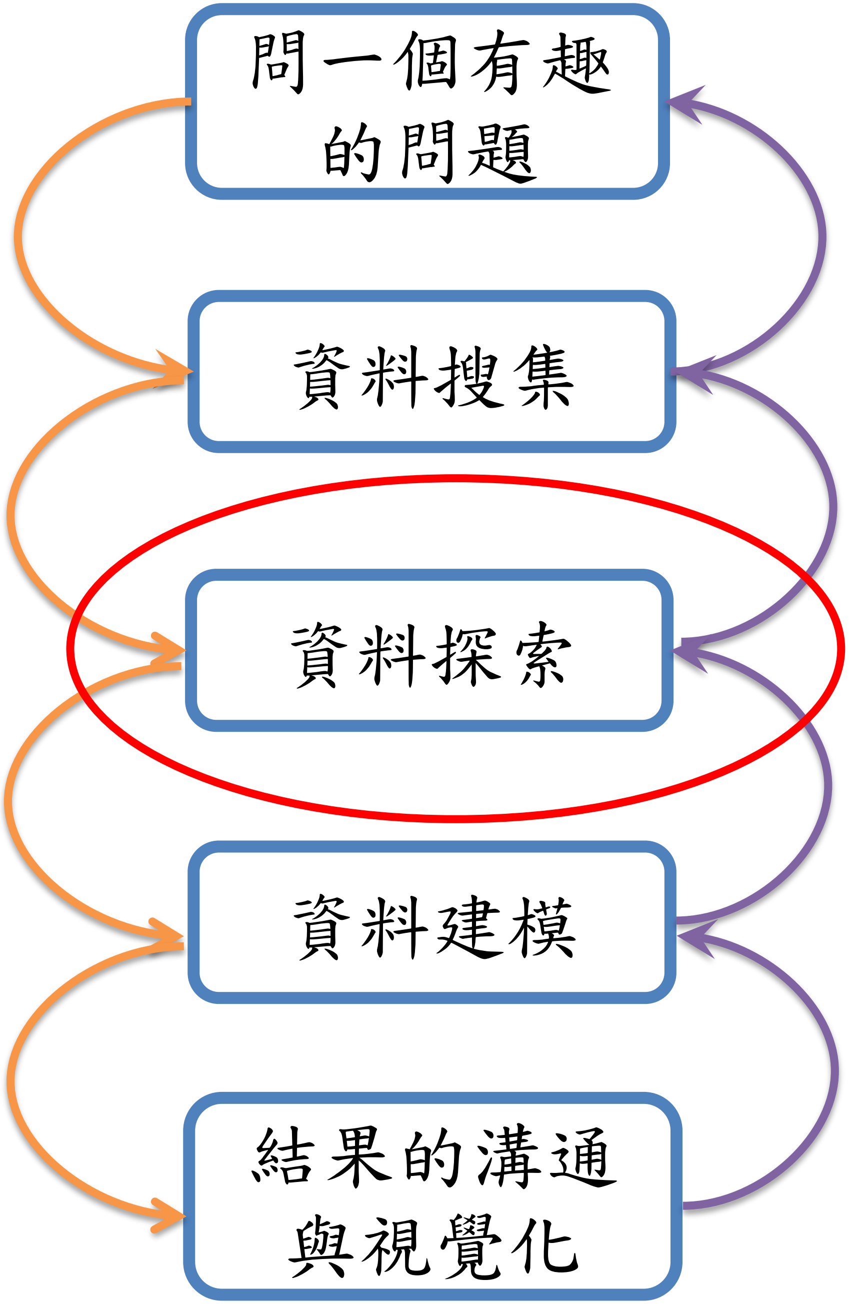
Lecture 2: Exploratory data analysis

BTBI30081

統計應用方法Applied Methods in Statistics

2025/2/26

分析流程



你的目的為何?
你想預測或估計什麼?

相關的資料在那裡?
資料是如何取樣(sample)的?
有沒有隱私權的問題?

資料視覺化
有沒有反常資料?
資料的pattern?

建立模型
模型參數估計
模型驗證

獲得什麼新知識?
結果合理嗎?
Can we tell a story?

Exploratory data analysis (EDA)

- 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]
- Seeing what the data can tell us beyond the formal modeling or hypothesis testing task. [Wikipedia]

Data visualization

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

-John Tukey (1915 - 2000)

- Data visualization is a powerful approach to detecting mistakes, biases, systematic errors and unexpected variability that are commonly found in data regardless of applications.
- Data visualization can provide a powerful way to communicate a data-driven finding.

Anscombe's quartet

Same mean, variance, correlation, and linear regression line

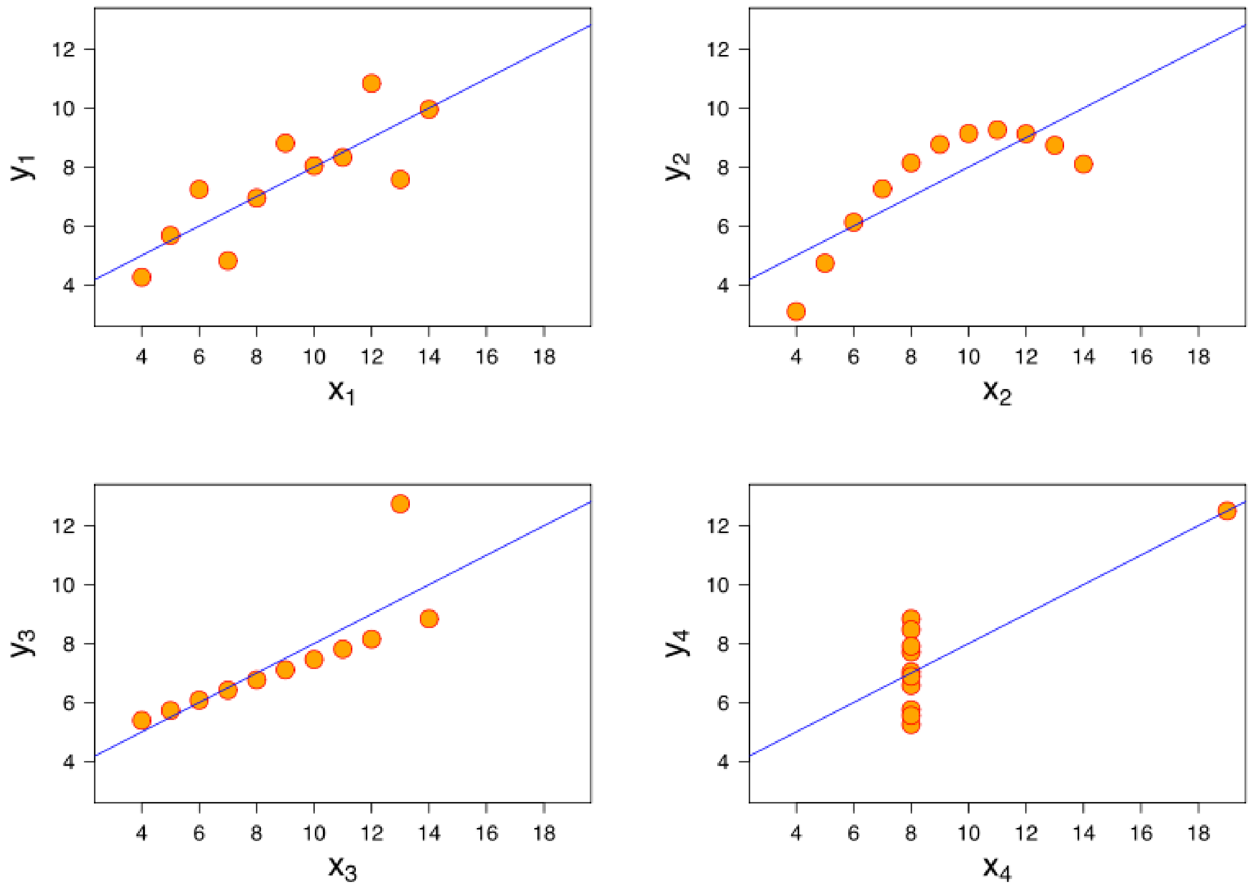
Anscombe's Quartet: Raw Data

	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 ≠ 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.

On the theory of scales and measurements [S. Stevens, 46]

Measurement scales

- Q - Interval (location of zero arbitrary)

  Cannot compare directly. Only differences (i.e., intervals) can be compared

  Dates: Jan 19; Location: (Lat, Long)

- Q - Ratio (zero fixed)

  Origin is meaningful, can measure ratios & proportions

  Measurements: Length, Mass, Temp, ...

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 2.1

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	[illegible]	[illegible]	[illegible]	-0.153
DE65	40	0	6.25	0	[illegible]	[illegible]	[illegible]	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	[illegible]	[illegible]	-0.025
SF67	48	1	1.98	0	-0.601	[illegible]	[illegible]	0.108
LD44	33	1	1.4	0	0.786	[illegible]	[illegible]	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	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	[illegible]	[illegible]	-0.153
DE65	40	0	6.25	0	-0.749	[illegible]	[illegible]	0.032
HG87	53	0	5.18	0	-0.426	[illegible]	[illegible]	0.429
HG88	37	1	1.09	100	-0.566	[illegible]	[illegible]	-0.336
AB22	37	0	5.8	90	-0.42	[illegible]	[illegible]	-0.048
HG91	30	1	1.03	0	-0.499	[illegible]	[illegible]	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

Attribute
Feature
Variable

Example: Gene expression microarray data (samplexprs.csv) – 78 items (patients), 4746 variables

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	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	[illegible]	[illegible]	[illegible]	[illegible]
AA01	43	0	12.53	80	[illegible]	[illegible]	[illegible]	[illegible]
GL73	52	1	2.13	0	[illegible]	[illegible]	[illegible]	[illegible]
AA10	49	0	11.16	80	[illegible]	[illegible]	[illegible]	[illegible]
HG86	54	0	5.89	50	[illegible]	[illegible]	[illegible]	[illegible]
DE62	40	0	6.97	50	[illegible]	[illegible]	[illegible]	[illegible]
AB26	41	0	8.17	10	[illegible]	[illegible]	[illegible]	[illegible]
SF57	41	1	2	0	[illegible]	[illegible]	[illegible]	[illegible]
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

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

Example: Gene expression microarray data (samplexprs.csv)

Resources for R Graphics

- CRAN Task View: Dynamic Visualizations and Interactive Graphics

  https://CRAN.R-project.org/view=DynamicVisualizations
- R Graphics 2nd Edition

  https://www.stat.auckland.ac.nz/~paul/RG2e/
- ggplot2

  https://ggplot2.tidyverse.org/
- Cookbook for R graphs

  http://www.cookbook-r.com/Graphs/

Univariate data (1 dimension)

Distributions

- The most basic statistical summary of a list of numbers is it's distribution.
- The simplest way to think of a distribution is as a compact description of many numbers.
    - For example, we have measured gene expression intensities of all patients in the study.

Displaying distributions

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

Stem-and-leaf plot

For data $x_1, x_2, \cdots, x_n$,

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 | 00111111112333444
4 | 55555666678888888999
5 | 0012222222333444444444
```

age

