山形大学
Yamagata University

IR活動における データマネジメントの実践
～データを深く理解するために～

山形大学 学術研究院 教授

藤原 宏司

2021.7.27

*IR*初級人材育成研修会@九州大学

今日の内容

- データマネジメントとは？
  IR（Institutional Research）とは？
- IRとデータマネジメント
- 実践例：出席情報のデータマネジメント
- まとめ

自己紹介：藤原 宏司 | Koji Fujiwara
専門はIR（Institutional Research）と応用統計学（Ph.D.）。
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。フロリダ州立大学大学院IR履修証明プログラム修了。

データマネジメントとは？

IR（Institutional Research）とは？

データマネジメントとは？

■ **データマネジメント*とは、分析に必要なデータを**

- **分析可能な形で「速やかに」準備**すること
  （含：データ収集、変形、修正等）
- **将来に備えて整理**すること
  （含：データの管理、更新、追加、修正等）

■ なぜ、IRに「データマネジメント」が必要なのか？

- **意思決定には、高品質なデータ・情報が必要だから**
  （Data Informed Decision Making）

→ **Quality in, quality out**（Garbage in, garbage out）

* Data Wrangling等も含めた広義なものとして考えています

IR（Institutional Research）とは？

■ **IRとは、米国の大学で発展してきた、大学経営や教育改善をSupportする機能**

ポイント

- 日本の大学で生まれた機能・考え方ではない
  - ロ **輸入された考え方**（ゆえに、**人によって解釈が違う…**）
  - ロ 解釈が異なる代表例： GPA、Research、Support

- **IRにおける「Research」**をどのように解釈するか
  - → 米国の大学におけるIR業務や、米国、英国における英単語「Research」の使われ方から考えると、

「ちょっとした調査 & データ収集・可視化機能」

と考えるのが**現実的**

IR（Institutional Research）とは？

- **IRは、Support機能**

  → IRは、裏方・黒子・助演（support act）であり、
  花形・主役（lead role）ではありません

  「IRが大学を変える！」といった考え方は、非現実的

  → IRは、大学執行部や他部署を「Support」しますが、
  大学執行部や他部署は、IRからの「Support」を
  「無条件」で受ける義務はありません

  例：Supportの例 → **意思決定等に必要な情報提供**

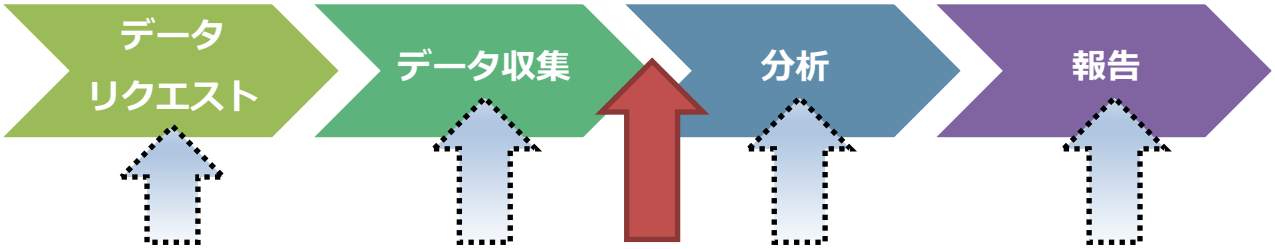
  ロ 英単語のニュアンスの違い

  Helping vs. Supporting（← 同じではない）

IRとデータマネジメント

データ収集と分析の間にある問題点

「情報提供」までの一般的な流れ



今回の問題点

データ収集と分析の間にある代表的な問題点

1. **データの内容に問題**があって、分析ができない
2. **データの形式に問題**があって、分析ができない

データマネジメントが必要になる

※「データが収集できない」「分析手法が分からない」等の問題は、
ここでは取り扱いません

データマネジメントが必要な例

【データの内容に問題がある例】表記ゆれ

氏名	大学名	住所
山本幸一	学校法人明治大学	東京都千代田区
鈴木　達哉	山形大学	山形市小白川町1丁目4番12号
浅野　茂	国立大学法人 山形大学	山形県山形市小白川町1-4-12
藤原 宏司	山形大 法人本部	山形市

