Reproducible Reporting with RMarkdown (R3)

Mar. 2022

Agenda

01

Preface

02

What's RMarkdown?

03

How to make reports with RMarkdown?

04

Report Exapmles

01

Preface

Motivation for Reporting Task

Robust Research:

- “Robust research is about doing small things that stack the deck in your favor to prevent mistakes.”
  —Vince Buffalo, author of Bioinformatics Data Skills (2015)

Reproducible Research:

- Reproducible research can be repeated by other researchers with the same results

Interactive Reporting:

- Apart from paper or PDF reports, allowing users to interact with the report allows them to ask questions about the data itself

Requirements for Reporting

Distributable:

- Self-contained report that makes it easier to get feedbacks

Viewable to Anyone:

- Viewable to non-engineers in order to eliminate information asymmetry
  e.g.) In most cases, JupyterLab can only be accessed by engineers

Automate Reporting:

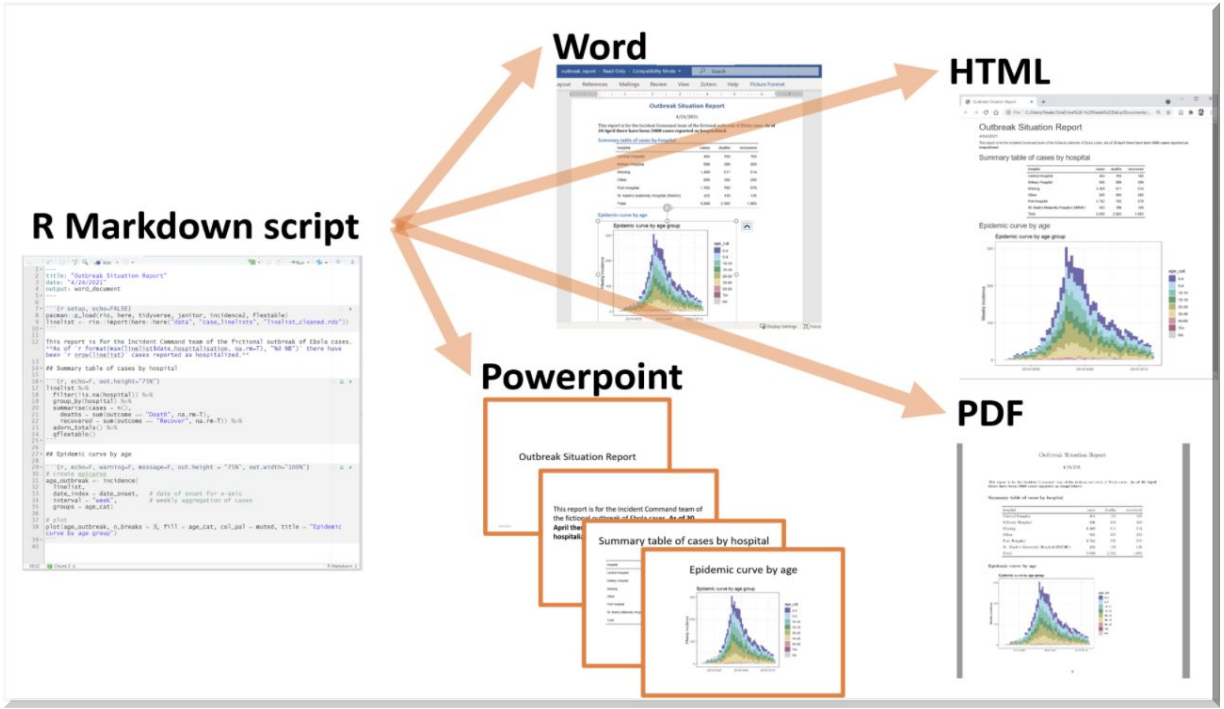
- Reporting task should be done at the same time as analysis task
- Preliminary analytical design becomes more important

02

What's RMarkdown?

What's RMarkdown?