```
The decimal point is at the |

-2 | 0
-1 |
-1 | 321000
-0 | 999998888888888877777776666666666666665555555555555
-0 | 44444444433221
 0 | 01
 0 | 689
 1 | 2
```

J00129

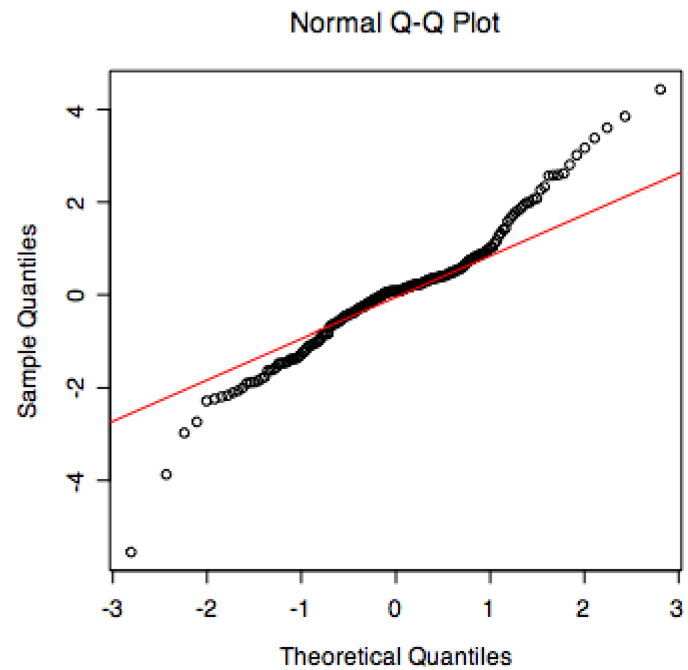
- RMD_example 2.2

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

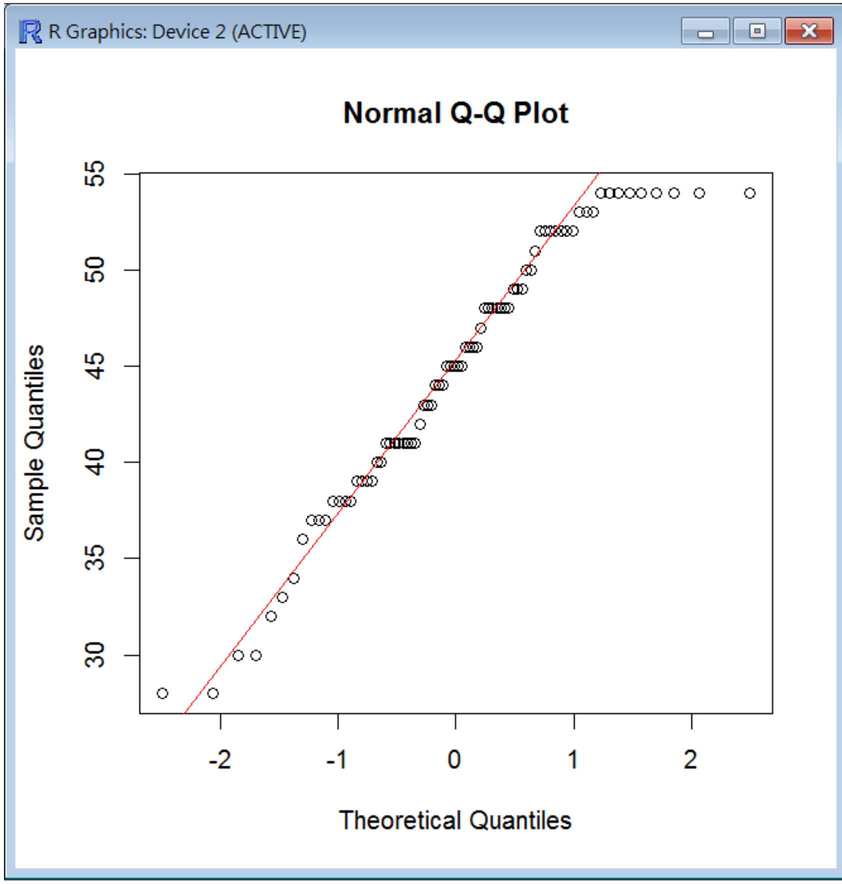
- “q” stands for quantile
- A graphical method for comparing two probability distributions by plotting their quantiles against each other



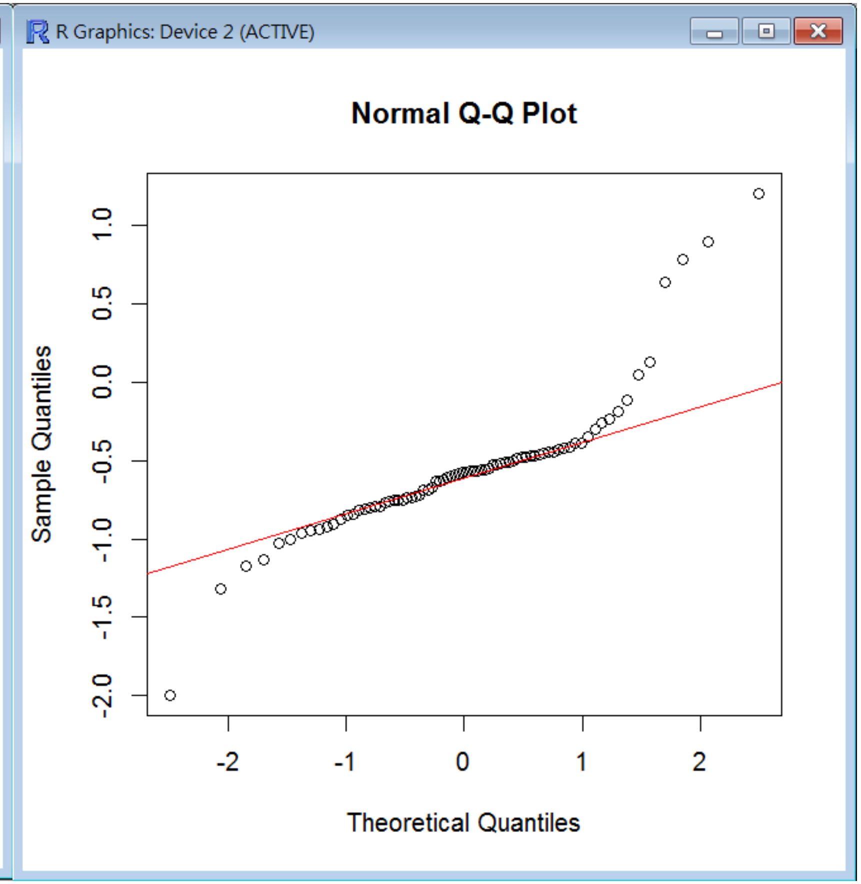
- If the two distributions being compared are similar, the points in the q–q plot will approximately lie on the line y = x

q-q plot

age



J00129



- RMD_example 2.3

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.

Data visualization with ggplot2

- ggplot2 is a powerful data exploration and visualization package that can create graphics in R.
- ggplot2 = grammar of graphics
- ggplot2 cheat sheet: https://rstudio.github.io/cheatsheets/html/data-visualization.html

- The idea of the "Grammar of Graphics" is to break the graph into components and handle the components of a graph separately.
- The ggplot2 package contains a set of functions that allow us to build the features of the graph in a series of layers for versatility and control.

- The main plotting functions in ggplot2:
  - ggplot() = create a "grammar of graphics" (gg) plot object
- Compared to functions in base R, the ggplot2 package can create elegant and complex plots. It can highly improve the quality and aesthetic of your graphs.

Learning **ggplot2**

- Example:

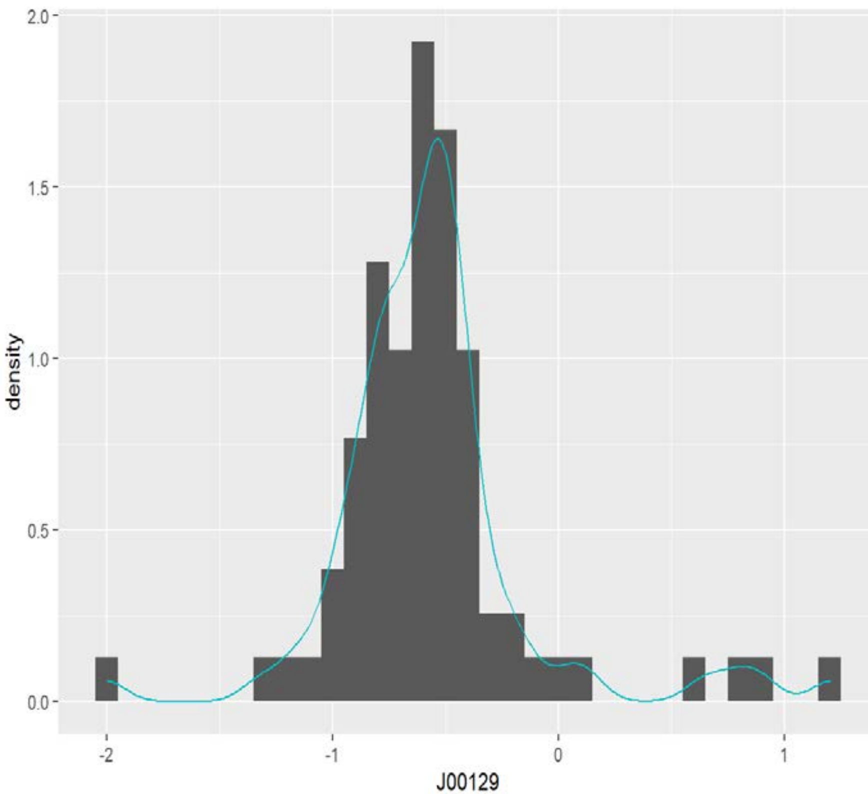
```
samplexprs %>%
ggplot(aes(x=J00129)) +
geom_histogram(aes(y=after_stat(density)), binwidth=0.05) +
geom_density(col="#00BFC4") +
xlab("J00129 gene expression intensity ratios") +
ggtitle("Histogram of J00129 gene expression")