主な原因

- データ入力担当者が定期的に変わる（引き継ぎが上手くいっていない）

【データの形式に問題がある例】複数選択式問題の回答データ

ID	併願候補大学	ID2
1	岩手大学;東北大学	00000001
2	岩手大学;東北大学;宮城教育大学	00000002
3	秋田大学	00000003
4	弘前大学;秋田大学	00000004
5	弘前大学;岩手大学;東北大学;秋田大学;福島大学	00000005

なぜ、データマネジメントが重要か？（1）

■ データマネジメントが、
データ分析の**成否**に大きく影響するから

例：下記のデータを用いて、併願候補大学を個別に集計したい

ID	併願候補大学	ID2
1	岩手大学;東北大学	00000001
2	岩手大学;東北大学;宮城教育大学	00000002
3	秋田大学	00000003
4	弘前大学;秋田大学	00000004
5	弘前大学;岩手大学;東北大学;秋田大学;福島大学	00000005
7	弘前大学;岩手大学;東北大学;秋田大学;福島大学	00000007
8	弘前大学;東北大学;宮城教育大学	00000008
9	弘前大学;東北大学;宮城教育大学;福島大学	00000009
10	弘前大学;宮城教育大学;秋田大学	00000010

「このまま」だと分析できない（データ形式の問題）
→ 「併願候補大学」列の処理

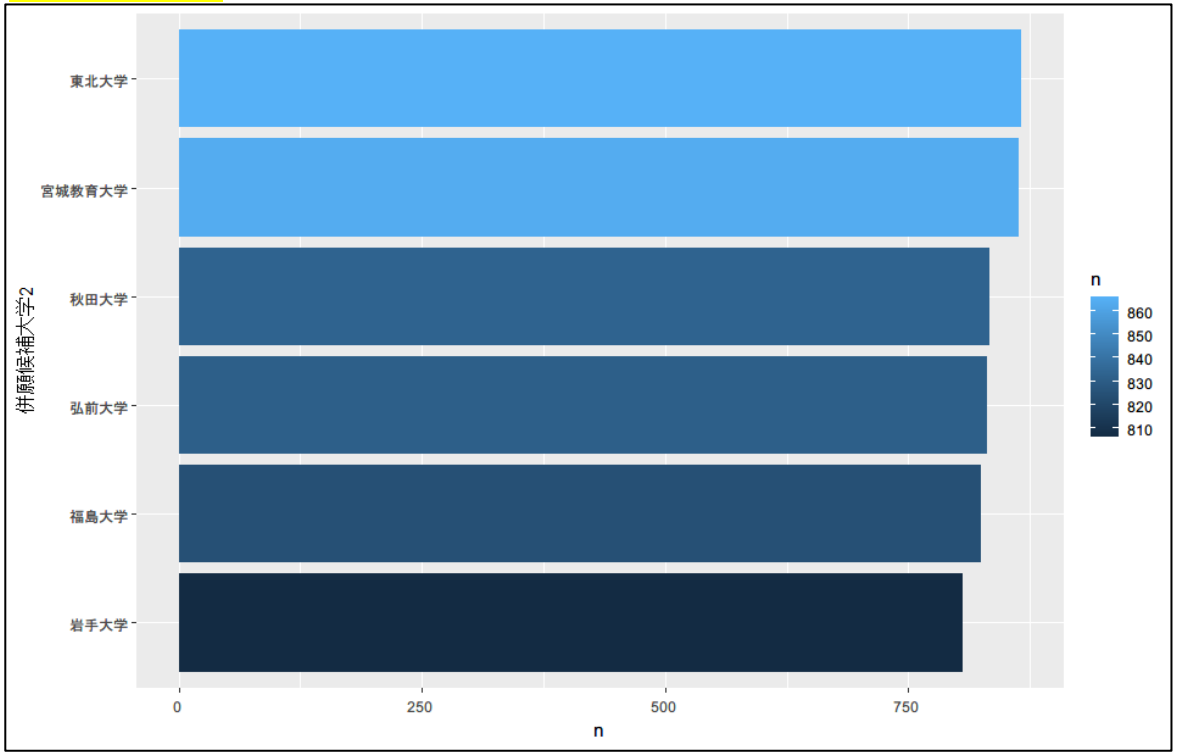
データマネジメント後のデータ

分析可能なデータ形式

ID2	併願候補大学2
00000001	岩手大学
00000001	東北大学
00000002	岩手大学
00000002	東北大学
00000002	宮城教育大学
00000003	秋田大学
00000004	弘前大学
00000004	秋田大学
00000005	弘前大学
00000005	岩手大学
00000005	東北大学
00000005	秋田大学
00000005	福島大学

（Rで処理）

可視化例



他の例：名寄せ

- イベント1のイベント2における**効果検証**を行いたい

データ数：イベント1 = 5,000人 & イベント2 = 2,000人

データ例（一部）

