

Exploratory Data Analysis

可重复性报告 - 作为报告草稿

目录

1 环境	1
1.1 R info	1
1.2 python info	2
2 分析	2
2.1 The Workflow	2
2.2 Import	3
2.3 Tidy	3
2.4 Understand	5
2.5 Communicate	7
3 总结	7
参考文献	7

1 环境

1.1 R info

```
xfun::session_info(  
  packages = c(  
    "readr", "tidyr", "stringr", "dplyr", "purrr",  
    "tidyverse", "lubridate", "mice",  
    "ggplot2", "ggdag", "showtext", "VIM"  
  ), dependencies = FALSE  
)
```

R version 4.1.0 (2021-05-18)

Platform: x86_64-pc-linux-gnu (64-bit)

Running under: Ubuntu 20.04.2 LTS

Locale:

LC_CTYPE=zh_CN.UTF-8 LC_NUMERIC=C
LC_TIME=zh_CN.UTF-8 LC_COLLATE=zh_CN.UTF-8
LC_MONETARY=zh_CN.UTF-8 LC_MESSAGES=zh_CN.UTF-8
LC_PAPER=zh_CN.UTF-8 LC_NAME=C
LC_ADDRESS=C LC_TELEPHONE=C
LC_MEASUREMENT=zh_CN.UTF-8 LC_IDENTIFICATION=C

Package version:

dplyr_1.0.6 ggdag_0.2.3 ggplot2_3.3.3 lubridate_1.7.10
mice_3.13.0 purrr_0.3.4 readr_1.4.0 showtext_0.9-2
stringr_1.4.0 tidyr_1.1.3 tidyverse_1.3.1 VIM_6.1.0

1.2 python info

// TODO

2 分析

2.1 The Workflow

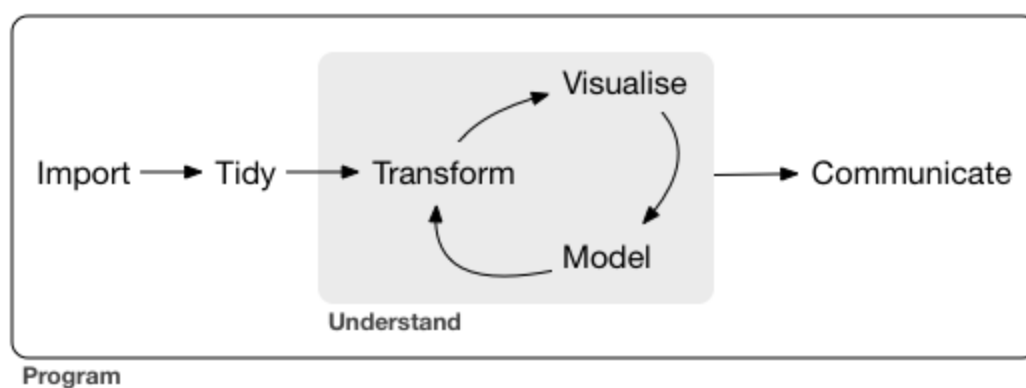


图 1: The Data Science Workflow¹

¹This picture is from [R for Data Science](#) by Hadley Wickham and Garrett Golemund, released under [CC BY-NC-ND 3.0 US](#).

2.2 Import

// 需要数据集的完整描述和获取方式

// TODO - R. Li

2.3 Tidy

```
raw_df <- read_csv("./data/investment/FDI_untidy.csv")

process <- function(raw_df) {
  simplified_df <- raw_df %>%
    filter(X1 %>% str_detect("^\\d")) %>%
    rename(时间 = X1)

  flipped_df <- simplified_df %>%
    pivot_longer(c(-时间), names_to = "observation", values_to = "val")

  stdize <- function(str) {
    str %>%
      str_replace(pattern = "(.*):(总计 | 一带一路)", replacement = "\\1/\\2/\\2") %>%
      str_replace(pattern = ":::", replacement = ":") %>%
      str_replace(pattern = "(.*):(.* 洲):*(.*)", replacement = "\\1/\\2/\\3")
  }

  sep_df <- flipped_df %>%
    mutate(observation = observation %>% stdize()) %>%
    separate(col = "observation", into = c("type", "地区", "国家"), sep = "/")

  df <- sep_df %>% spread(key = "type", value = "val")
}

raw_df %>%
  process() %>%
  write_csv("./data/investment/FDI_tidy.csv")

cont <- raw_df %>%
  filter(X1 == "状态") %>%
  as_vector() %>%
  [. == "继续"] %>%
```

```

names()
raw_df %>%
  select(X1, all_of(cont)) %>%
  process() %>%
  write_csv("./data/investment/FDI_tidy_cont.csv")

raw_df <- read_csv(
  file = "./data/investment/FDI_tidy_cont.csv",
  col_types = cols(
    时间 = col_date(format = "%m/%Y")
  ),
  guess_max = 50000
)

df0 <- raw_df %>%
  filter(!is.na(国家))

# 对外直接投资：非金融类：累计 为一带一路数据所特有
OBOR_col <- " 对外直接投资：非金融类：累计"

df <- df0 %>%
  filter(国家 != " 一带一路" & 国家 != " 总计") %>%
  select(-all_of(OBOR_col))

df <- df %>%
  filter(month(时间) == 12) %>%
  mutate(年份 = as.integer(year(时间)), .keep = "unused", .before = 1) %>%
  filter(年份 >= 2002)

df <- df %>%
  select(names(df) %>% str_subset(pattern = " 投资（和其他）*$", negate = TRUE)) %>%
  filter(!is.na(`对外直接投资：截至累计`))

df %>% write_csv(file = "./data/investment/FDI_useful.csv")

df1 <- df0 %>%
  filter(国家 == " 一带一路" & !is.na(.[OBOR_col])) %>%
  select(时间, all_of(OBOR_col)) %>%
  mutate(

