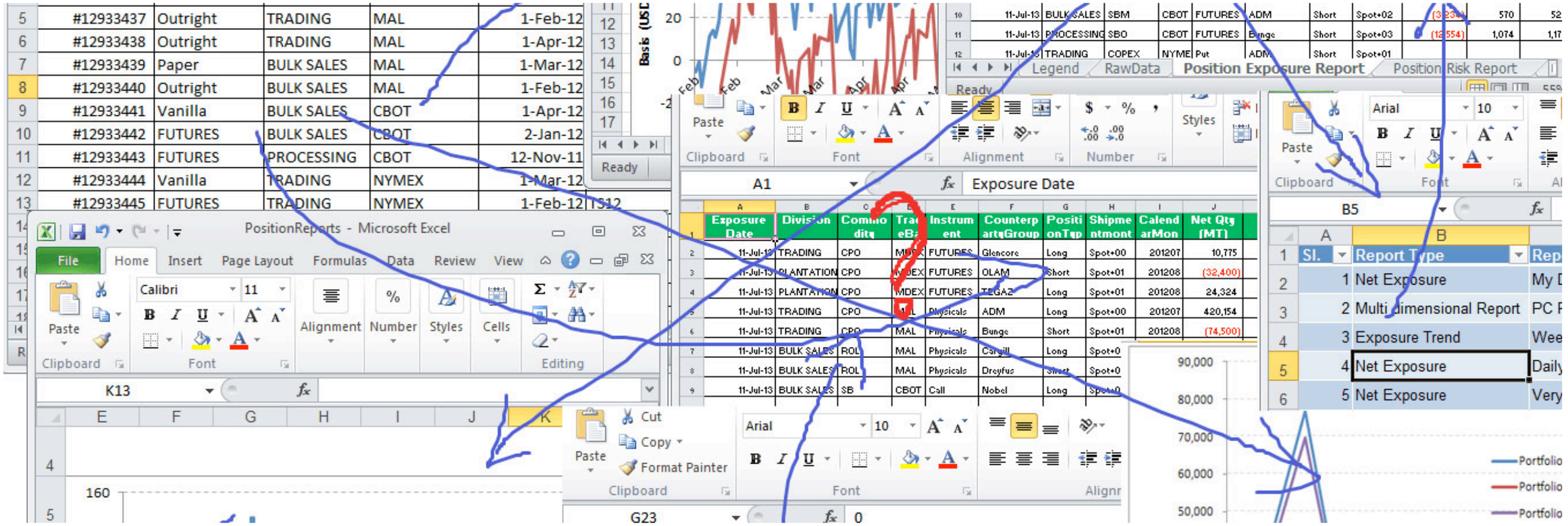
BASICs in R
Introduction

Tuesday 22nd January 2019

Anne Bech Risum
Peter Skou
Morten Arendt Rasmussen

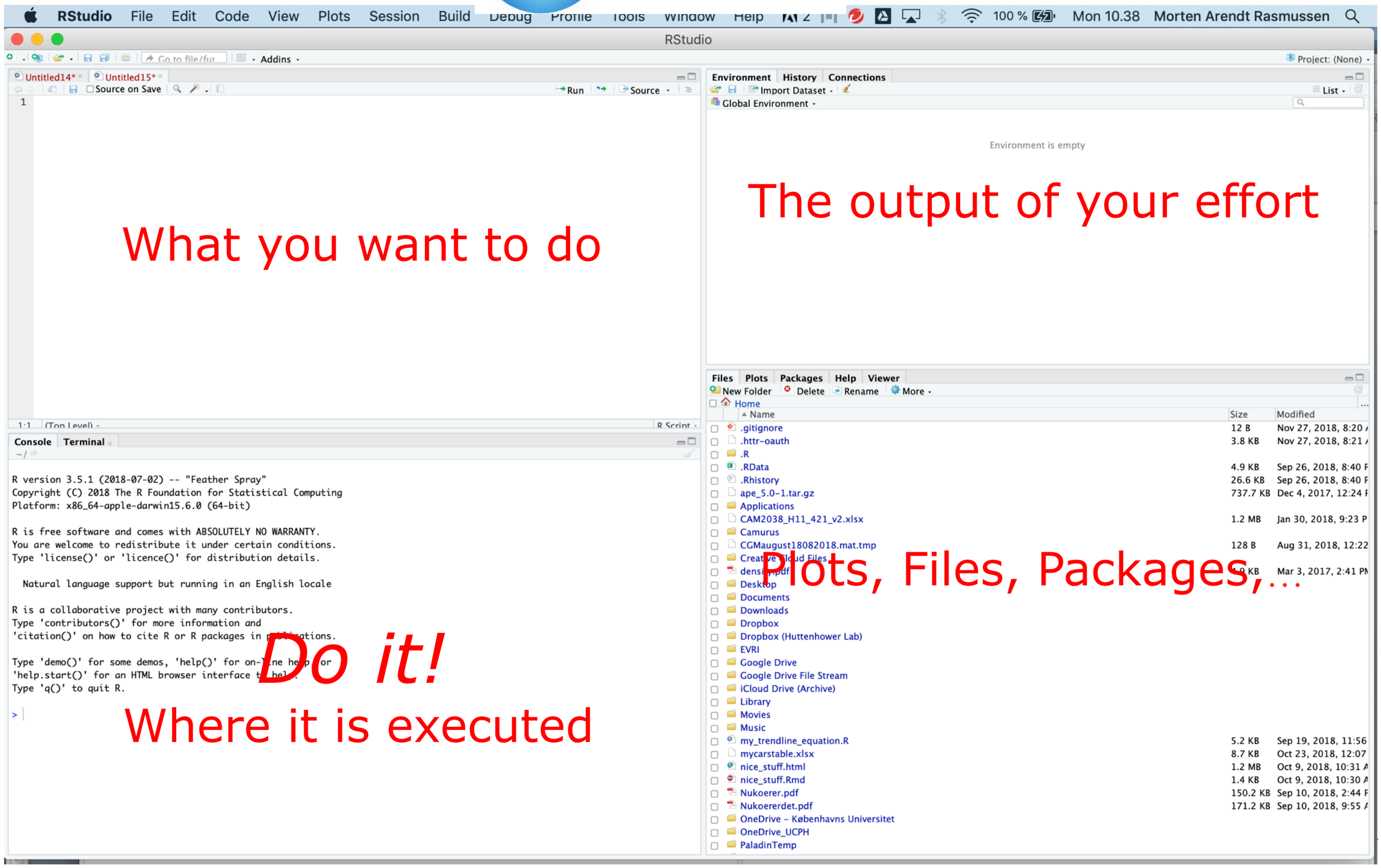
MOTIVATION

- Digitalization on KU
- Reproducibility
- Get *quicker insight* and *More knowledge* out of your data


#12933437 Outright TRADING MAL 1-Feb-12
#12933438 Outright TRADING MAL 1-Apr-12
#12933439 Paper BULK SALES MAL 1-Mar-12
#12933440 Outright BULK SALES MAL 1-Feb-12
#12933441 Vanilla BULK SALES CBOT 1-Apr-12
#12933442 FUTURES BULK SALES CBOT 2-Jan-12
#12933443 FUTURES PROCESSING CBOT 12-Nov-11
#12933444 Vanilla TRADING NYMEX 1-Mar-12
#12933445 FUTURES TRADING NYMEX 1-Feb-12
PositionReports - Microsoft Excel
Exposure Date
Position Exposure Report
Position Risk Report
Net Exposure
Multi dimensional Report
Exposure Trend
Portfolio