イベント1		イベント2	
名前	生年月日	名前	生年月日
藤原　宏司	YYMMDD	藤原　宏司	YYMMDD
高橋　〇〇	S490906	高橋　〇〇	S490906
高崎　××	S511213	高﨑　××	S511213
山崎　△△	S530426	山﨑　△△	S530426
斉藤　□□	S600919	齋藤　□□	S600919
山形　三郎	H021114	山形　三郎	S540628
鈴木　達哉	S1M1D1	浅野　茂	S2M2D2

「このまま」だと分析できない（データにおける内容の問題）
→ 「名寄せ」情報の追加

名寄せ情報の追加

分析可能なデータ

イベント1	イベント2	合致率	名寄せ結果
藤原宏司_YYMMDD	藤原宏司_YYMMDD	100.0%	同じ
高橋〇〇_S490906	高橋〇〇_S490906	94.4%	同じ
高崎××_S511213	高﨑××_S511213	94.4%	同じ
山崎△△_S530426	山﨑△△_S530426	94.4%	同じ
斉藤□□_S600919	齋藤□□_S600919	94.4%	同じ
山形三郎_H021114	山形三郎_S540628	71.5%	別人
鈴木達哉_S1M1D1	高﨑××_S511213	62.4%	別人

データマネジメント

- 「合致率」列の作成（Rで計算）
- 「名寄せ結果」列の作成：94.4%を閾値として、それ以上を「同一人物」とした

データ分析

→ **「名寄せ結果」列を集計・可視化**
（ここまで来ると簡単）

他の例：住所データの処理

■ 学生がどこから通っているか、「**市町村ごとに**」集計したい
（対象学生数：**9,000人**）

データ例（一部）

大学名	住所
山形大学	山形県山形市小白川町1-4-12
山形県立保健医療大学	山形県山形市上柳260番地
山形県立米沢栄養大学	山形県米沢市通町6-15-1
東北芸術工科大学	山形県山形市上桜田3-4-5
東北公益文科大学	山形県酒田市飯森山3-5-1
東北文教大学	山形県山形市片谷地515
山形県立米沢女子短期大学	山形県米沢市通町6丁目15-1
東北文教大学短期大学部	山形県山形市片谷地515
羽陽学園短期大学	山形県天童市大字清池1559

