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

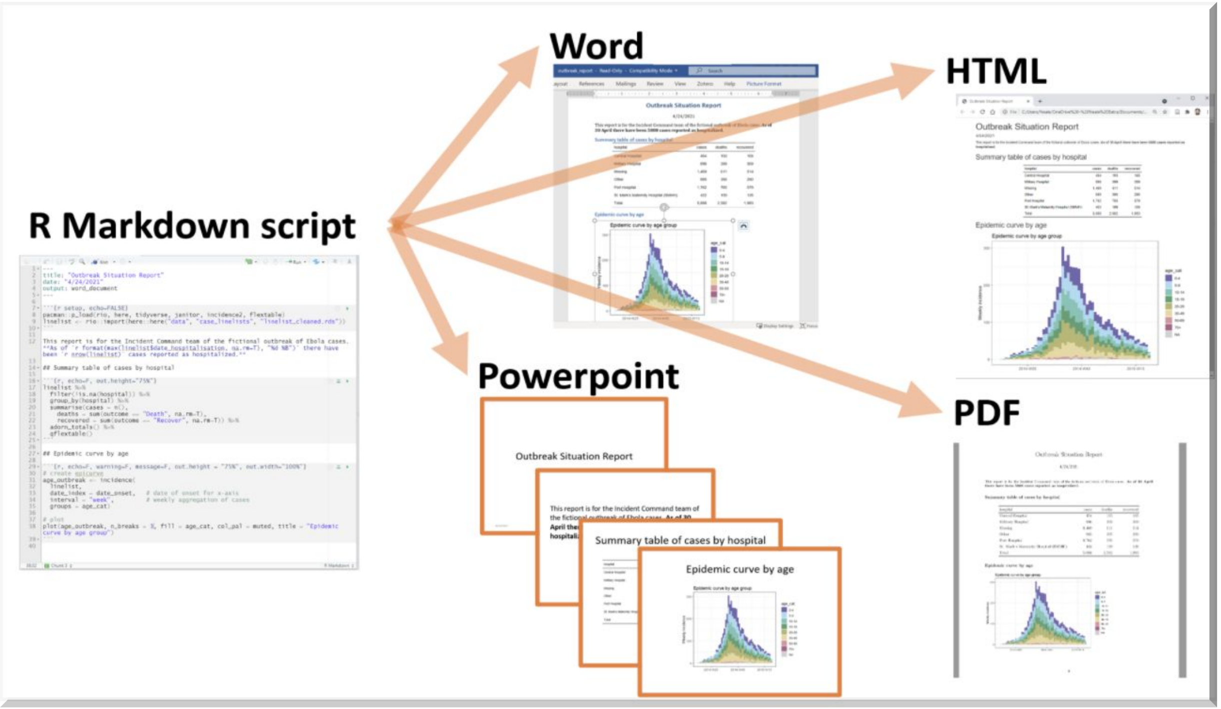
02

…

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, …)

