Hands-on data exploration using R

Sebastian Sauer

last update: 2018-11-16

Setup

Overview

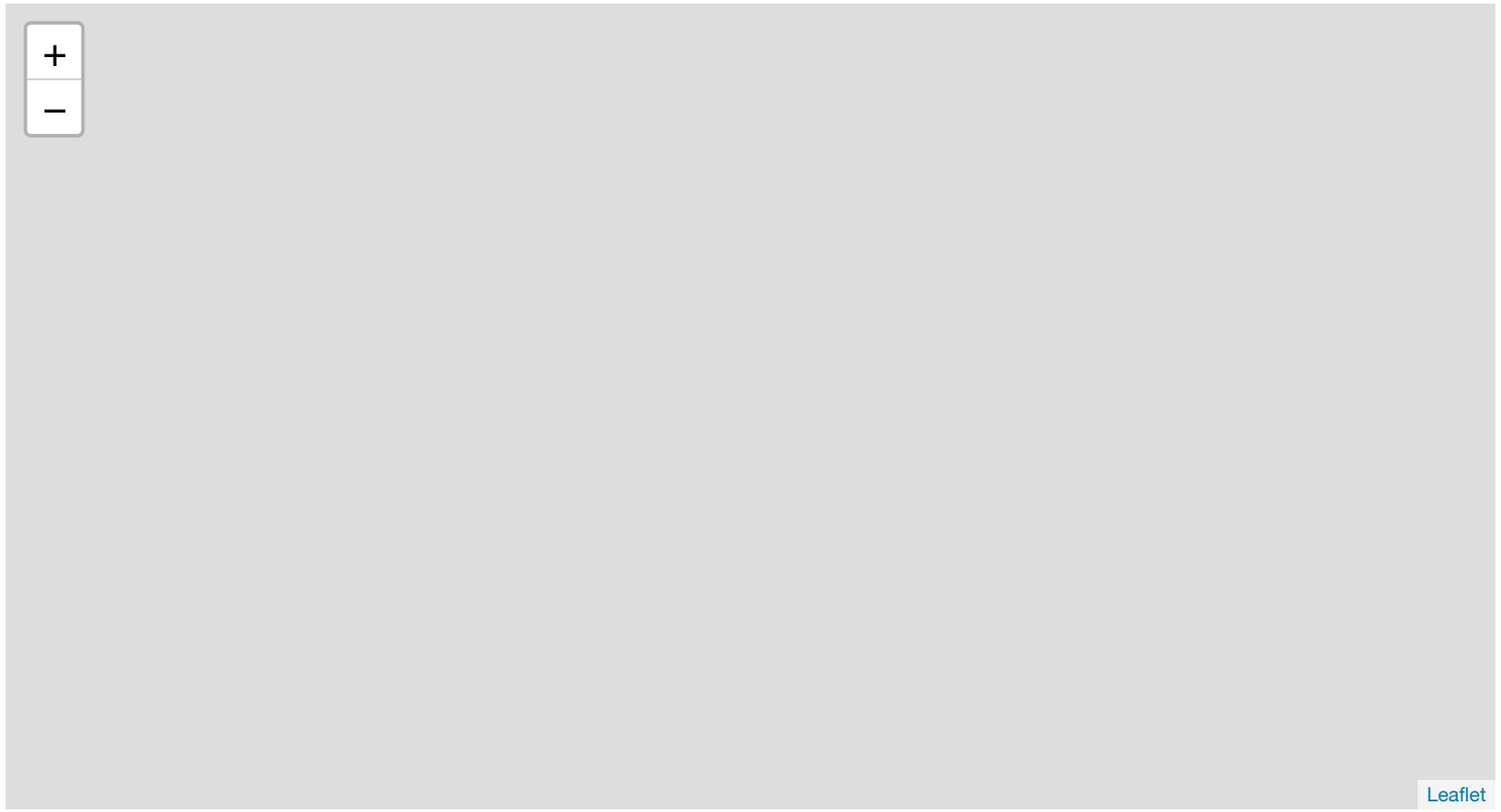
- Setup
- Tidyverse 101
- Data diagrams 101
- Case study

whoami

```
```

- R enthusiast
- Data analyst/scientist
- Professor at FOM Hochschule

The lights are on



Upfront preparation

Please install the following software upfront:

- R
- RStudio Desktop

Starting RStudio will start R automatically.

Please also make sure:

- Your OS is up to date
- You have internet access during the course
- You reach the next power socket (maybe better bring a power cable)

You, after this workshop

Well, kinda off…

Learning goals

- Understanding basic tidyverse goals
- Applying tidyverse tools
- Visualizing data
- Basic modeling

We'll use the following R packages

```
```

Please install each missing package prior to the workshop from within R:

```
```

Load each package after each start of Rstudio

```
```

Tip: Use to see loaded packages and for installed packages.

Simpler: Check the "packages pane" in RStudio.

Data we'll use:

- is a toy dataset built into R (no need for installing).
- Data come from 1974 motor sports magazine describing some automotive.
- Columns: e.g., horsepower, weight, fuel consumption

Load the dataset:

Get help:

Data we'll use:

- is a dataset from R package (package must be installed).
- Data come from flights leaving the NYC airports in 2013.
- Columns: e.g.., delay, air time, carrier name

Load the dataset:

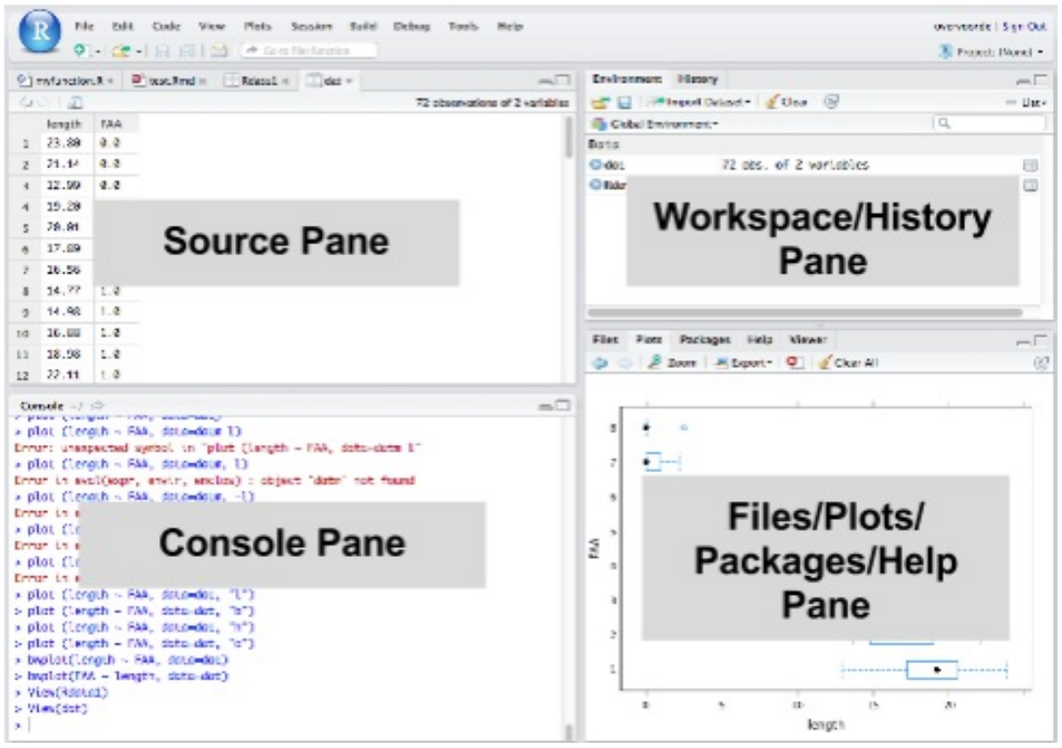
```
```

Get help:

```
```

Load the data each time you open RStudio (during this workshop).