```

- The first step in learning ggplot2 is to be able to break a graph apart into components.
- The main three components to note are:
  1. **Data**: The samplexprs in the example. We refer to this as the data component.
  2. **Geometry**: The plot in the example is a histogram. This is referred to as the **geometry** component. Other possible geometries are scatter plots, smooth densities, qq-plots, and boxplots.
  3. **Aesthetic mapping**: The J00129 values and plot attributes (x and y variables) are used to display the histogram. These are the **aesthetic mappings** component. How we define the mapping depends on what **geometry** we are using.

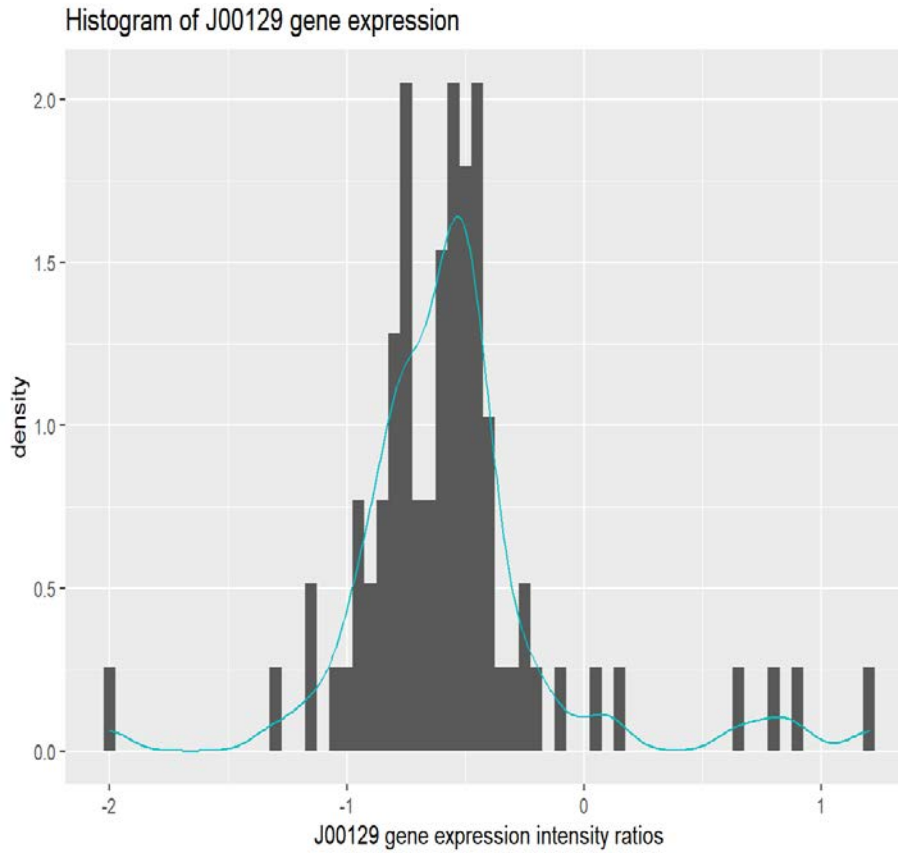
- We also note that:
  4. **Labels**: a title, a legend, and the style.

- Construct the ggplot2 plot piece by piece: RMD_example 2.5

Histogram



binwidth=0.1



binwidth=0.05

Display a smooth density estimate

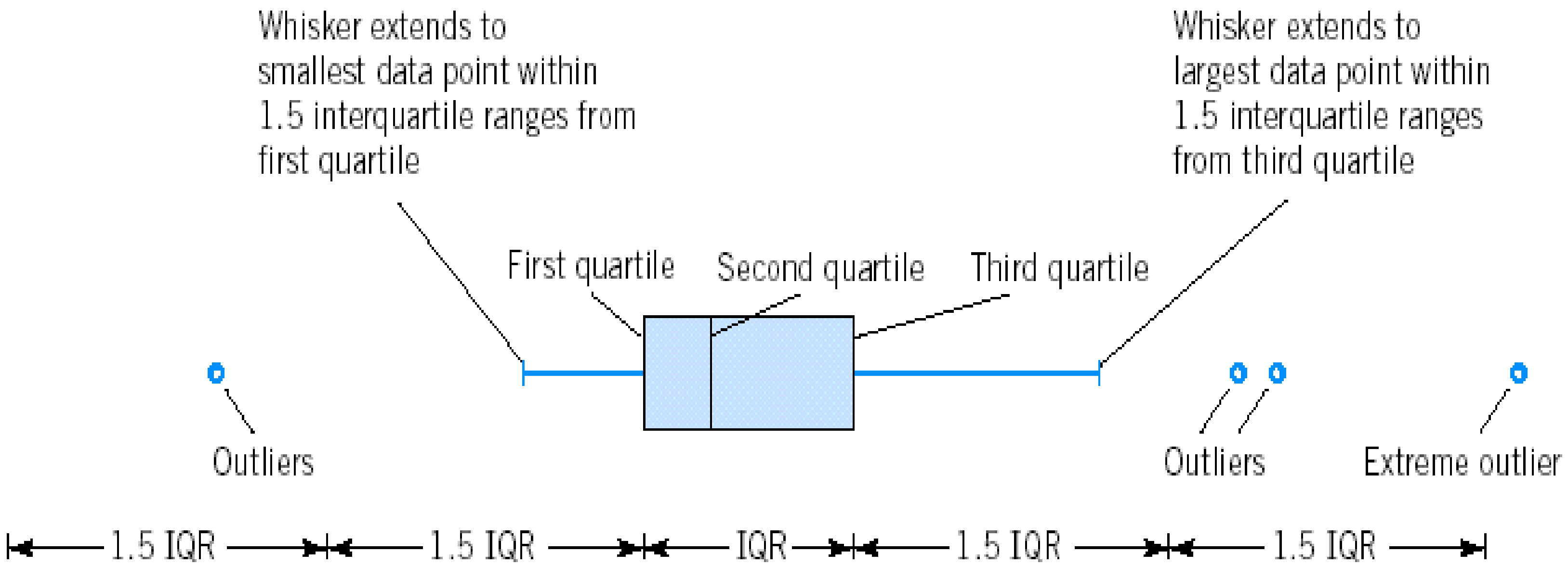
- RMD_example 2.4

Quartile

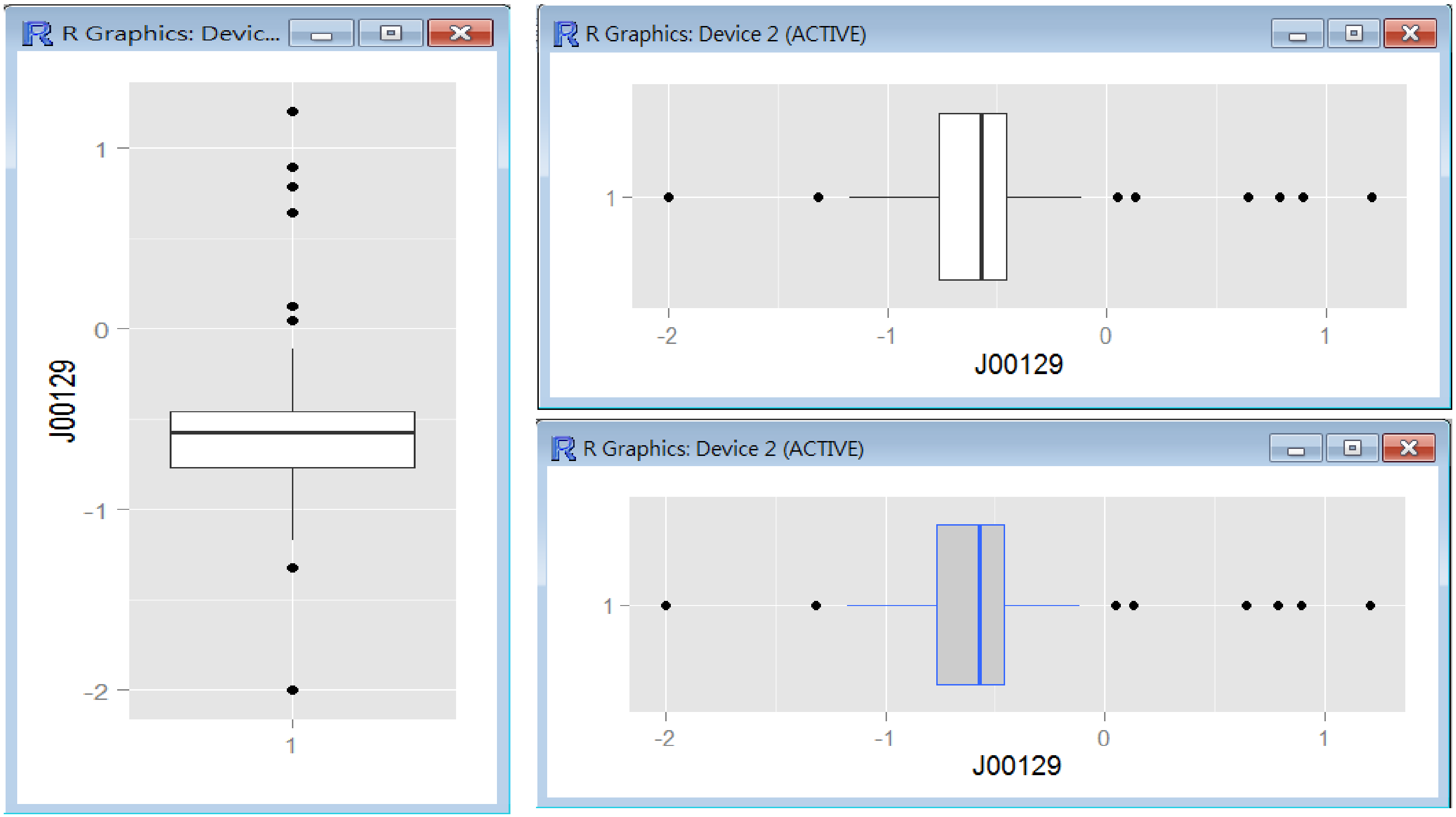
- First quartile (Q1) = 25-th percentile
- Second quartile (Q2) = 50-th percentile
- Third quartile (Q3) = 75-th percentile
- Interquartile range (IQR) = $|Q3 - Q1|$

Boxplot

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



Boxplot

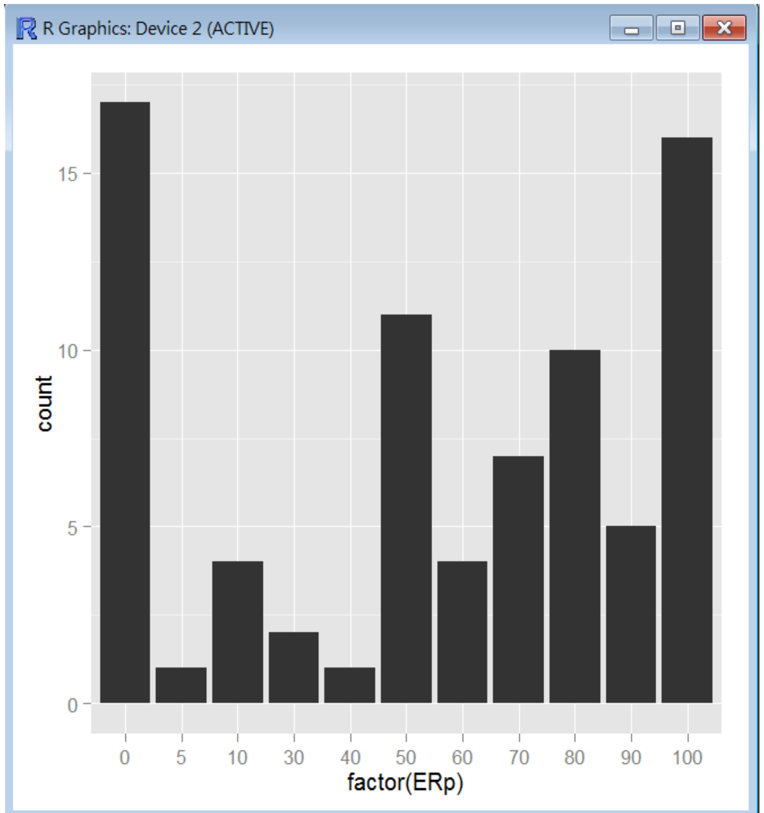


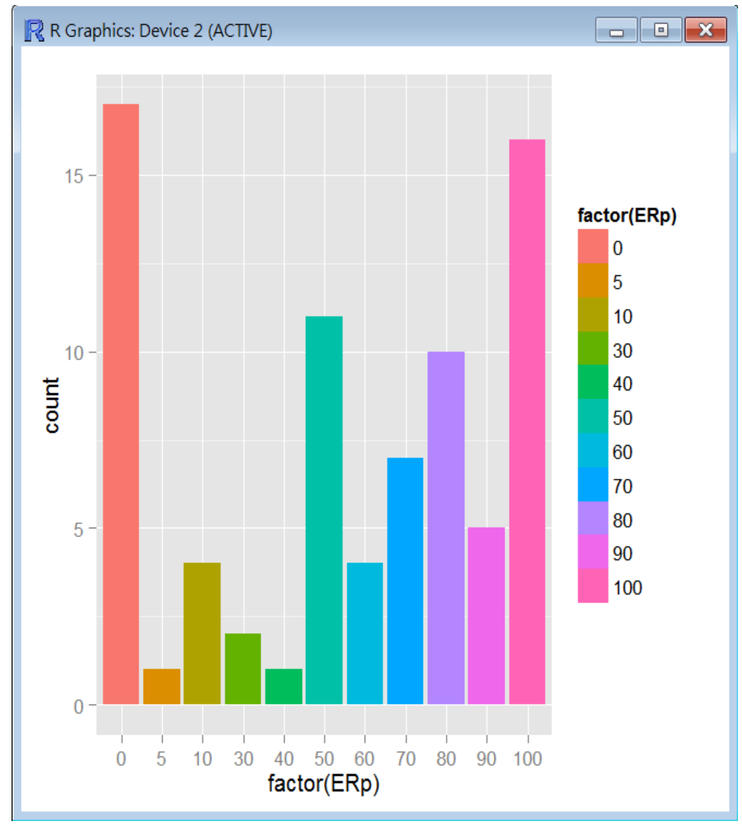
- RMD_example 2.5

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

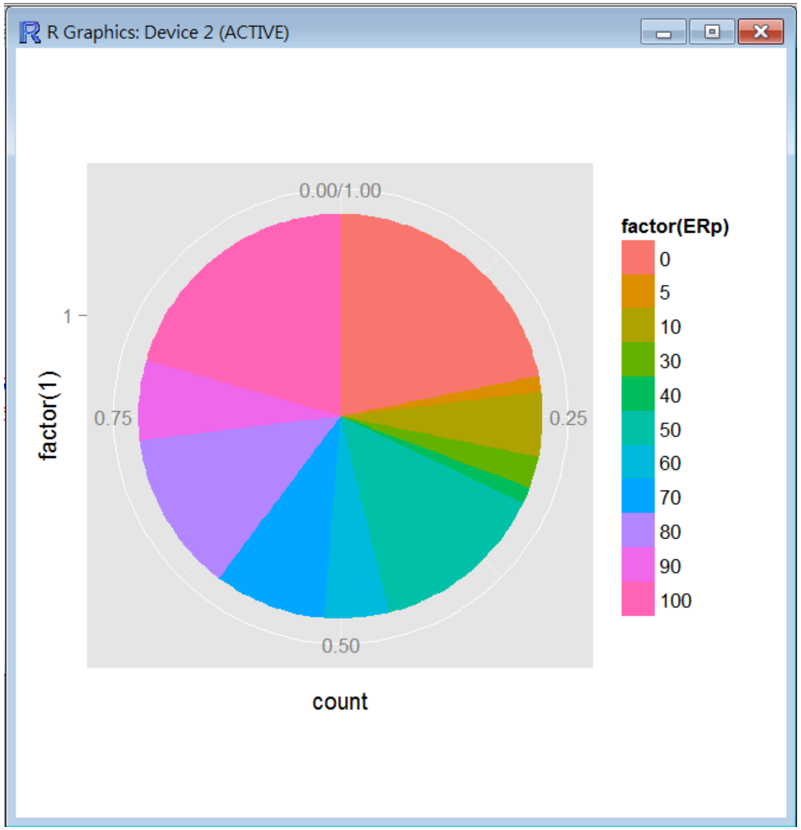




- RMD_example 2.6

Pie chart

For categorical data with proportions



- RMD_example 2.7

Bivariate data (2 dimensions)

Example

- In a jar, we have 4 green balls, 3 blue balls and 2 red balls.
- We draw 2 balls from the jar.
- When sampling with and without replacement, what are the probabilities obtaining two green balls?