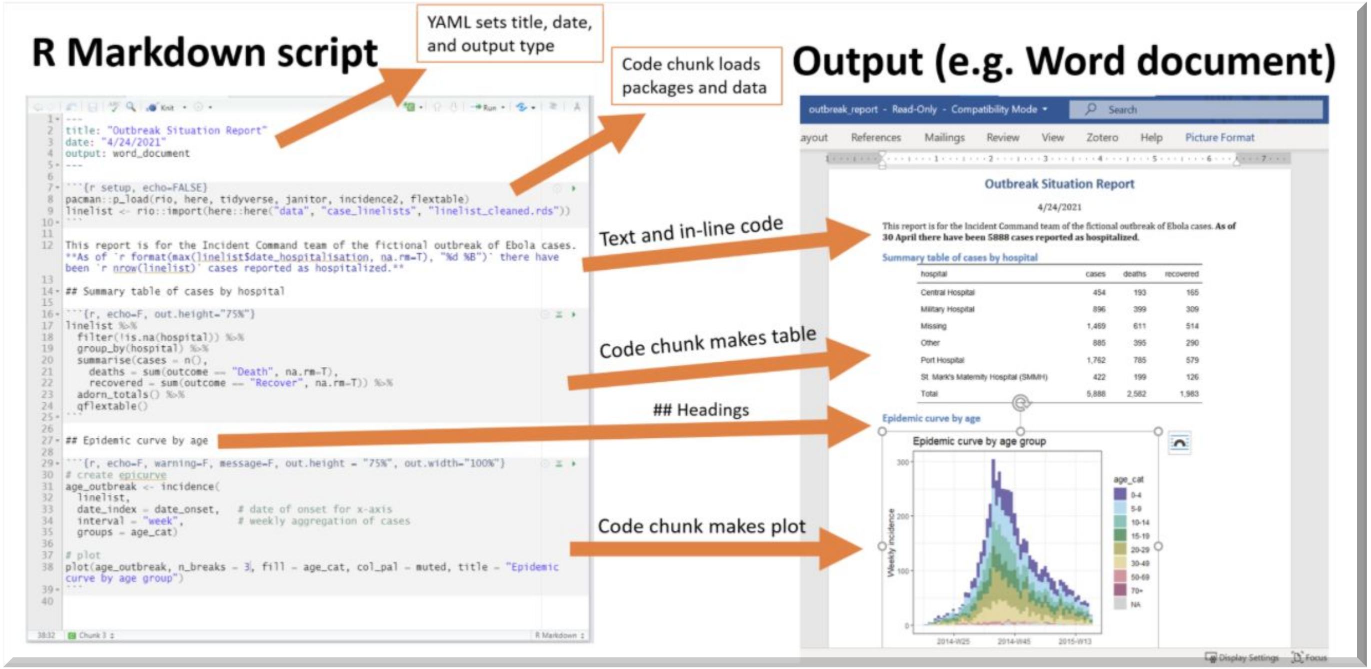

R Markdown script
Word
HTML
Powerpoint
PDF
Outbreak Situation Report
Summary table of cases by hospital
Epidemic curve by age

~~~
---
title: "Outbreak Situation Report"
date: "4/24/2021"
output: word_document
---

```{r setup, echo=FALSE}
pacman::p_load(rio, here, tidyverse, janitor, incidence2, flextable)
linelist <- rio::import(here::here("data", "case_linelists", "linelist_cleaned.rds"))
```

This report is for the Incident Command team of the fictional outbreak of Ebola cases. **As of `r format(max(linelist$date_hospitalisation, na.rm=T), "%d %B")` there have been `r nrow(linelist)` cases reported as hospitalized.**

## Summary table of cases by hospital

```{r, echo=F, out.height="75%"}
linelist %>%
  filter(!is.na(hospital)) %>%
  group_by(hospital) %>%
  summarise(cases = n(),
    deaths = sum(outcome == "Death", na.rm=T),
    recovered = sum(outcome == "Recover", na.rm=T)) %>%
  adorn_totals() %>%
  qflextable()
```

## Epidemic curve by age

```{r, echo=F, warning=F, message=F, out.height = "75%", out.width="100%"}
# create epicurve
age_outbreak <- incidence(
  linelist,
  date_index = date_onset,    # date of onset for x-axis
  interval = "week",          # weekly aggregation of cases
  groups = age_cat)

# plot
plot(age_outbreak, n_breaks = 3, fill = age_cat, col_pal = muted, title = "Epidemic curve by age group")
```
~~~

[Source]

What’s RMarkdown?

