



山形大学
Yamagata University

Exploratoryによる 作業内容の見える化と共有

山形大学 学術研究院 教授

藤原 宏司

2022.2.18

Exploratory データサイエンス勉強会 #22

- この発表内容は、発表者個人の見解に基づくものであり、発表者が所属する組織の公式見解ではありません。

Outline

1. 自己紹介
2. 業務の背景
3. 問題点：業務の属人化・継続性
4. 手作業の問題点
5. 我々の課題（やりたいこと）
6. Exploratory導入の理由
7. 公開データ（学校基本調査）の可視化
8. まとめ

自己紹介

- 藤原 宏司 (ふじわら こうじ)
- 専門分野:
 - 米国における大学経営 (Higher Education Administration)
 - IR (Institutional Research*) → 後述します。
 - 応用統計学 (Ph.D.)
- 略歴：米国の大学・短期大学等で、統計解析、IR* および大学評価対応業務に従事。2016年8月から現職。現職では、BI (Business Intelligence) を活用したIRシステムの構築等を担当。「山形大学 IR担当者向け実践プログラム (IR履修証明プログラム)」プログラムディレクター。
- Exploratory歴：< 3ヶ月 (R (S言語) は20年+)

* Investor Relationsや統合型リゾートを意味する「IR」とは違うもの

山形大学 Quick Fact (AY2021)

Current Enrollment

8,720



Undergraduate: 7,398 (% Women: 35.6%)
Graduate: 1,237 (% Women: 22.2%)
Special Program: 40
Non-Degree Seeking: 45

International Student

234



% Women: 34.6%

New First Year Student

1,690



% Women: 37.5%
% In-State: 24.7%

Degrees Awarded

2,153 (AY2020)



Baccalaureate: 1,636
Masters: 468
Doctoral: 49

山形大学 Quick Fact (AY2021)