Multiple random variables

- We always consider multiple random variables at once.
- The ball drawing example
  - $X_1 = 1, 2 \text{ or } 3$ if the first ball is green, blue or red, respectively. X_2 represents the color for the second ball.
  - We are interested in $\Pr(X_1 = 1, X_2 = 1)$.

Joint and marginal distributions

- Let X and Y be discrete random variables.
  - **Joint distribution**:
$$P_{X,Y}(x,y) = \Pr(X = x \text{ and } Y = y)$$
$$\Pr(a \le X \le b, c \le Y \le d) = \sum_{x=a}^{b} \sum_{y=c}^{d} P_{X,Y}(x,y)$$
  - **Marginal distributions**:
$$P_X(x) = \Pr(X = x) = \sum\nolimits_y P_{X,Y}(x,y)$$
$$P_Y(y) = \Pr(Y = y) = \sum\nolimits_x P_{X,Y}(x,y)$$

Ball drawing example

- When sampling with replacement, the joint and marginal distributions of X_1 and X_2

		X_2			
	P_{X_1,X_2}	1	2	3	P_{X_1}
X_1	1	0.19849	0.14713	0.09815	0.44377
	2	0.14865	0.10986	0.07527	0.33378
	3	0.09882	0.07388	0.04975	0.22245
	P_{X_2}	0.44596	0.33087	0.22317	

- When sampling without replacement, the joint and marginal distributions of X_1 and X_2

		X_2			
	P_{X_1,X_2}	1	2	3	P_{X_1}
X_1	1	0.16810	0.16719	0.10972	0.44501
	2	0.16609	0.08259	0.08424	0.33292
	3	0.11083	0.08274	0.02850	0.22207
	P_{X_2}	0.44502	0.33252	0.22246	

Independence

- Random variables X and Y are independent if:
$$P_{X,Y}(x, y) = P_X(x)P_Y(y)$$
- In other words/symbols:
$$\Pr(X = x \text{ and } Y = y) = \Pr(X = x)\Pr(Y = y)$$
- Define the conditional probability
$$P_{X|Y}(x|y) = \Pr(X = x|Y = y)$$

  If X and Y are independent, then
$$P_{X|Y}(x|y) = P_X(x)$$

Ball drawing example

- When we sample with replacement, two balls' results are independent.

		X_2		
	$P_{X_1} \times P_{X_2}$	1	2	3
X_1	1	0.19790	0.14683	0.09904
	2	0.14885	0.11044	0.07449
	3	0.09920	0.07360	0.04964

- In sampling without replacement, two balls' results are not independent.