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



R Markdown script

YAML sets title, date, and output type

Code chunk loads packages and data

Text and in-line code

Code chunk makes table

Headings

Code chunk makes plot

Output (e.g. Word document)

~~~
---
title: "Outbreak Situation Report"
date: "4/24/2021"
output: word_document
---

```{r setup, echo=FALSE}
pacman::p_load(rio, here, tidyverse, janitor, incidence2, flextable)
linelist <- rio::import(here::here("data", "case_linelists", "linelist_cleaned.rds"))
~~~

This report is for the Incident Command team of the fictional outbreak of Ebola cases.
**As of `r format(max(linelist$date_hospitalisation, na.rm=T), "%d %B")` there have
been `r nrow(linelist)` cases reported as hospitalized.**

Summary table of cases by hospital

```{r, echo=F, out.height="75%"}
linelist %>%
  filter(!is.na(hospital)) %>%
  group_by(hospital) %>%
  summarise(cases = n(),
    deaths = sum(outcome == "Death", na.rm=T),
    recovered = sum(outcome == "Recover", na.rm=T)) %>%
  adorn_totals() %>%
  qflextable()
```

Epidemic curve by age

```{r, echo=F, warning=F, message=F, out.height = "75%", out.width="100%"}
# create epicurve
age_outbreak <- incidence(
  linelist,
  date_index = date_onset,   # date of onset for x-axis
  interval = "week",         # weekly aggregation of cases
  groups = age_cat)

# plot
plot(age_outbreak, n_breaks = 3, fill = age_cat, col_pal = muted, title = "Epidemic
curve by age group")
```
```

Outbreak Situation Report

4/24/2021

This report is for the Incident Command team of the fictional outbreak of Ebola cases. **As of 30 April there have been 5888 cases reported as hospitalized.**

Summary table of cases by hospital

hospital	cases	deaths	recovered
Central Hospital	454	193	165
Military Hospital	896	399	309
Missing	1,469	611	514
Other	885	395	290
Port Hospital	1,762	785	579
St. Mark's Maternity Hospital (SMMH)	422	199	126
Total	5,888	2,582	1,983

Epidemic curve by age

Knit
Run
38:32
Chunk 3
R Markdown
outbreak_report - Read-Only - Compatibility Mode
Search
ayout
References
Mailings
Review
View
Zotero
Help
Picture Format
Epidemic curve by age group
Weekly incidence
300
200
100
0
2014-W25
2014-W45
2015-W13
age_cat
0-4
5-9
10-14
15-19
20-29
30-49
50-69
70+
NA
Display Settings
Focus

[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:

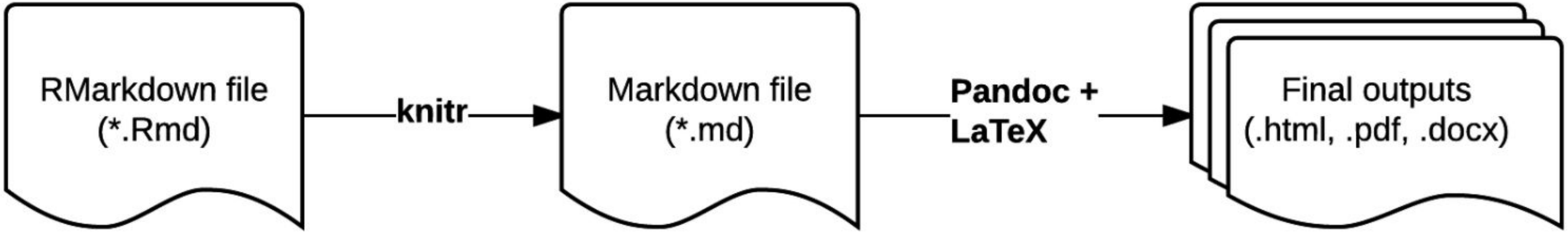
- Interactive report (only html format)
- Ingenuities that can aggregate information, such as Tab / Table display

# 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)



RMarkdown file
(*.Rmd)

**knitr**

Markdown file
(*.md)

**Pandoc +**
**LaTeX**

Final outputs
(.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)