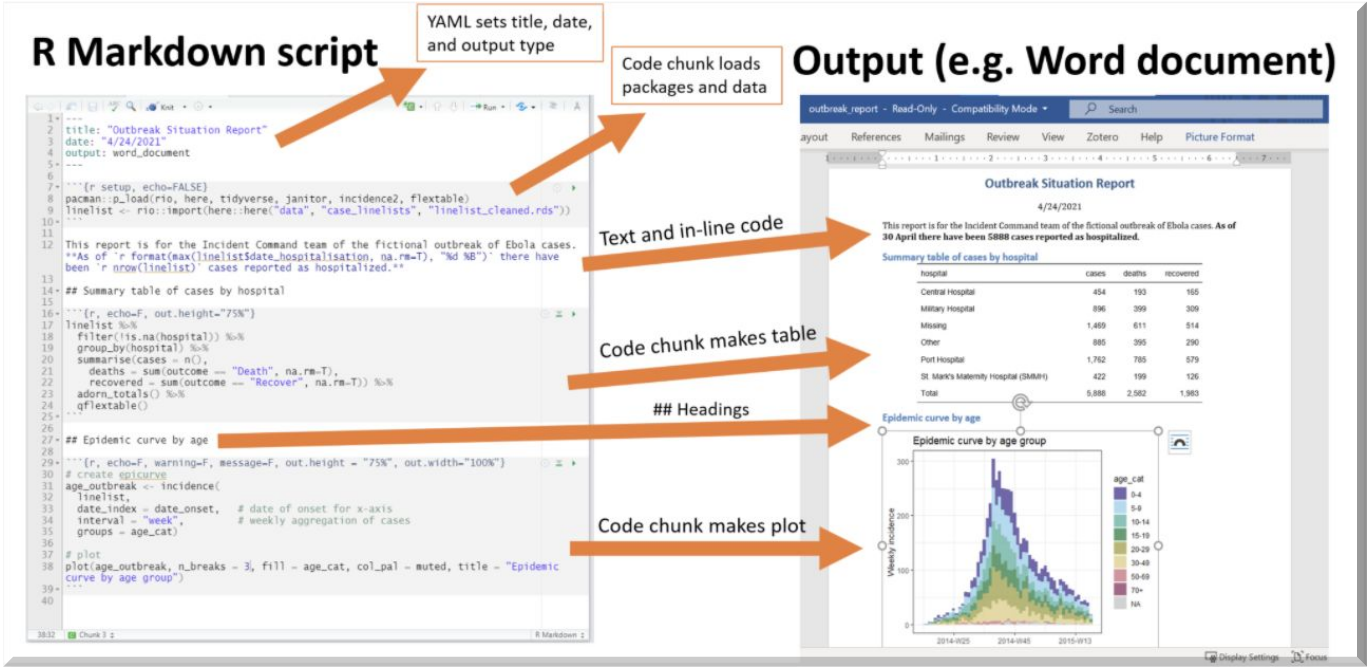
RMarkdown is a widely-used tool for creating automated, reproducible, and share-worthy outputs, such as reports (html, pdf, docx, ...)



[Source]

What's RMarkdown?

RMarkdown is a widely-used tool for creating automated, reproducible, and share-worthy outputs, such as reports (html, pdf, docx, ...)



[Source]

Benefits of RMarkdown

Easy to Deliver:

- We can distribute html reports. Other formats (pdf, docx) are available
- Can be attached to ESA (up to 10MB by default)

Easy to Reproduce:

- Easy to recreate the report if the report data is updated
- We can easily create reports if know Markdown

Rich Expression:

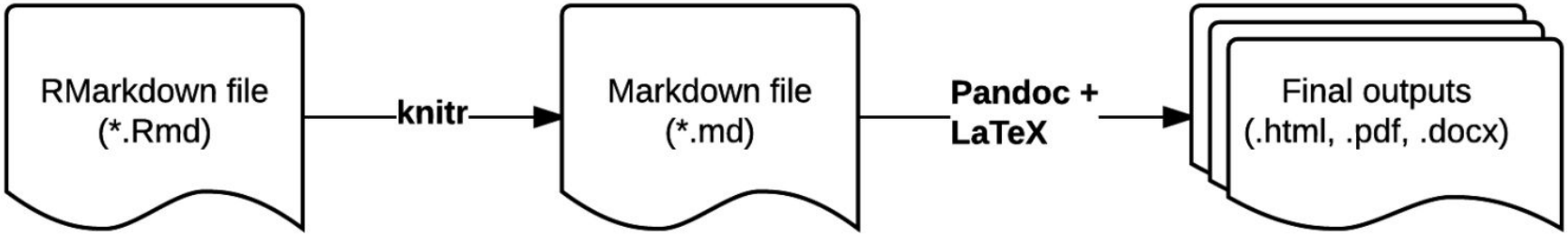
- Interactive report (only html format)
- Ingenuities that can aggregate information, such as Tab / Table display

