PROFINIT

Effective Data Science Reporting

How to communicate your DS results

Dominik Matula

2023

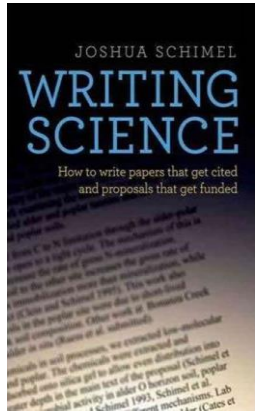
Data Science Reporting

Why to bother?

PROFINIT

As a scientist, you are a professional writer.

Joshua Schimel, Writing Science



Outline

PROFINIT

A. Theory

1. Report types
2. Report structure
3. Communicating data
4. Report as a code
5. Tips

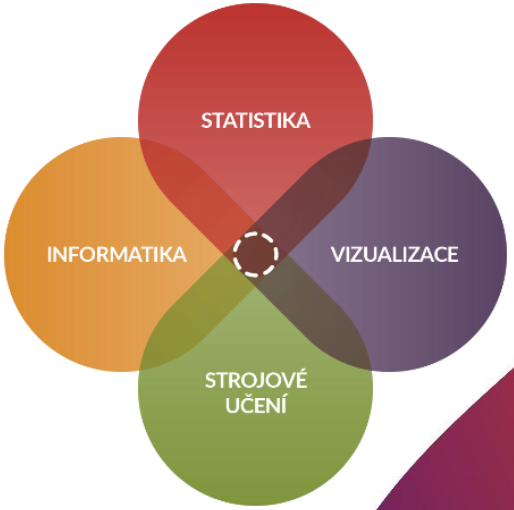
B. Examples & Discussion

Before we start

Data science report

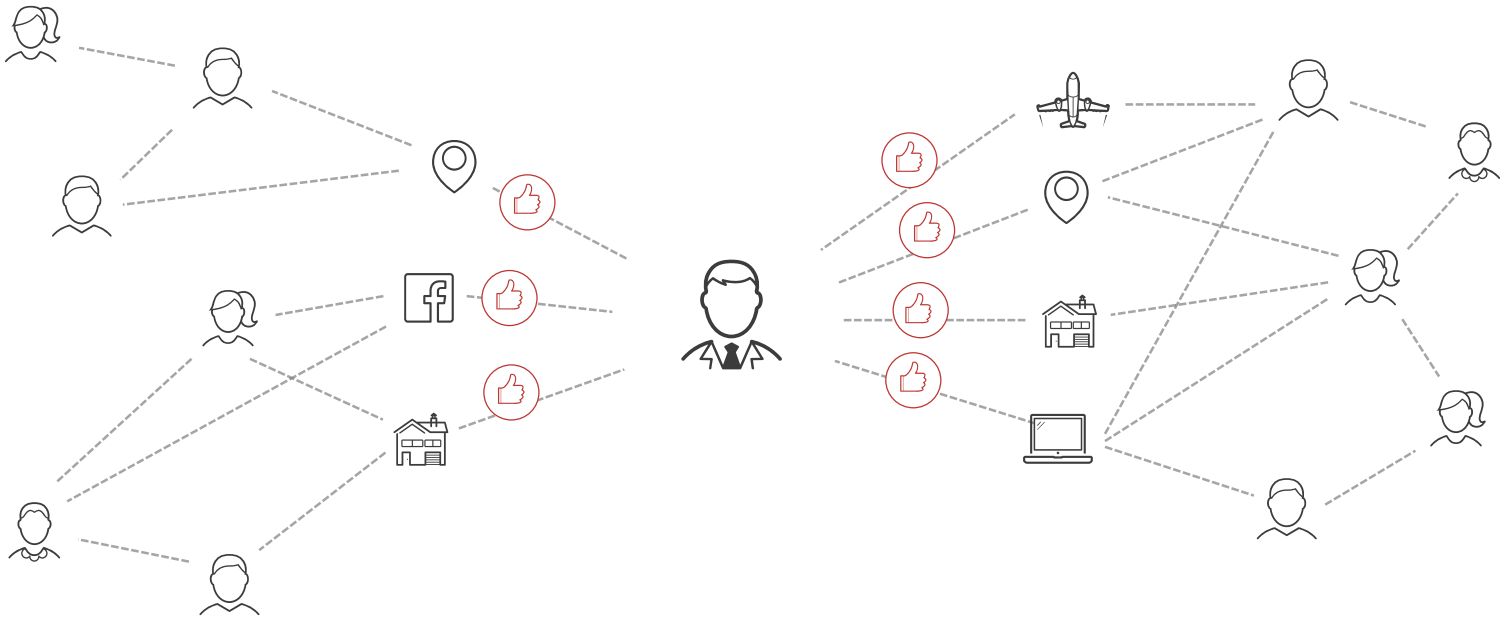
PROFINIT

- **Have you ever produced a DS report?**
  - Yes, you have!
- **What‘s your favorite tool for DS reporting?**
  - Python (Jupyter, Quarto)
  - R (Rmarkdown, Quarto)
  - MS Office
  - PowerBI
  - Tableau
  - …
- **Examples -** which report is better and why?



Examples based on a research project

Pseudosocial networks



Google amazon NETFLIX facebook Spotify

1

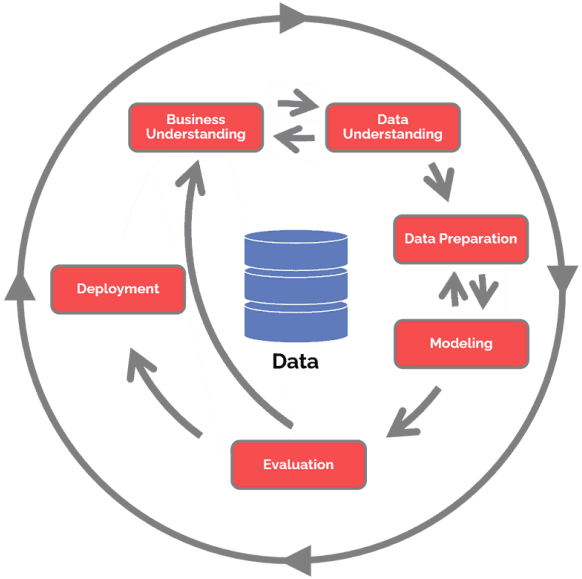
Report types

Report types based on PURPOSE

What story am I going to tell?