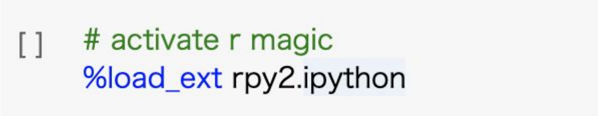
```
library(rmarkdown)

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

Call R from Python

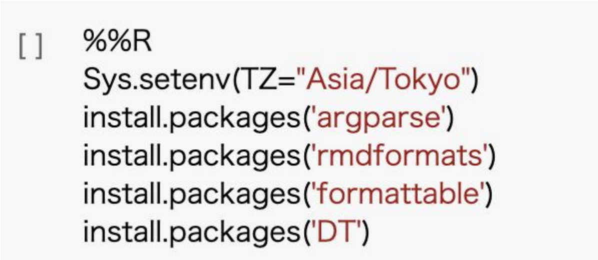
- We can easily call R from Python

```
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')
```

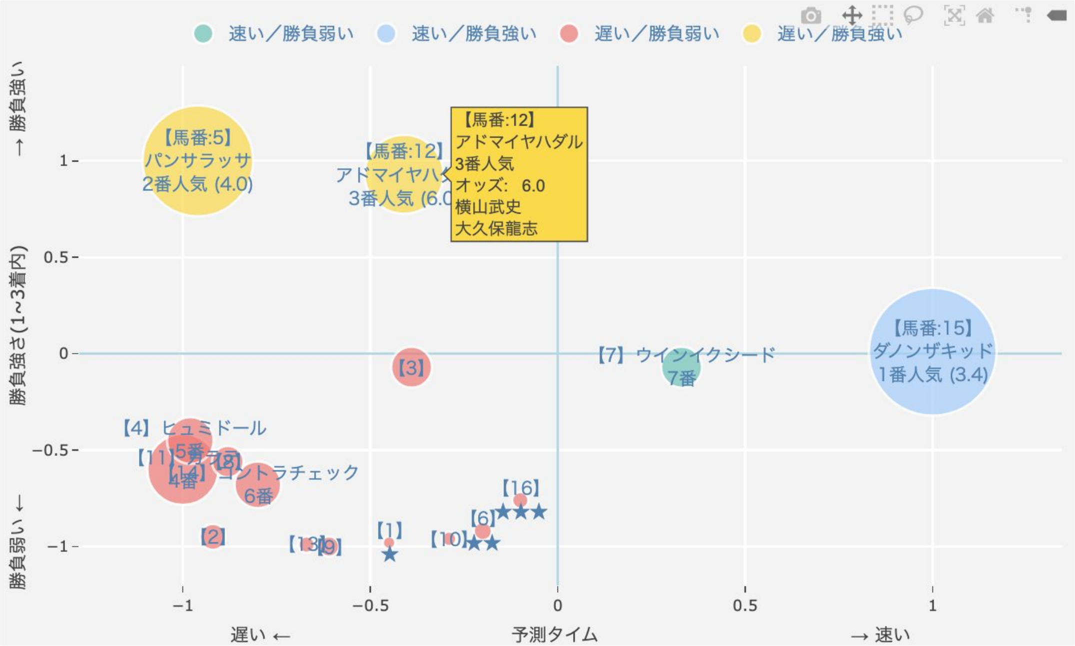
04

Report Examples

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

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

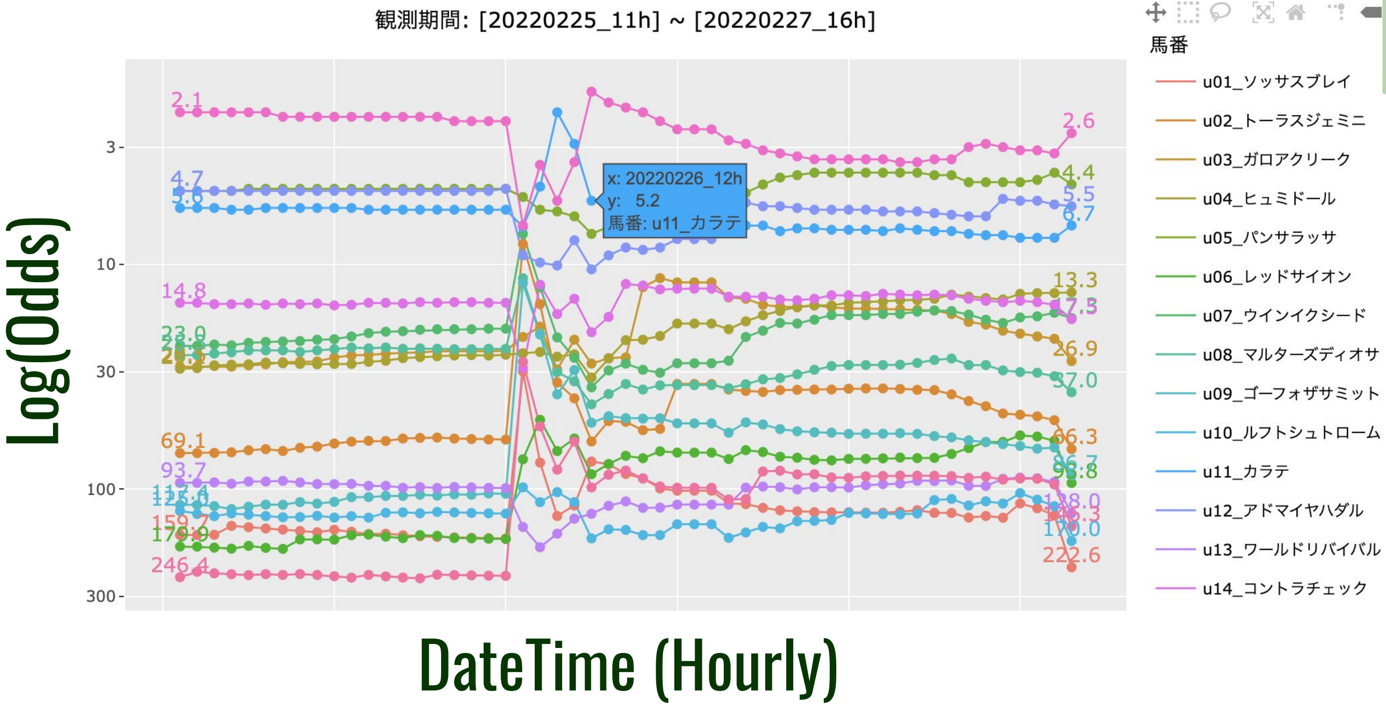
Prediction Chart (Speed/Clutch)