		X_2		
	$P_{X_1} \times P_{X_2}$	1	2	3
X_1	1	0.19804	0.14797	0.09900
	2	0.14816	0.11070	0.07406
	3	0.098826	0.07384	0.04940

Continuous random variables

- Continuous random variables have joint density function, say $f_{X,Y}(x, y)$, and
  $$\Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)dydx$$
- The marginal density functions are obtained by integration:
  $$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$
  $$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$
- X and Y are independent if
  $$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Functions of independent random variables

- Let $X_1, \cdots, X_N$ be independent random variables with means $\mathrm{E}(X_i) = \mu_i$ and variances $\mathrm{var}(X_i) = \sigma_i^2, i = 1, \cdots, N$. Then the linear combination
$$Y = a_1 X_1 + \cdots + a_N X_N$$
has mean
$$E(Y) = a_1 \mu_1 + \cdots + a_N \mu_N$$
variance
$$\mathrm{var}(Y) = a_1^2 \sigma_1^2 + \cdots + a_N^2 \sigma_N^2$$

What if the random variables are not independent?

- The covariance between random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = \mathrm{E}\{\,[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\}$$

- The (Pearson) correlation between X and Y is

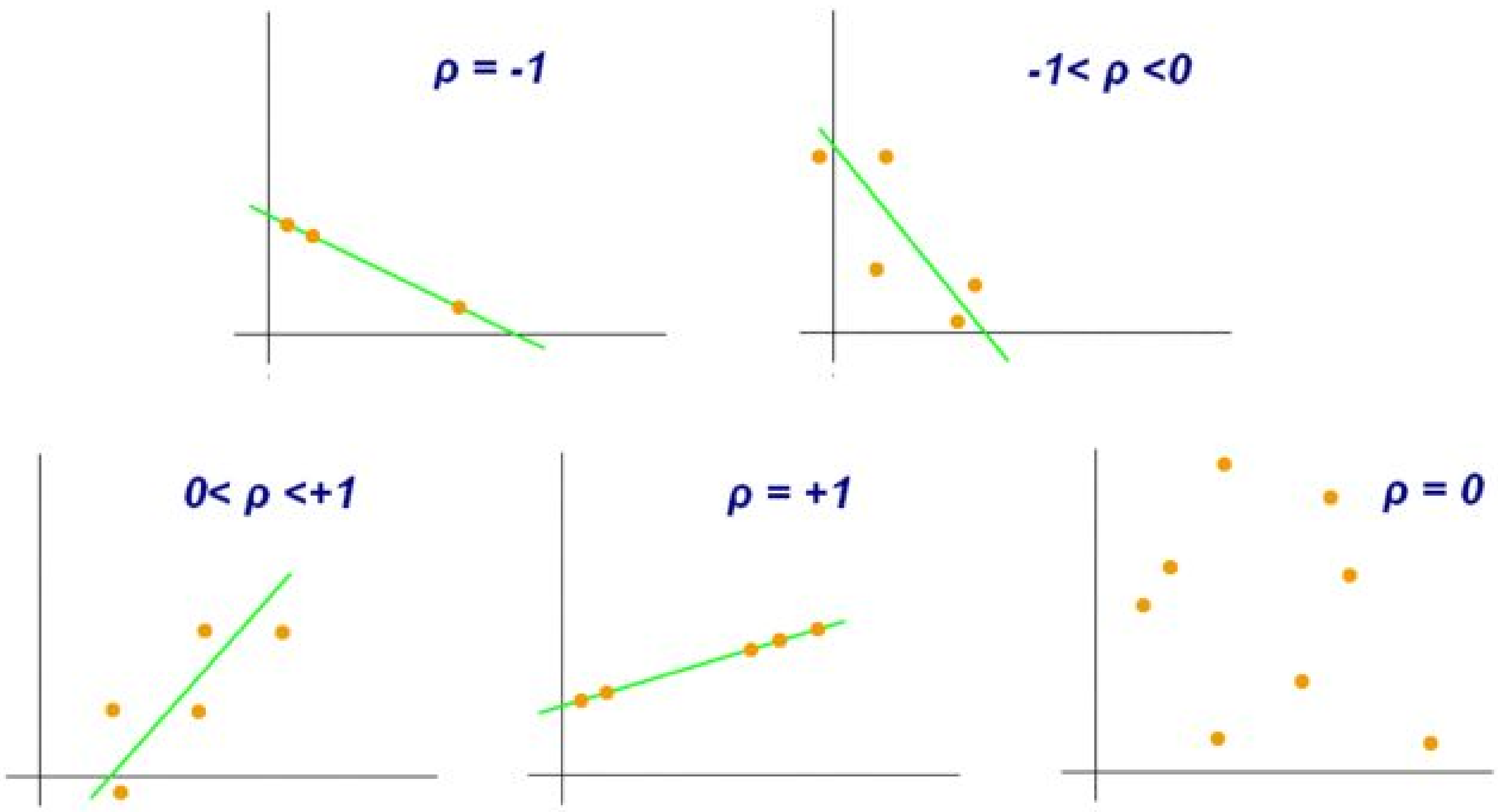
$$\rho = \mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

- If $\Pr(X = x_i, Y = y_i) = 1/N$ (each value takes equal probability), the correlation is

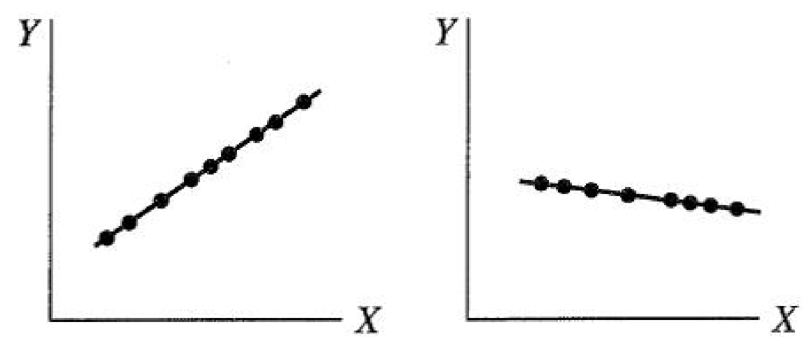
$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

which is the sample correlation.

- ρ (and r) measures the strength of the **linear** relationship between X and Y



- ρ (and r) assumes values between -1 and 1:
  - -1: perfect negative linear relationship
  - 0: no linear relationship