Job Placement Rate
Undergraduate Alumni

99.5%



Student-Faculty Ratio

11:1



Faculty

788

% Women: 15.7%



Staff

1,455

% Women: 59.9%



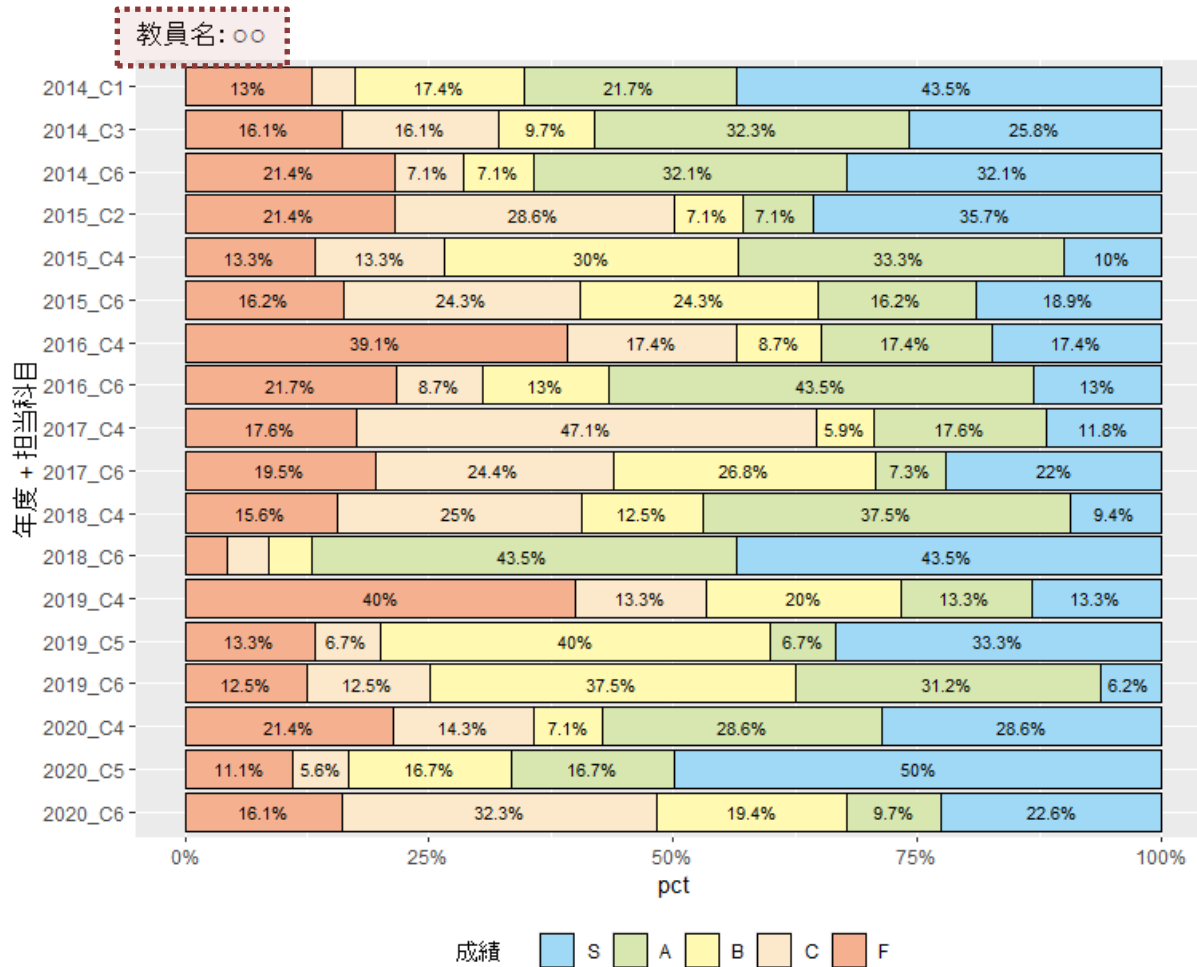
業務の背景（1）

- **Institutional Research（IR）**とは、大学における計画立案、政策形成および意思決定等のサポートをするために必要な「**情報提供**」を目的とした、**学内外データの収集・分析機能**のこと
 - アメリカの大学で生まれ、発展してきた考え方
 - 発表者は、上記で説明した「**IRに関する業務**」を行う**チーム（IRオフィス）**に属しています
 - スタッフ数：4名
（教員2名、専門員1名、事務補佐員1名）

業務の背景 (2)

- 【再掲】 Institutional Research (IR) オフィスは、学内外に存在する様々なデータを収集・分析・可視化することがミッション
- しかし、学内外のデータは「Tidy」ではないものが多く、分析前処理 (i.e., Data Wrangling) のスキルが無いと、分析自体ができないこともある
 - Tidyなデータ = 分析可能な状態になっているデータのこと
- また、日本の大学には職員の定期異動があり、業務の属人化・継続性に問題を抱えている大学も？
→ 教職員におけるデータ処理スキルには、バラツキがある？

分析例：担当科目ごとの成績分布 × 所属教員分



× 500名分

※ 架空のデータを使用しています。

業務の属人化・継続性に関する問題点

データリクエスト

- 教員ごとに「担当している科目群の成績分布グラフ」を作成
→ 教員数 = 500名

どのように作業をするか

- 「その時」の**担当者が使えるツール**で作業
(一般的には、**エクセル**を使って作業することが多い印象)
- その時の担当者が、**R**や**VBA**等を使えた場合は、**作業を自動化**
→ **担当者のスキル次第**
→ **そのようなスキルを持つ担当者が異動した後は？**
→ **グラフ作成業務が継続・再現**できるかは、**次の担当者のスキル次第**

グラフをエクセル（手作業）で作成した場合

その前に

- 「**十分な時間と根性** + **Tidyなデータ**」があれば、例に挙げた「数百のグラフ」は、誰でも作成可能
 - Tidyなデータ = 分析可能な状態になっているデータのこと

しかしながら

- 前述の通り、学内外のデータは、**Tidyなデータではないことが多い**
 - **Tidyなデータを用意することが必要（分析前処理: Data Wrangling）**

エクセル（手作業）の問題点

業務の継続に必要なこと

- 日本の大学のような、異動を前提とした組織において、業務を滞りなく継続させていくためには、作業内容の見える化と共有が必要

エクセル（手作業）の問題点

- 作業内容が保存されないので、他の人が行った作業の理解・把握が困難
- 結果、このような2択になることが多いのでは？
1. 作業内容を理解しないまま作業を行う
(作業のブラックボックス化 → 質の悪い伝言ゲームへの危険性)
 2. 作業工程の作り直し (前任者、他者の知見を活かせない)

我々の課題（やりたいこと）

- データ分析業務に関するツールを、可能な限り**統一**
 - 分析前処理、分析・可視化、レポートニング等

現在、我々が使用しているツール

MS Word	STATA
MS Excel	Tableau
Excel VBA	R
MS Access	RStudio
MS SQL Server	MariaDB
MS Power BI Desktop	Exploratory
MS Power BI Report Server	
MS Power Query	
MS Power Automate	
MS Windows Server	

問題点

- メンバーによって、使えるツールが異なる
- ゆえに、他メンバーが行った作業内容の理解・再現が困難

つまり

- 作業内容の見える化・共有ができておらず、業務継続のリスクがある
- 業務の属人化も問題

分析例のグラフをRで作る場合

問題点

1. Rに関する専門知識が必要となるため、
Rが使える人にしか「その業務」が行えない
→ 業務の属人化
2. 他のスタッフに、Rを勉強する時間がない
→ 業務は他にもたくさんある
3. Rを勉強しても、すぐに別部署への人事異動がある
→ モチベーションの問題
(手作業を続ける理由にはならないと思いますけども…)
→ 個人的には、人事異動制度が日本におけるIT化・DX化の妨げになっている「理由の一つ」と感じています