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

Odds information will be updated hourly until just before the race

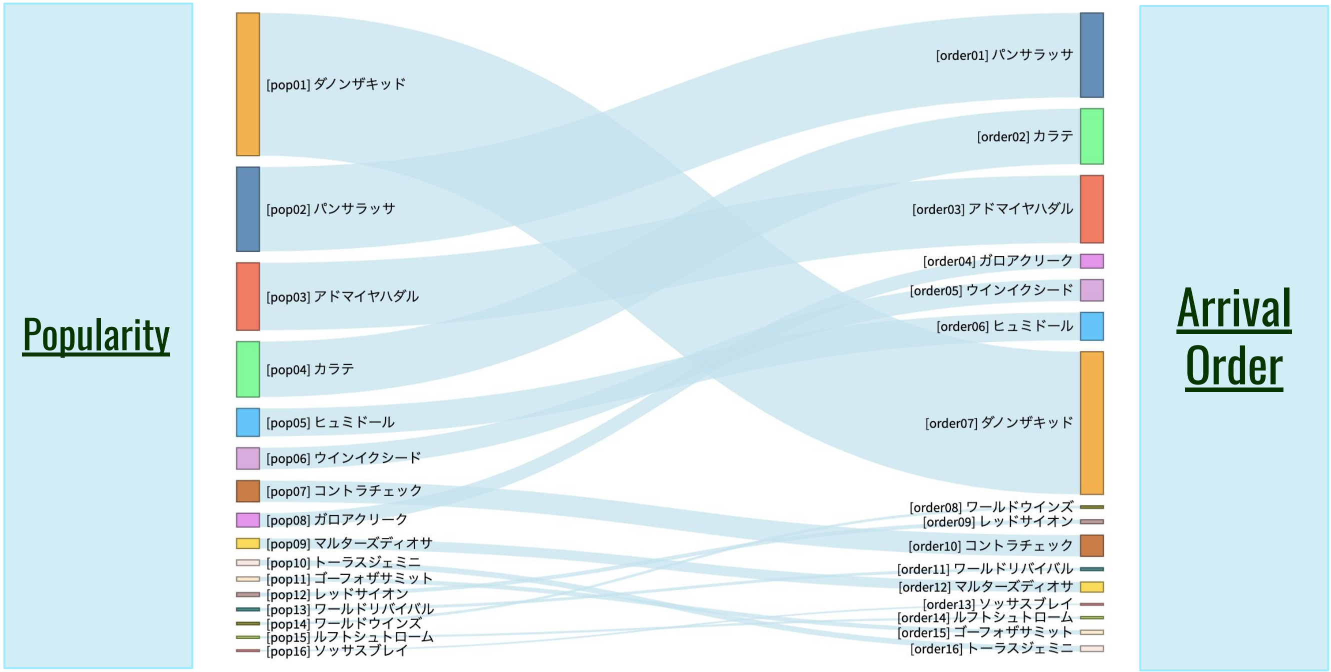
Monitoring Odds (Hourly)



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

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

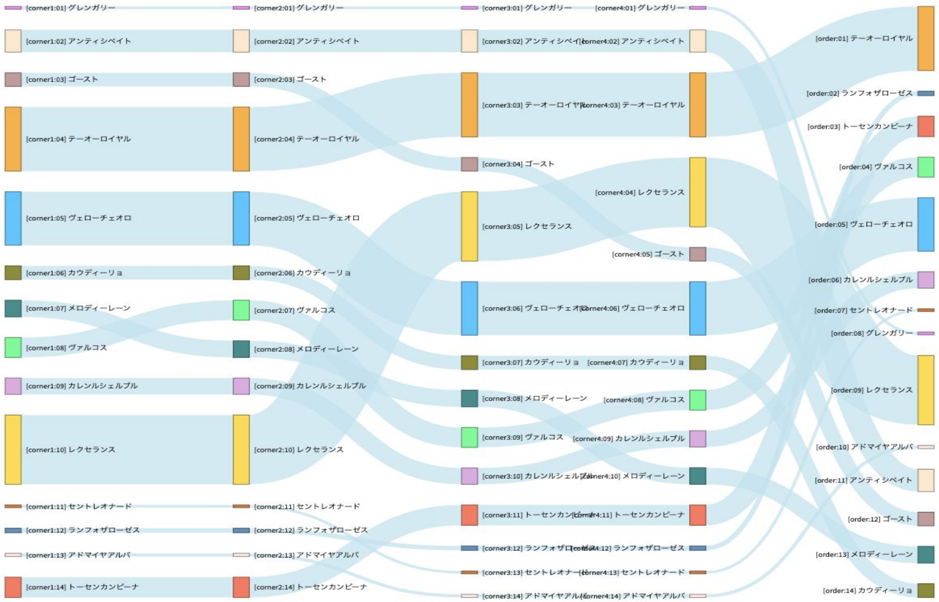
Popularity vs Arrival Order



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

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

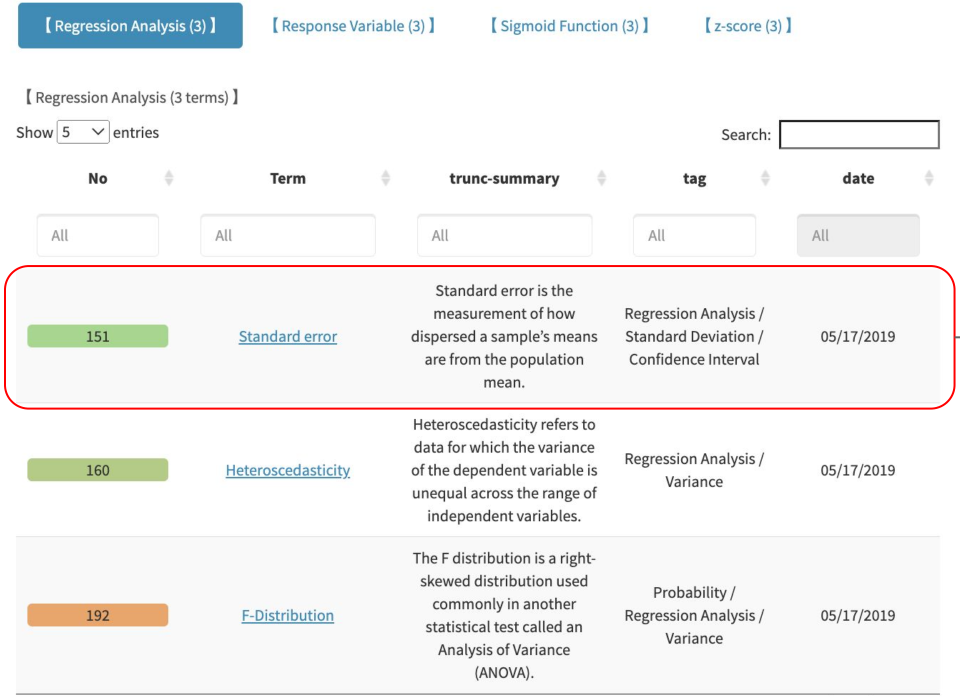
The Order of Passing Corners