03

How to make reports with RMarkdown

RMarkdown Workflow: [Link]

- RMarkdown is an enhanced version of Markdown that lets you embed R code into the docs
- .Rmd → .md → Document(.html/.pdf/.docx)



Creation example by R

- Prepare 2 type of files for report contents(.Rmd) and rendering(.R)

Report contents(.Rmd)

````
# Let's embed some R code
```{r}
library(dplyr)
library(readr)
gm <- read_csv('data/gapminder.csv')
```

The mean life expectancy is
`r mean(gm$lifeExp)` years.

The years surveyed in this data include:
`r unique(gm$year)`.
````

Rendering(.R)

```
library(rmarkdown)

rmarkdown::render(
        input=input_file_Rmd,
        output_format="html_document",
        output_file=output_file,
        encoding='utf-8',
        quiet=F)
```

Call R from Python

- We can easily call R from Python

```
import os
os.system("Rscript ./render.R --arg1=xxx --arg2=yyy, ...")
```

- We can also use R with Google Colaboratory, just adding below

```
[ ]  # activate r magic
     %load_ext rpy2.ipython
```

```
[ ]  %%R
     Sys.setenv(TZ="Asia/Tokyo")
     install.packages('argparse')
     install.packages('rmdformats')
     install.packages('formattable')
     install.packages('DT')
```

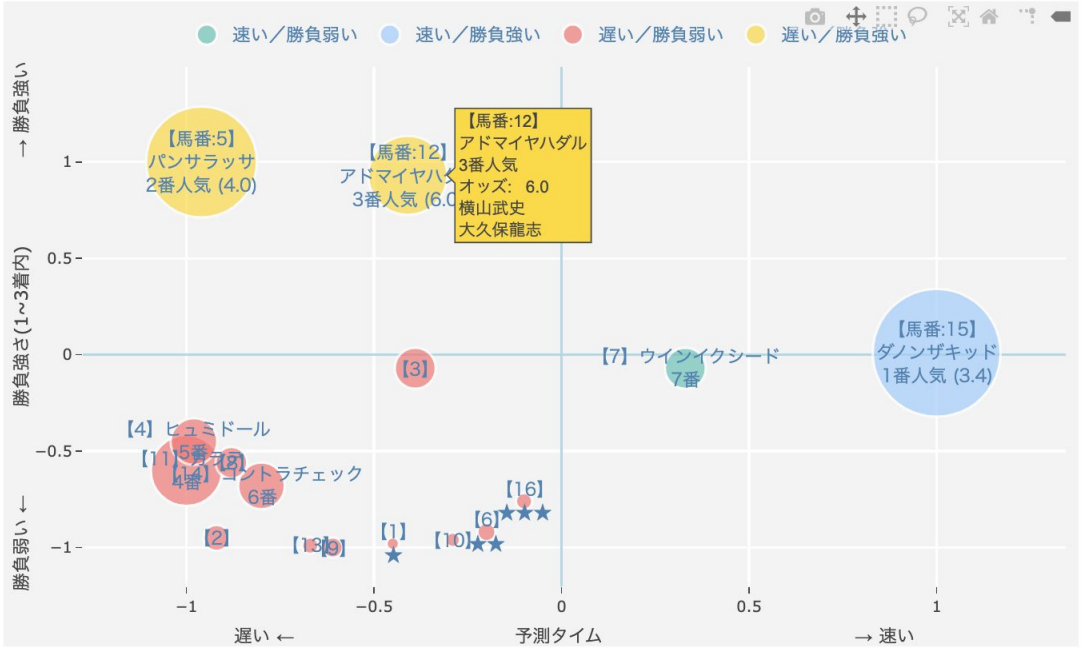
04

Report Examples

Ex.1) Horse Racing Prediction(1/4): [Link]

Visualize the prediction scores of spped (x-axis) and clutch (y-axis)