  - $+1$: perfect positive linear relationship
- The value of ρ (and r) is independent of the units used to measure the variables.
- ρ^2 (and r^2) is not a measure of the magnitude of the slope of the regression line for example,



(a) $r^2 = 1, \hat{\beta}_1 > 0$ (b) $r^2 = 1, \hat{\beta}_1 < 0$

- Notice that when two random variables X and Y are independent, then $\operatorname{corr}(X, Y) = 0$. However, for two random variables with $\operatorname{corr}(X, Y) = 0$, it’s not necessarily true that X and Y are independent.
- The Pearson correlation is meaningful when the random variables are **continuous**. For discrete random variables, other measurements (e.g., Chi-square statistic, Kendall's tau) are used.

Functions of non-independent random variables

- The linear combination

$$Y = a_1 X_1 + \cdots + a_N X_N$$

has mean

$$E(Y) = a_1 \mu_1 + \cdots + a_N \mu_N$$

variance

$$\mathrm{var}(Y) = a_1^2 \sigma_1^2 + \cdots + a_N^2 \sigma_N^2 + 2 \sum_{i<j} \sum a_i a_j \, \mathrm{cov}(X_i, X_j)$$

Random sample

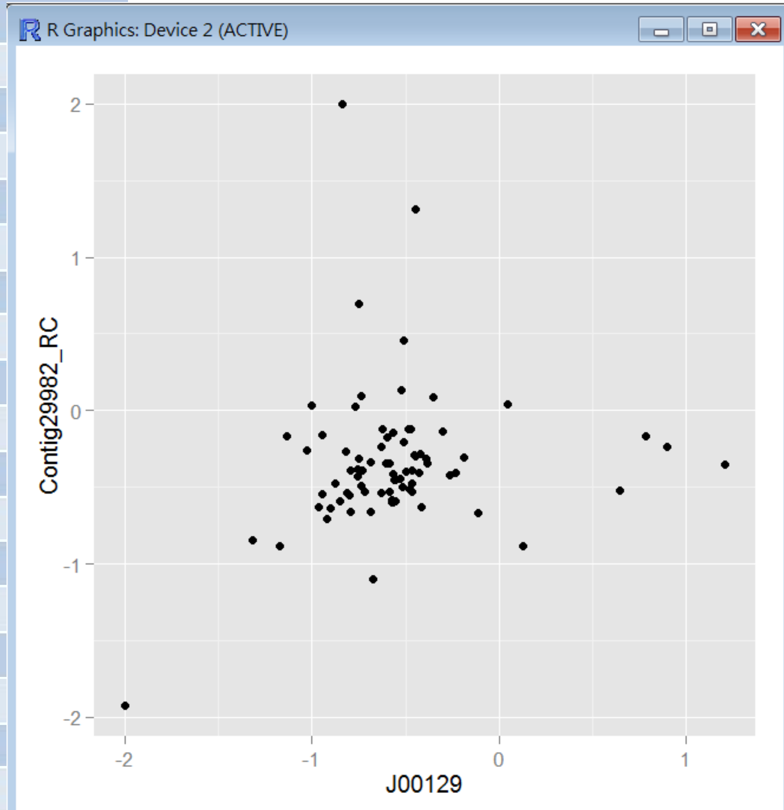
- Independent random variables $X_1, \cdots, X_N$ with the same distribution (i.e., $\mathrm{E}(X_i) = \mu,\ \mathrm{var}(X_i) = \sigma^2, \forall i$) are called **i.i.d.** (**i**ndependently and **i**dentically **d**istributed) or a **random sample**.

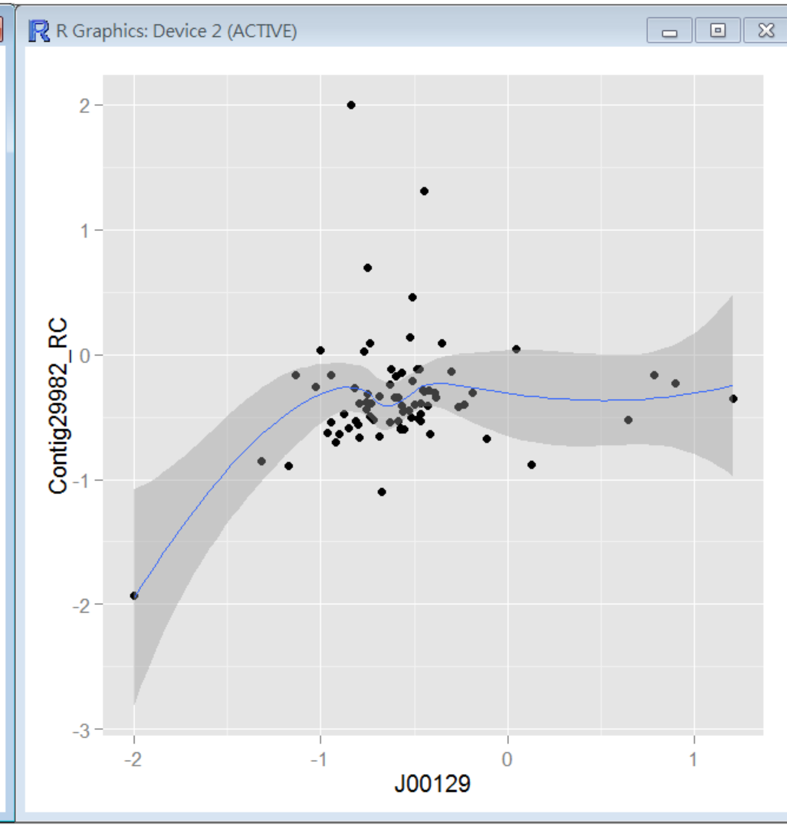
Displaying correlations

- 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 2.8

exprs_sig.csv

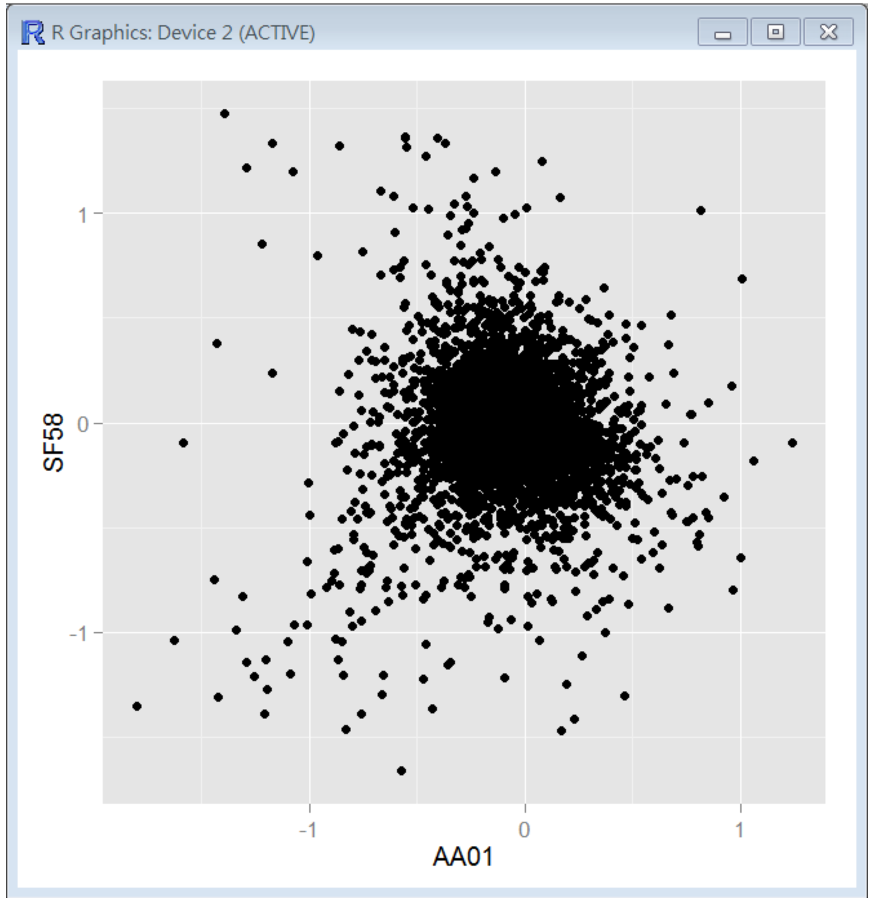
4741 items (genes), 78 variables (patients)

	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 2.9

	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

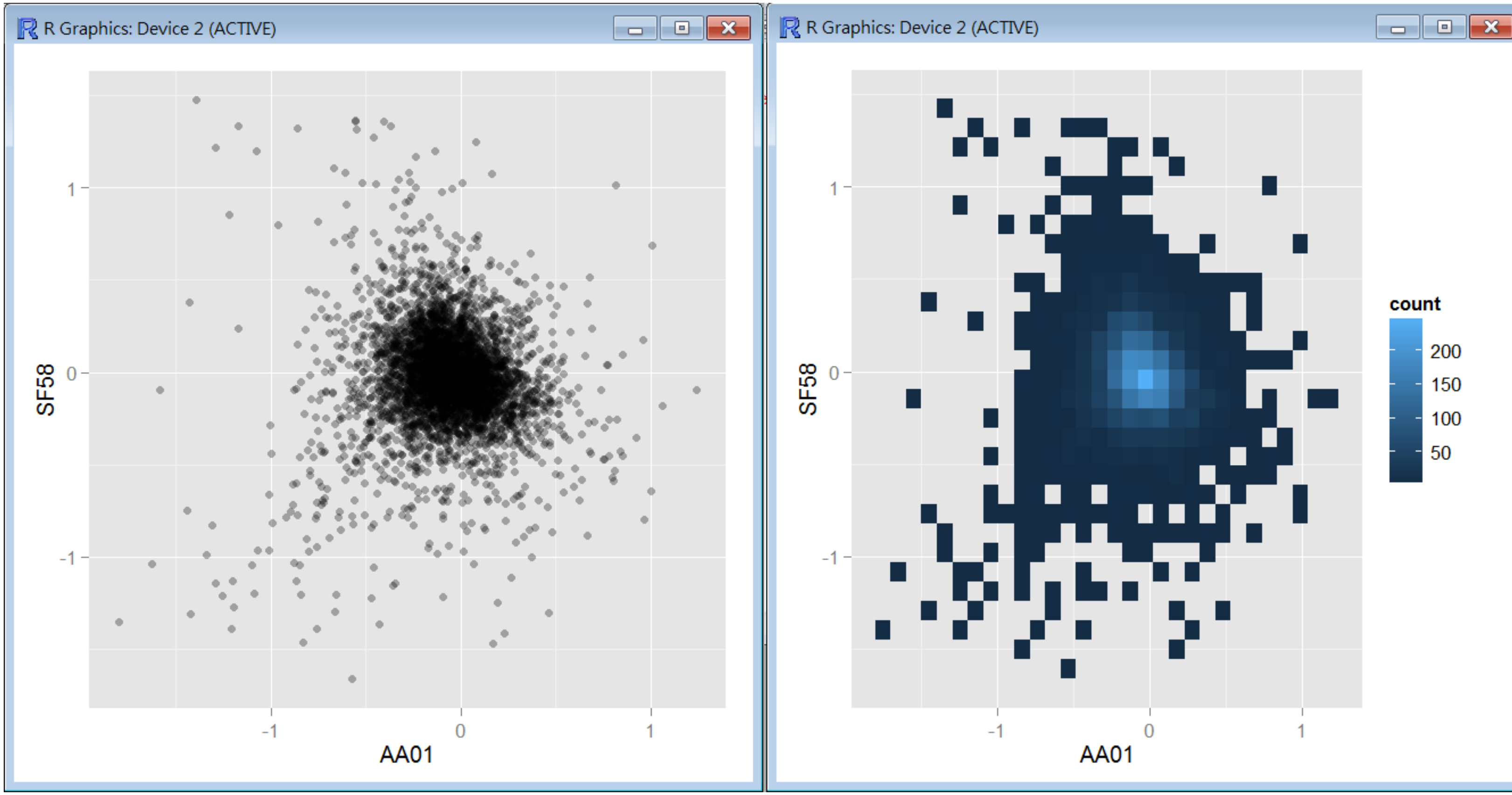
Scatterplot for big data



For large datasets with overplotting

- RMD_example 2.10

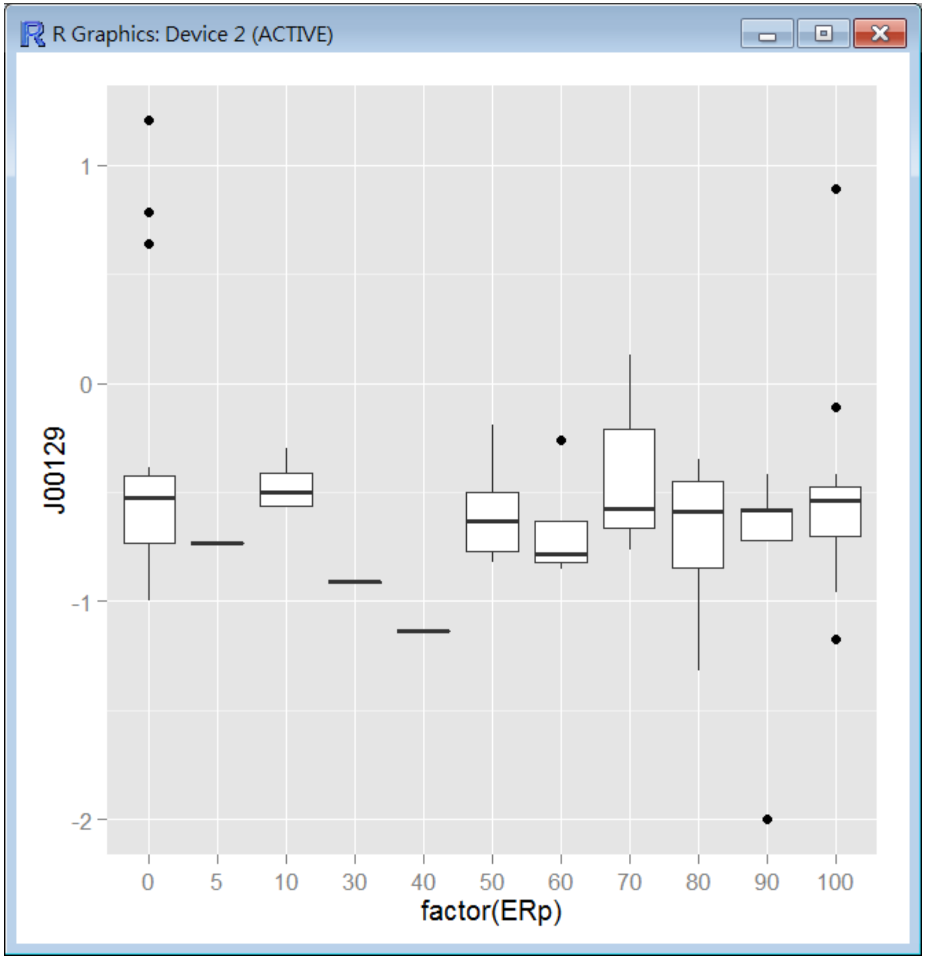