Excel Hell

RStudio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
100 % Mon 10.38 Morten Arendt Rasmussen
RStudio
Addins
Project: (None)
Untitled14* Untitled15*
Source on Save
Run
Source
1
1:1 (Top Level)
R Script
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
.gitignore 12 B Nov 27, 2018, 8:20
.httr-oauth 3.8 KB Nov 27, 2018, 8:21
.R
.RData 4.9 KB Sep 26, 2018, 8:40 P
.Rhistory 26.6 KB Sep 26, 2018, 8:40 P
ape_5.0-1.tar.gz 737.7 KB Dec 4, 2017, 12:24 P
Applications
CAM2038_H11_421_v2.xlsx 1.2 MB Jan 30, 2018, 9:23 P
Camurus
CGMaugust18082018.mat.tmp 128 B Aug 31, 2018, 12:22
Creative Cloud Files
Desktop
Documents
Downloads
Dropbox
Dropbox (Huttenhower Lab)
EVRI
Google Drive
Google Drive File Stream
iCloud Drive (Archive)
Library
Movies
Music
my_trendline_equation.R 5.2 KB Sep 19, 2018, 11:56
mycarstable.xlsx 8.7 KB Oct 23, 2018, 12:07
nice_stuff.html 1.2 MB Oct 9, 2018, 10:31 A
nice_stuff.Rmd 1.4 KB Oct 9, 2018, 10:30 A
Nukoerer.pdf 150.2 KB Sep 10, 2018, 2:44 P
Nukoererdet.pdf 171.2 KB Sep 10, 2018, 9:55 A
OneDrive - Københavns Universitet
OneDrive_UCPH
PaladinTemp
Console Terminal
~/
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
What you want to do
The output of your effort
Do it!
Where it is executed
Plots, Files, Packages, …

Packages

R comes with some functionality, but not everything is covered.

Therefor we need additional functions.

There comes in the form of ***packages*** which is installed directly from within R.

> *install.packages('ggplot2')*

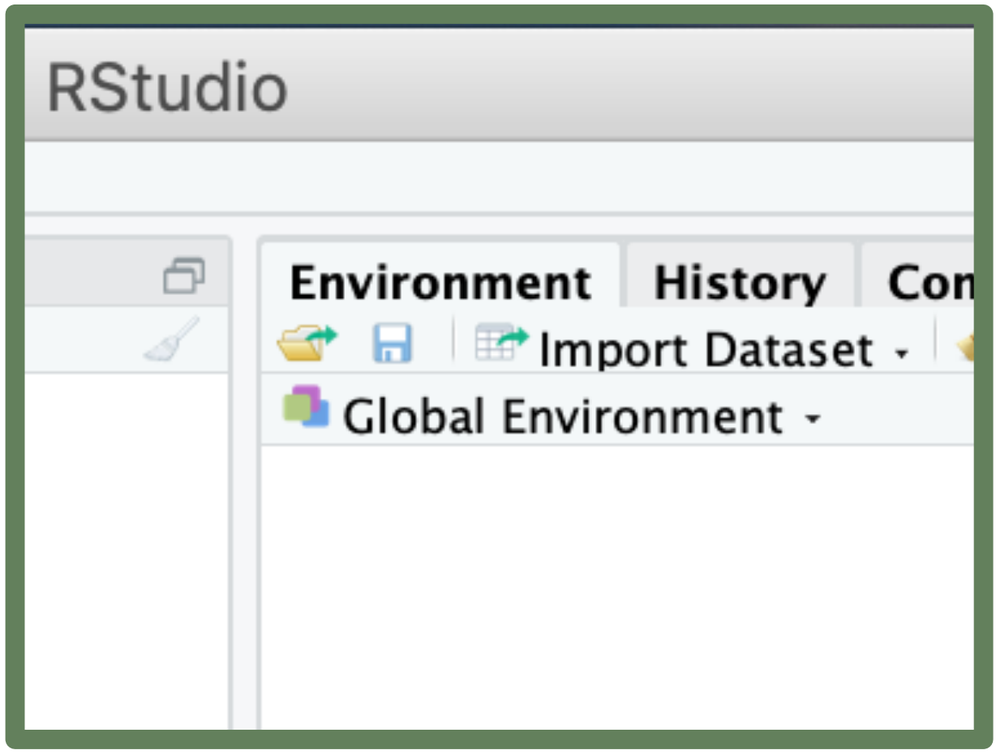
From
CRAN (http://cran.r-project.org/)
(13713 available packages)

> *devtools::install_github('vqv/ggbiplot')*

From
github (https://github.com/)

Data import

Use the *point-n-click* method in Rstudio



…. Eventually copy paste the commands produced into the script

Or

Use the **rio** package

```
> library(rio)
> X <-
import('myExcelFile.xlsx')
```

data.frame()

You should store your data in a data.frame

A data frame is as an excel sheet with first row being variable names.

To be aware of:

Avoid using space, leading numerics and repetitions in names.

Some useful functions to look at the data.frame

> head() / tail()
> View()
> rownames()
> str()
> colnames()

Descriptive stats

Calculate descriptive stats – *a single number describing a distribution of numbers* – splitted according to grouping information.

Functions

> mean(), sd(), min(), max(), range(), length()

> aggregate()

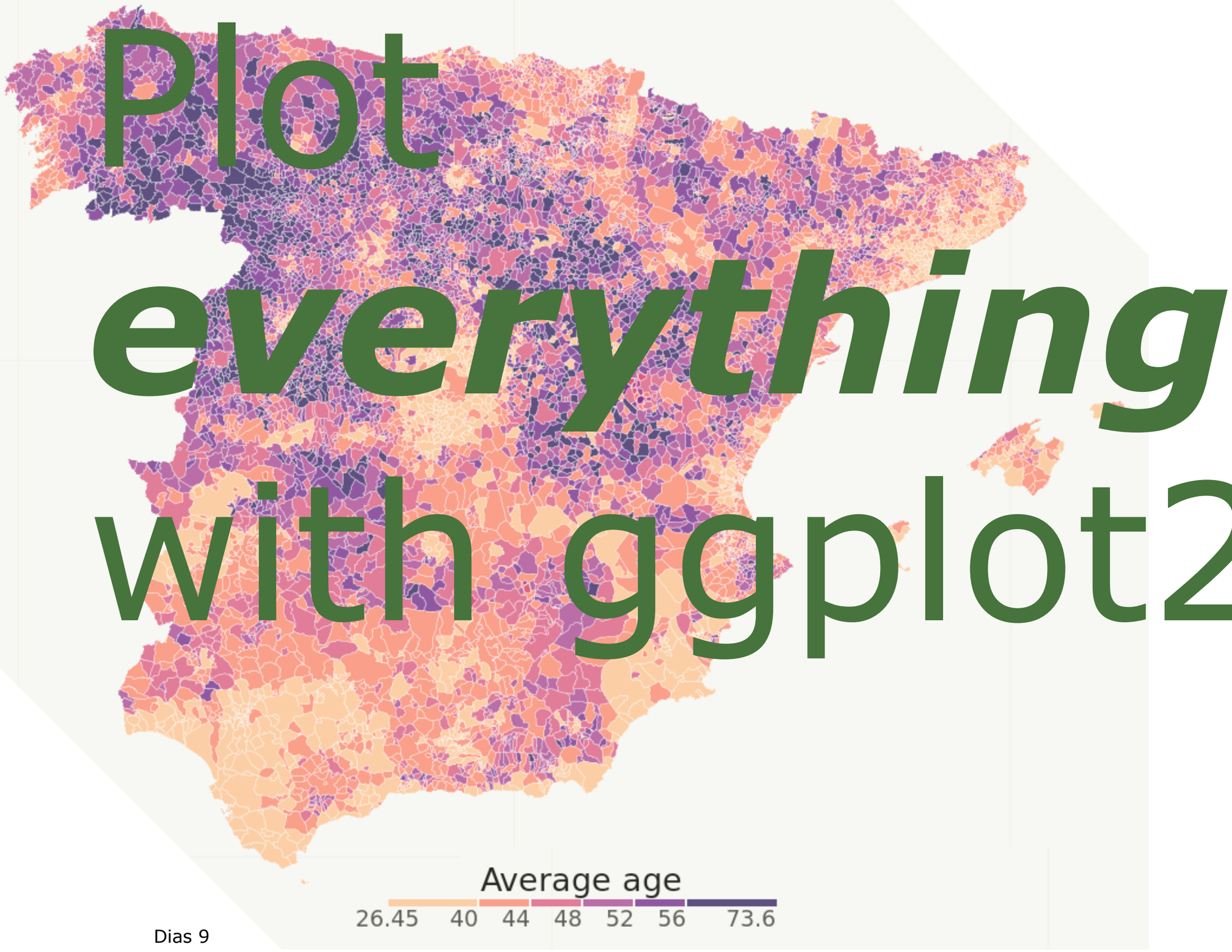
Overlook of a design

How many observations are there for a combination of (design) variables

> table()