注意

- 「大学名」 ＝ 「学生氏名」と読み替えてください
- 住所データの収集方法がそもそも「間違っている」点については不問とします

Rで処理

大学名	住所	都道府県	市町村	番地以降
山形大学	山形県山形市小白川町1-4-12	山形県	山形市	小白川町1-4-12
山形県立保健医療大学	山形県山形市上柳260番地	山形県	山形市	上柳260番地
山形県立米沢栄養大学	山形県米沢市通町6-15-1	山形県	米沢市	通町6-15-1
東北芸術工科大学	山形県山形市上桜田3-4-5	山形県	山形市	上桜田3-4-5
東北公益文科大学	山形県酒田市飯森山3-5-1	山形県	酒田市	飯森山3-5-1
東北文教大学	山形県山形市片谷地515	山形県	山形市	片谷地515
山形県立米沢女子短期大学	山形県米沢市通町6丁目15-1	山形県	米沢市	通町6丁目15-1
東北文教大学短期大学部	山形県山形市片谷地515	山形県	山形市	片谷地515
羽陽学園短期大学	山形県天童市大字清池1559	山形県	天童市	大字清池1559

なぜ、データマネジメントが重要か？（2）

- データマネジメントが、データ分析および提供する情報の**quality**に大きく影響するから
  - Quality in, quality out
    （データマネジメントの質 → 情報提供の質）

- データ分析の質を上げるには？
  → **多様なデータや分析項目を準備**する必要がある
  → ．．．（後述）

- 多様なデータや分析項目が準備できなかったら？
  → データから得られる示唆の数が減る
  → データに対する理解が浅くなる

実践例：出席情報のデータマネジメント

あるクラスの出席情報（ランダム生成：N = 1,000）

ID	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15
1	出	出	出	欠	出	出	出	出	出	出	出	出	出	欠	出
2	出	出	欠	出	出	欠	出	出	出	欠	欠	出	出	欠	出
3	出	出	出	出	出	欠	欠	出	出	出	出	欠	出	出	出
4	出	出	出	出	出	出	出	出	欠	出	出	出	出	出	欠
5	欠	出	出	欠	出	出	欠	欠	出	出	出	出	出	出	出
6	出	出	出	出	出	出	出	出	出	出	出	出	欠	出	出
7	出	出	出	出	出	出	出	出	出	欠	出	欠	出	欠	欠
8	出	出	出	出	出	出	出	出	出	出	欠	出	出	出	欠
9	出	欠	欠	出	欠	欠	欠	出	出	出	出	出	出	出	出
10	出	出	出	欠	出	出	出	出	欠	欠	出	出	出	欠	出
11	出	出	出	欠	出	欠	出	出	出	欠	出	出	欠	出	出
12	出	出	出	出	出	出	出	欠	出	出	出	欠	出	出	出
13	欠	出	出	出	出	出	欠	出	出	出	欠	欠	出	出	出
14	出	出	欠	出	出	出	出	出	出	出	出	出	出	欠	欠
15	出	欠	出	出	欠	出	出	出	出	出	出	出	出	欠	出
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
991	出	出	欠	出	出	出	欠	出	出	出	出	出	出	出	出
992	欠	出	出	欠	出	出	欠	出	出	出	出	出	出	出	出
993	出	出	欠	出	出	出	欠	出	出	欠	出	出	出	欠	出
994	欠	出	出	出	出	出	出	出	出	出	出	出	欠	出	出
995	出	出	出	出	出	出	出	出	出	出	出	出	出	出	出
996	欠	欠	出	出	出	出	出	欠	出	出	欠	出	出	出	欠
997	出	出	出	出	出	欠	欠	出	欠	欠	欠	出	欠	欠	出
998	出	欠	欠	出	出	出	欠	出	出	出	出	出	欠	出	出
999	欠	出	出	欠	出	出	出	出	出	出	出	出	出	出	出
1000	欠	出	欠	出	出	出	欠	出	出	欠	出	出	欠	出	出