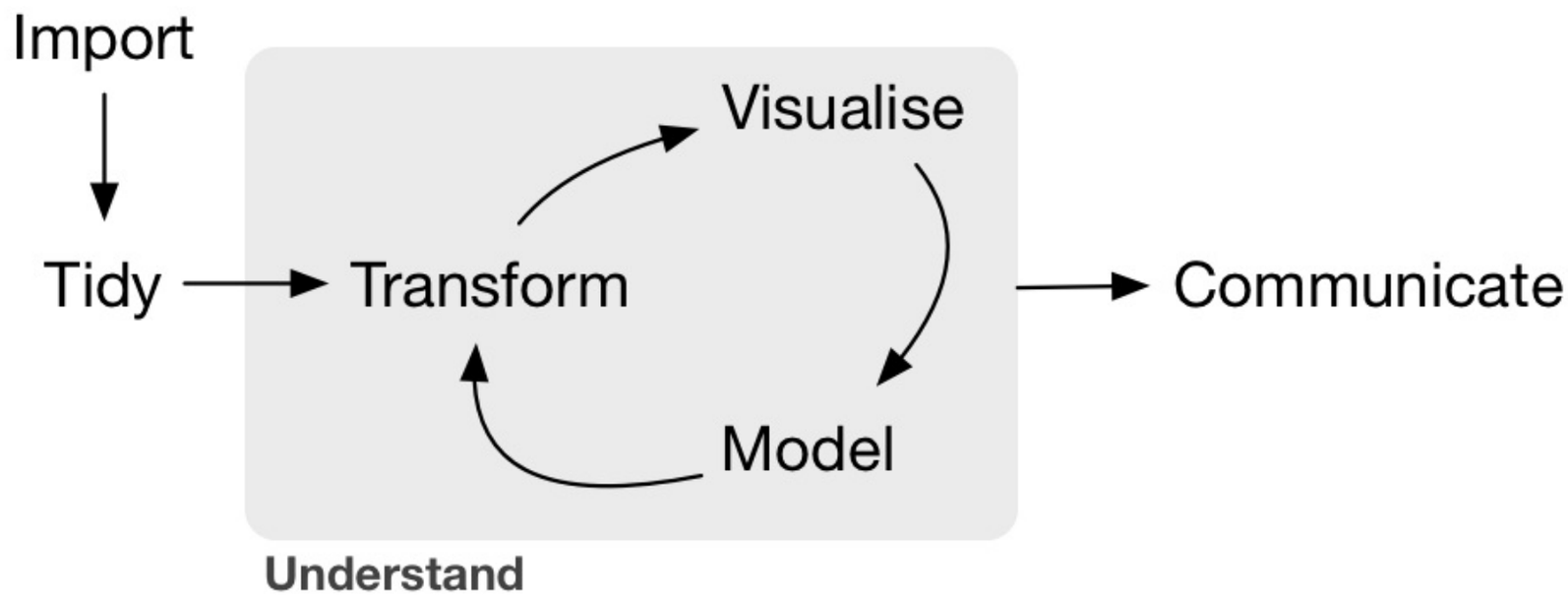
RStudio running



The tidyverse


broom
dplyr
reshape2
%>%
Leci n'est pas un pipe.
test that
purrr
rmarkdown
R Studio
knitr
Shiny
packrat
lubridate
ggplot2
haven
stringr
tidyr
readr

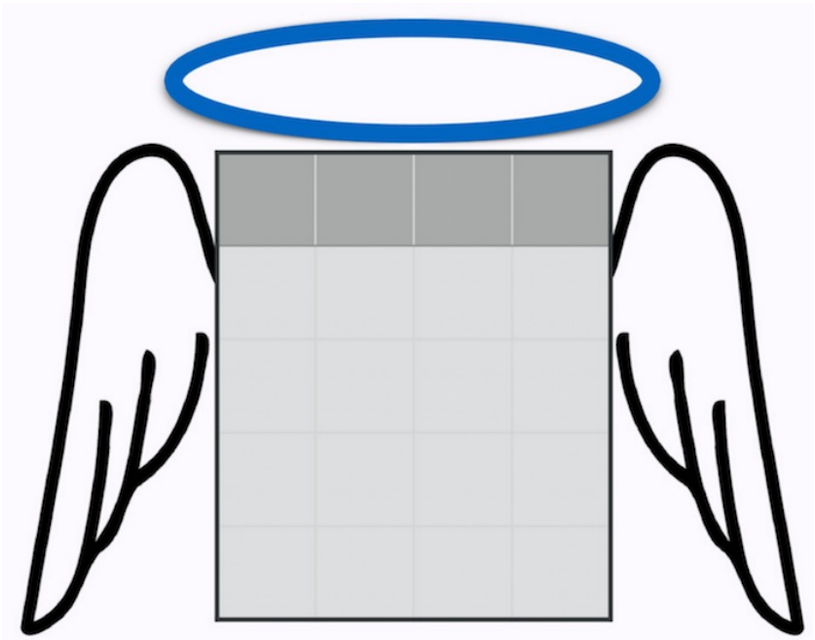
The data analysis (science) pipeline



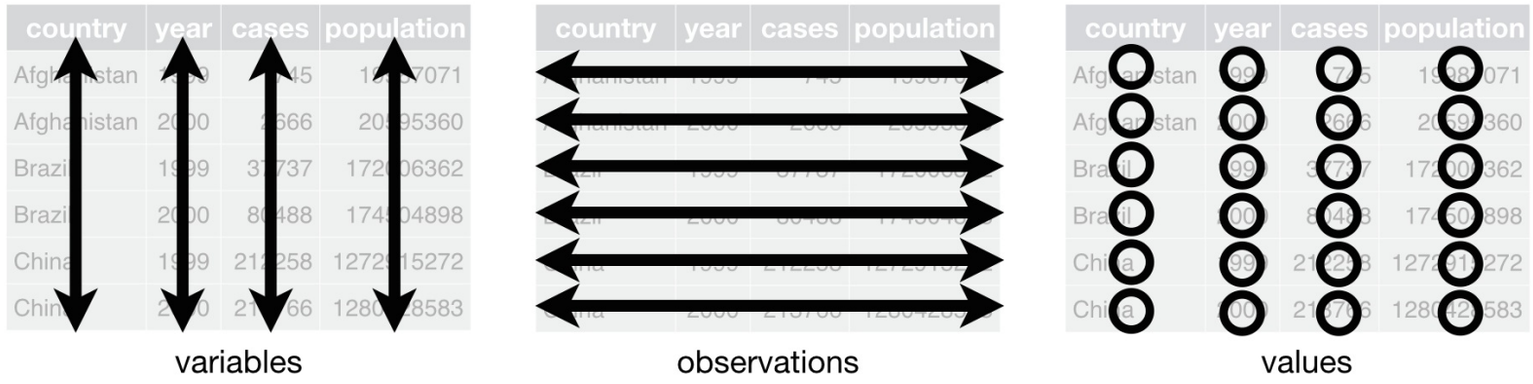
Get the power of the ~~uni~~ tidyverse

But I love the old way ...