CAT – Food Science – mortenr@food.ku.dk

Plot *everything* with ggplot2



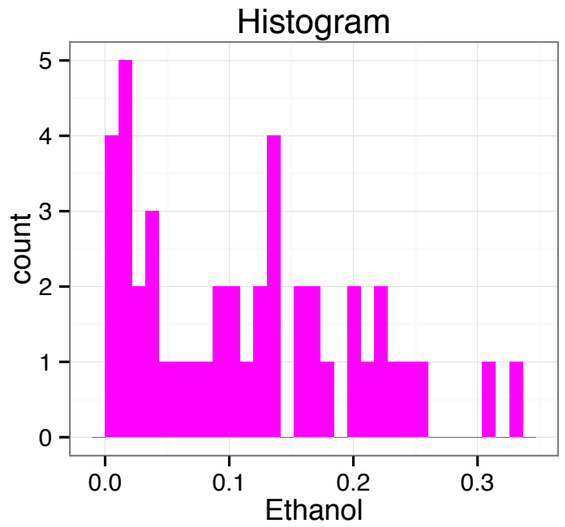
PLOTS

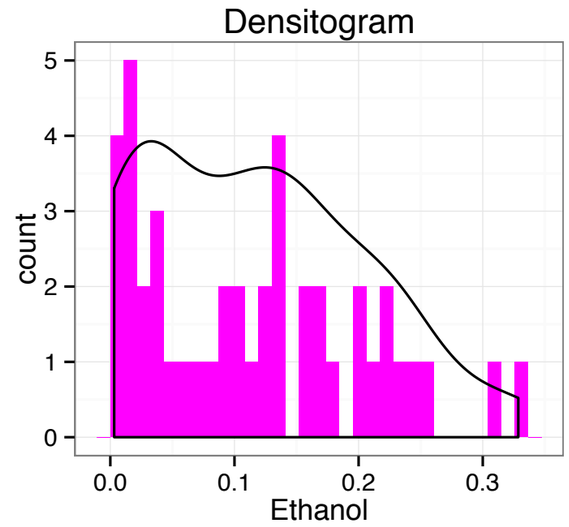
- Histogram, Densitogram, Boxplot, Jitterplot, Violinplot
    - Continuous data
    - Describes distribution
- Lineplot, Stem plot
    - Continuous data
    - Ordinal in e.g. time
- Bar chart, pie chart
    - Categorical data or compositional data
- Scatterplot
    - Bi variate continuous data
- Spline plot
    - Smooth relation between two (or more) variables

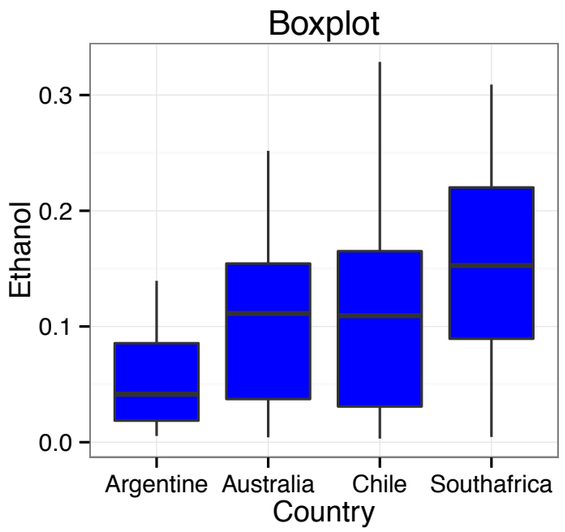
Histogram, Densitogram, Boxplot, Jitterplot, Violin plot,

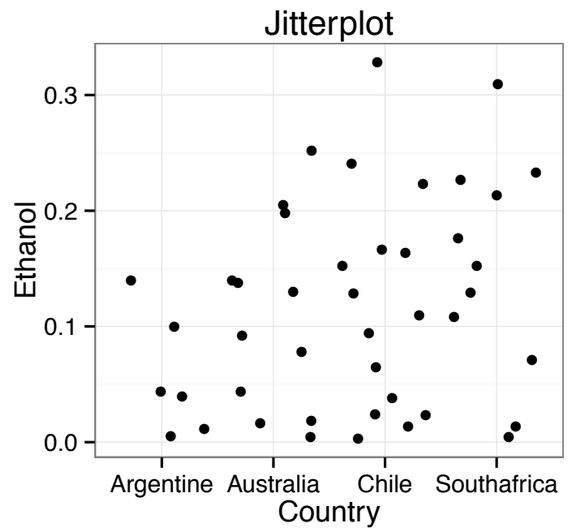
- Jitter shows raw data
- Boxplot and violin can handle many observations
  Densitogram and Violin may cheat in the representation of raw data