我々が必要なツール

- 他の方が行った作業内容（工程）を、

簡単に理解・再現できるツール

→ その候補が「Exploratory」

- 導入した主な理由

- Rがベースになっている
- データ処理の過程が、分かりやすく記録できる
- 作業内容を共有、分担できる（プロジェクトの共有）
- 学内限定情報をオンプレミスで公開できる
（コラボレーションサーバーの利用）

Exploratoryを使って
公開データ（学校基本調査）を
可視化してみた



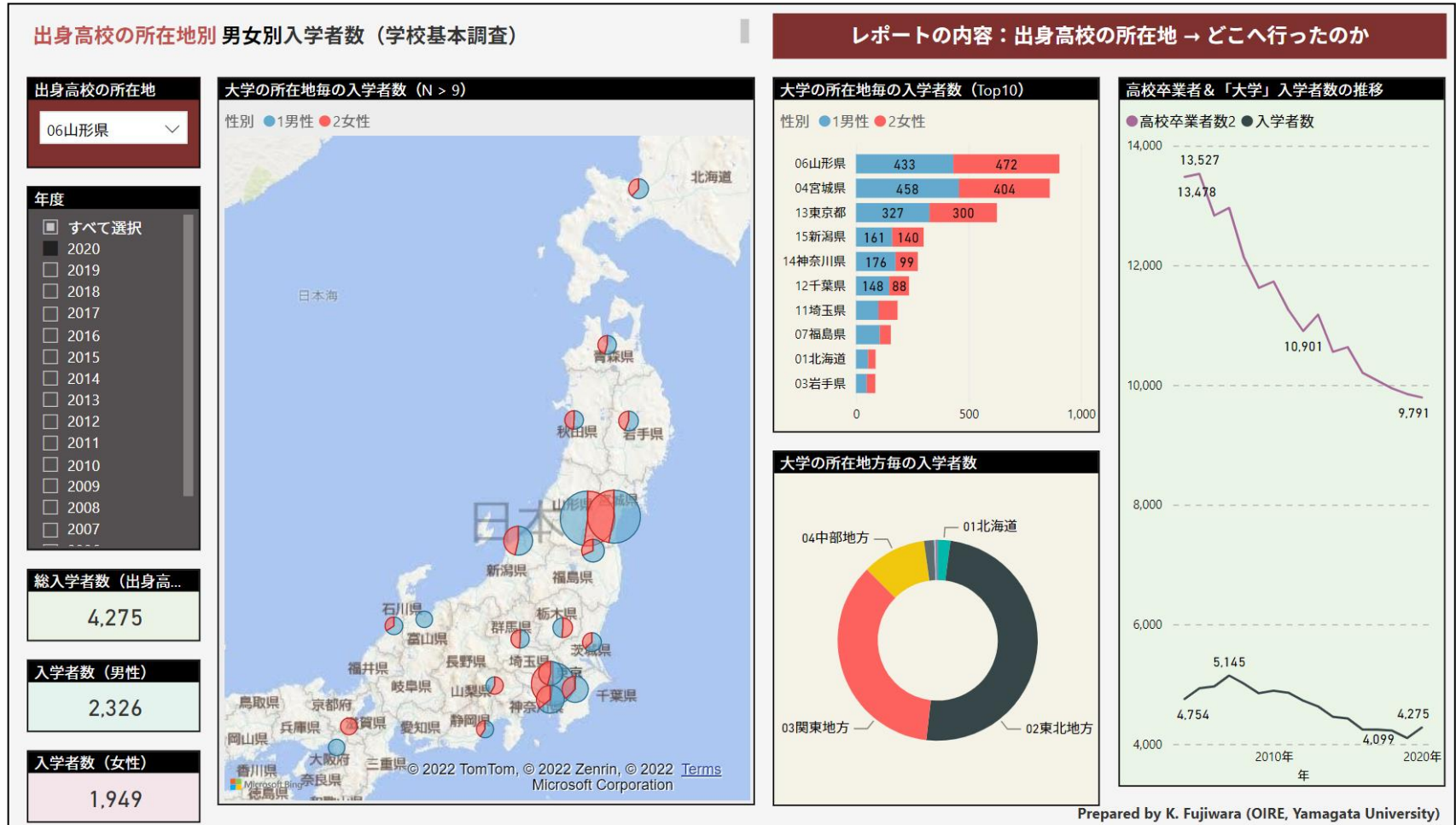
実例：公開データ（学校基本調査）の可視化

データ例：出身高校の所在地県別入学者数（2020年度）

		16 出身高校の所在地 県別 入学者数（8-1）																													
1 計																															
4	出身高校の所在地																													出身高校の所在地	
	大学の所在地	計	北海道	青森	岩手	宮城	秋田	山形	福島	茨城	栃木	群馬	埼玉	千葉	東京	神奈川	新潟	富山	石川	福井	山梨	長野	岐阜	静岡	愛知	大学の所在地					
6	令和2年度	625,003	21,415	5,018	4,681	10,052	3,379	4,275	7,283	14,784	9,155	9,332	35,148	29,698	77,773	43,804	9,400	4,932	5,939	4,076	5,079	9,382	9,695	17,271	39,200	令和2年度					
7	北海道	19,488	14,320	364	272	173	184	87	91	161	99	55	241	209	519	269	115	64	64	22	52	81	59	165	261	北海道					
8	青森	3,455	470	1,881	280	95	207	59	46	27	18	9	28	19	46	21	30	4	3	2	5	16	1	32	15	青森					
9	岩手	2,623	105	222	1,283	257	175	86	64	28	42	12	23	15	47	29	25	3	3	-	6	18	6	18	33	岩手					
10	宮城	12,009	185	660	869	5,679	525	862	1,034	204	183	125	235	120	270	156	156	46	49	12	37	63	11	101	50	宮城					
11	秋田	2,048	45	107	133	73	816	61	45	60	67	50	34	34	74	27	69	17	9	1	7	19	11	56	69	秋田					
12	山形	2,808	80	70	144	567	85	905	183	73	119	31	38	23	55	31	115	4	6	-	9	44	10	55	54	山形					
13	福島	3,334	34	57	96	200	58	155	1,482	276	225	84	63	48	60	36	105	7	2	1	15	48	4	50	30	福島					
14	茨城	7,302	102	57	43	83	42	65	238	3,050	214	117	432	695	575	147	104	41	35	22	47	97	33	148	98	茨城					
15	栃木	4,947	45	71	77	94	59	84	408	523	2,143	307	284	66	98	35	72	13	18	9	43	72	12	57	21	栃木					
16	群馬	7,083	107	67	81	73	42	53	112	243	457	2,856	631	142	293	154	203	70	30	14	45	327	21	125	50	群馬					
17	埼玉	30,023	361	162	166	292	128	185	492	1,240	1,218	1,392	10,811	1,840	5,164	909	668	127	67	30	240	681	58	451	203	埼玉					
18	千葉	27,609	341	191	150	299	141	236	470	2,018	590	350	2,047	10,039	4,828	912	502	101	64	44	243	476	82	682	216	千葉					