```

```
年份 = as.integer(year(时间)),
月份 = as.integer(month(时间)),
.keep = "unused", .before = 1) %>%
arrange(年份, 月份)

df1 %>% write_csv(file = "./data/investment/FDI_OBOR.csv")
```

2.4 Understand

我们的数据模型非常简单，如图所示：

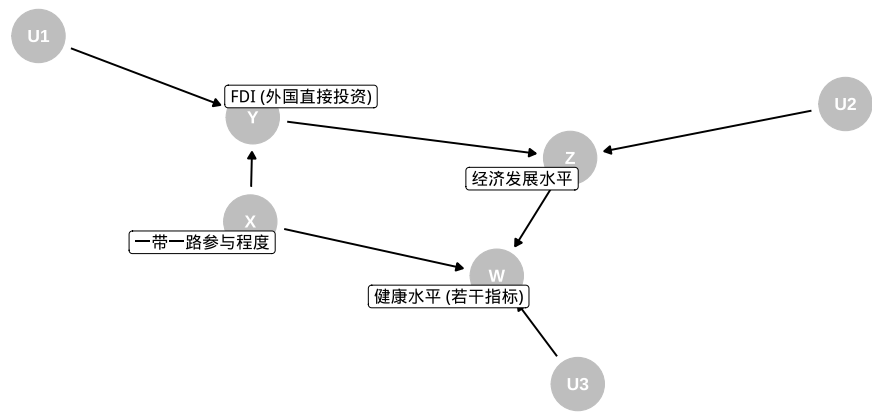


图 2: 数据模型示意图

此图是有向无环图 (Directed acyclic graph, DAG)，边代表因果作用。

我们利用 (Chernozhukov et al., 2021)^[1] 的方法进行分析。

首先注意到数据集中存在许多缺失数据：

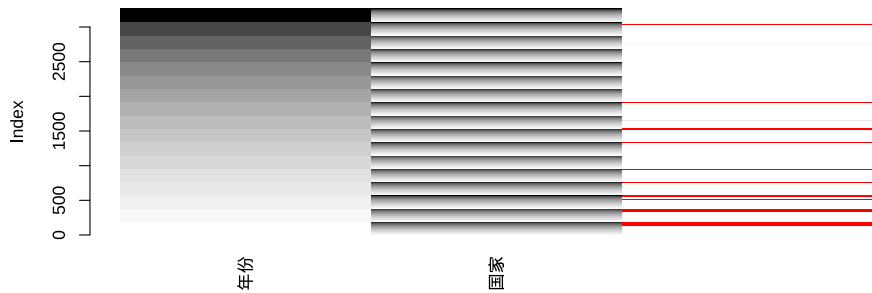


图 3: 缺失数据示意图

使用 linear regression with bootstrap 进行缺失数据填补。

```

fdi <- read_csv(
  file = "./data/investment/FDI_useful.csv",
  col_types = cols(
    年份 = col_double(),
    国家 = col_factor()
  )
) %>% unite(col = 国家, 地区, 国家)

country_name <- fdi[[" 国家"]] %>% unique()

fdi_na <- fdi %>%
  tidyr::complete(年份, 国家) %>%
  rename(对外直接投资 = `对外直接投资：截至累计`)

fdi_lg <- fdi_na %>%
  mutate(lg = log(对外直接投资), .keep = "unused")

fill_a_country <- function(.dt, .cn) {
  res <- .dt %>%
    filter(国家 == .cn) %>%
    mice(method = "norm.boot", m = 1, maxit = 3) %>%
    complete()
  if (any(is.na(res$lg))) {
    non_na <- !(res$lg %>% is.na())
    res$lg <- res$lg[non_na][1]
  }
  return(res)
}

fdi_filled <- country_name %>% map(~fill_a_country(fdi_lg, .x))

result <- fdi_filled %>%
  reduce(rbind) %>%
  mutate(对外直接投资 = exp(lg), .keep = "unused") %>%
  separate(col = 国家, into = c(" 地区", " 国家"), sep = "_")

result %>% write_csv("./data/investment/FDI_filled.csv")

```

2.5 Communicate

// Use **echarts**, maybe [pyecharts](#)?

// TODO - H. Fan

3 总结

参考文献

- [1] CHERNOZHUKOV V, WÜTHRICH K, ZHU Y. An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls[J]. Journal of the American Statistical Association, Taylor & Francis, 2021, 0(ja): 1–44.