Nice data



Tidy data



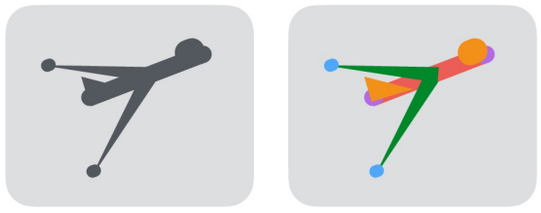
More Details

Dataset

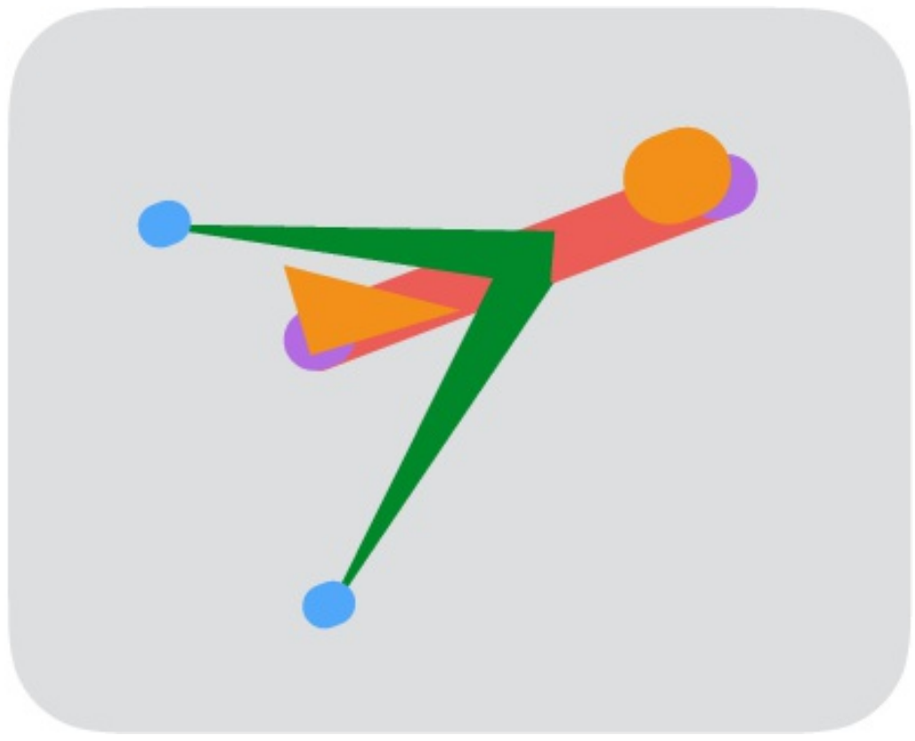
Data wrangling

Two tidyverse principles

Knock-down principle



Pipe princriple

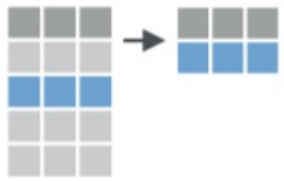


Atoms of the knock-down principle

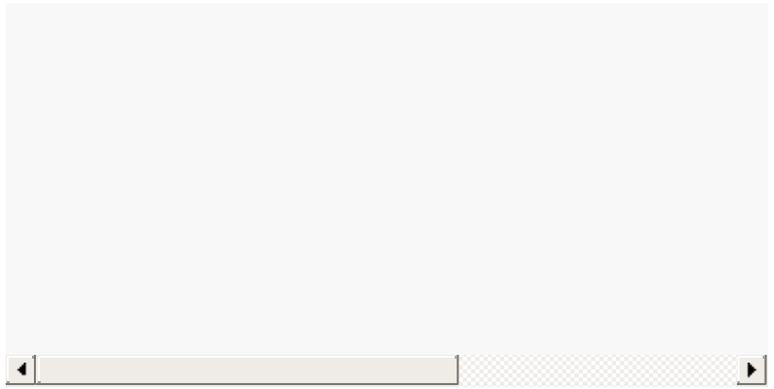
-
-
-
-
- ...

Filtering rows with

Extract rows that meet logical criteria.



Filter table such that only rows remain where equal



- exercises

Filter the automatic cars.

Filter the automatic cars with more than 4 cylinders.

Filter cars with either low consumption or the the super. thirsty ones

- solutions to exercises

Select columns with

Extract columns by name.

Select the columns and . Discard the rest.

- exercises

Select the first three columns.

Select the first and third column.

Select all columns containing the letter "c".

- solutions to exercises

Add or change a column with

Apply vectorized functions to columns to create new columns.

Define weight in kg for each car.

- exercises

Compute a variable for consumption (gallons per 100 miles).

Compute two variables in one mutate-call.

- solutions to exercises

Summarise a column with

Apply function to summarise column to single value.

Summarise the to their mean.

- exercises

Compute the median of consumption.

Compute multiple statistics at once.

- solution to exercises

Group with

Create "grouped" copy of table. dplyr functions will manipulate each group separately and then combine the results.

Group cars by (automatic vs. manual). Then summarise to mean in each group.

- exercises

Compute the median consumption, grouped by cylinder.

Compute the median consumption, grouped by cylinder and .