19	東京	151,714	2,258	617	517	1,183	409	627	1,469	4,837	2,405	2,325	17,332	13,295	51,659	22,437	1,846	569	501	273	1,594	2,197	547	3,348	1,991	東京					
20	神奈川	46,894	659	209	248	388	158	275	522	1,018	714	680	1,939	2,101	10,133	17,035	799	246	159	94	614	939	169	2,104	654	神奈川					
21	新潟	6,605	83	60	66	131	177	301	329	109	140	200	71	41	113	50	3,484	214	125	31	40	288	30	85	50	新潟					
22	富山	2,568	33	4	14	8	8	12	7	8	17	30	22	16	42	24	88	946	371	114	12	124	127	43	228	富山					
23	石川	6,793	86	17	12	26	22	38	20	40	43	102	45	36	70	47	301	990	2,730	419	22	310	198	187	247	石川					
24	福井	2,454	11	1	1	-	-	2	5	4	4	6	8	4	13	9	24	118	179	1,235	6	43	83	37	256	福井					
25	山梨	4,323	81	27	55	28	22	31	48	96	60	64	120	69	475	266	75	61	27	18	1,282	318	46	346	96	山梨					
26	長野	4,008	50	7	15	13	13	10	21	77	44	111	96	77	169	95	120	112	49	38	152	1,621	115	181	300	長野					
27	岐阜	5,051	40	5	8	5	5	9	2	16	4	6	14	18	27	17	21	101	40	91	9	133	2,158	141	1,377	岐阜					
28	静岡	8,281	109	19	23	30	5	27	38	105	58	79	117	117	214	224	53	39	28	58	139	115	167	4,876	716	静岡					
29	愛知	43,014	220	28	20	60	19	11	17	79	37	47	50	104	139	83	64	339	248	293	85	657	4,557	2,060	27,862	愛知					
30	三重	3,374	17	-	1	4	1	6	4	2	5	5	4	3	17	11	5	10	8	19	7	25	107	87	713	三重					
31	滋賀	7,630	66	5	7	23	3	6	8	18	11	16	27	27	95	38	22	57	81	105	14	47	175	132	459	滋賀					
32	京都	34,470	362	29	37	104	24	24	61	108	73	136	134	158	497	224	134	262	373	443	94	268	303	561	1,206	京都					
33	大阪	56,154	366	31	20	63	15	13	23	90	47	94	97	97	273	123	62	156	328	355	69	131	211	346	661	大阪					
34	兵庫	28,145	237	22	7	34	9	12	9	58	20	30	37	43	136	77	45	87	168	152	53	47	107	183	371	兵庫					
35	奈良	5,140	47	10	2	12	8	6	4	26	8	14	17	14	41	16	16	34	33	28	13	42	36	73	135	奈良					
36	和歌山	1,868	10	-	-	-	-	2	-	2	1	-	3	3	11	7	4	4	4	7	-	1	5	8	16	和歌山					
37	鳥取	1,572	10	-	1	2	-	2	2	6	3	4	2	2	19	7	1	7	10	15	2	4	23	31	53	鳥取					
38	島根	1,692	8	1	2	2	2	-	2	4	3	3	6	7	15	7	6	5	13	10	3	12	13	25	34	島根					
39	岡山	9,566	48	1	4	11	2	4	1	6	8	7	12	7	35	8	7	9	14	27	10	16	26	62	46	岡山					
40	広島	16,437	60	3	3	5	5	4	6	23	14	7	15	21	58	26	7	20	26	37	23	22	46	96	142	広島					