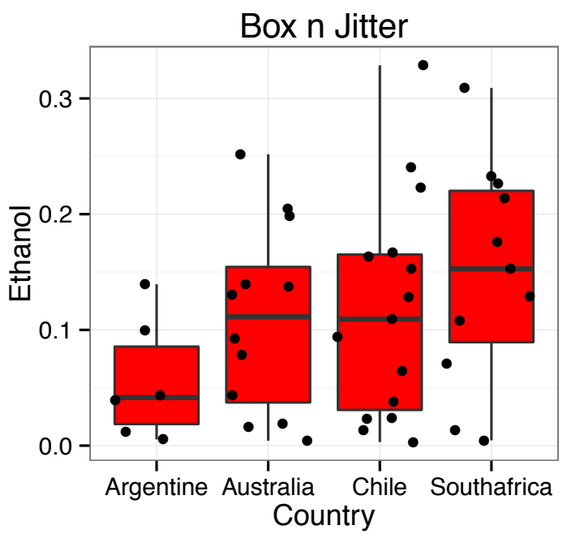
Continuous data
Describes distribution

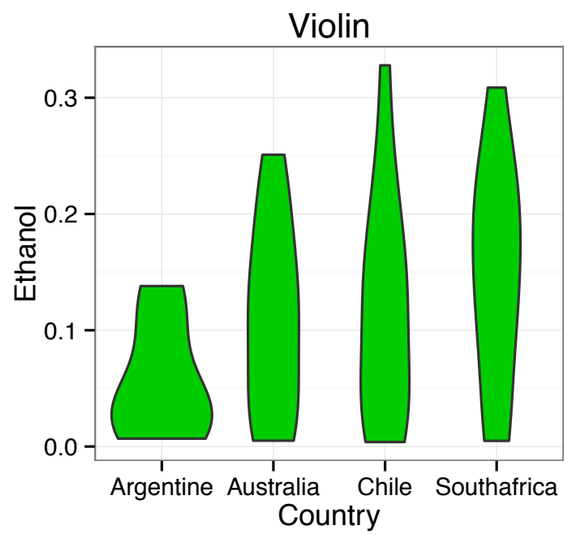






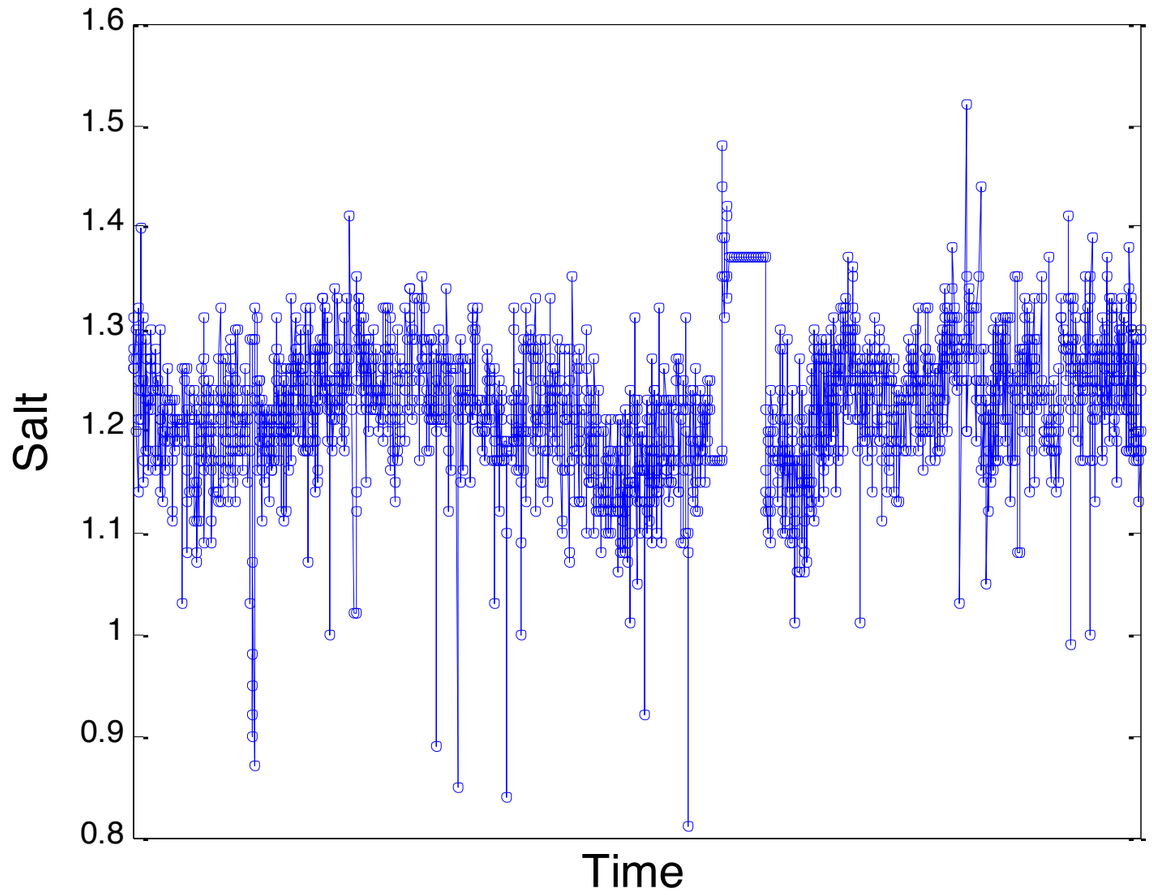






Lineplot

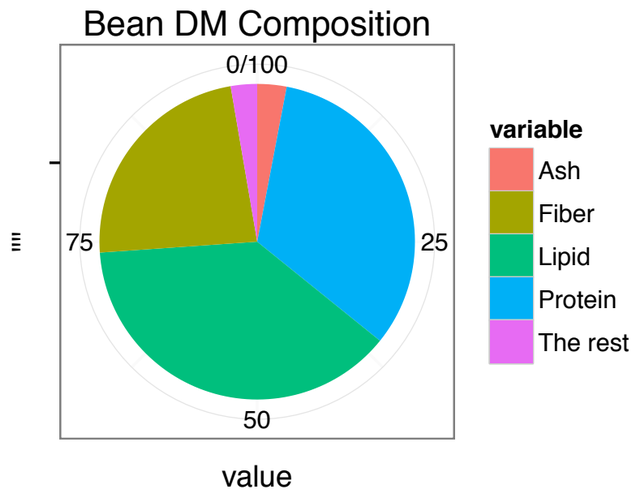
- Is used in e.g. Statistical Process Control (SPC)
- Excellent for capturing drift (systematical change over time) in the laboratory of in the production.

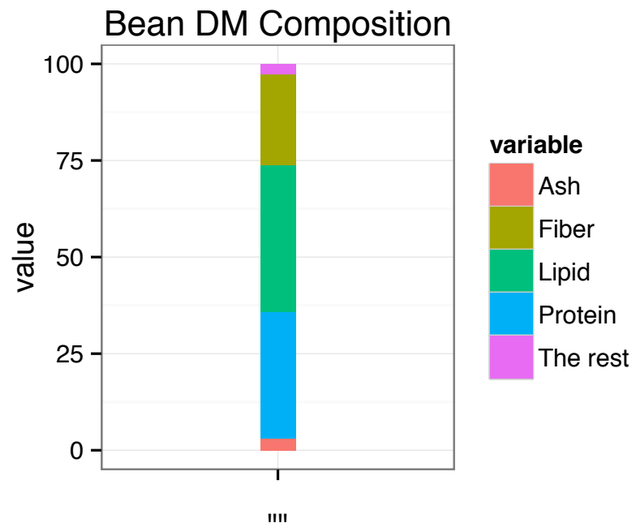


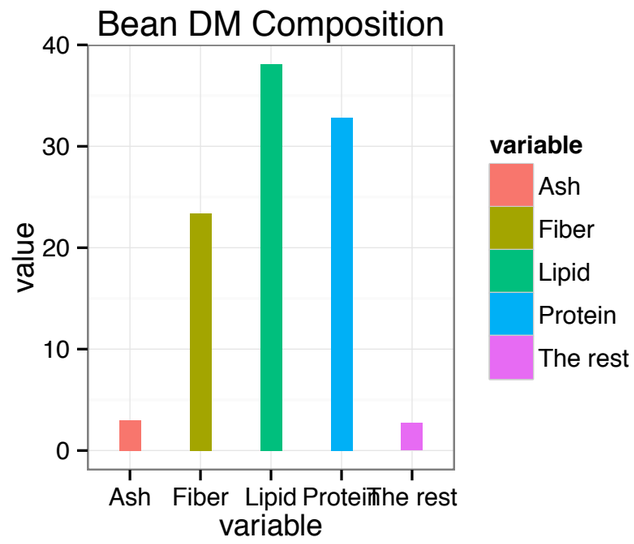
Continuous data
Ordinal in e.g. time

Bar charts
Pie charts

- OBS: Do not replace a boxplot with a bar chart!