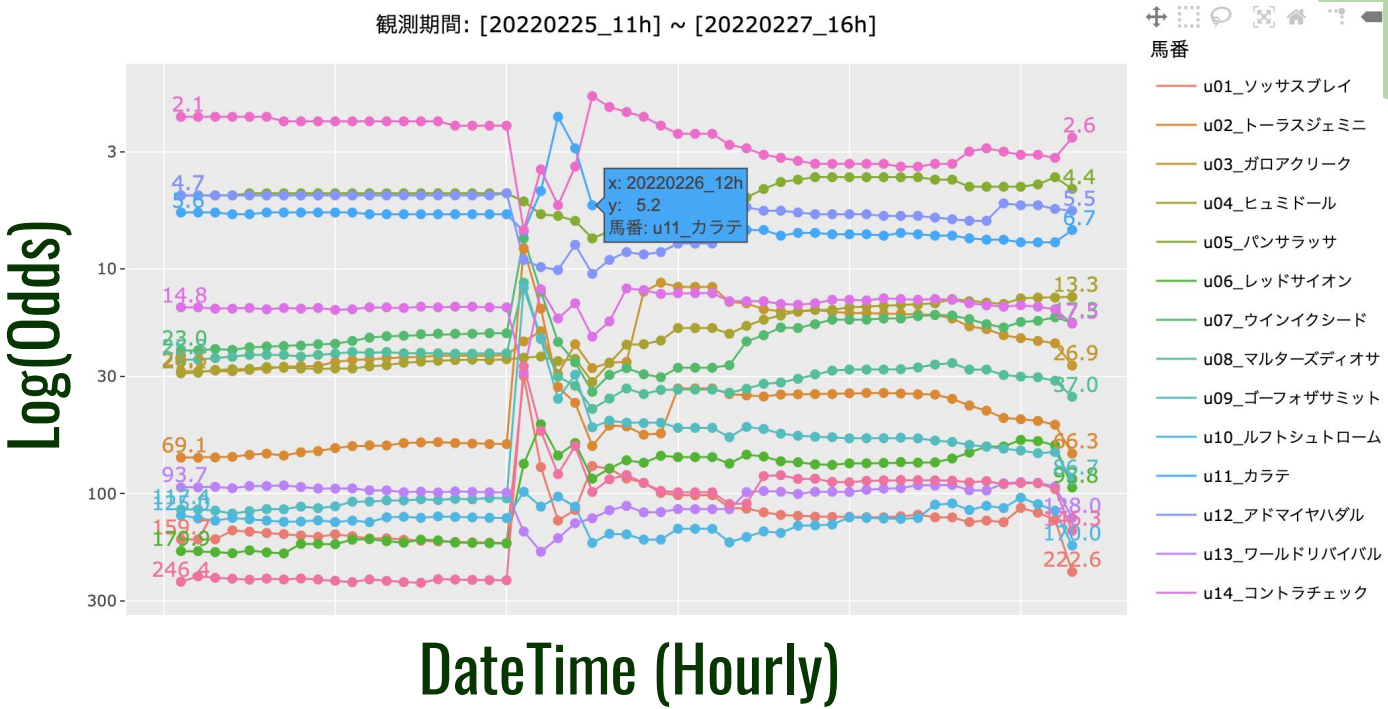
Prediction Chart (Speed/Clutch)



Ex.1) Horse Racing Prediction(2/4): [Link]