- exercises

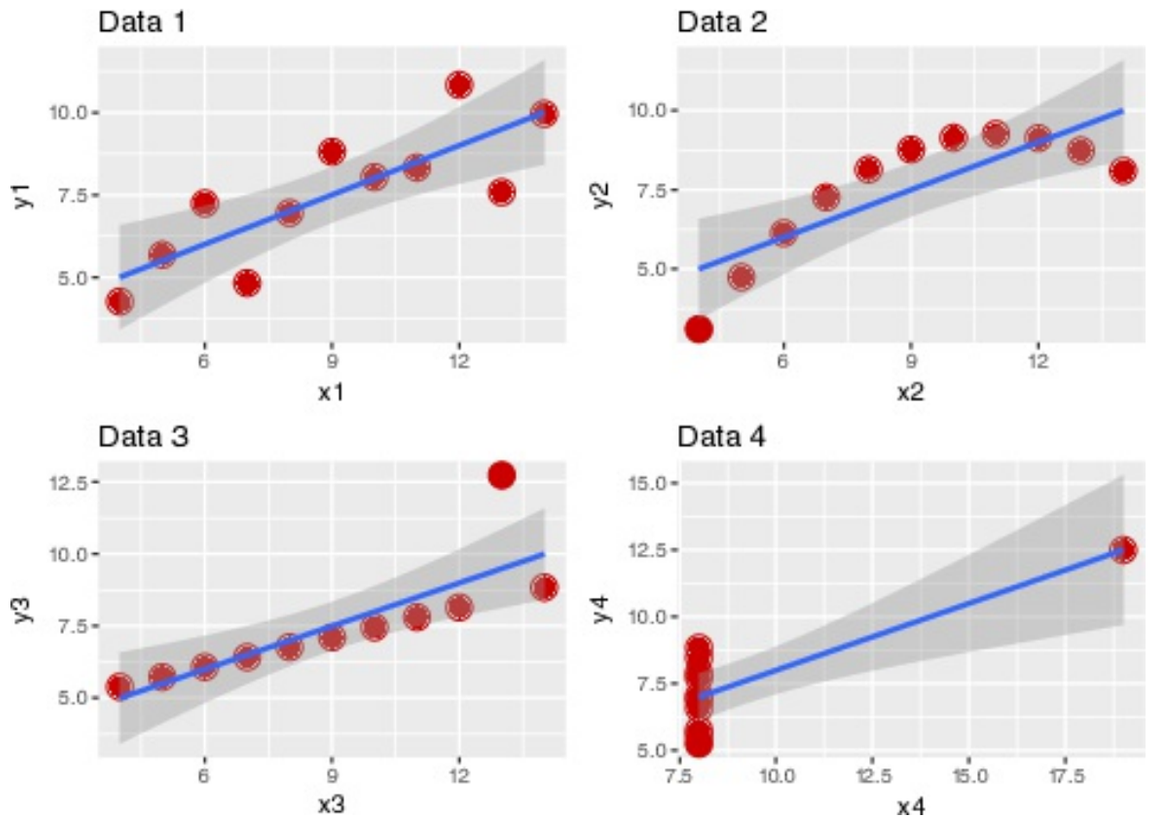
Enter the pipe

Life without the pipe operator

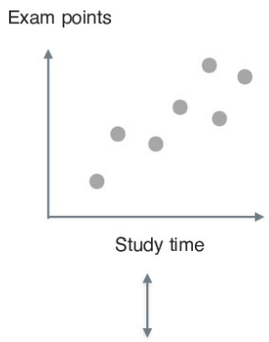
Life with the pipe operator

Data diagrams

Why we need diagrams

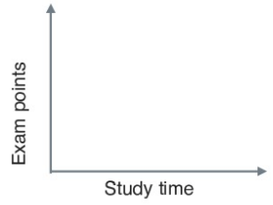


Anatomy of a diagram



	Study time	Exam points
Anna	10	30
Berta	30	60
Carla	20	90
Dora	50	120

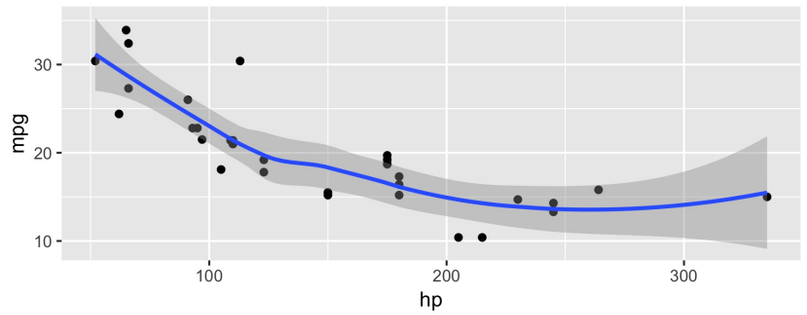
Data



Axes

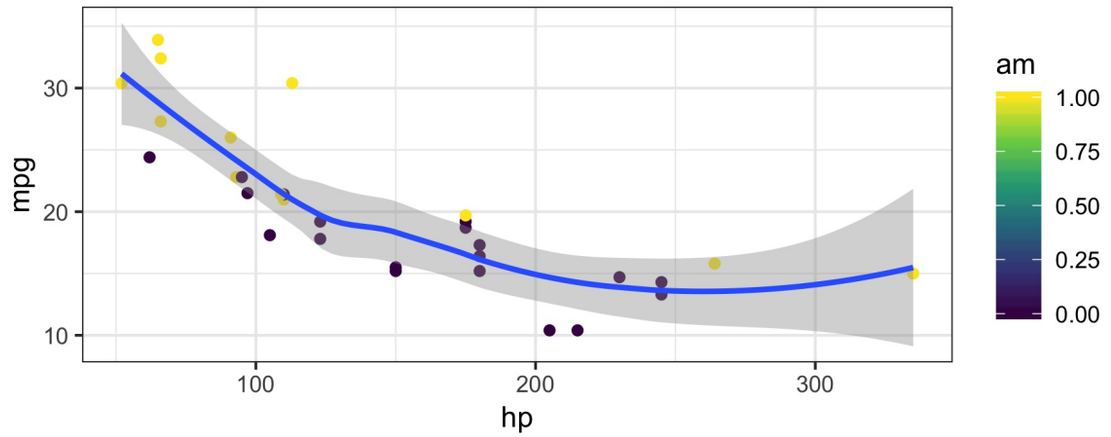
Geom

First plot with



Notice the in contrast to the pipe .

Groups and colors

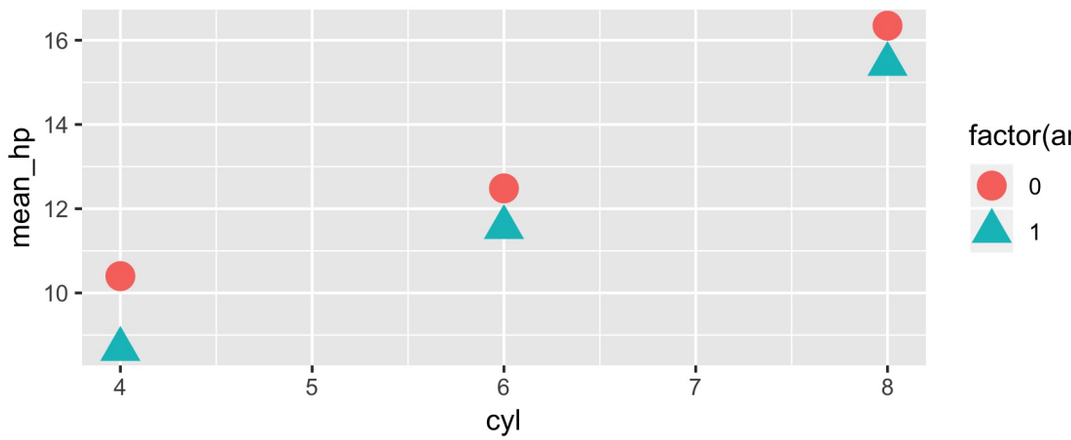


Diagrams - exercises

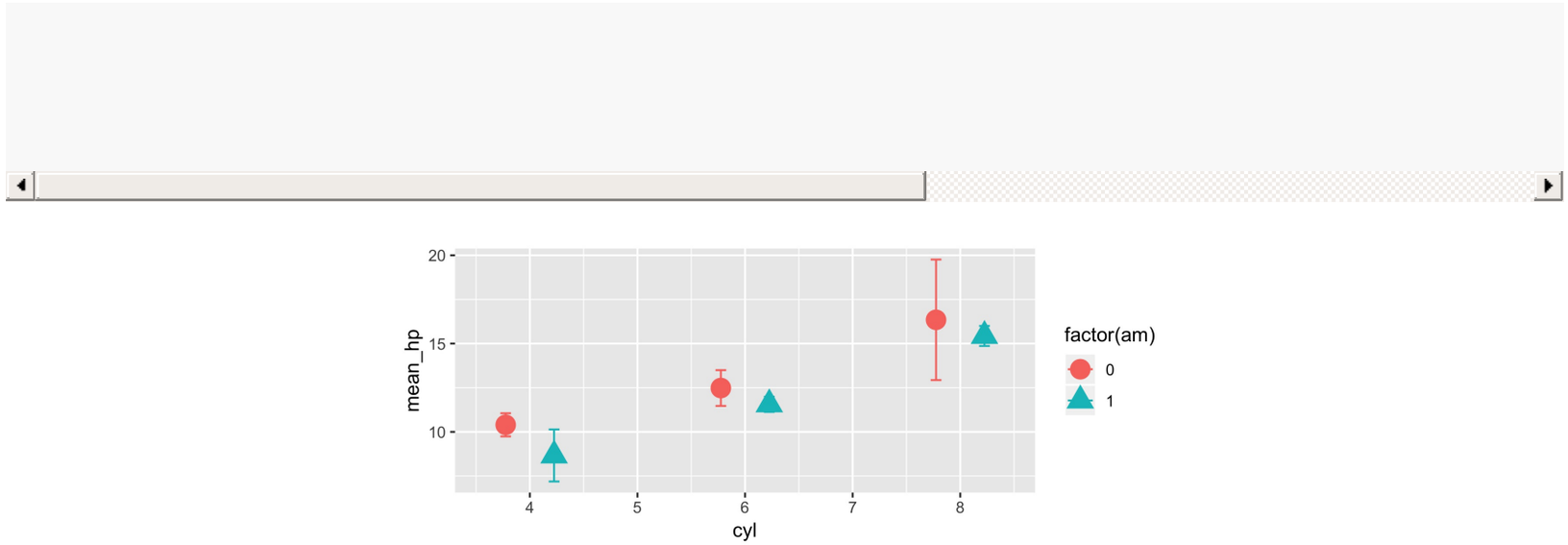
Plot the mean and the median for each cylinder group (dataset).