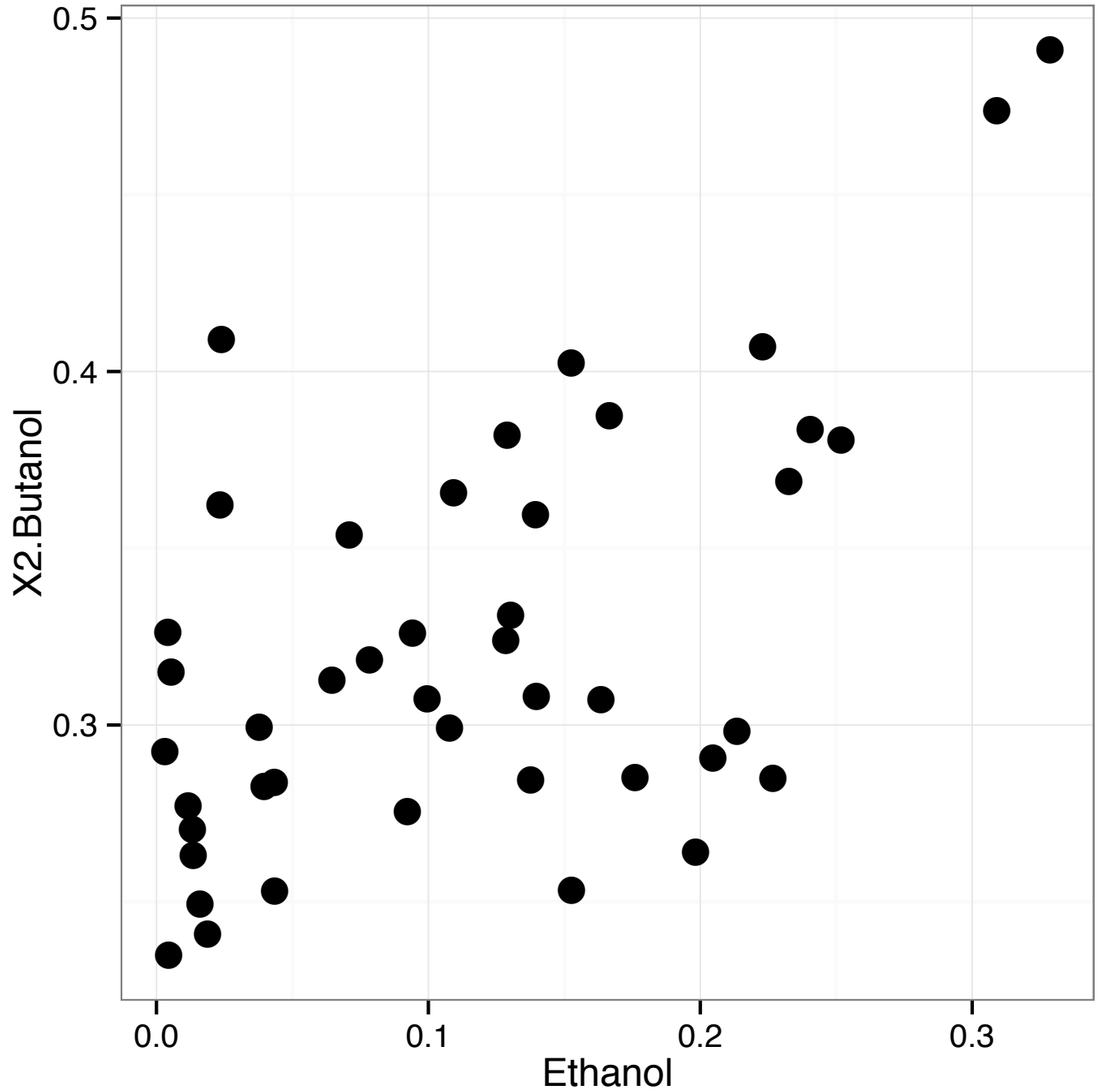






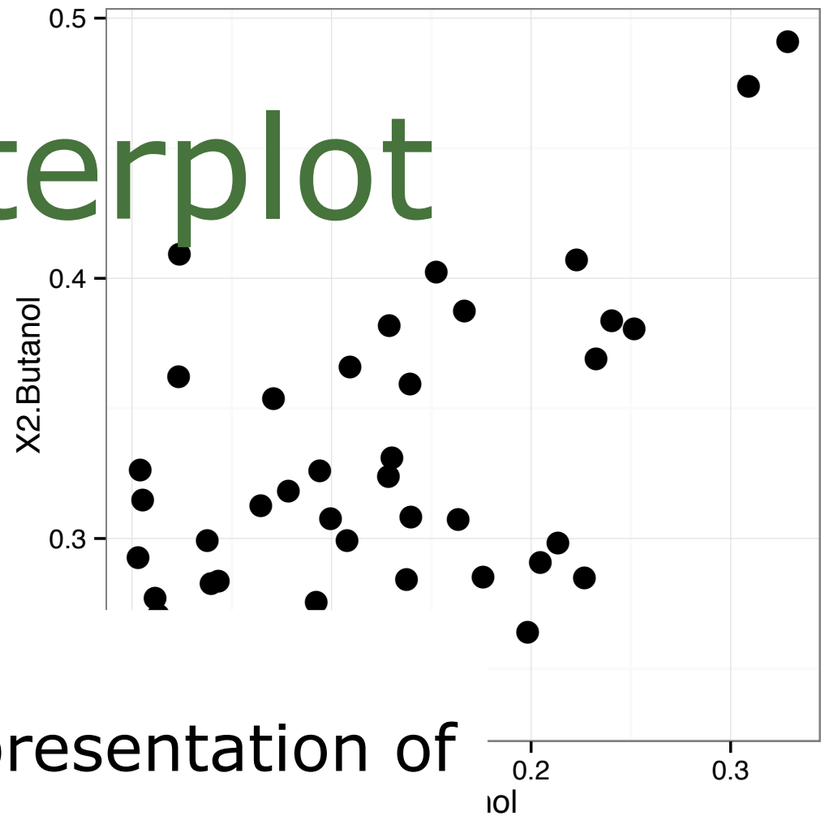
Compositional data
Categorical data

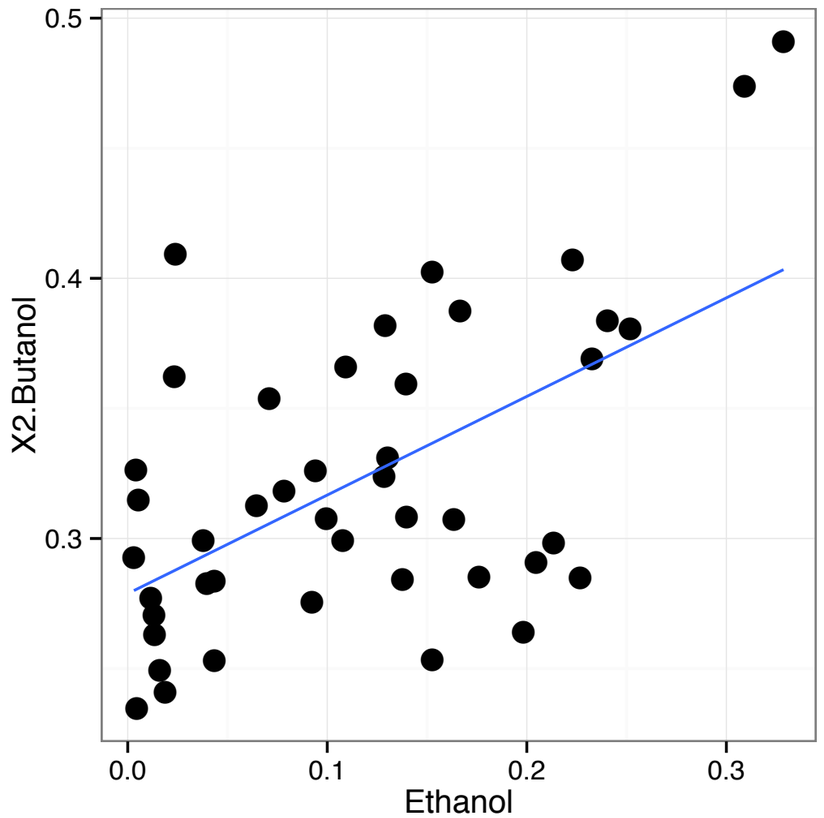
Scatterplot



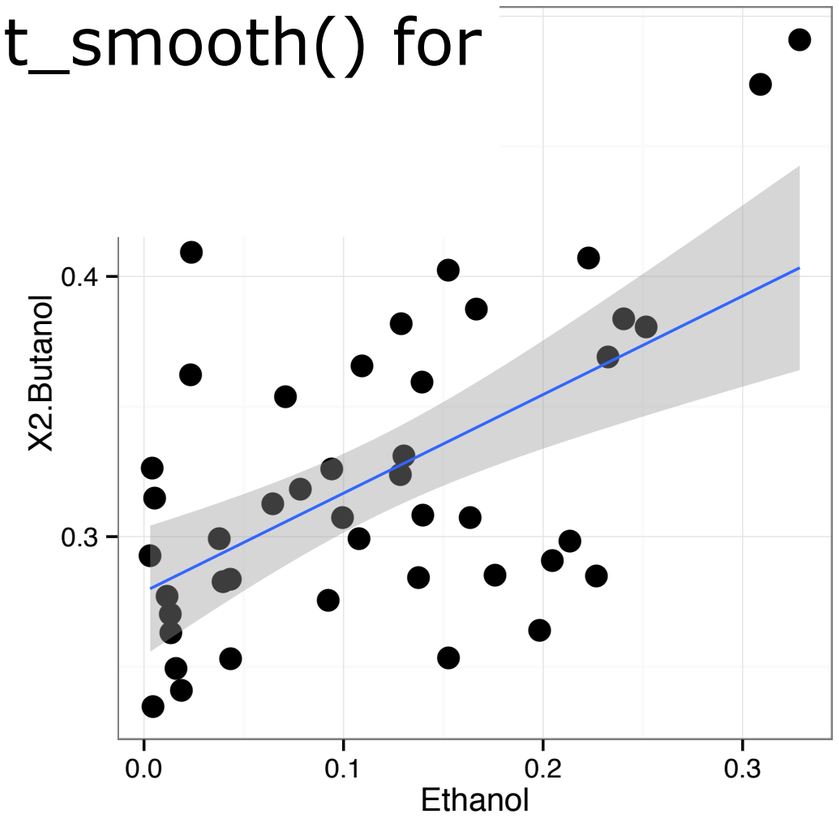
- Two response variables versus each other
- Very usefull for Multivariate data

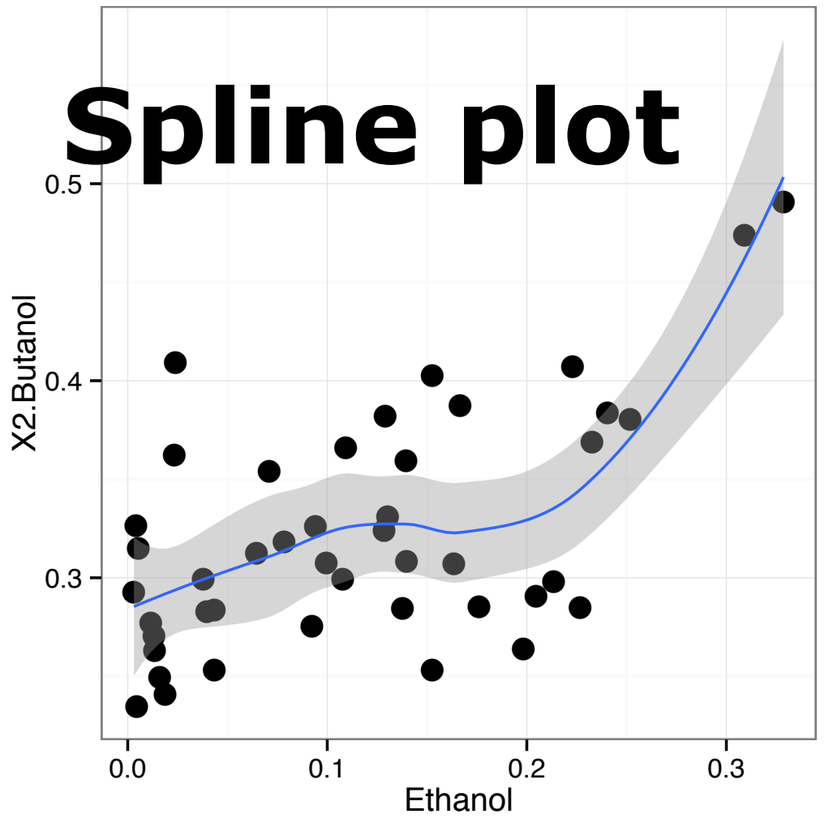
Scatterplot





- Same data!
- Different representation of the relation
- Check: +stat_smooth() for details





t.test Linear models and ANOVA

T-test

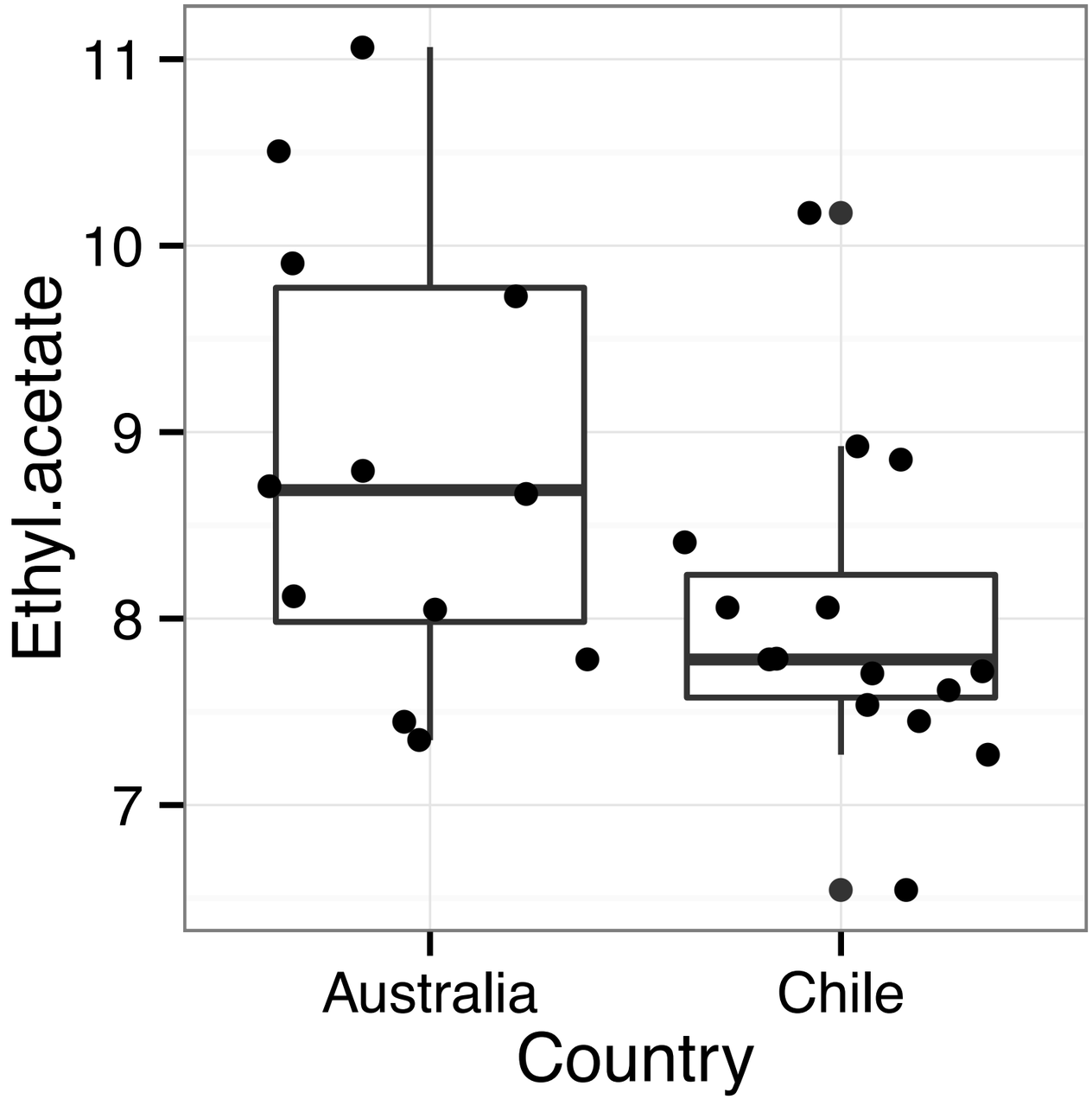
Above ALL

statistical software should be able to do a **t.test**

… to compare the means of two distributions

Function

> t.test()



”Account” of the Variation ANOVA table

Source	$SumSq$	Df	$MeanSq$	F_{obs}	$Pr(F_{df_{Source},df_e} > F_{obs})$
Factor(A)	SS_A	$df_A = k_A - 1$	$MS_A = SS_A/df_A$	$F_A = MS_A/MS_e$	p_A
Factor(B)	SS_B	$df_B = k_B - 1$	$MS_B = SS_B/df_B$	$F_B = MS_B/MS_e$	p_B
Residuals	SS_e	$df_e = n - df_A - df_B - 1$	$MS_e = SS_e/df_e$		
Total	SS_{tot}	$df_{tot} = n - 1$			

k_A and k_B describes the number of levels within factor A and B respectively

Model assumptions

in R
m <- aov()
plot(m)

$$X_i = \alpha(A_i) + \beta(B_i) + \gamma(A_i, B_i) + e_i$$

$$where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent$$

$$for\ j = 1, ..., n$$

We assume that the residuals:

1) Comes from the **same** Normal distribution
2) That they are **independent**

This can be checked via inspection of the residuals.