PROFINIT

- **A: Data description**
  - Let‘s take a look & briefly describe this new dataset
- **B: Exploratory analysis**
  - Let‘s dig into data to find some stories, relations, distributions etc.
- **C: Explanatory analysis**
  - Let‘s answer some hypothesis I have.
- **D: Model overview / Monitoring report**
  - I have a model/s and a need to inform about its performance, outputs, …



CRISP-DM

Do not mix them. Split the report (at least using paragraphs/chapters)

Report types based on PURPOSE

Let‘s practice

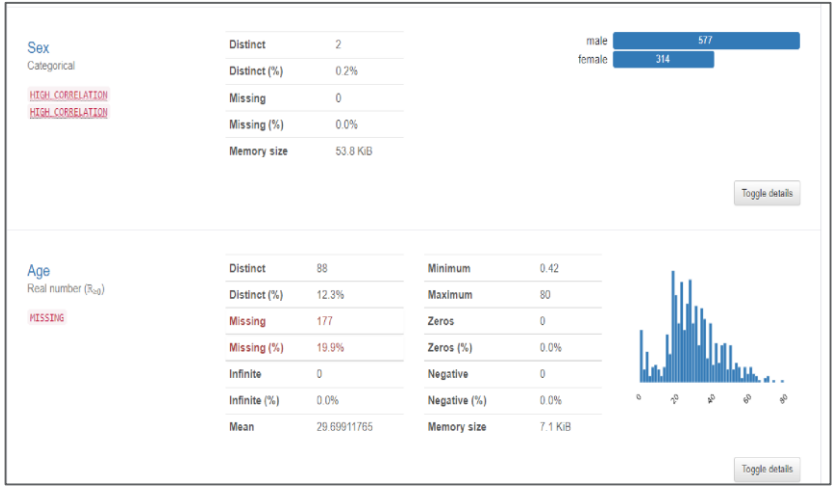
PROFINIT

Titanic Dataset Overview

- # observations
- # variables
- Do I have an outcome?
- Dataset structure
  - Row per passenger?
  - Denormalized data?
- Missings
- Variables distributions
- Outliers
- …

Include sanity checks!

Id	Tgt	Cls	Name	Sex	Age	Sib	Pch	Ticket	Fare	Cabin	Embarked
1	0	3	Owen Harris	M	22	1	0	A/5 21171	7.25		S
2	1	1	John Bradley	F	38	1	0	PC 17599	71.2833	C85	C
4	1	1	Jacques Heath	F	35	1	0	113803	53.1	C123	S
5	0	3	William Henry	M	35	0	0	373450	8.05		S



(Pandas) ydata-profiling (docs, example - Titanic)

Report types based on PURPOSE

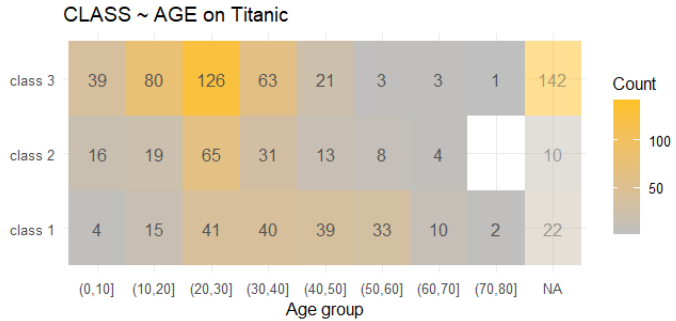
Let‘s practice

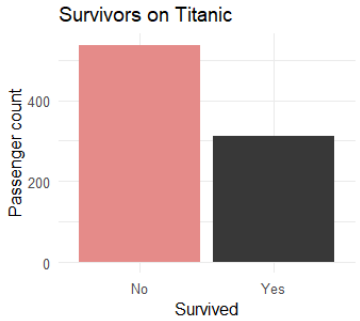
PROFINIT

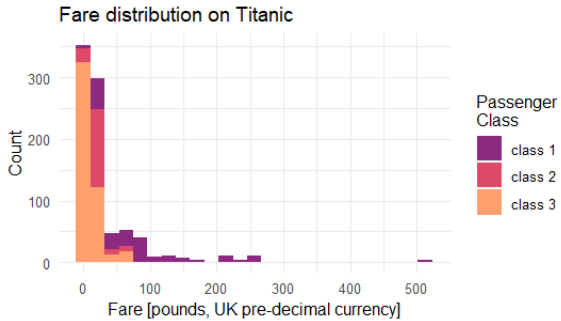
Titanic Dataset Exploration

- Charts / tables
- Outcome ~ variable relations
  - Is the mortality rate lower among specific subgroup?
  - Variable ~ Variable relations
- NA‘s - missing at random?
- **Goal**: generate hypotheses (stories)

Id	Tgt	Cls	Name	Sex	Age	Sib	Pch	Ticket	Fare	Cabin	Embarked
1	0	3	Owen Harris	M	22	1	0	A/5 21171	7.25		S
2	1	1	John Bradley	F	38	1	0	PC 17599	71.2833	C85	C
4	1	1	Jacques Heath	F	35	1	0	113803	53.1	C123	S
5	0	3	William Henry	M	35	0	0	373450	8.05		S







Report types based on PURPOSE

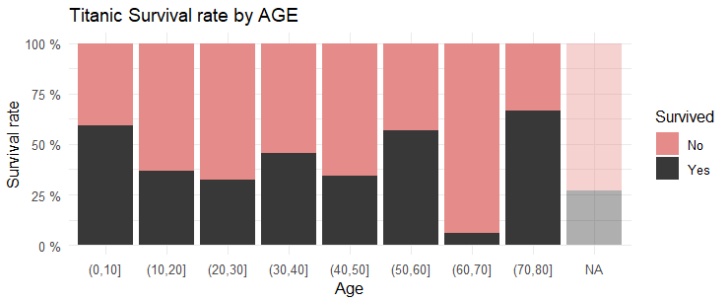
Let‘s practice

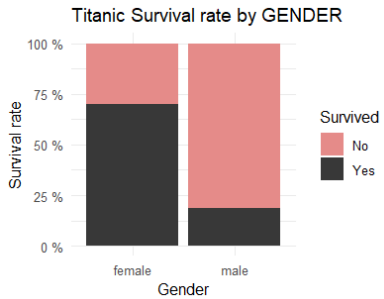
PROFINIT

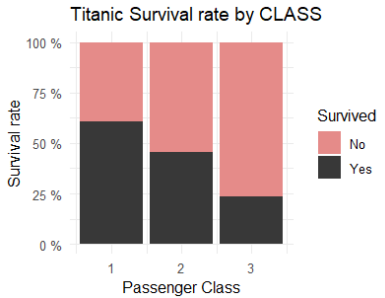
Titanic Dataset Explanation

- Selecting a subpopulation
- Variables transformation
- Feature engineering
- Outcome modelling
  (or some unsupervised technique, eg. clustering)
- Model evaluation

Id	Tgt	Cls	Name	Sex	Age	Sib	Pch	Ticket	Fare	Cabin	Embarked
1	0	3	Owen Harris	M	22	1	0	A/5 21171	7.25		S
2	1	1	John Bradley	F	38	1	0	PC 17599	71.2833	C85	C
4	1	1	Jacques Heath	F	35	1	0	113803	53.1	C123	S
5	0	3	William Henry	M	35	0	0	373450	8.05		S