■ 1年分のデータファイル「1つ」に、シートが「8つ」含まれています

レポート例 (with Power BI)



■ 17年分 (2004-2020) のデータをまとめて可視化しています

一例：レポート作成までの流れ

1. データの入手

- a. 17年分のエクセルファイル（17個）をダウンロード
- b. 各年度ごとのエクセルファイルには8つのシートがある

2. Data Wrangling

- a. 各ファイルの各シートごとに、データを**ロング形式**に変形して保存（MS Excel + MS Power Query）
- b. 作業結果の結合・加工処理

3. MS Power BIでレポート作成

簡単そうに見えるけど…

公開データの問題点

- **年度**によって入力フォーマットが異なる場合がある
- **同じファイル**でも、**シート**によって入力フォーマットが異なる場合がある（次項参照）



- 単純なマクロでは処理できないため、汎用性の高い処理方法を考える必要がある
- **一つのアルゴリズムで、全てのデータを処理したい**
(年度・シートごとに、データの処理方法を変えるのは非効率)

シートによって入力フォーマットが違う例

同じファイルにある、異なるシート（セルB4の内容に注目）

	A	B	C	D	E	F	G	H
1								
2								
3	1	計						
4		出身高校の 所在地		計	北海道	青森	岩手	宮城
5		大学の所在地						
6	令和2年度			635,003	21,415	5,018	4,681	10,052

出身高校の 所在地

[空空]出身高校の[空×9]所在地

	A	B	C	D	E	F
1						
2						
3	2	国立				
4		出身高校の 所在地		計	北海道	青森
5		大学の所在地				
6	令和2年度			98,365	4,475	1,292

**出身高校の
所在地**

出身高校の[改行]所在地

Rによるデータ処理の効率化

問題点

- ・ 敷居が高い
- ・ 作業する人を選んでしまう

```
1-----
2  rn(list=ls(all=TRUE))
3
4  options(encoding = "UTF-8")
5
6  library(tidyverse)
7  library(openxlsx)
8  library(readr)
9  library(janitor)
10-----
11
12 # エクセルデータの指定
13 myfile <- "hi0016.xlsx"
14
15 # エクセルのシートをリストに一括読み込み
16 dfs <- myfile %>%
17   excel_sheets() %>%
18   set_names() %>%
19   map(read_excel, path = myfile)
20
21 # 関数の定義
22 myfunction1 <- function(dfs, rownum) {
23   # 空白セルをNAに変換 + 全部NAの列を削除 + 最終列を削除 + 1列目を列名にする + 列名から空白を削除 + 列名から改行を削除
24   d02 <- dfs %>%
25     mutate(across(everything(), ~ifelse(is.na(.), NA, as.character(.)))) %>%
26     select(where(~!all(is.na(.)))) %>%
27     select(-last_col()) %>%
28     row_to_names(row_number = rownum) %>%
29     rename_all(~str_replace_all(.x, "[[:blank:]]", "")) %>%
30     rename_all(~str_replace_all(.x, "[\\r\\n]", ""))
31
32   # 1列目から空白を削除 + 1列目の列名を変更
33   d03 <- d02 %>%
34     mutate(出身高校の所在地 = str_replace_all(出身高校の所在地, "[[:blank:]]", "")) %>%
35     rename(大学の所在地 = 出身高校の所在地)
36
37   # 2列目が空白の行を削除 + 1行目の削除
38   d04 <- d03 %>%
39     drop_na(2) %>%
40     slice(-1)
41
42   # LONG型データの作成
43   out <- d04 %>%
44     pivot_longer(
45       cols = c(2:last_col()),
46       names_to = "出身高校の所在地",
47       values_to = "入学者数"
48     ) %>%
49     mutate(入学者数 = as.integer(入学者数)) %>%
50     filter(出身高校の所在地 != "計")
51 }
52
53 # LONG型データの出力 (全部)
54 df1 <- map_dfr(dfs, myfunction1, rownum = 1, .id = "シート名")
55
```