Fix for **same** distribution: transformation of the respons varible
Fix for **Independence**: *Subject for another day…*

ANOVA- *Cook book*

1. **Visualize data** (w. two factors: connect the points).

2. Consider data **transformation**

3. Specify a model and some relevant hypotheses.

4. Calculate ANOVA and maybe adjust the model (remove terms)

5. Check the model assumptions (plot(aov())) – maybe repeat from pt 2

6. Compare contrasts via **confidence intervals** for differences and maybe a **test** for whether the difference could be =0

7. **Comment on the results**

Useful aov() and lm() functions

```
model<-aov()                    # calculates model

anova(model)                    # display an ANOVA table
        = summary(model)
drop1(model)                    # tests each factor versus
                                the full model
model.tables(model)             # returns estimates or
                                means
coef(model)                     # display the model
                                parameter estimates
confint(model)                  # calculates confidence
                                intervals for the model
                                parameters
plot(m)                         # makes plos for validating
                                the model assumptions
```

PCA

Correlated Multivariate Data

GC-Aroma profiles, Sensorical attribute data, Spectral data,… are all examples of **Multivariate data.**

PCA might be the most powerful tool to initially look at such data, to gain insight on outliers, grouping, structure between variables etc.

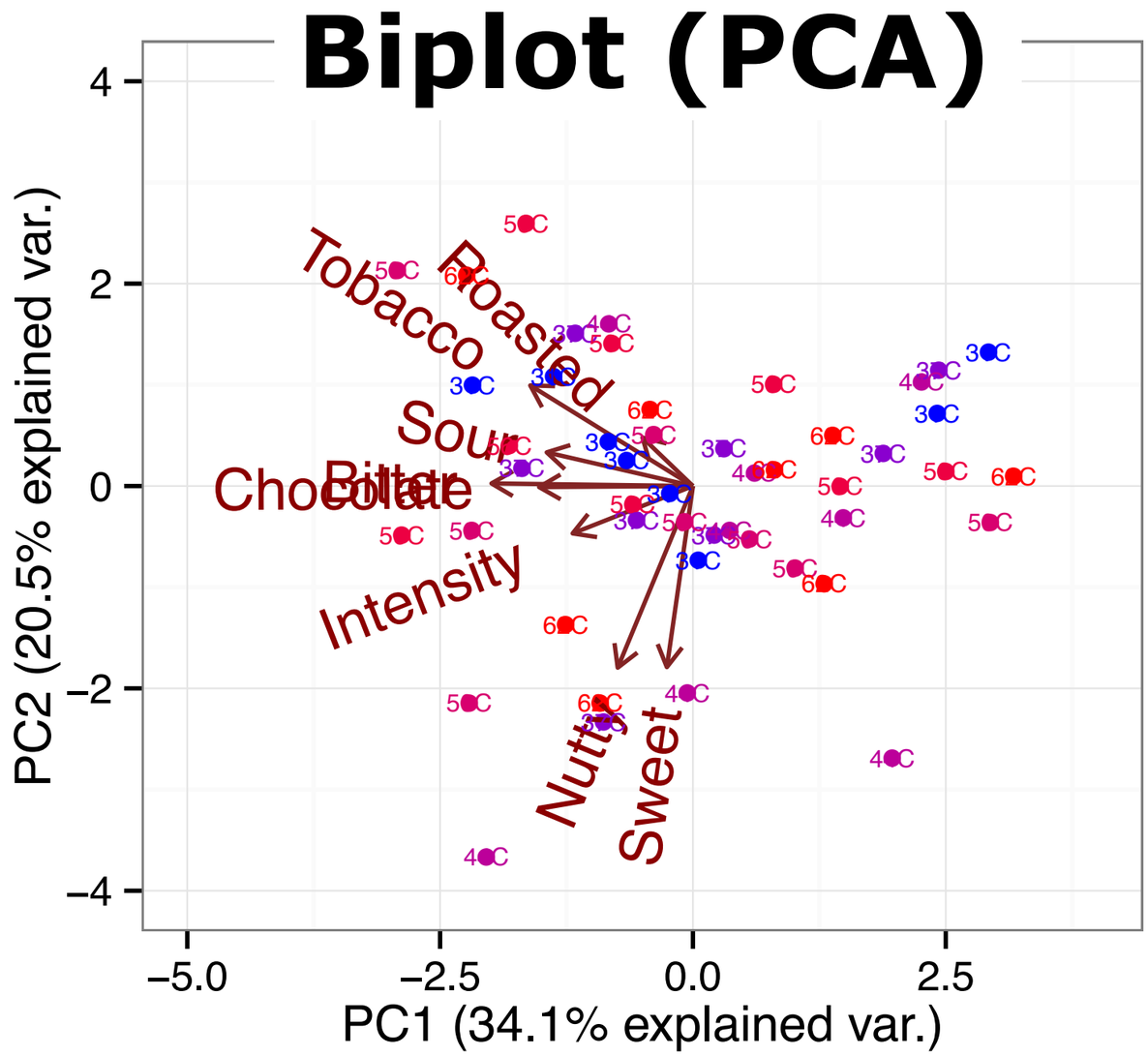
In R it is easily estimated and plottet.

Model

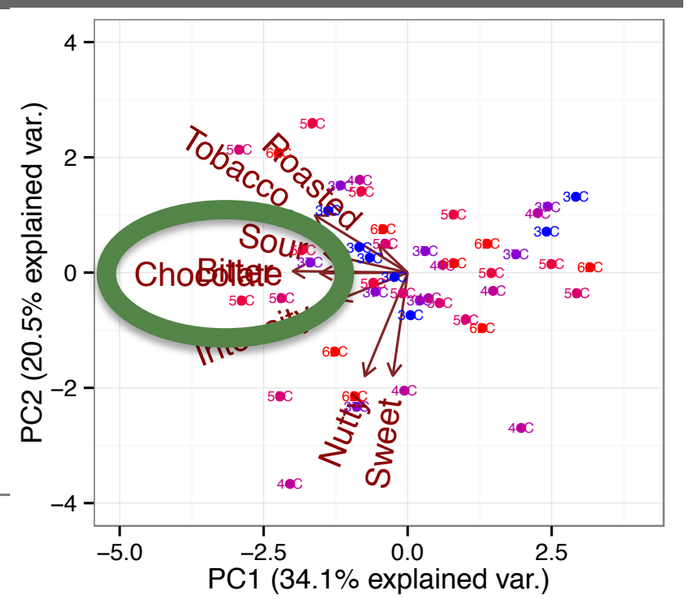
> prcomp() [*remember scaling*]