Scatterplot for big data



- Alpha aesthetic makes the points more transparent
- Heatmap shows the density
  - RMD_example 2.10

ERp	J00129
100	-0.795
0	-0.509
100	-0.961
0	-0.749
0	-0.426
100	-0.566
90	-0.42
0	-0.499
80	-0.465
50	-0.189
70	-0.739
0	-0.601
0	0.786
50	-0.819
80	-0.448
0	1.206
80	-0.391
50	-0.234
50	-0.75
10	-0.299
0	-0.455
100	-1.173
90	-0.721
100	-0.416
50	-0.688
80	-0.352
5	-0.734
100	-0.112
30	-0.919

Box plots for different ERp

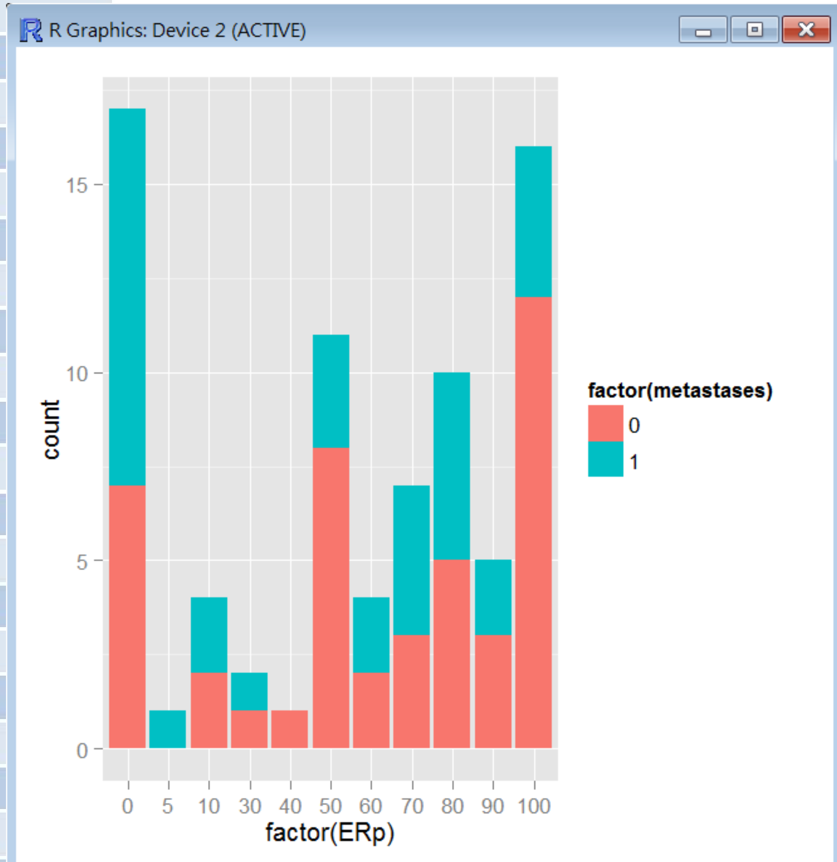


RMD_example 2.11

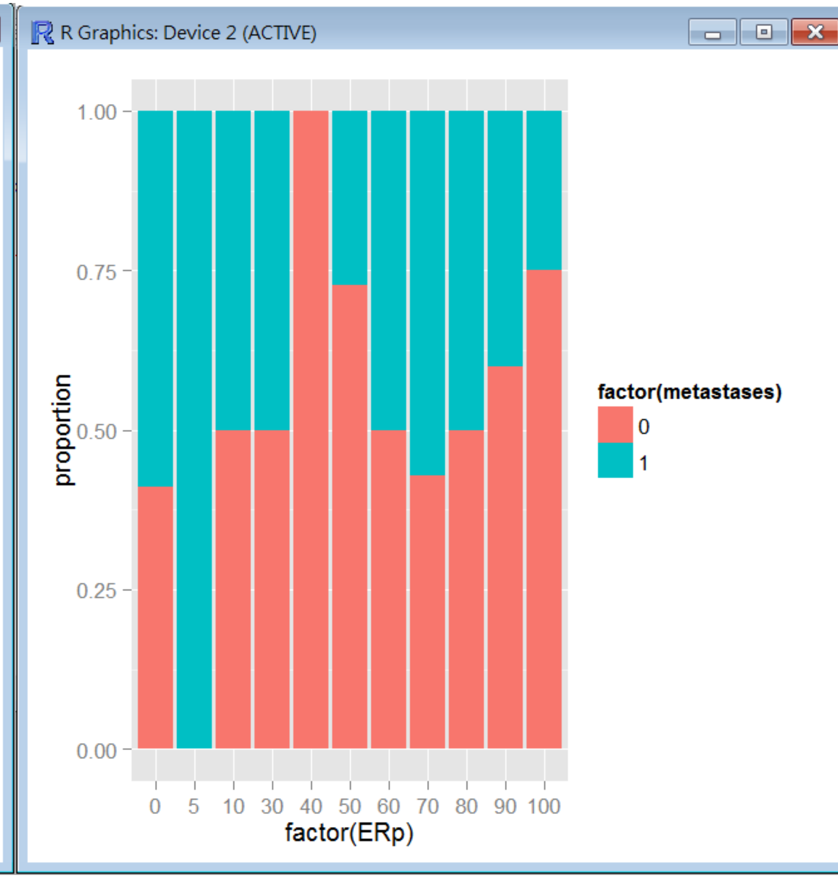
Stacked bar chart

ERp	metastases
100	0
0	1
100	0
0	0
0	0
100	1
90	0
0	1
80	1
50	1
70	1
0	1
0	1
50	0
80	0
0	1
80	0
50	0
50	0
10	0
0	1
100	0
90	0
100	0
50	1
80	1
5	1
100	0
30	1

For counts

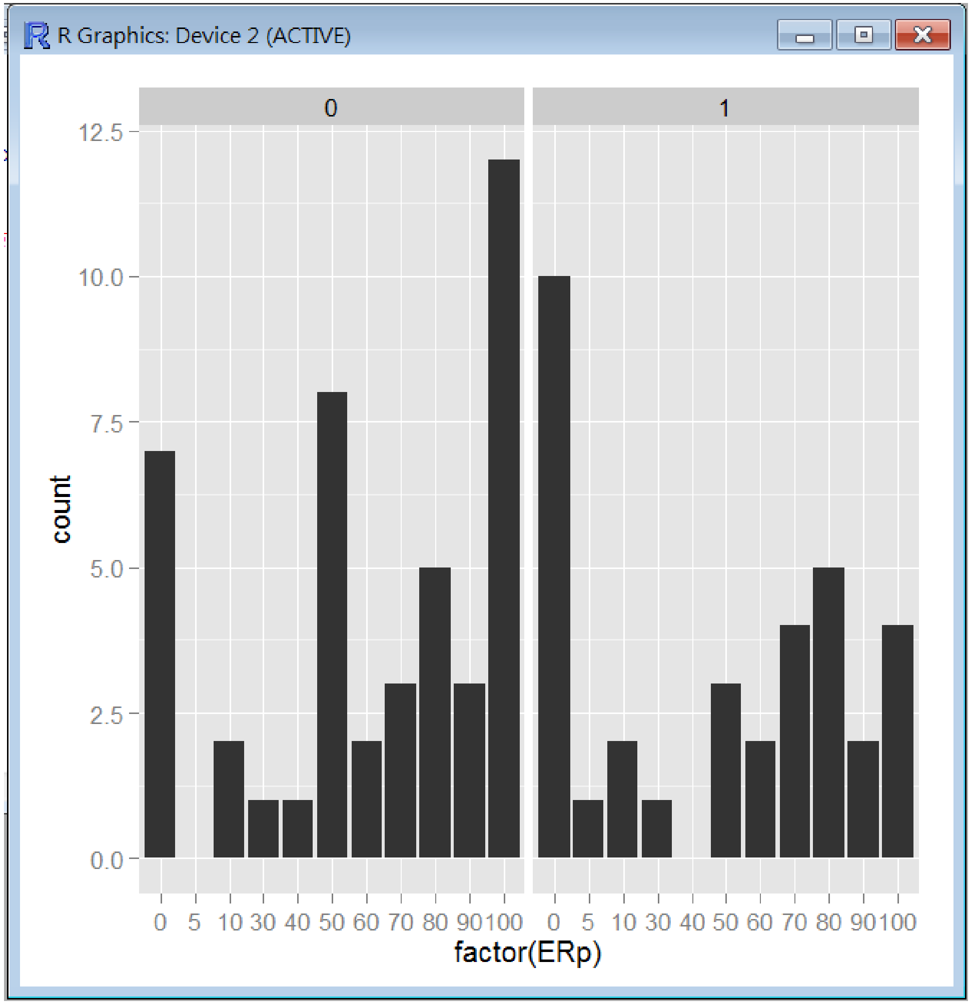


For proportions



- RMD_example 2.12

Faceting bar charts

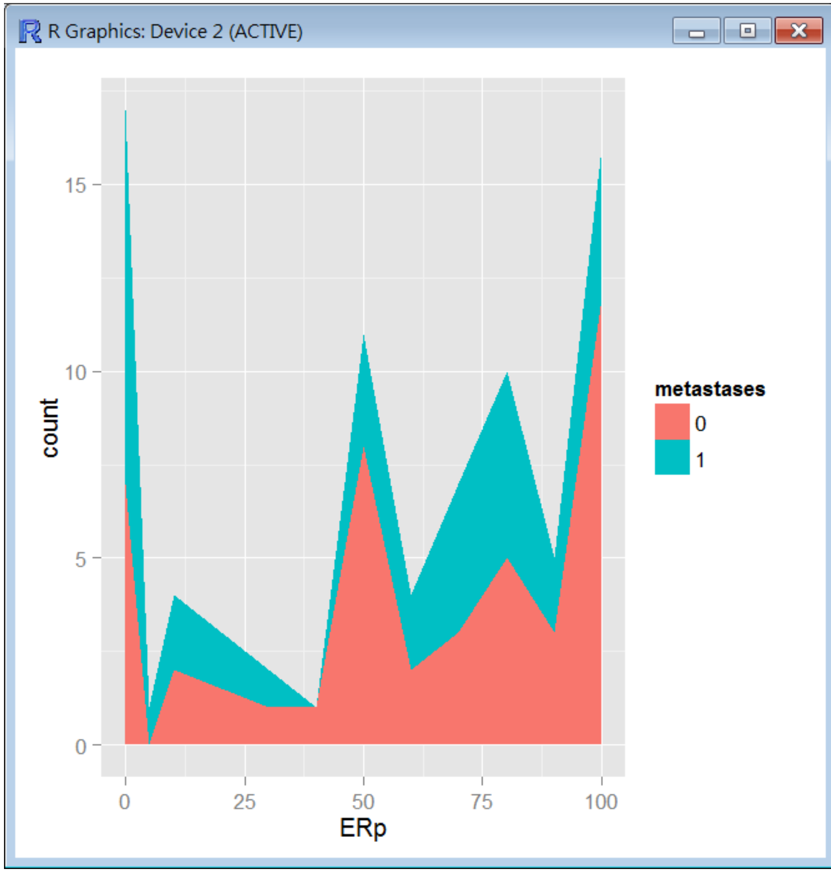


Bar charts of ERp for different metastases

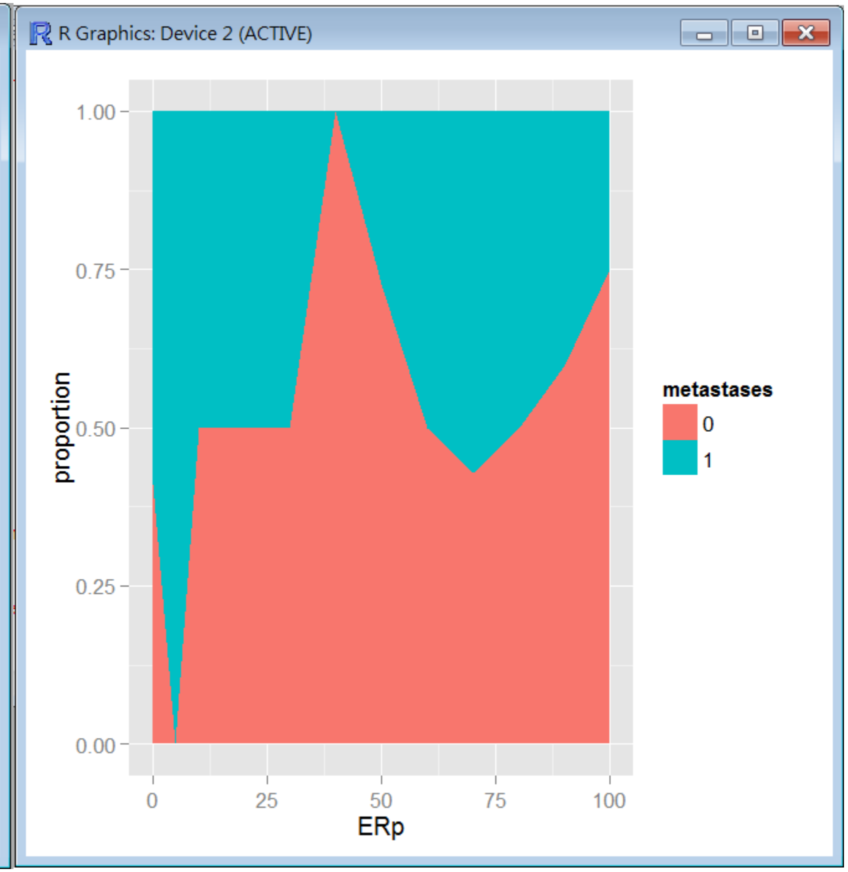
- RMD_example 2.13

Stacked area chart

For counts



For proportions

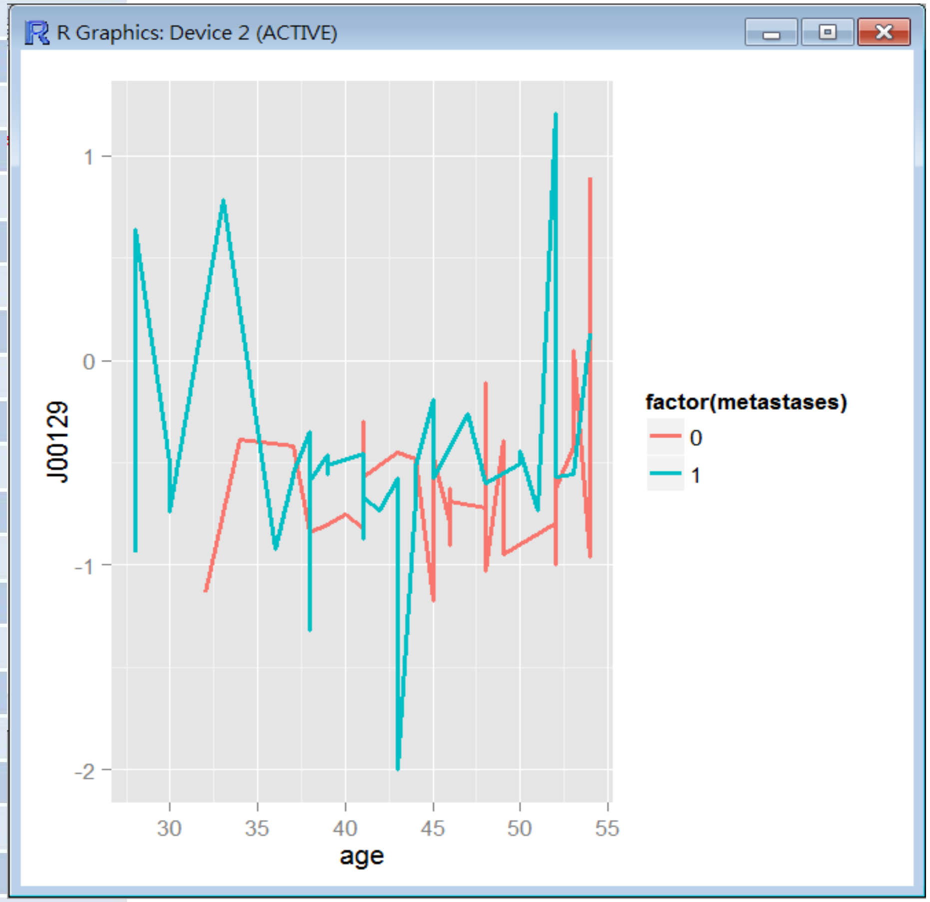


Treat ERp as continuous!

- RMD_example 2.14

Time series plot

age	J00129	metastases
52	-0.795	0
50	-0.509	1
54	-0.961	0
40	-0.749	0
53	-0.426	0
37	-0.566	1
37	-0.42	0
30	-0.499	1
39	-0.465	1
45	-0.189	1
30	-0.739	1
48	-0.601	1
33	0.786	1
41	-0.819	0
43	-0.448	0
52	1.206	1
49	-0.391	0
54	-0.234	0
40	-0.75	0
41	-0.299	0
41	-0.455	1
45	-1.173	0
48	-0.721	0
48	-0.416	0
44	-0.688	1
38	-0.352	1
51	-0.734	1
48	-0.112	0
36	-0.919	1



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

- RMD_example 2.15

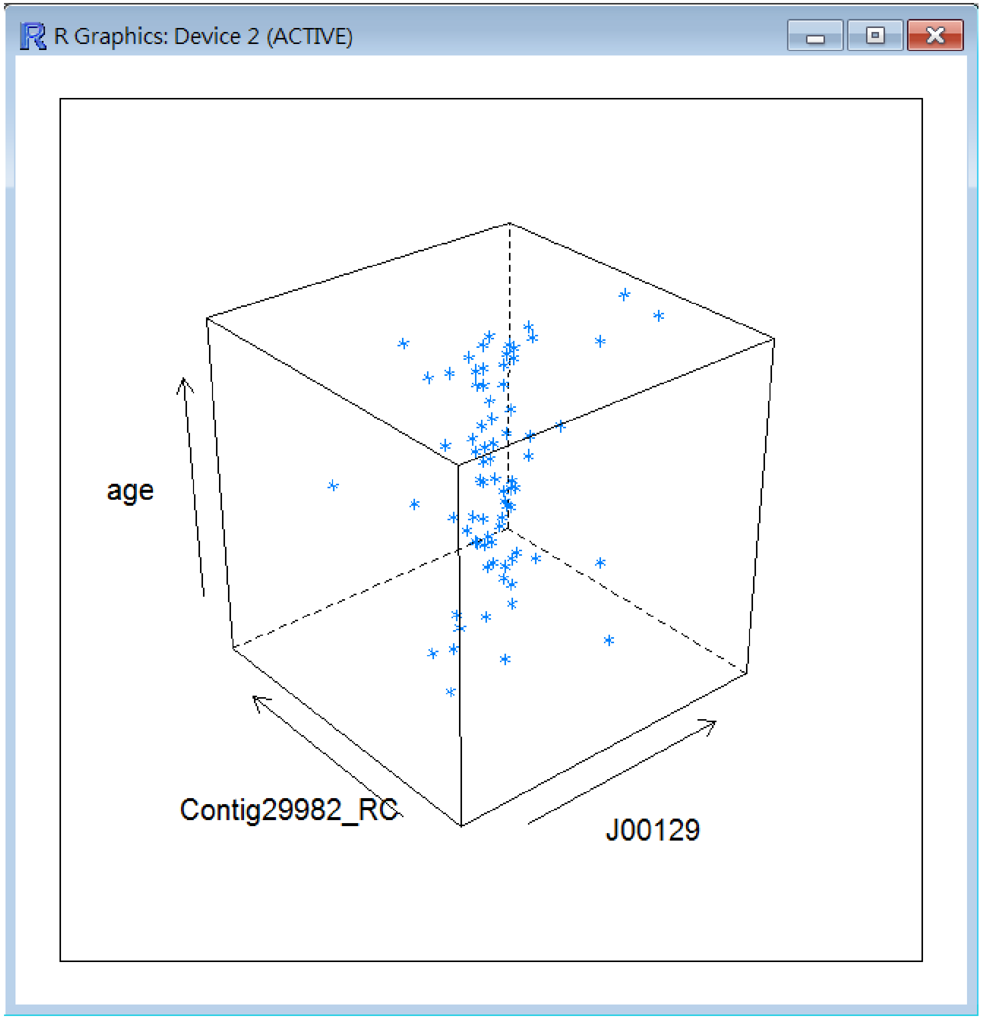
Multivariate data ($\geq$3 dimensions)

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

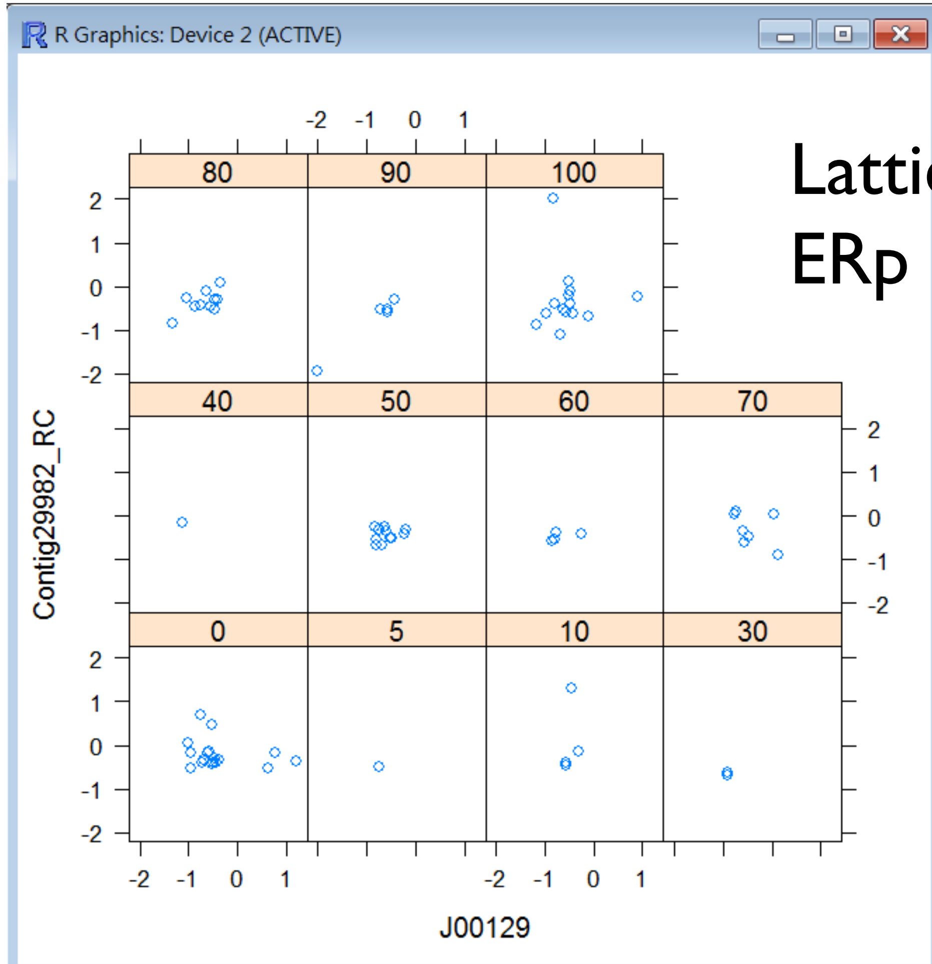
age	J00129	Contig29982_RC
52	-0.795	-0.387