Report types based on PURPOSE

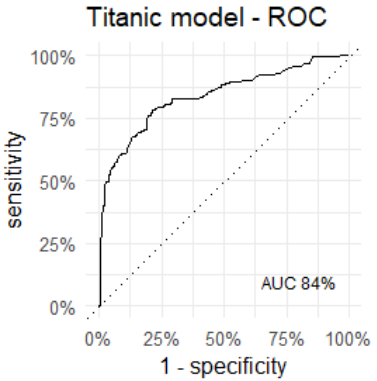
Let‘s practice

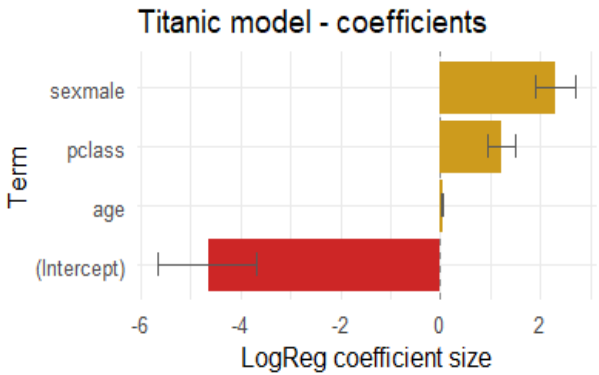
PROFINIT

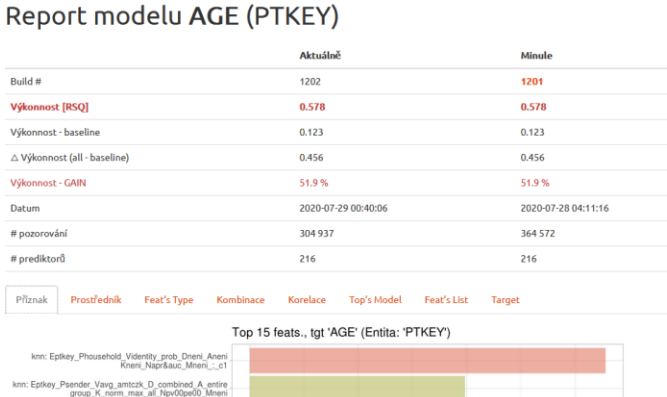
Titanic Model Overview

- Collecting artifacts
  - Performance metrics
  - Feature importances, coeficients, shap values, ..
- Comparision to a reference model
- Automate it!

Id	Tgt	Cls	Name	Sex	Age	Sib	Pch	Ticket	Fare	Cabin	Embarked
1	0	3	Owen Harris	M	22	1	0	A/5 21171	7.25		S
2	1	1	John Bradley	F	38	1	0	PC 17599	71.2833	C85	C
4	1	1	Jacques Heath	F	35	1	0	113803	53.1	C123	S
5	0	3	William Henry	M	35	0	0	373450	8.05		S







Report types based on LEVEL OF DETAIL

Typical types of DS reports

PROFINIT

› Knowledge report

– **To:**
  - Co-worker / future me
  - To be studied on his/her own

–  **Tips**:
  - High detail
  - Include methodology (to be able to reproduce it)
  - Open questions / not followed paths

› Business report

– **To:**
  - Customer, management, public
  - To be presented / walked through

–  **Tips:**
  - Strong story line
  - Fancy & straightforward graphics
  - Low detail
  - Call to action

Report types based on RESOURCES

Who & how will be consuming the story?

PROFINIT

- **Amount of prior knowledge?**
  - Future me x Coworker x Business x General public

- **Amount of time?**
  - Paper on its own x Paper walked through x Presentation x …

- **Amount of interaction?**
  - One man show x Discussion x Interactive exploration
  - Online / Offline

2

Report Structure

Structure of a report

PROFINIT

Academic paper

- **Abstract**
  - A brief synopsis of the paper – what did I do?
- **Introduction**
  - Context+purpose of the study – what‘s the problem?
- **Methods**
  - How did I solve the problem?
- **Results**
  - What did I find out?
- **Discussion**
  - What does it mean? (an interpretation in a context)
- **References, Acknowledgements, …**

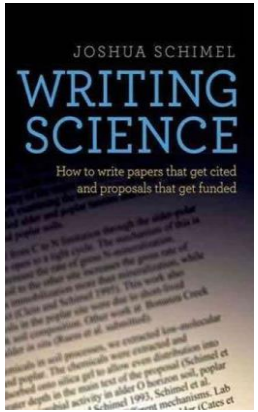
Data Science report

- **Introduction**
  - What do we do and why?
- **Data**
  - What data do we use
  - Some basic stats & sanity checks
- **Methodology**
  - This describes your work
    (data analyses, problems, findings, model building, …)
- **Summary**
  - Lessons learned in a brief structured form.
- **Next steps**
  - Call to action; open questions

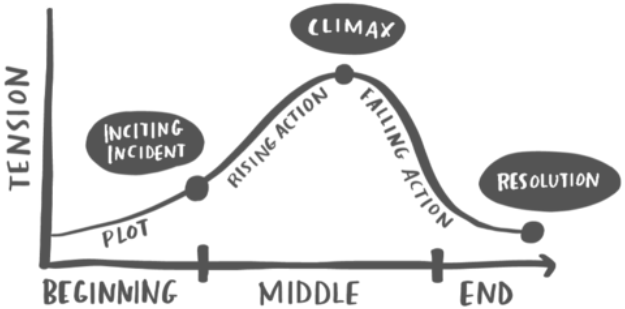
Structure of a story

PROFINIT

Pick a storyline & follow it



- **Common structures of a story:**
  - **OCAR**
    - **O**pening = set up story, describe data
    - **C**hallenge = the problem you‘re going resolve
    - **A**ction = gradual revealing the evidences
    - **R**esolution = summary, conclusions, call to action
  - **ABDCE** (Action – Background – Development – Climax – Ending)
  - **Journal article** (Lead – Development)



Structure of a story

Let‘s practice – what story should I pick?

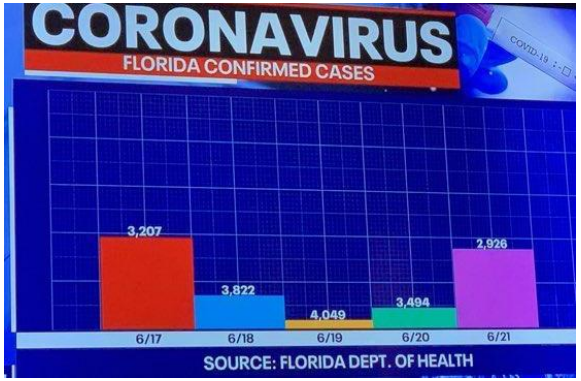
PROFINIT

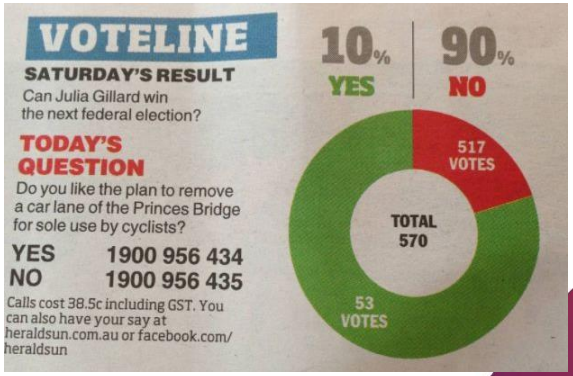
- **A: Data description**
  - „Let me introduce you my dataset.“
  - It‘s all right, there are no problems (or are there any?)
  - Do I have enough data to tell a story…
- **B: Exploratory analysis**
  - Sometimes it‘s part of an assignemnt
  - Outcome-related (if any outcome)
  - Do some explorations & pick the most promissing ones
- **C: Explanatory analysis**
  - Describe your hypothesis, show relevant data and proceed with feature engineering and modelling
- **D: Model overview**
  - Let me show you the model is all right..
  - It‘s strengths and weaknesses
  - This is how it defeats the champion