Now add a visualization for the variation in the data.

Diagrams - solutions to exercises 1



Diagrams - solutions to exercises 2



Case study Why are flights delayed?

Know thy data

Don't forget to load it from the package via:

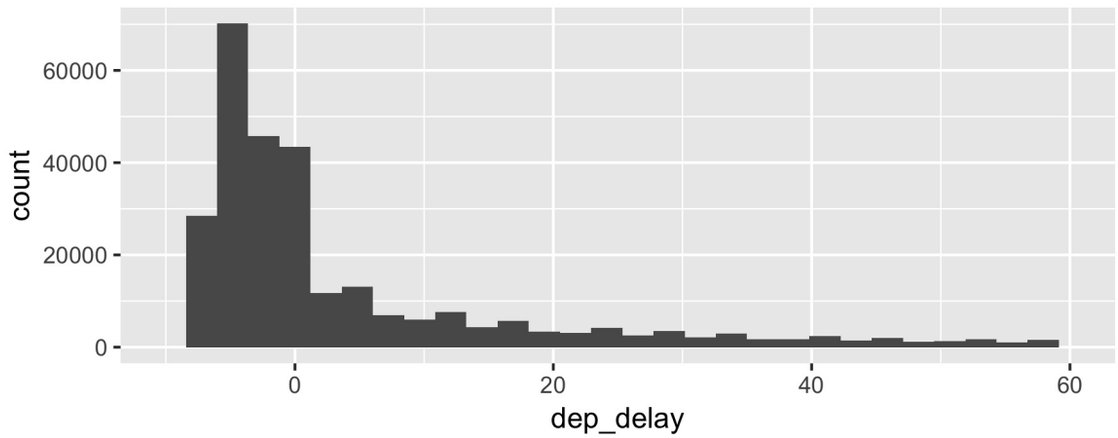
A look to the help page:

Glimpse data

Data sanity - quantitative variables

Data sanity - qualitative variables

Distribution - quantitative variables



Note the long right tail ("anomaly")

Deal with missing data - omit

Deal with missing data - replace by mean

Deal with missing data - tidy approach

Use to disambiguate.

Descriptive statistics for delay

Descriptive statistics by origin

Start modeling

Delay as a function of origin?

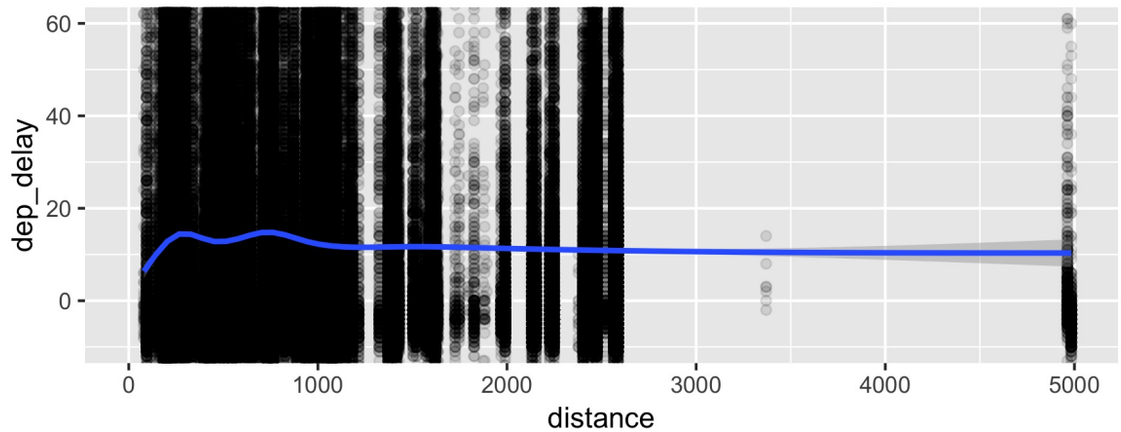
More Rish:

Linear models

Some as above, stated differently.

Does predicts ?

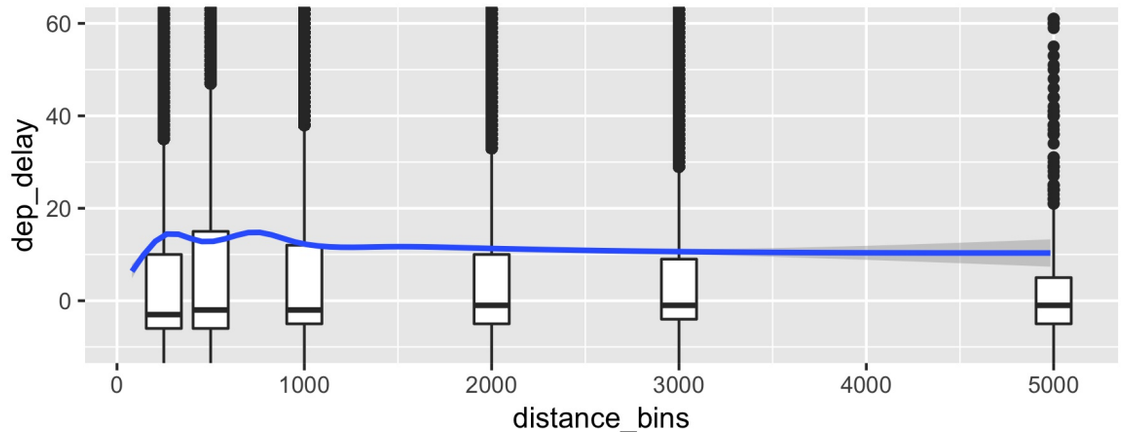
Does predicts ?



Alternative visualization (binned data, code)

Use for binning and recoding of data values.

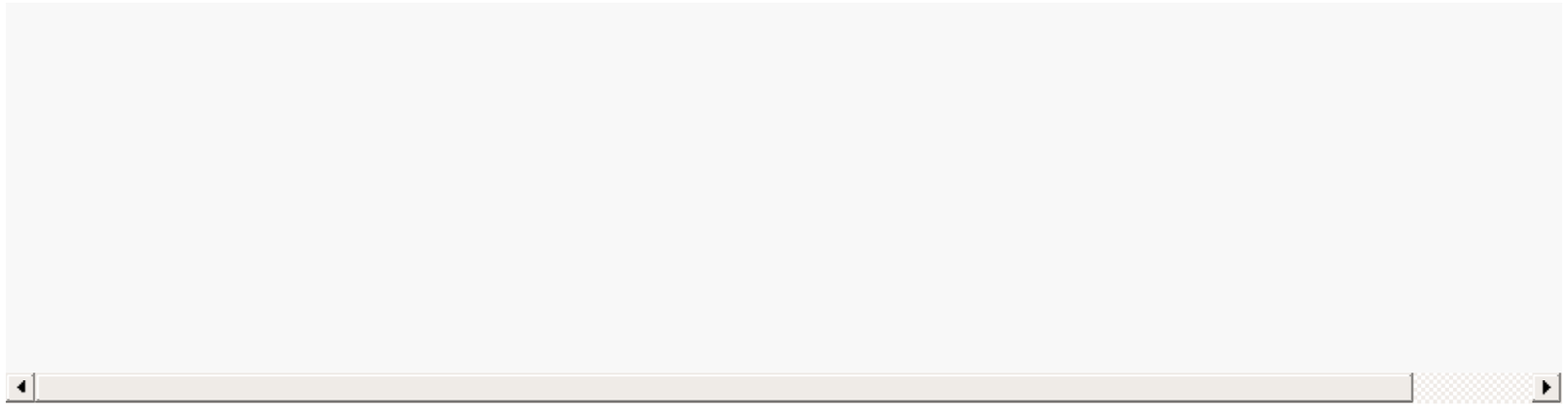
Alternative visualization (binned data)