Odds information will be updated hourly until just before the race

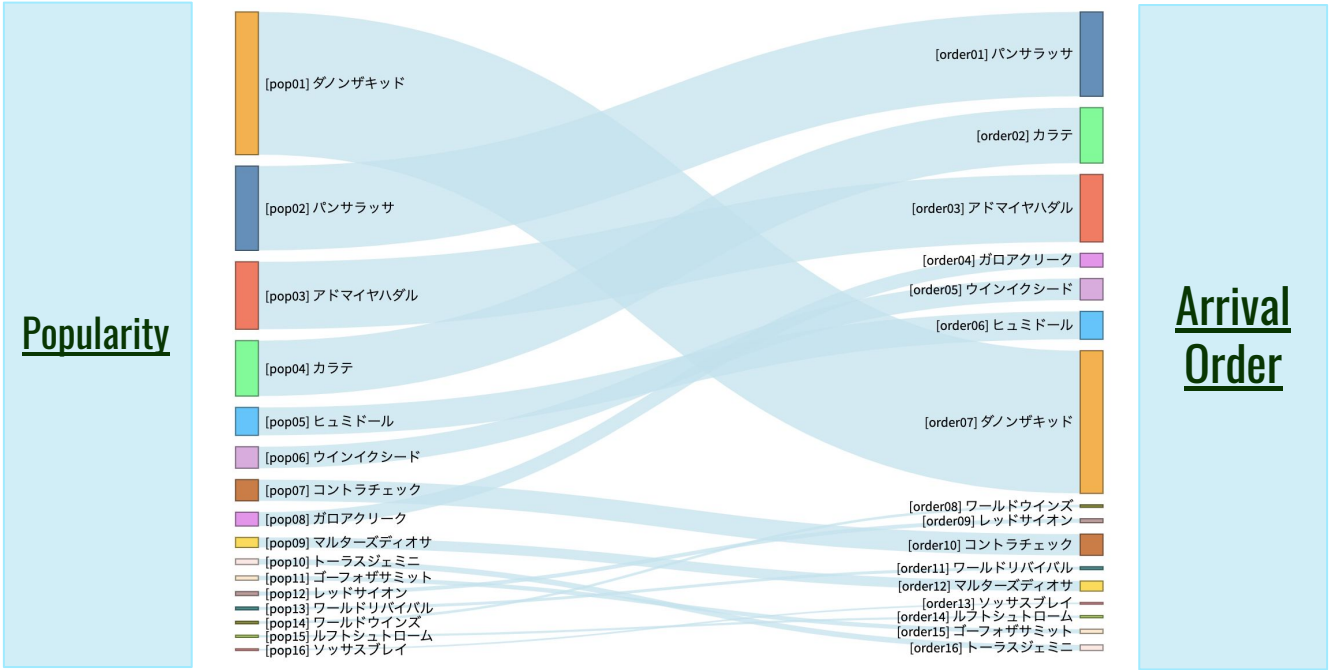
Monitoring Odds (Hourly)



Ex.1) Horse Racing Prediction(3/4): [Link]

Visualize the discrepancy between popularity and arrival order as a back test

Popularity vs Arrival Order



Ex.1) Horse Racing Prediction(4/4): [Link]

Visualize the order of passing corners, as the race development can be imagined

The Order of Passing Corners

Ex.2) Machine Learning Glossary: [Link]

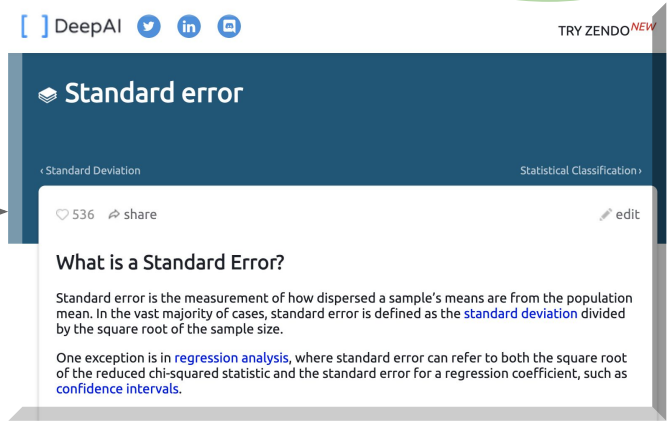
Create a table report for each category by scraping the glossary

【Regression Analysis (3)】 【Response Variable (3)】 【Sigmoid Function (3)】 【z-score (3)】

【Regression Analysis (3 terms)】

Show 5 entries Search:

No	Term	trunc-summary	tag	date
All	All	All	All	All
151	Standard error	Standard error is the measurement of how dispersed a sample's means are from the population mean.	Regression Analysis / Standard Deviation / Confidence Interval	05/17/2019
160	Heteroscedasticity	Heteroscedasticity refers to data for which the variance of the dependent variable is unequal across the range of independent variables.	Regression Analysis / Variance	05/17/2019
192	F-Distribution	The F distribution is a right-skewed distribution used commonly in another statistical test called an Analysis of Variance (ANOVA).	Probability / Regression Analysis / Variance	05/17/2019



Ex.3) Leaflet Map: [Link]

Scraping the addresses of medical institutions and put it on the map

【05_小児科 (79件)】 【06_産婦人科 (22件)】 【07_眼科 (35件)】 【08_耳鼻科 (25件)】 【09_歯科 (253件)】
【10_精神科 (20件)】

References

- [R Markdown: The Definitive Guide]
- [Reproducible Reporting with RMarkdown]
- [SportSciData]
- [Smart and Interactive Documents]
- [40 Reports with R Markdown]
- [Tools for Reproducible Research]


End of Document