50	-0.509	0.459
54	-0.961	-0.631
40	-0.749	0.699
53	-0.426	-0.406
37	-0.566	-0.596
37	-0.42	-0.286
30	-0.499	-0.402
39	-0.465	-0.533
45	-0.189	-0.309
30	-0.739	0.093
48	-0.601	-0.177
33	0.786	-0.164
41	-0.819	-0.267
43	-0.448	-0.296
52	1.206	-0.353
49	-0.391	-0.31
54	-0.234	-0.404
40	-0.75	-0.316
41	-0.299	-0.137
41	-0.455	-0.288
45	-1.173	-0.887
48	-0.721	-0.527
48	-0.416	-0.633
44	-0.688	-0.659
38	-0.352	0.088
51	-0.734	-0.493
48	-0.112	-0.67
36	-0.919	-0.704



- RMD_example 2.16

ERp	J00129	Contig29982_RC
100	-0.795	-0.387
0	-0.509	0.459
100	-0.961	-0.631
0	-0.749	0.699
0	-0.426	-0.406
100	-0.566	-0.596
90	-0.42	-0.286
0	-0.499	-0.402
80	-0.465	-0.533
50	-0.189	-0.309
70	-0.739	0.093
0	-0.601	-0.177
0	0.786	-0.164
50	-0.819	-0.267
80	-0.448	-0.296
0	1.206	-0.353
80	-0.391	-0.31
50	-0.234	-0.404
50	-0.75	-0.316
10	-0.299	-0.137
0	-0.455	-0.288
100	-1.173	-0.887
90	-0.721	-0.527
100	-0.416	-0.633
50	-0.688	-0.659
80	-0.352	0.088
5	-0.734	-0.493
100	-0.112	-0.67
30	-0.919	-0.704

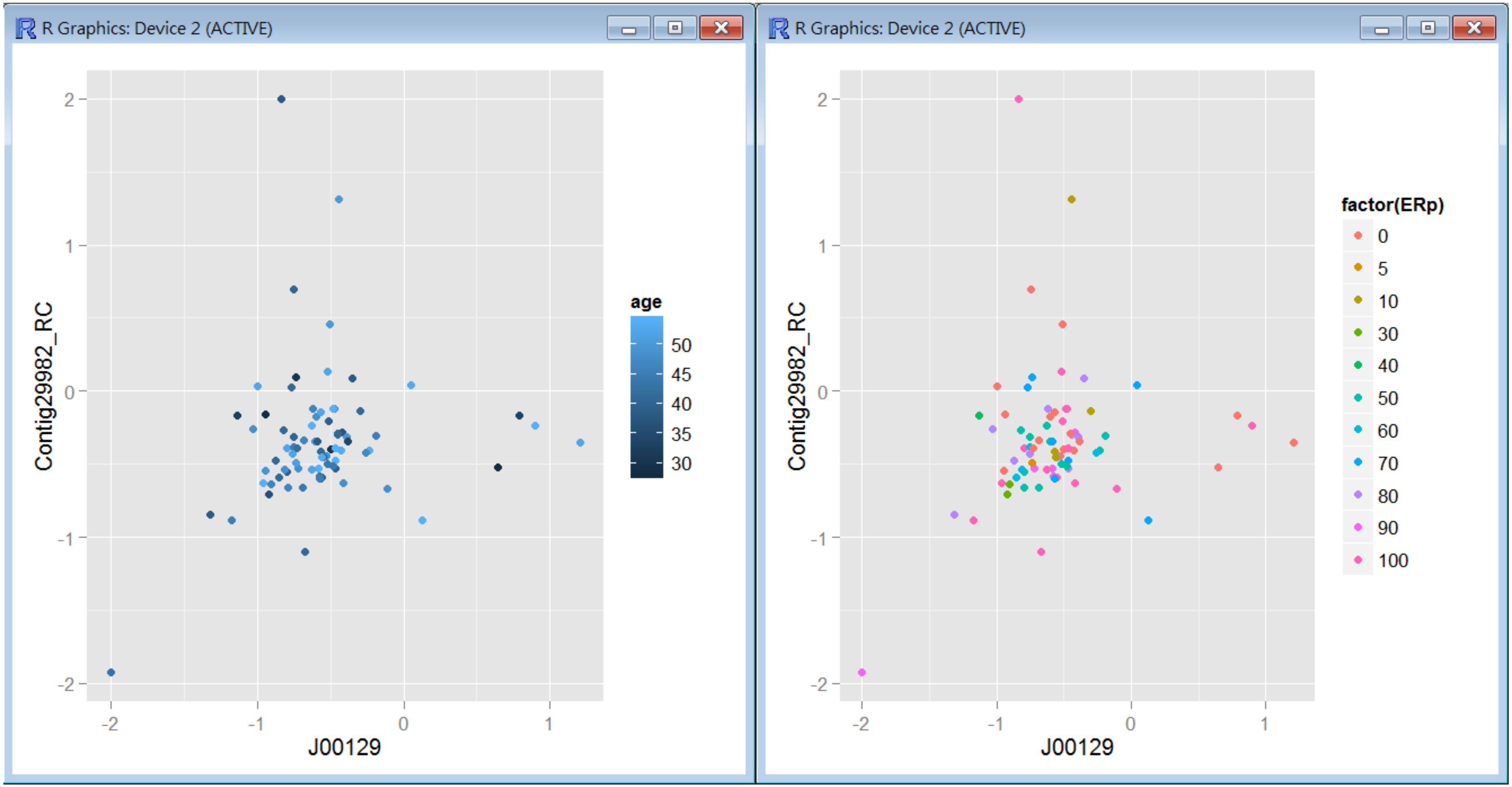
Lattice in the 3rd dim



Lattices in ERp

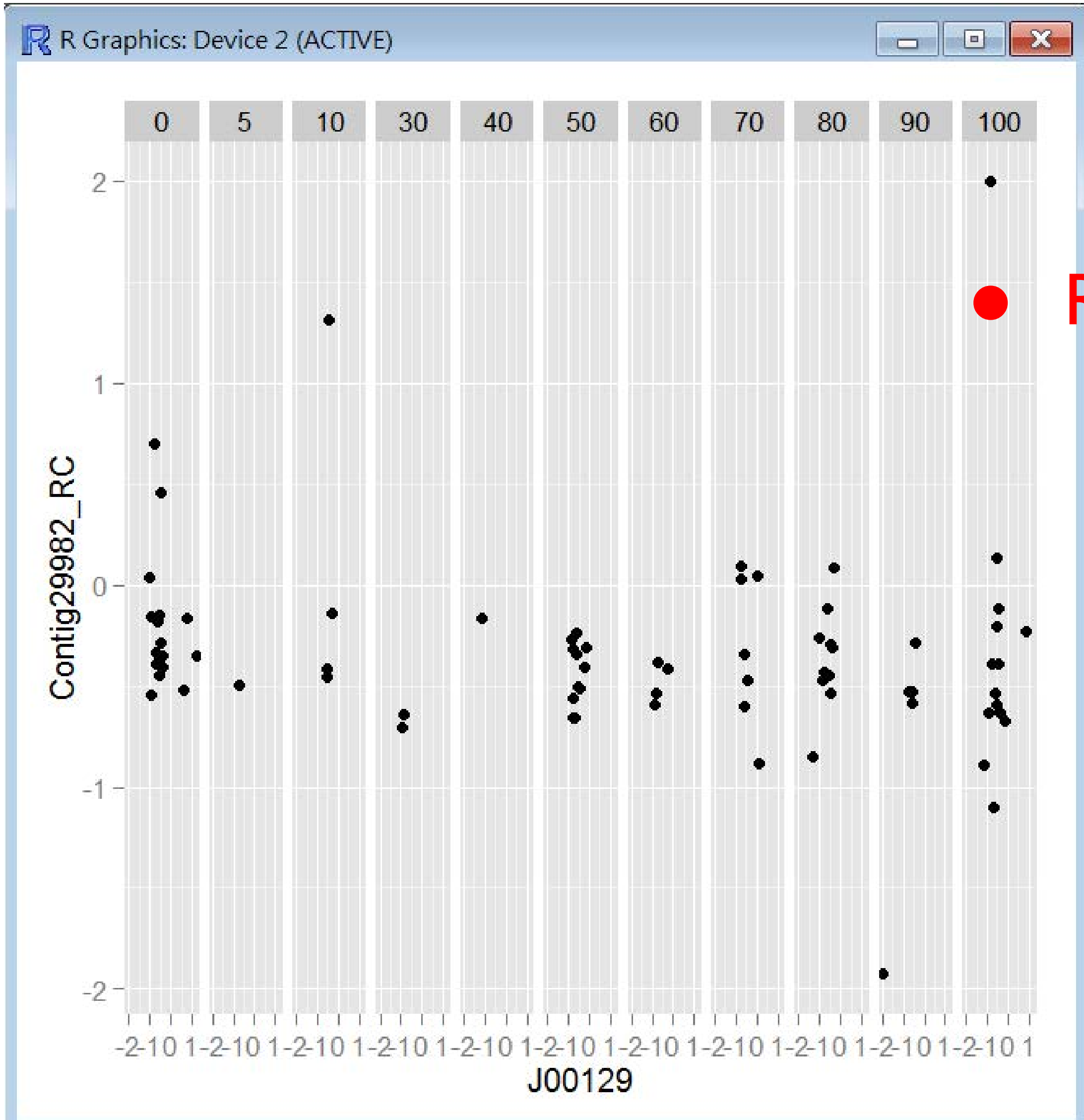
- RMD_example 2.16

Map the 3rd dim to colors

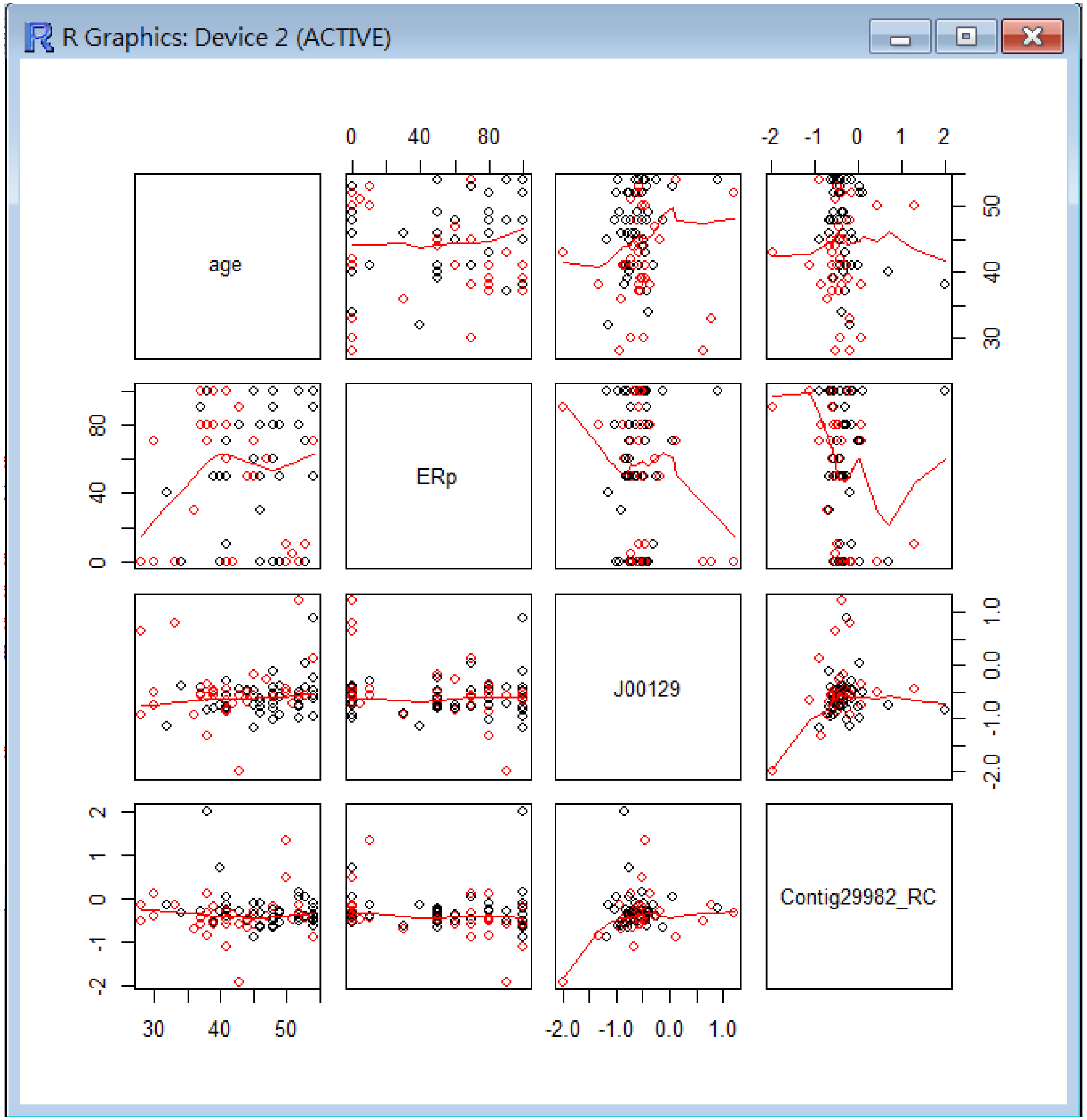


- RMD_example 2.16

Lay out panels in the 3rd dim



- RMD_example 2.16



Scatterplot matrices

Color in metastases
Add smooth lines

- RMD_example 2.17

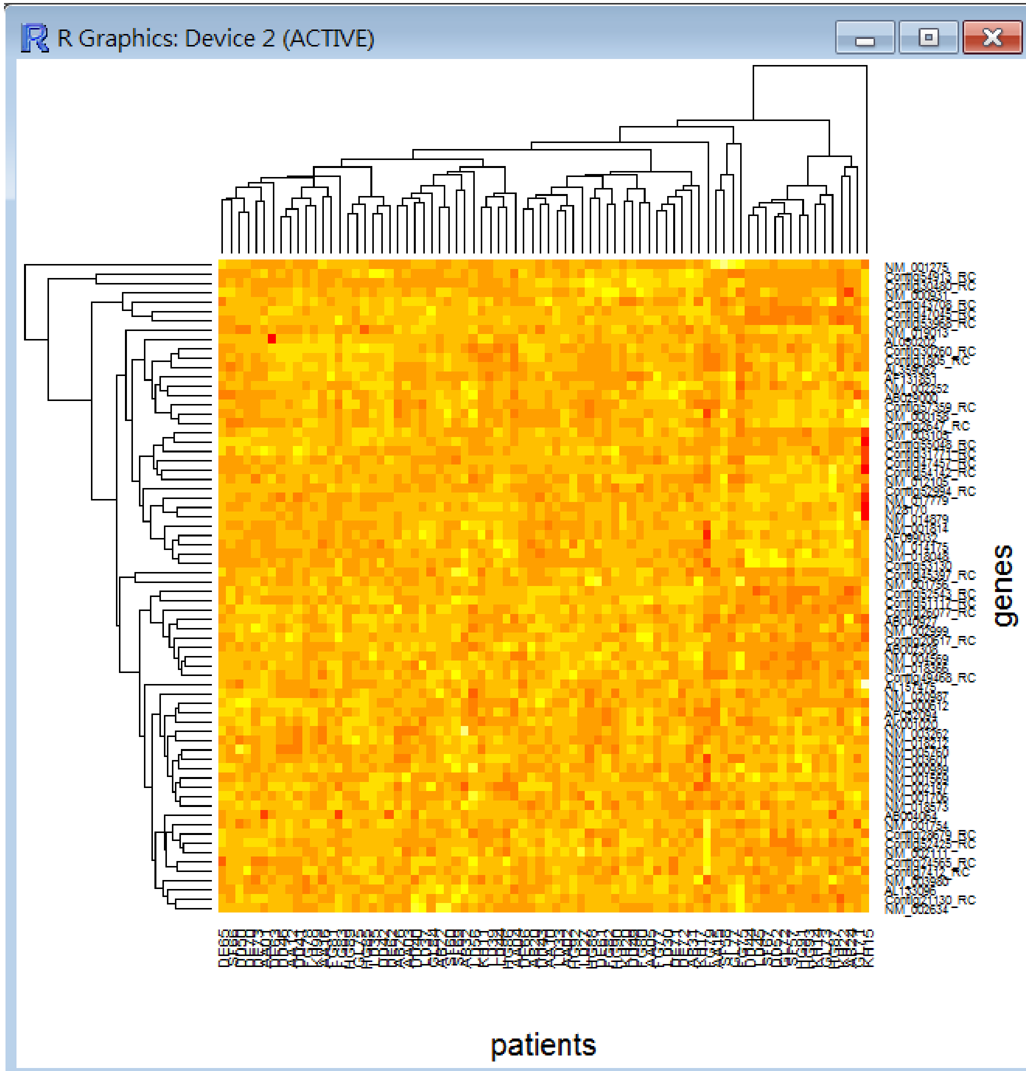
Heatmap

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

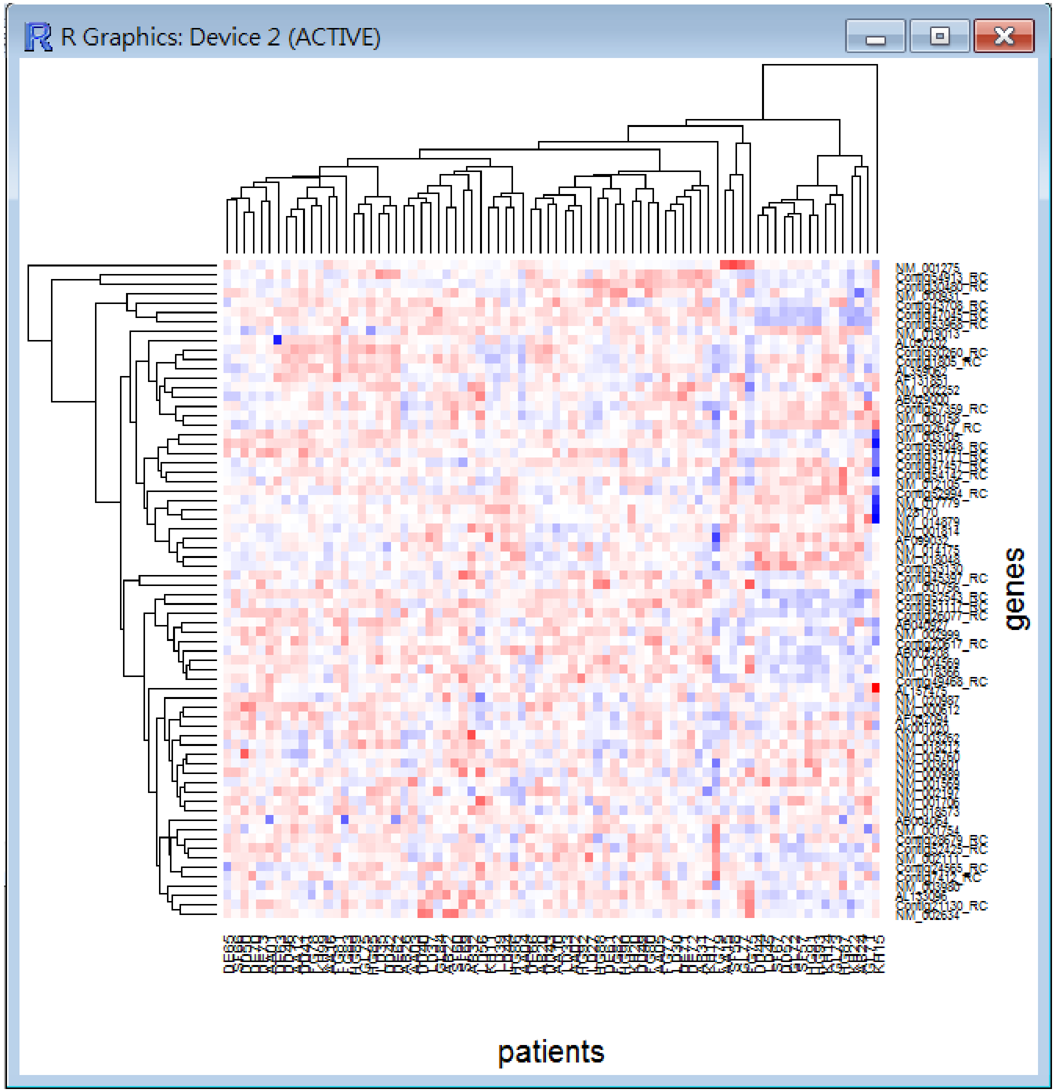
A graphical representation of data where the individual values contained in a matrix are represented as colors. [wikipedia]



- RMD_example 2.18

Heatmap

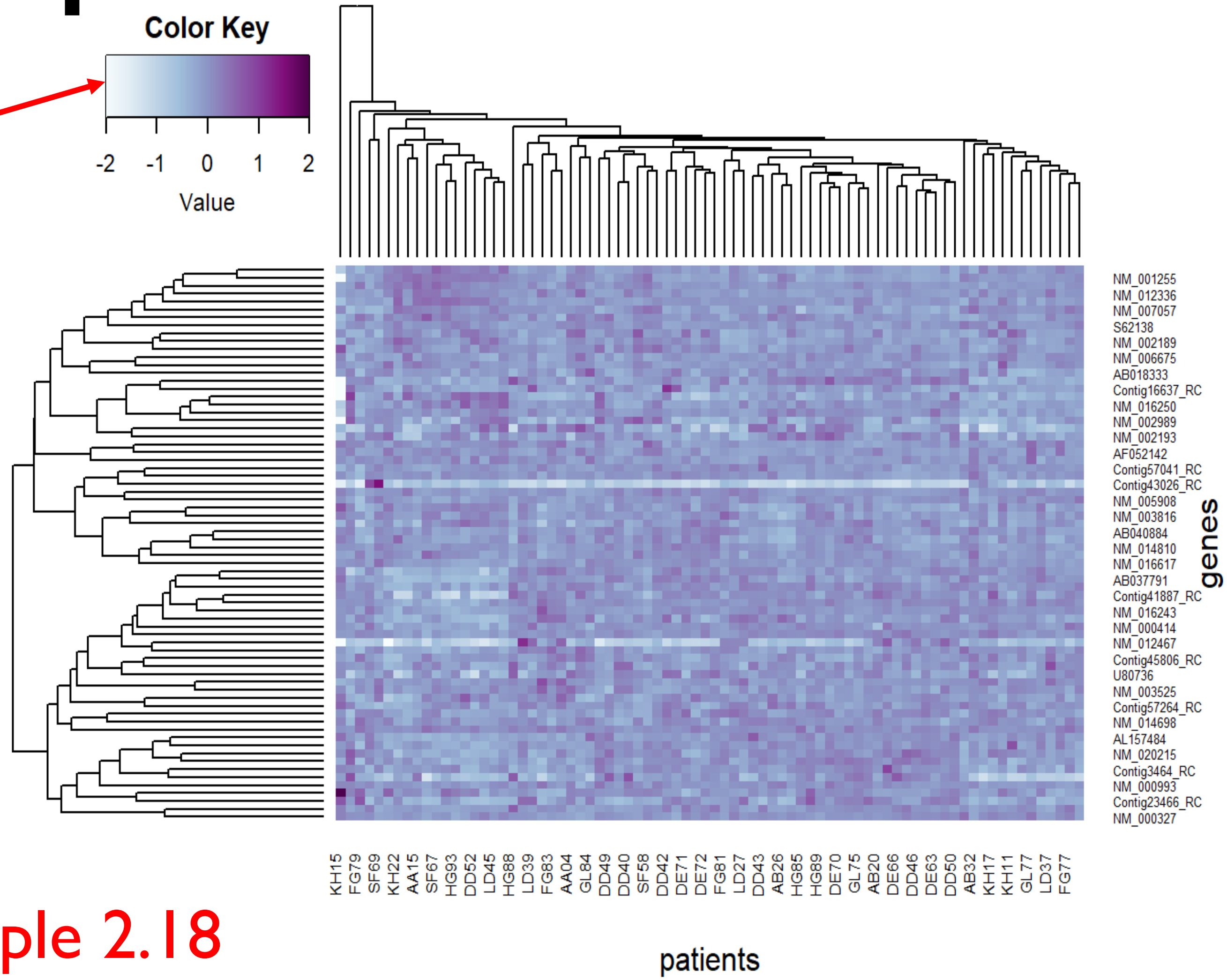
With different color schemes



- RMD_example 2.18

Heatmap

With
color key



- RMD_example 2.18