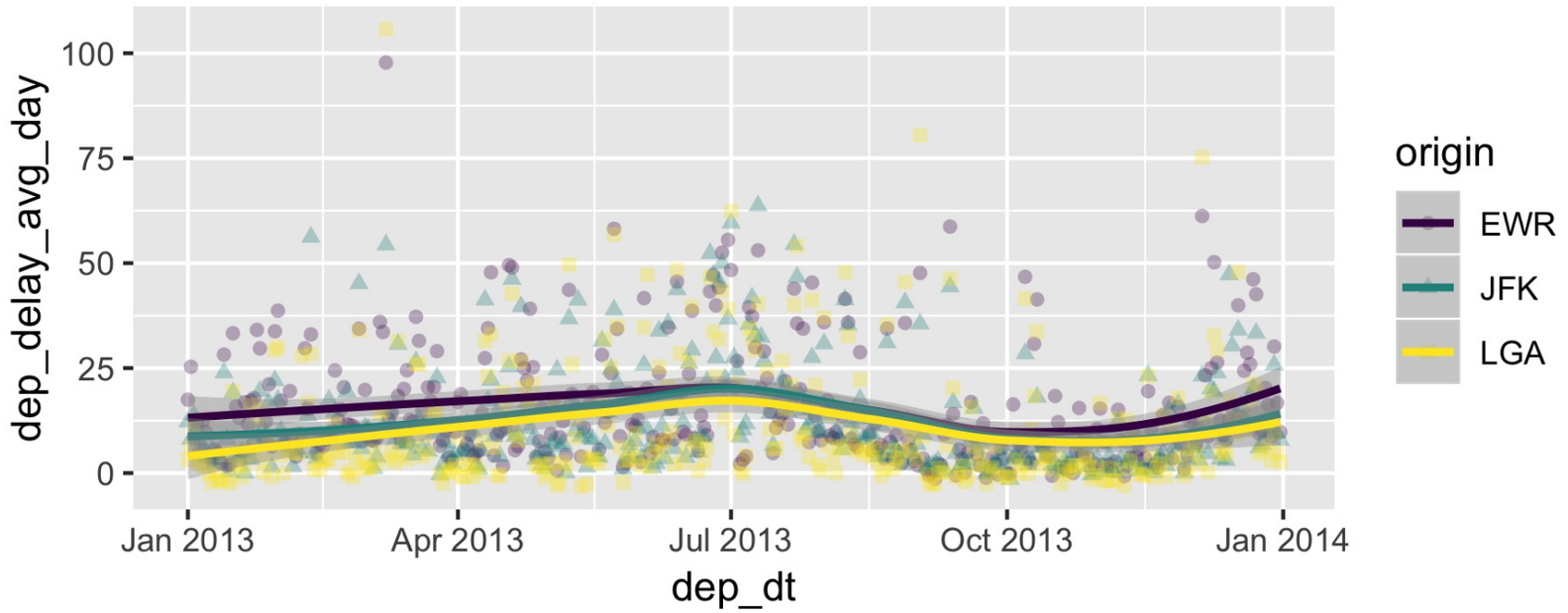
Correlation of distance and delay

Delay as a function of distance

Delay per month (code)

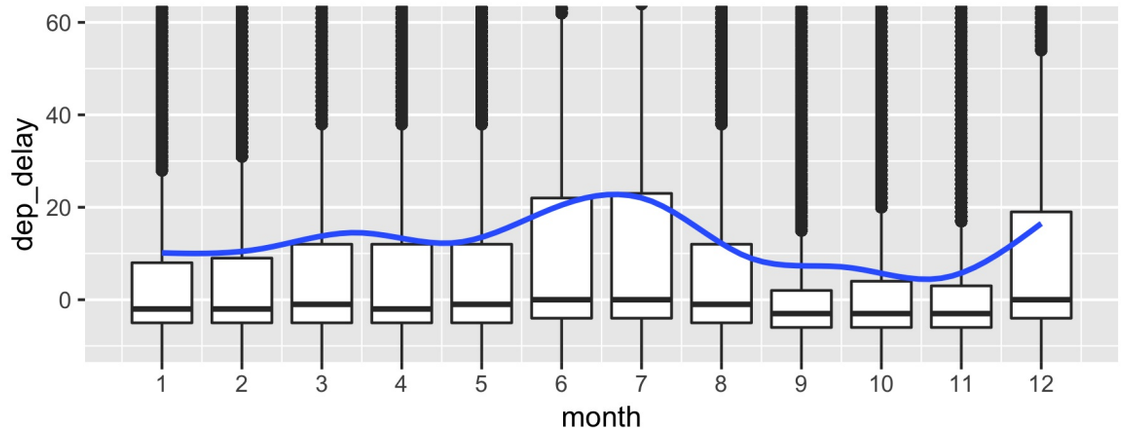


Delay per month (output)



Delay per month - boxplot (code)

Delay per month - boxplot (output)



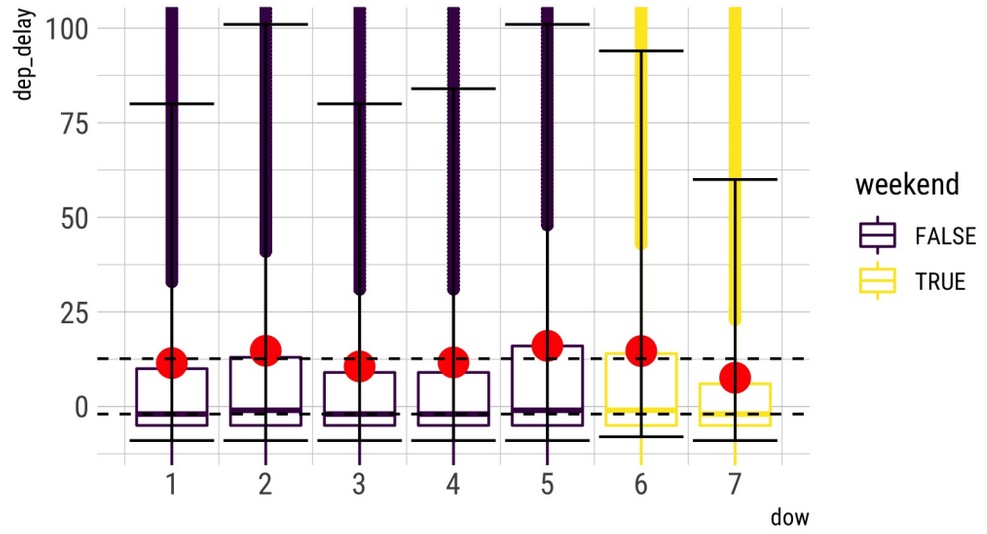
Is it the weekends? (code)

Is it the weekends? (data processed)

dow	delay_m	delay_md	q_05	q_95
1	11.477476	-2	-9	80
2	14.718728	-1	-9	101
3	10.588355	-2	-9	80
4	11.643321	-2	-9	84
5	16.043451	-1	-9	101
6	14.653974	-1	-8	94
7	7.594406	-2	-9	60

Is it the weekends? (code)