1st Corner 2nd 3rd 4th Arrival Order

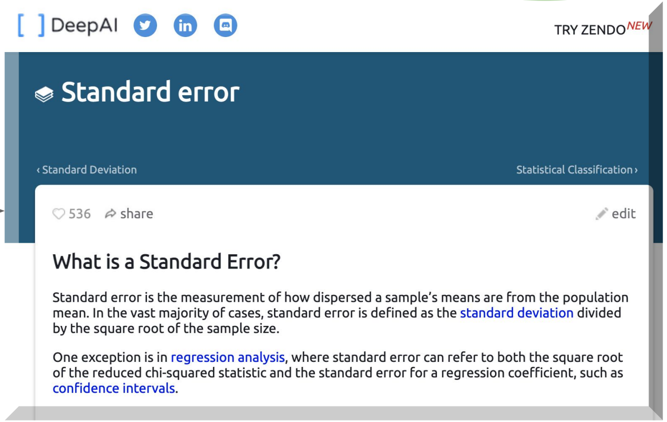
Ex.2) Machine Learning Glossary: [Link]

Create a table report for each category by scraping the glossary


【 Regression Analysis (3) 】
【 Response Variable (3) 】
【 Sigmoid Function (3) 】
【 z-score (3) 】
Show 5 entries
Search:
All
All
All
All
All

【 Regression Analysis (3 terms) 】

No	Term	trunc-summary	tag	date
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


[ ] DeepAI
TRY ZENDO NEW
Standard error
‹ Standard Deviation
Statistical Classification ›
536
share
edit

What is a Standard Error?

Standard error is the measurement of how dispersed a sample's means are from the population mean. In the vast majority of cases, standard error is defined as the standard deviation divided by the square root of the sample size.

One exception is in regression analysis, where standard error can refer to both the square root of the reduced chi-squared statistic and the standard error for a regression coefficient, such as confidence intervals.

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