ExploratoryによるData Wranglingの例

学校基本調査_16出身高校の所在地別入学者数 - hi0016_1_Total_a

学校基本調査_16出身高... 検索

データフレーム

出身高校の所在地	計	北海道	青森	岩手	宮城	秋田	山形	福島
1 大学の所在地	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
4 青森	3,455	470	1,881	280	95	207	59	46
5 岩手	2,623	105	222	1,283	257	175	86	64
6 宮城	12,009	185	660	869	5,679	525	862	1,034
7 秋田	2,048	45	107	133	73	816	61	45
8 山形	2,808	80	70	144	567	85	905	183
9 福島	3,334	34	57	96	200	58	155	1,482
10 茨城	7,302	102	57	43	83	42	65	238
11 栃木	4,947	45	71	77	94	59	84	408
12 群馬	7,083	107	67	81	73	42	53	112
13 埼玉	30,023	361	162	166	292	128	185	492
14 千葉	27,609	341	191	150	299	141	236	470
15 東京	151,714	2,258	617	517	1,183	409	627	1,469
16 神奈川	46,894	659	209	248	388	158	275	522
17 新潟	6,605	83	60	66	131	177	301	329
18 富山	2,568	33	4	14	8	8	12	7
19 石川	6,793	86	17	12	26	22	38	20
20 福井	2,454	11	1	1	0	0	2	5
21 山梨	4,323	81	27	55	28	22	31	48
22 長野	4,008	50	7	15	13	13	10	21
23 岐阜	5,051	40	5	8	5	5	9	2
24 静岡	8,281	109	19	23	30	5	27	38
25 愛知	43,014	220	28	20	60	19	11	17
26 三重	3,374	17	0	1	4	1	6	4
27 滋賀	7,630	66	5	7	23	3	6	8
28 京都	34,470	362	29	37	104	24	24	61

ステップ

1. Excel - ローカル
2. カスタムコマンド
3. カスタムコマンド
4. テキストデータの加工
5. 列名を変更
6. カスタムコマンド
7. 行を選択
8. ホワイト聖からロング聖へ (Gather)
9. カスタムコマンド
10. フィルタ (1)
11. 計算を作成 (Mutate) (2)
12. 計算を作成 (Mutate) (1)

- シートごとにデータフレームを作成する必要があったが、簡単に操作できた (データフレームの複製 → 読み込むシートを変更)

必要なステップ

ステップ

- 1. Excel - ローカル
hi0016.xlsx, sheetIndex = 1, skip = 3
前回のインポート: 2021-10-23 9:48:59 AM
- 2. カスタムRコマンド
select(where(~!all(is.na(.x)))) %>% ...
- 3. カスタムRコマンド
rename(出身高校の所在地 = 1)
- 4. テキストデータの加工
出身高校の所在地 = str_rem...
- 5. 列名を変更
`大学の所在地` = 出身高校の...
- 6. カスタムRコマンド
drop_na(2)

- 7. 行を選択
-1:-1
- 8. ワイト型からロング型へ (Gather)
`出身高校の所在地`, `入学者数`, -大...
- 9. カスタムRコマンド
mutate(入学者数 = as.numeric(入学...
- 10. フィルタ (1)
(is.na(出身高校の所在地) | 出...
- 11. 計算を作成 (Mutate) (2)
row_number = row_number()
出身高校の所在地 = fct_reorder(...
- 12. 計算を作成 (Mutate) (1)
大学の所在地 = fct_inorder(...