But story needs to follow your data…

PROFINIT

… and not vice versa





See e.g. wtf.viz for reference

2

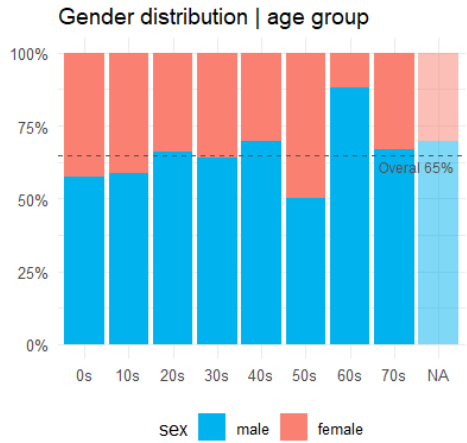
Communicating data

Communicating data

PROFINIT

- Your job is to **communicate data**
- Choice of a channel has a big impact
- **Text**
  - Good at describing a background/context
  - Slow & tedious to retrieve an information
- **Tables**
  - Structured information, precise, side information
  - Hard to spot trends
- **Charts**
  - High signal-to-noise ratio, a quick glance at trends
  - May be missleading (un/deliberately), imprecise

age_cut	count	male_pct	mean_fare	surv_pct
(0,10]	59	0.576	31.152	0.593
(10,20]	114	0.588	29.785	0.368
(20,30]	232	0.659	27.793	0.323
(30,40]	134	0.642	44.245	0.455
(40,50]	73	0.699	51.854	0.342
(50,60]	44	0.500	70.626	0.568
(60,70]	17	0.882	65.701	0.059
(70,80]	3	0.667	38.875	0.667
NA	174	0.695	18.424	0.270

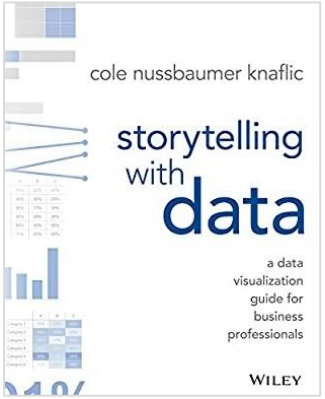


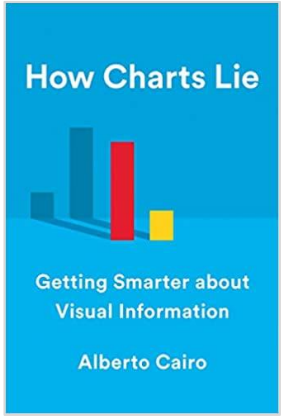
Graphs in a report

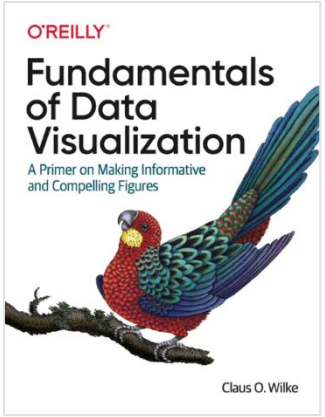
PROFINIT

- Charts are **the most efficient** way to communicate data
- But they could be harmful, too.
  - See viz.wtf for reference

- **4+1 basic rules for charts in report**
  1) Pick relevant graphs only
  2) Use familiar graph types
  3) Be consistent
  4) Link, annotate & describe
  5) Write down what should reader see on the chart.





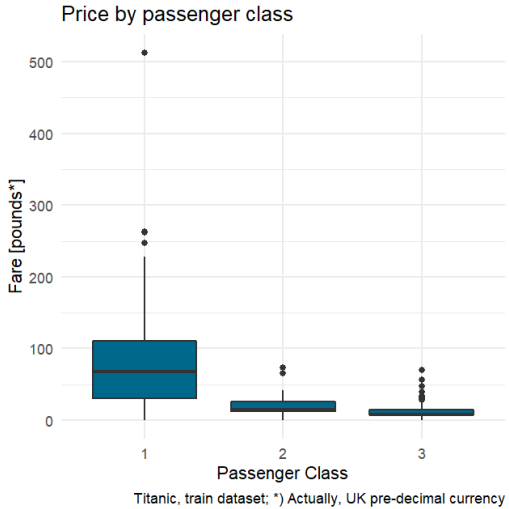


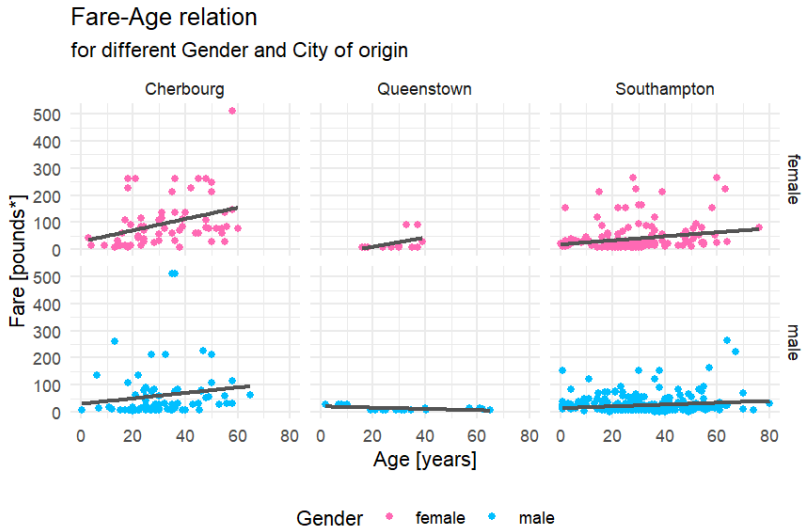
Graphs in a report

1) Pick relevant graphs only

PROFINIT

- **Include all pictures to tell the story. Neither less, nor more.**
  - Don‘t force your audience to re-do your investigation.
  - Don’t overengineer your charts