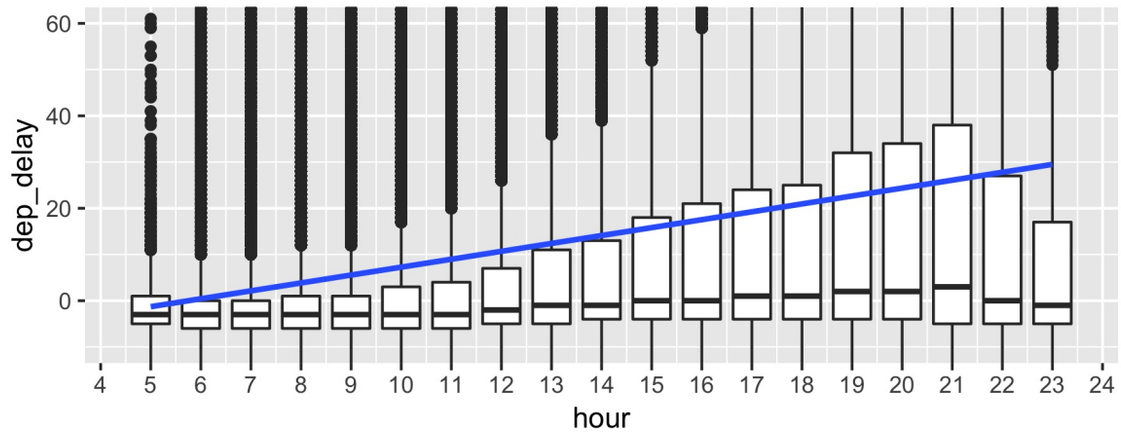
Is it the weekends? (output)



Delay per time of the day (code)

Let's check whether delays add up during the day, a popular opinion among travellers.

Delay per time of the day (output)



Delay as function of month, hour, origin, and weekday

Geoplotting

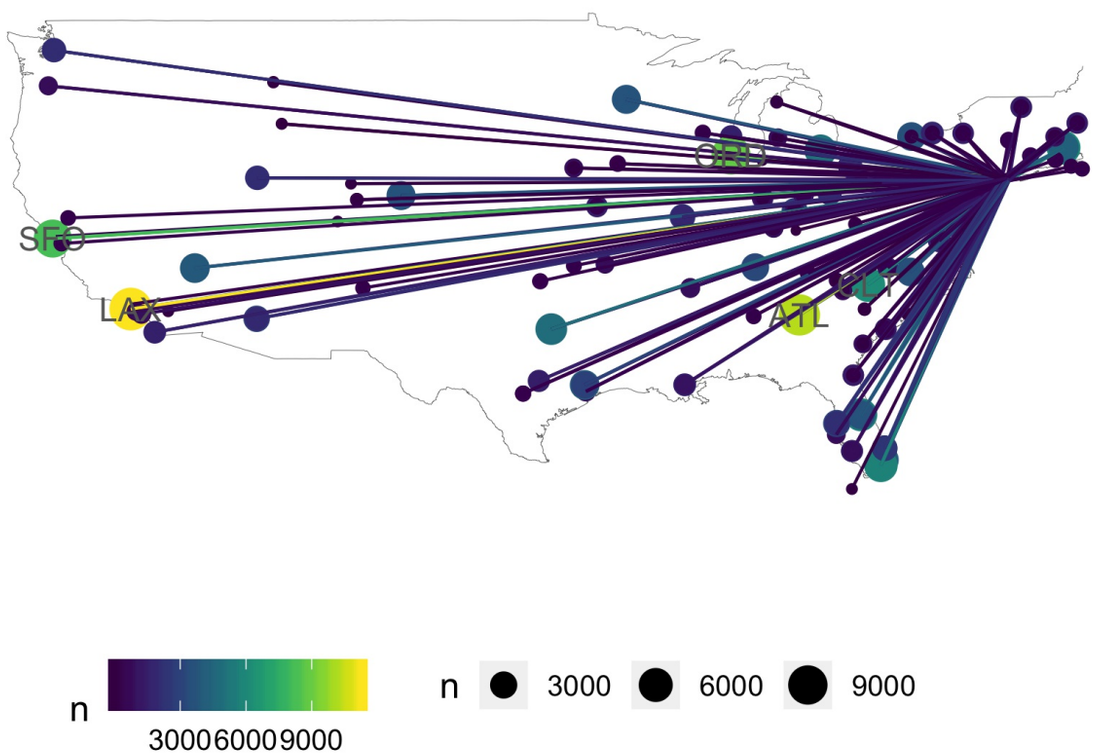
Join airport data

Dataframe for plotting (code)

Dataframe for plotting (output)

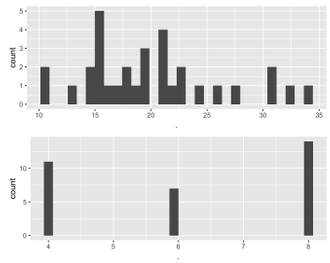
Geo plot flights (code)

Geo plot flights (output)