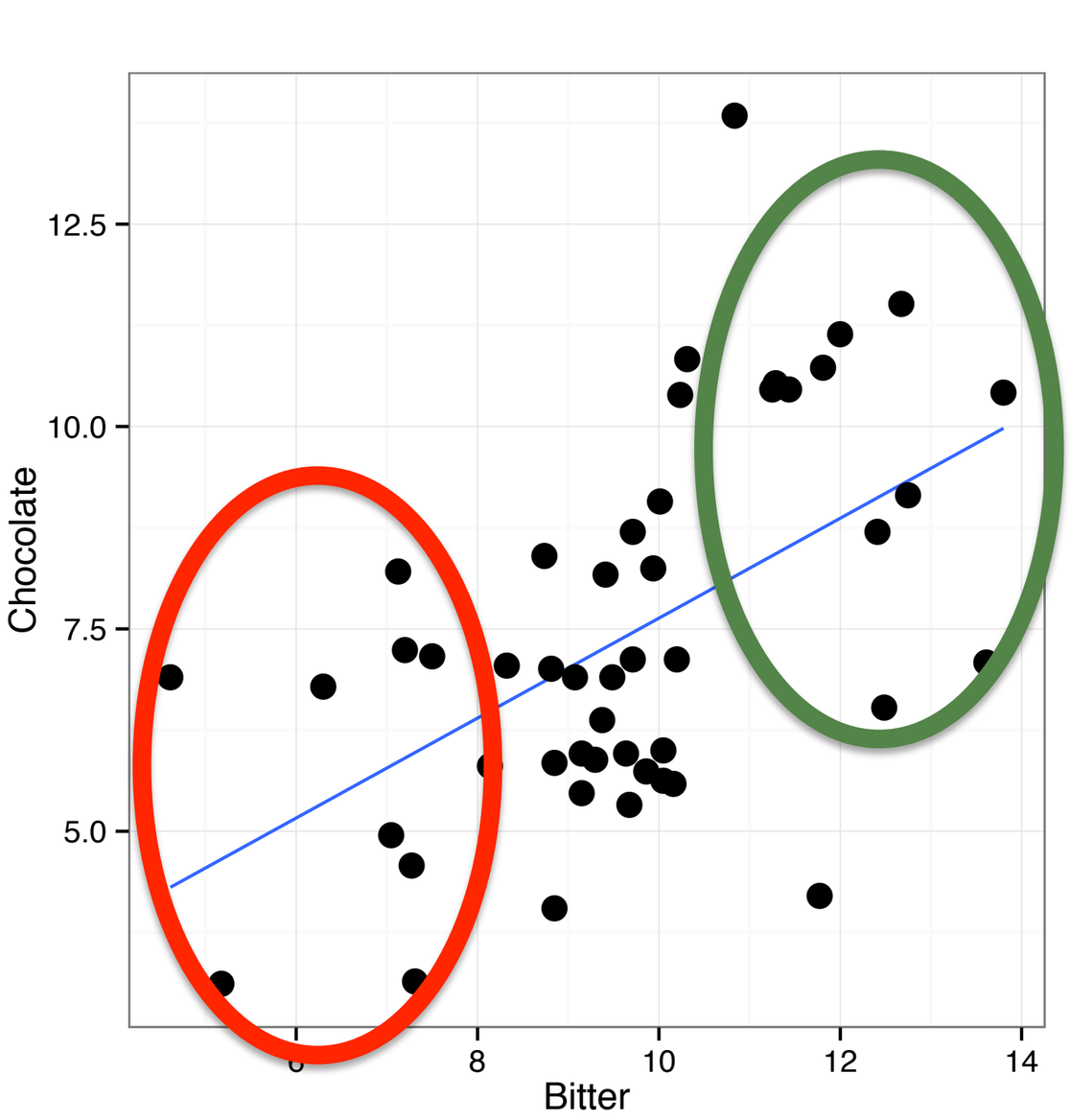
Plot

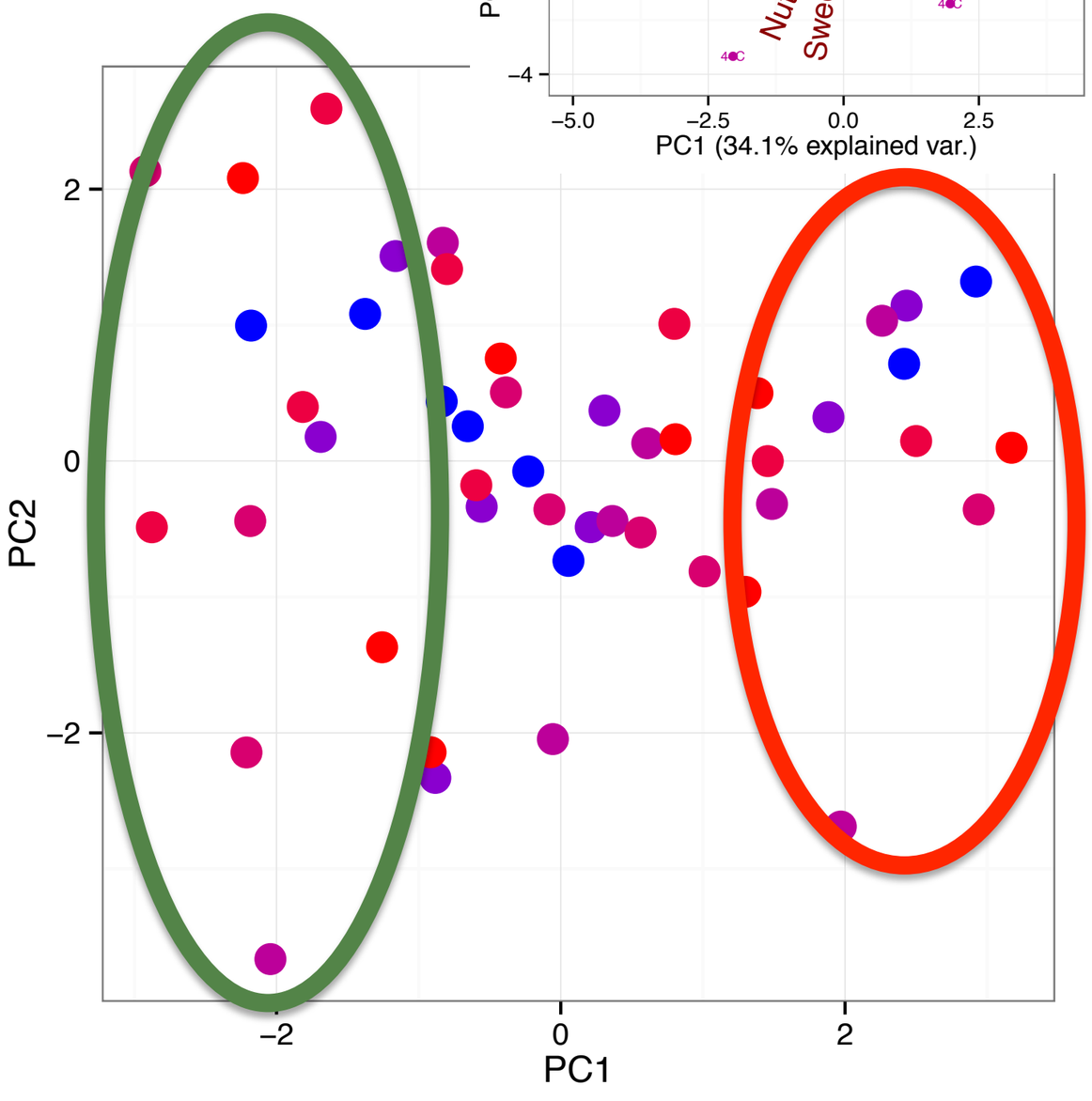
> ggbiplot() [from the ggbiplot package]



Which samples drives the association?







PCA versus Correlation analyses

Number of variable	Correlation Analyses	PCA
1	Doesn't make sense	Doesn't make any sense
2	**Spot on!**	= The exact same
p>2	Number = p(1-p)/2	***Generalized***

Correlation analyses – everything in one analyses