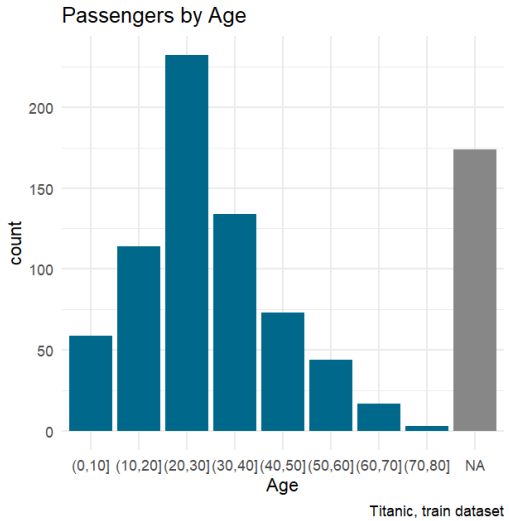


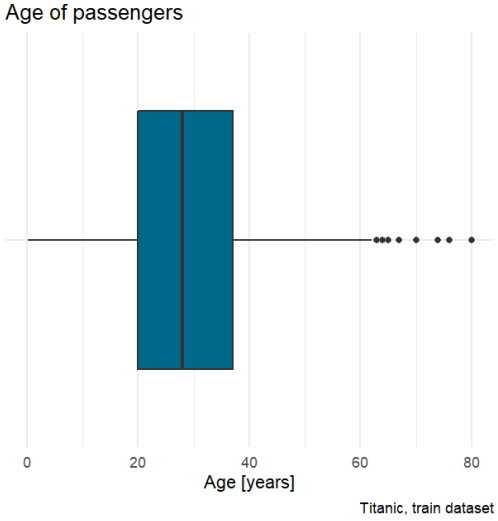
Graphs in a report

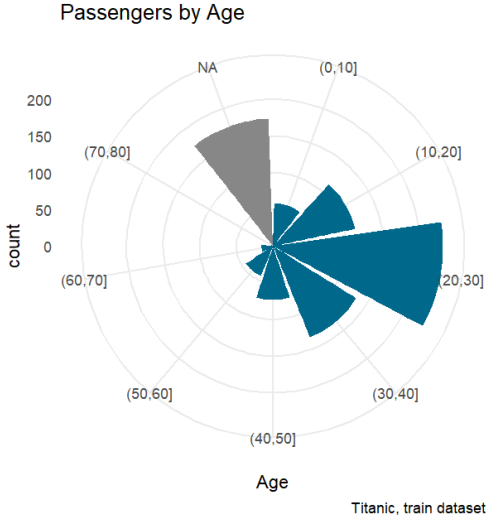
2) Prefere familiar graph types

PROFINIT

- Use as few graph types as possible (if that makes sense)
- Prefer a graph types with high signal-to-noise ratio
- Is your reader comfortable with the plot type? (include ‚*how to*‘ othrewise)





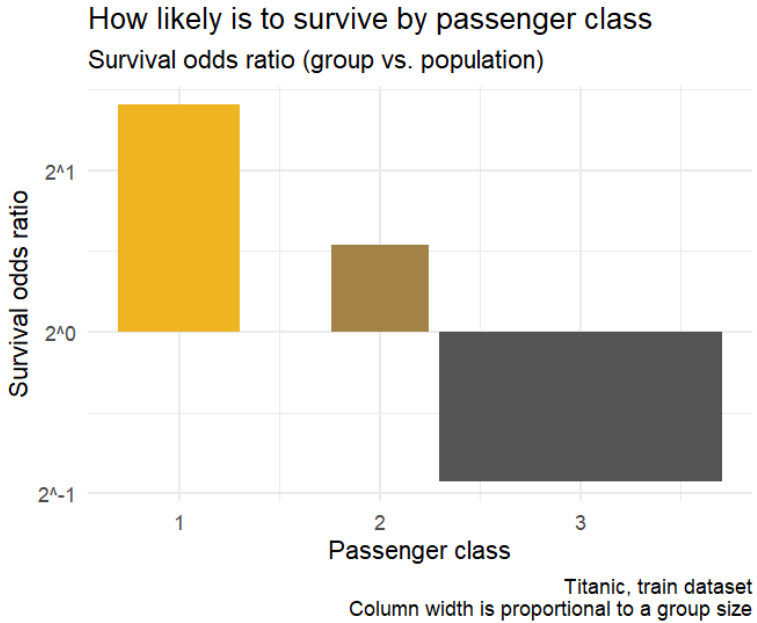


Graphs in a report

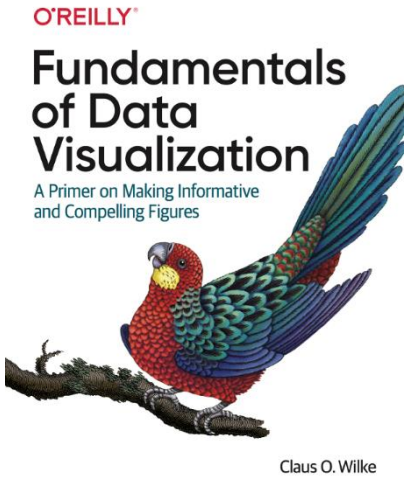
2) Prefere familiar graph types

PROFINIT

› If not possible, take your time to **describe how to read the plot**



(book, online)

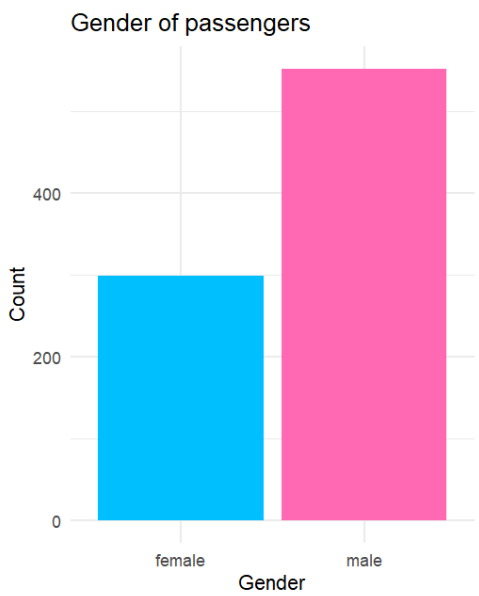


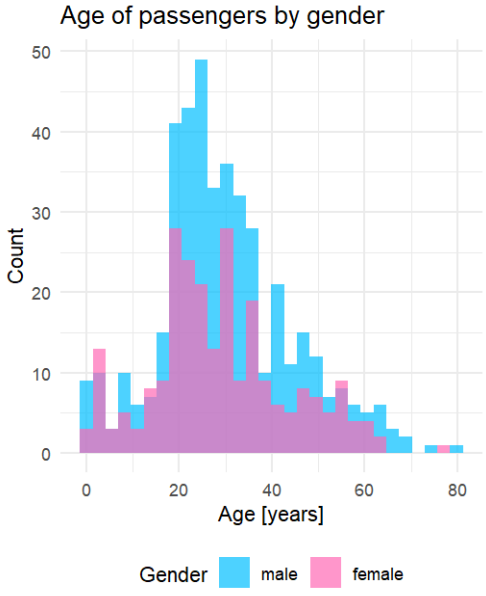
Graphs in a report

3) Be consistent

PROFINIT

- **Use scales wisely to tell the story**
  - (color, shape, transparency, linetype, …)
- Follow general **conventions**
  - May vary among cultures
- Be ware of defaults
  - `F < M` (english) vs. `M < Ž` (czech)
  - Changing no. levels among plots
  - Same palette for multiple uses