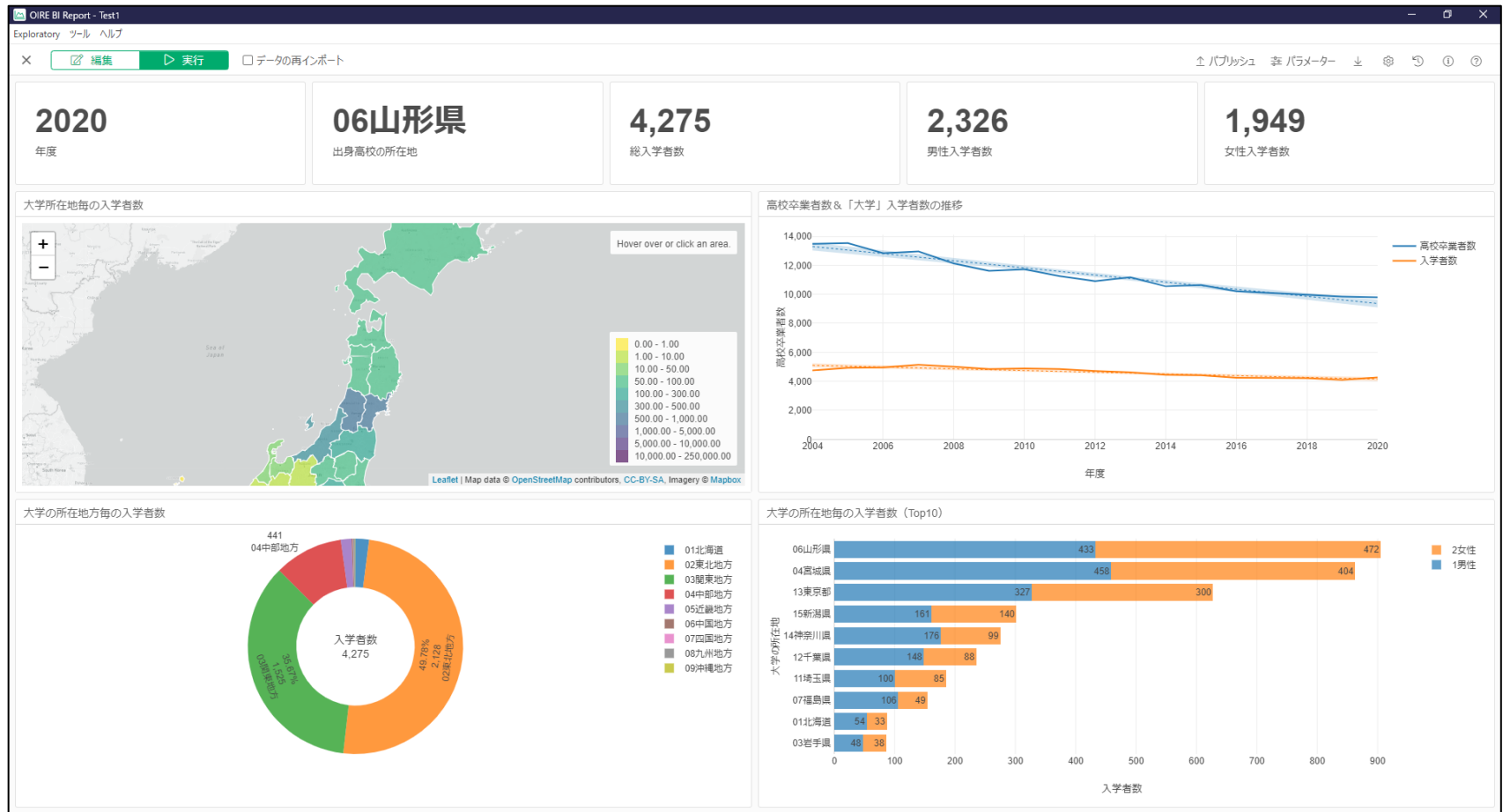
Pros

- ✓ 12ステップで、Tidyなデータの作成が可能
- ✓ 作業時間 = 10分未満 (発表者は初心者)
- ✓ メニューが分かりやすい
- ✓ 実行結果が確認しやすい
- ✓ 「カスタムRコマンド」がとても便利
- ✓ カスタマーサポート