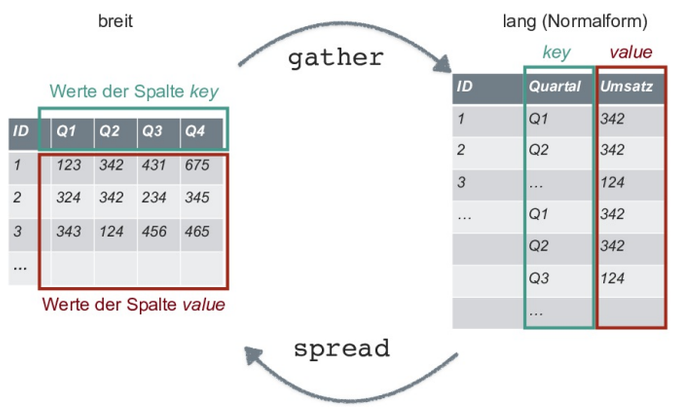
More advanced stuff

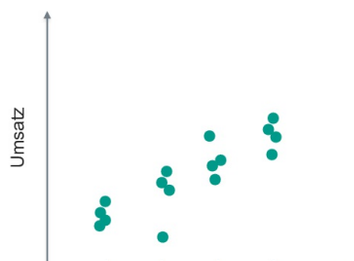
Map columns to function with



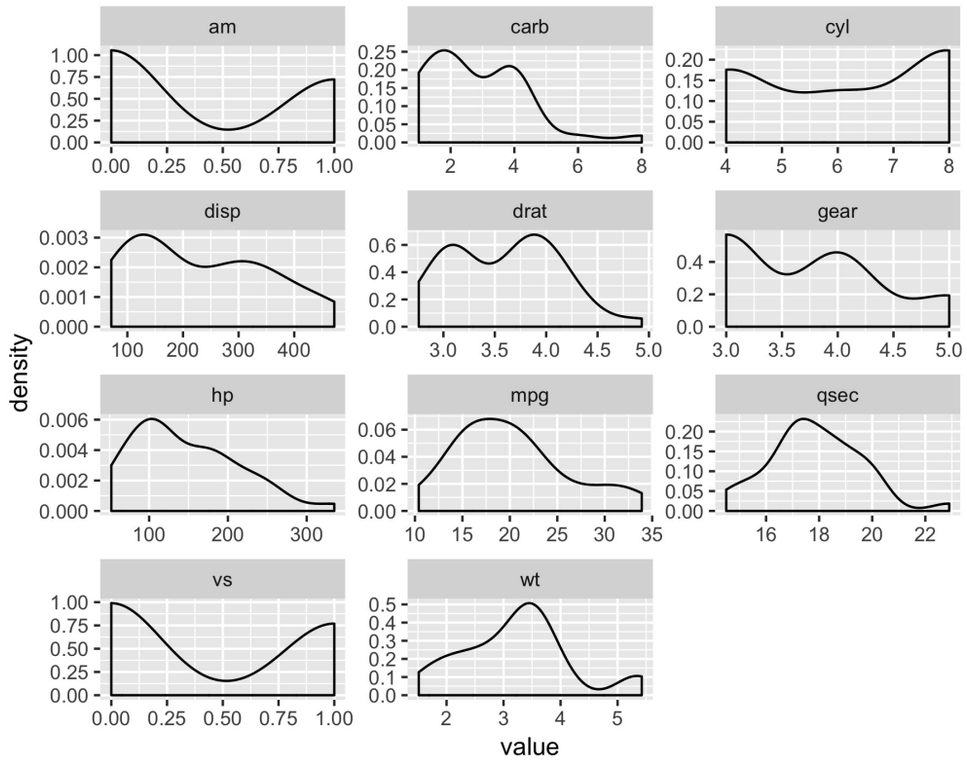
Map TWO columns to function with

Reshape (transform) dataframe

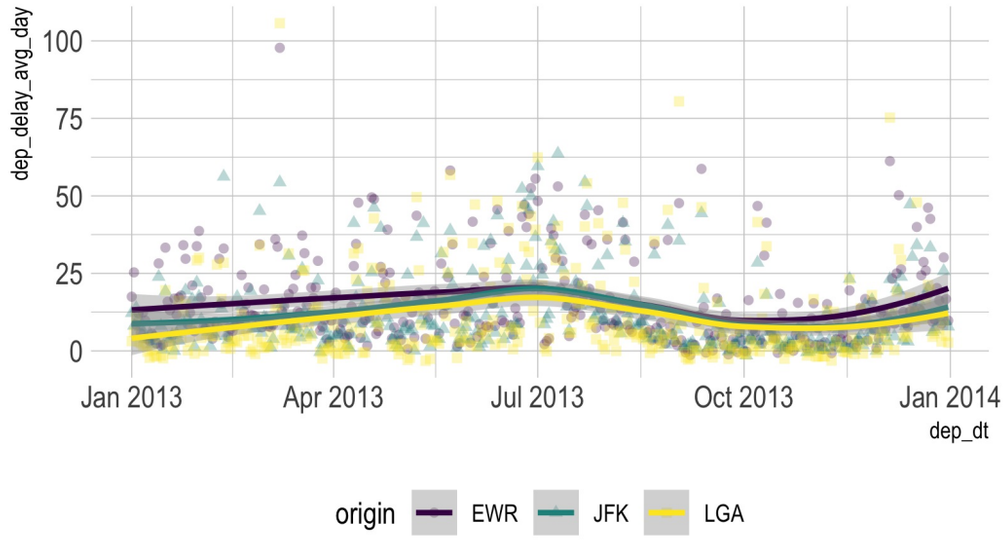




Transform dataframe for plotting

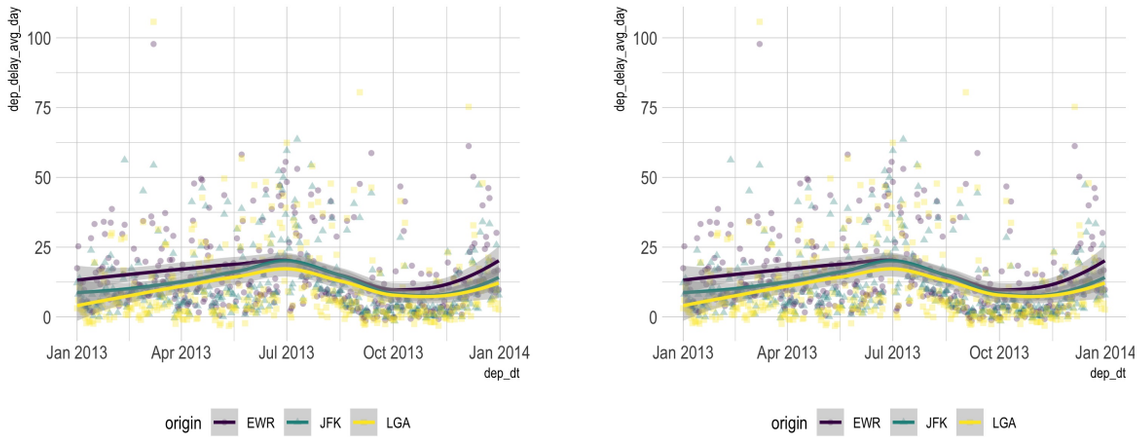


niceties: Themes



May may need to install fonts upfront; see .

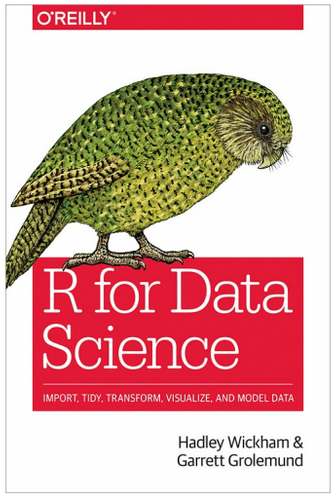
ggplot niceties: Combining plots



Resources

Modern Dive -- An Introduction to Statistical and Data Sciences via R Chester Ismay and Albert Y. Kim

R for Data Science

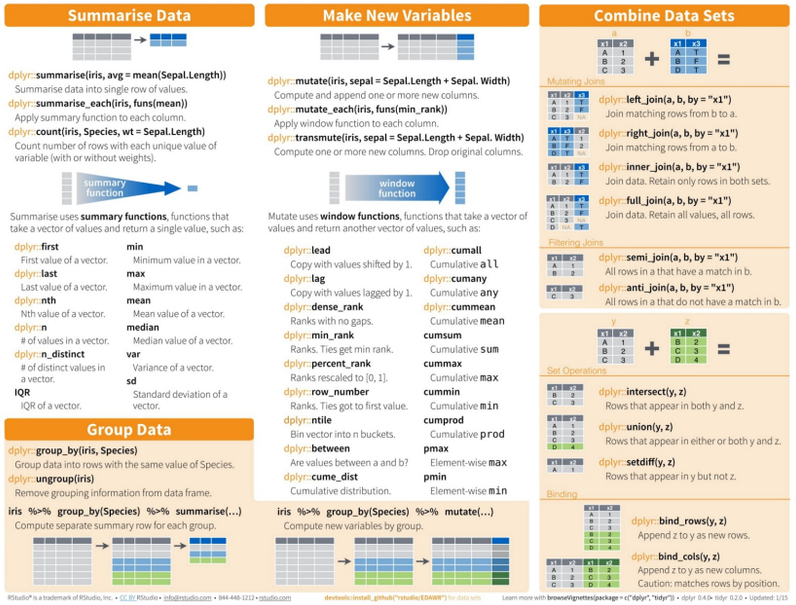


R for Data Science

Moderne Datenanalyse mit R

Moderne Datenanalyse mit R

Cheatsheets



https://www.rstudio.com/resources/cheatsheets/

Disclaimer: There may be issues at times

StackOverflow is your friend

Wrap-up

That was quick, but it was a start

Thank you

Sebastian Sauer

sebastiansauer

https://data-se.netlify.com/

sebastian.sauer@fom.de

Sebastian Sauer

Get slides here

: Get PDF of slides here

: Get Rmd source code of slides here

Credit to

Built using R, RMarkdown, Xaringan. Thanks to the R community and the tidyverse developers.

Thanks to Yihui Xie and Antoine Bichat, among others, for Xaringan inspiration.

Thanks to FOM Hochschule for supporting me.

Images:

- Data Transformation with dplyr Cheat Sheet, by RStudio
- Modern Dive, Chester Ismay and Albert Y. Kim
- Kermit typping
- RStudio running
- tidyverse
- Process of data analysis
- Yoda
- kid waves
- Nice data
- tidy data
- Magrittr pipe
- Jump car
- Overhead locker

Icons from FontAwesome

SessionInfo

See for package version (same folder as this presentation).

This document is made reproducible using checkpoint with day set to 2018-09-30.