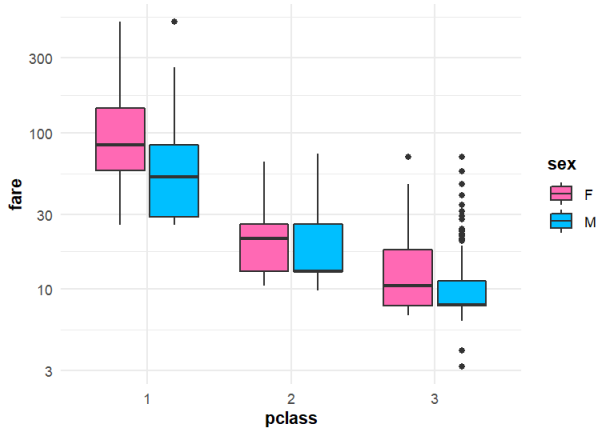


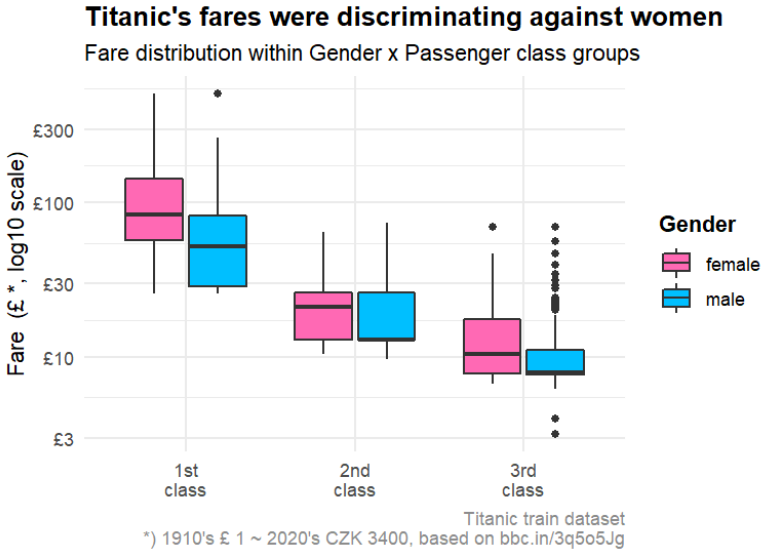


PROFINIT

4) Annotate, link & describe

- **Add appropriate descriptions** (title, axes, scales, data source)
  - The message must be obvious even if the author is absent.
  - Use human readable strings





- **Refer each graph in text**
  - Provide all insights a reader should gain from the plot !!

Interactive charts

PROFINIT

- **… are useful. Sometimes.**
  - to let readers explore the data on their own
  - to visualise high dim data, include pop-up labels etc.
- **… are dangerous. Sometimes.**
  - Putting the burden of storytelling on readers
  - May be time-consuming
  - Misuse just because it is so cool

plotly (available for both Python and R)

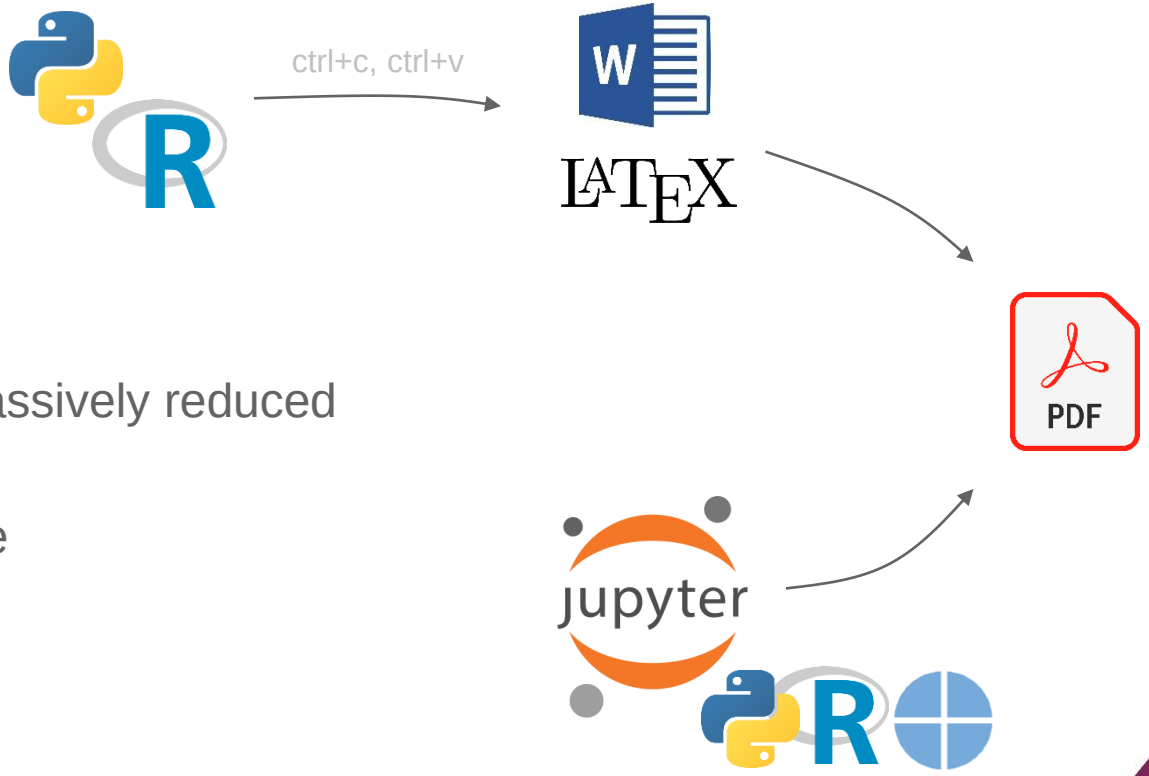
- Very easy integration into Jupyter/Rmarkdown/Quarto reports

2

Report as a code

Report as a code

PROFINIT



Some benefits:

- Manual work massively reduced
- Reproducibility
- Less error-prone
- Version control
- IDE features
- …

Jupyter Notebooks

PROFINIT

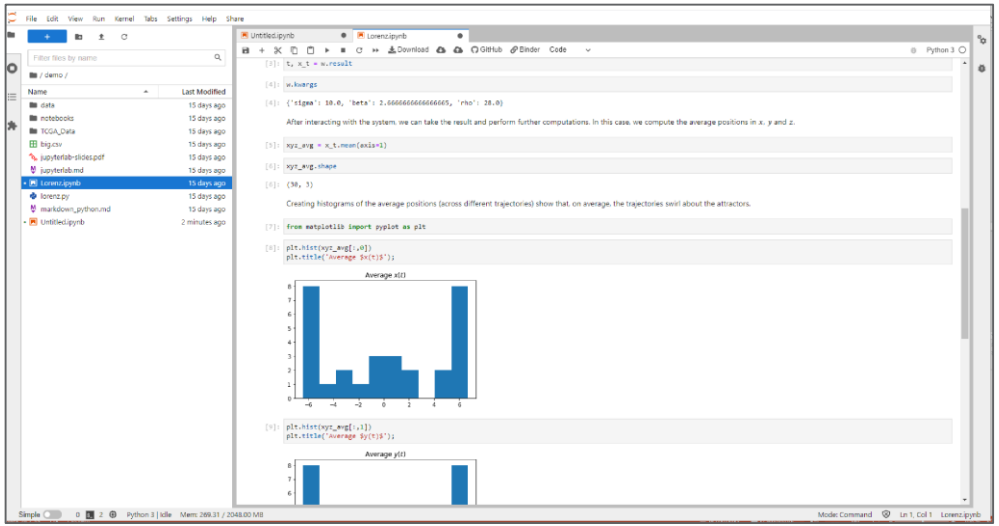
› *„Open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text*