Cons

- 自動処理はできますか？

ダッシュボード by Exploratory



- パーツを予め作成しておいて、キャンバスに配置していく形式
- キャンバスに配置するだけなら、作業時間は5分未満

フィルター（パラメーター）の操作

レイアウト ×

出身高校の所在地 ⊗

06山形県 × ▽

年度 ⊗

2020 × ▽

実行

自動実行



レイアウト ×

出身高校の所在地 ⊗

25滋賀県 × ▽

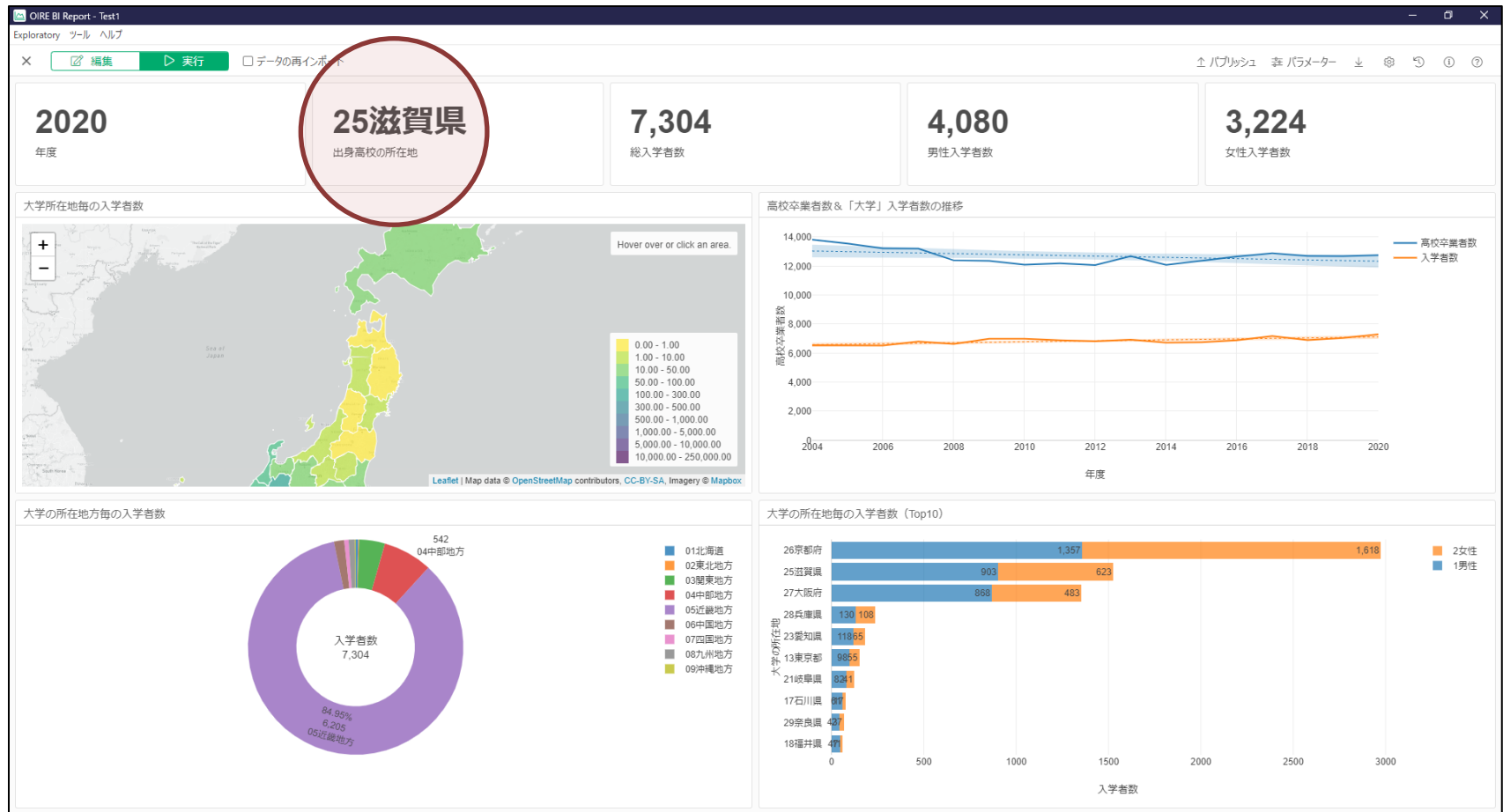
年度 ⊗

2020 × ▽

実行

自動実行

ダッシュボード by Exploratory



■ パラメーター変更後（山形県 → 滋賀県）結果が反映されるまでに、ちょっと時間がかかった。（僕のPC環境の問題かも？）

OS : Windows10 Pro

RAM: 16.0 GB

CPU: Intel(R) Core(TM) i7-10510U

HDD: SSD (1TB)

まとめ

- 本発表では、他の人が行った作業内容（工程）を、簡単に理解・再現できるツールとして導入した「Exploratory」の使用体験を共有
 - a. 初心者でも操作しやすいツールという印象（使う人を選ばないツール）
 - b. カスタマーサービスが素晴らしい！！（Special thanks to 村里さん & 白戸さん）
- 【個人的な意見】 さらに活用していくためには
 - 「R」の基本（i.e., dplyr、stringr等）を勉強すると、作業効率がもっと高まると思いました

THANK YOU!

ANY QUESTIONS, COMMENTS OR SUGGESTIONS?

藤原 宏司 | Koji Fujiwara, Ph.D.

kfujiwara@cc.yamagata-u.ac.jp