jupyter.org

› No. 1 choice for Python-based research
  – Other kernels available, incl. R

› Web-based IDE
  – Pycharm and VSCode compatible

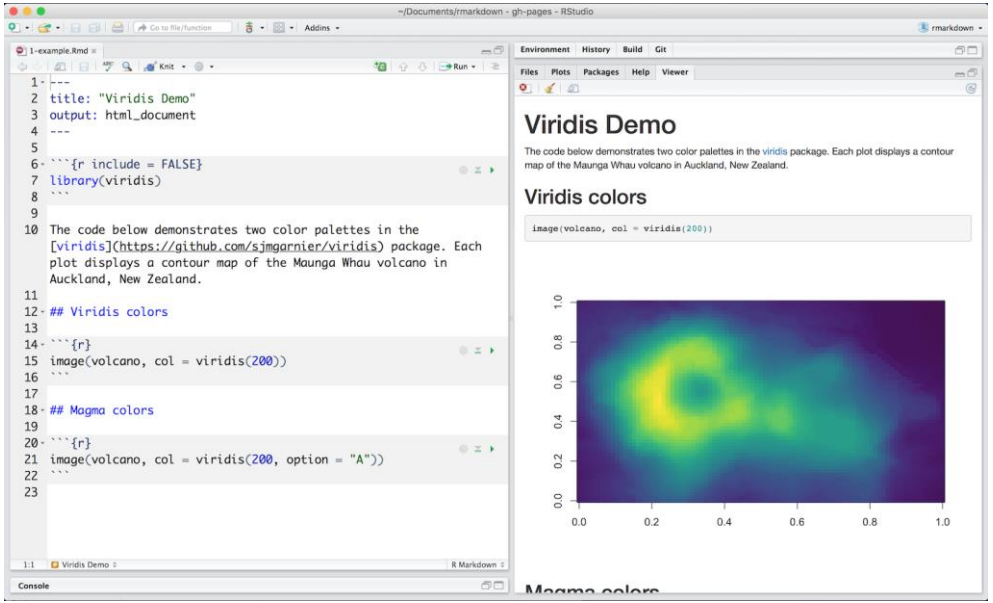
› Various extensions & flavours
  – **nbconvert**
  – **pretty-jupyter**



RMarkdown

PROFINIT

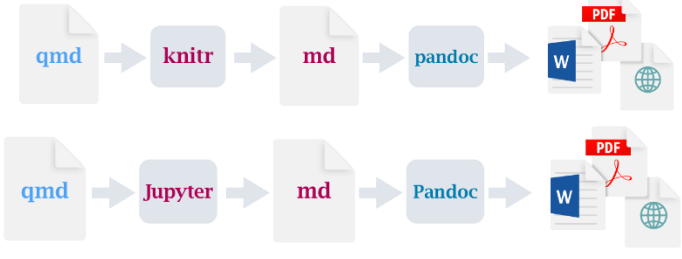
- No. 1 choice for R-based research
  - Other kernels available, incl. Python, …
- Simmilar to JupyterNotebooks
  - Text and Code chunks
  - Ready to be ‚knitted' into .md, .html, .pdf, .docx and more.
- Various flavours, e.g.
  - Bookdown (e.g. this book)
  - PkgDown
  - PageDown

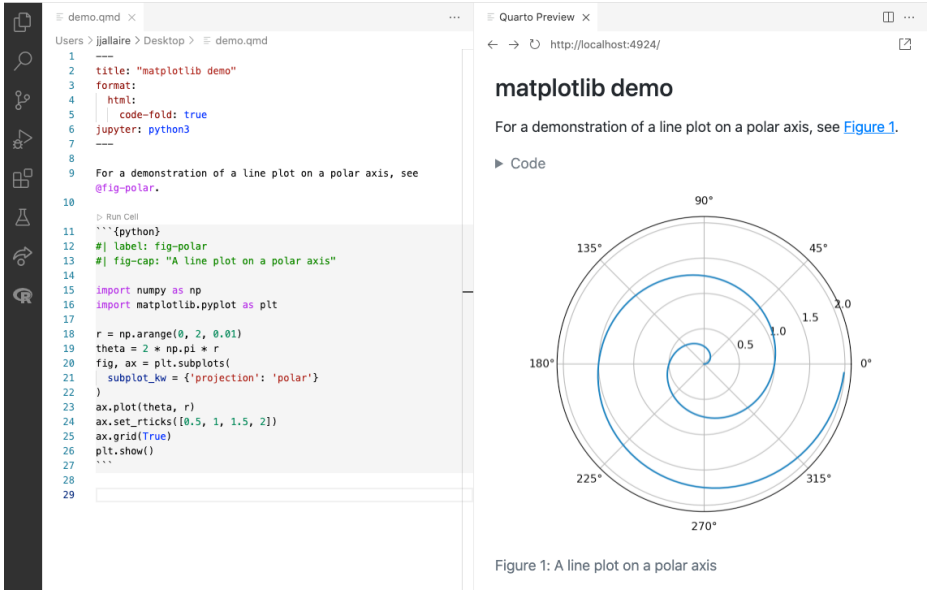


Quarto

PROFINIT

- Recent product of Posit.co (Rstudio PBC)
  - Next gen Rmd/Jupyter
  - Document design based on Rmd
  - Works with Python, R, Julia and Observable
  - Commandline utility + IDE extensions





Quarto vs Jupyter

PROFINIT

- Notebook wars by Yi
- Text file vs nested JSON
  - GIT, code reviews
- Not stored output vs Stored output
  - Out-of-order execution
- Quarto extras
  - Multiple output formats, prettier outputs (but pretty-jupyter!)
  - More customizable
  - Cross referencing, bibliographies
  - Can render ipynb, too ☺

Don‘t show the code

producing your report

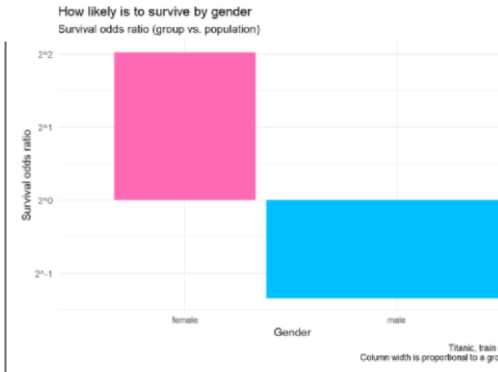
- The code should **stay behind the report**, not inside
  - Final version shouldn‘t contain code cells
  - If anybody needs the code, he/she can read the source file

Use **`jupyter nbconvert notebook.ipynb --no-input`**

(or `echo = FALSE` knit-option in .Rmd)

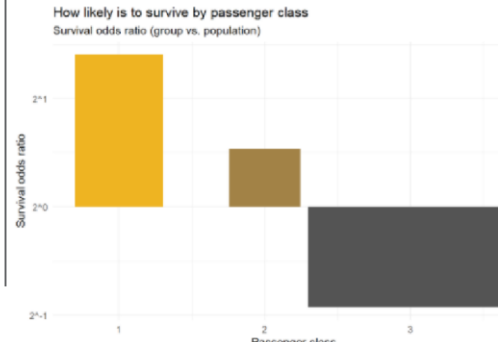
Use **`--template pj`** to prettier converted reports

(check out github.com/JanPalasek/pretty-jupyter




```
class_stats <-
  tt %>%
  group_by(pclass) %>%
  summarise(
    n = n(),
    p_surv = mean(survived),
    gr_odds = p_surv/(1-p_surv),
    odds_ratio = gr_odds/odds
  )


class_stats %>%
  ggplot(aes(x = pclass, y = odds_ratio, fill = pclass)) +
  geom_bar(stat = "identity", aes(width = n * 2.5/nrow(tt))) +
  # scale_fill_manual(values = c("female" = "hotpink", "male" = "deepskyblue1")) +
  scale_fill_gradient(high = "gray33", low = "goldenrod2") +
  scale_y_continuous(trans = "log2", breaks = 2^(-2:2), labels = str_c("2^",-2:2)) +
  labs(
    x = "Passenger class",
    y = "Survival odds ratio",
    title = "How likely is to survive by passenger class",
    subtitle = "Survival odds ratio (group vs. population)",
    caption = str_c(capt, "\nColumn width is proportional to a group size")
    ) +
  guides(fill = FALSE)
```



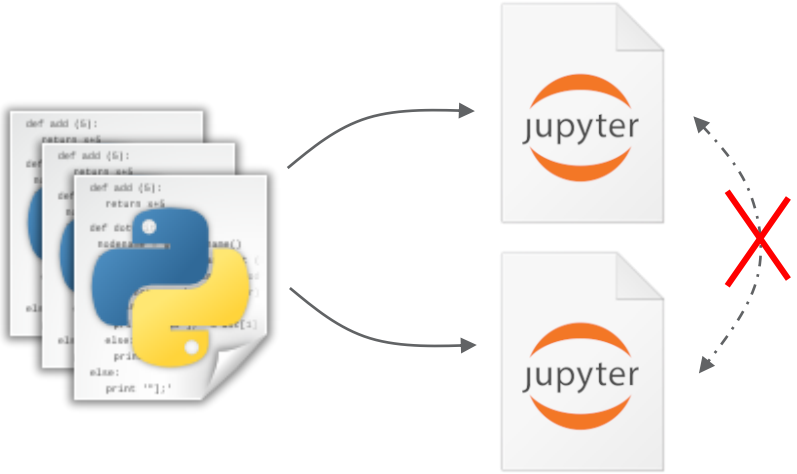
›

Reuse the code

producing your report

PROFINIT

- **The report is a code. Follow general SWENG best practices**
  - E.g., don‘t repeat yourself (and others)
  - Use functions etc., define them elsewhere!



Reuse the code

producing your report

PROFINIT

.Rmd/.ipynb are parameterizable

- **Reproducibility is of the first concern**
  - Watch out seeds of randomness.
  - **Don‘t get paralysed**.
    It‘s better to fix the data and have a new version of the report if needed.

```
In the dataset, we have `r ncol(data)` columns and `r
nrow(data)` rows. There are `r n_distinct(data$group)`
groups. `r ifelse("sex" %in% names(data), "There is a
gender column:", "")`.
```{r, eval="gender" %in% colnames(data)}
# some code
```
```

2

Few more tips

Few more tips

PROFINIT

- Producing efficient reports is not trivial
- Few more tips
  - **Do not overhelm your audience**
  - **Shape your story to fit your audience**
  - **Getting better every day**

Few more tips

Don‘t overwhelm your audience

PROFINIT

- **Don‘t force your audience to re-do all your work!**
  - That was your job. You know the data and the goal.
  - Do not include all possible charts/tables.

Use a collapsable content for extra details (e.g., <details></details>)

- **Report is not evidence of your hard work!**
  - Research is full of misleading paths. Think – is it beneficial to include them?

Use GIT to track your work instead. It can help you with text tracking, too.

- **Be as wordy as needed. But not more!**
  - Link to textbooks, don‘t write them
  - Typically, the use of bullets is enough, no need for long paragraphs

Few more tips

Choose your story

PROFINIT

- **Pick a starting point known by your audience**

- **Do not hesitate to cut-off not-followed options**
  - You know the data and the goal. You are telling the story.
  - Same applies when, e.g., setting thresholds, selecting population etc.

Comment your actions to ensure the reader you are aware of your decision.
It‘s great to provide a list of the next possible steps in the conclusion.

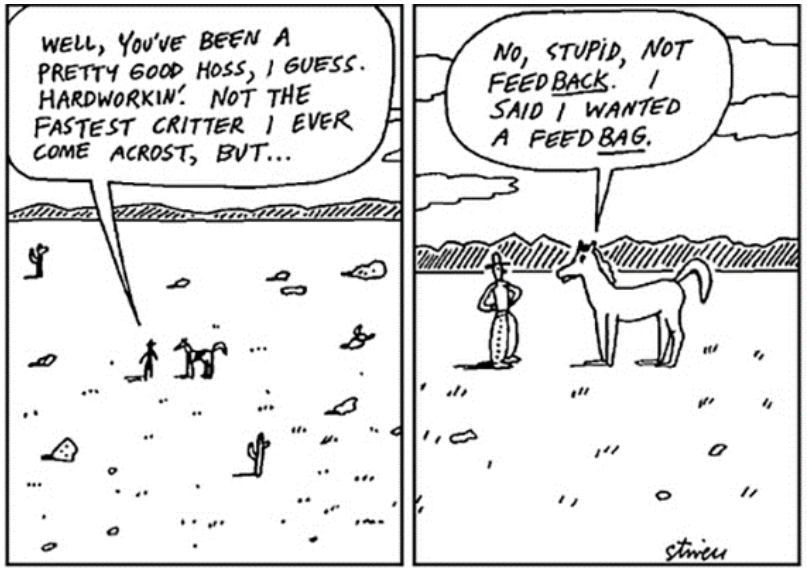
- **Follow just one storyline at a time**
  - Typically, it‘s better to dedicate a report to each story you‘d like to tell.

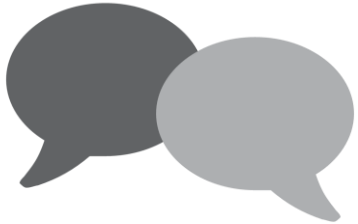
Few more tips

Getting better every day

PROFINIT

- **Watch out reproducibility**
  - Re-run all cells once in a while
  - There is a shortcut for that. ☺
- **Watch out for feedback**
  - Write → Present → Get feedback → Adjust.
- **Know your IDE by heart**
  - Actually, there is a shortcut for pretty much everything





Questions?

Literature

PROFINIT

- Joshua Schimel: **Writing science**
- Cole N. Knaflic: **Storytelling with data**
- Alberto Cairo: **How charts lie**
- Clause O. Wilke: **Fundamentals of data visualization**
- Yihui Xie: **Rmarkdown**
- Wes McKinney**: Python for Data Analysis**
- …

Thank you
for your attention

PROFINIT

Profinit EU, s.r.o., Tychonova 2, 160 00 Praha 6
Tel.: + 420 224 316 016, web: www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU

Facebook
facebook.com/Profinit.EU

Youtube
Profinit EU