✓ ID = 学生番号, C01 = 1回目の授業における出席状況, … , C15 = 15回目の授業における出席状況

この出席情報データから何を求めますか？

■ 代表例：**出席回数**、**欠席回数**、**出席率**、**欠席率**

```
# A tibble: 1,000 x 6
      ID 授業回数 出席回数 欠席回数 出席率 欠席率
   <dbl>    <int>    <int>    <int>  <dbl>  <dbl>
 1     1       15       13        2   0.87   0.13
 2     2       15       10        5   0.67   0.33
 3     3       15       12        3   0.8    0.2
 4     4       15       13        2   0.87   0.13
 5     5       15       11        4   0.73   0.27
 6     6       15       14        1   0.93   0.07
 7     7       15       11        4   0.73   0.27
 8     8       15       13        2   0.87   0.13
 9     9       15       10        5   0.67   0.33
10    10       15       11        4   0.73   0.27
# ... with 990 more rows
```

→ 分析例：これらの「**指標**」と「**成績**」を**比較**する

ここで分析を止めていませんか？

他の分析項目を考えよう

- **例：欠席のタイミング**
  a. **一番最初に休んだ**授業回は？
  b. どこの時点で、**通算欠席**が「*x*回 = 5回」に達する？

初回欠席回

```
# A tibble: 967 x 2
      ID 初回欠席回
   <dbl> <chr>
 1     1 C04
 2     2 C03
 3     3 C06
 4     4 C09
 5     5 C01
 6     6 C13
 7     7 C10
 8     8 C11
 9     9 C02
10    10 C04
# ... with 957 more rows
```

通算欠席（５回）到達回

```
# A tibble: 168 x 2
      ID 通算欠席5到達回
   <dbl> <chr>
 1     2 C14
 2     9 C07
 3    16 C13
 4    21 C15
 5    33 C09
 6    41 C14
 7    48 C15
 8    49 C09
 9    89 C10
10    94 C15
# ... with 158 more rows
```

上記を求めるには、データマネジメントが必要

データの型：ワイド型 vs. ロング型

- **ワイド型**：
  → 1人1行の横に広いデータ型
  → 直感的に理解しやすい

ワイド型の例

ID	C01	C02	C03	C04	C05
1	出	出	出	欠	出
2	出	出	欠	出	出
3	出	出	出	出	出

ロング型の例

ID	授業回	出席状況
1	C01	出
1	C02	出
1	C03	出
1	C04	欠
1	C05	出
2	C01	出
2	C02	出
2	C03	欠
2	C04	出
2	C05	出
3	C01	出
3	C02	出
3	C03	出
3	C04	出
3	C05	出

- **ロング型**：
  → 1人複数行の縦に長いデータ型
  → データ分析・可視化ツール等で扱いやすいデータ型

ロング型のデータが作れると？

ID	授業回	出席状況
1	C01	出
1	C02	出
1	C03	出
1	C04	欠
1	C05	出
1	C06	出
1	C07	出
1	C08	出
1	C09	出
1	C10	出
1	C11	出
1	C12	出
1	C13	出
1	C14	欠
1	C15	出
2	C01	出
2	C02	出
2	C03	欠
2	C04	出
2	C05	出
2	C06	欠
2	C07	出
2	C08	出
2	C09	出
2	C10	欠
2	C11	欠
2	C12	出
2	C13	出
2	C14	欠
2	C15	出

① フィルター処理

出席状況 ＝ “欠”

② 「欠席回数」列の追加

③ IDごとに連番を振る

ID	授業回	出席状況	欠席回数
1	C04	欠	1
1	C14	欠	2
2	C03	欠	1
2	C06	欠	2
2	C10	欠	3
2	C11	欠	4
2	C14	欠	5

一番最初に休んだ授業回 → 欠席回数 ＝ 1

ID	授業回	出席状況	欠席回数
1	C04	欠	1
2	C03	欠	1

通算欠席５回目の授業回 → 欠席回数 ＝ 5

ID	授業回	出席状況	欠席回数
2	C14	欠	5

他に、何を知りたい？何が分かりそう？

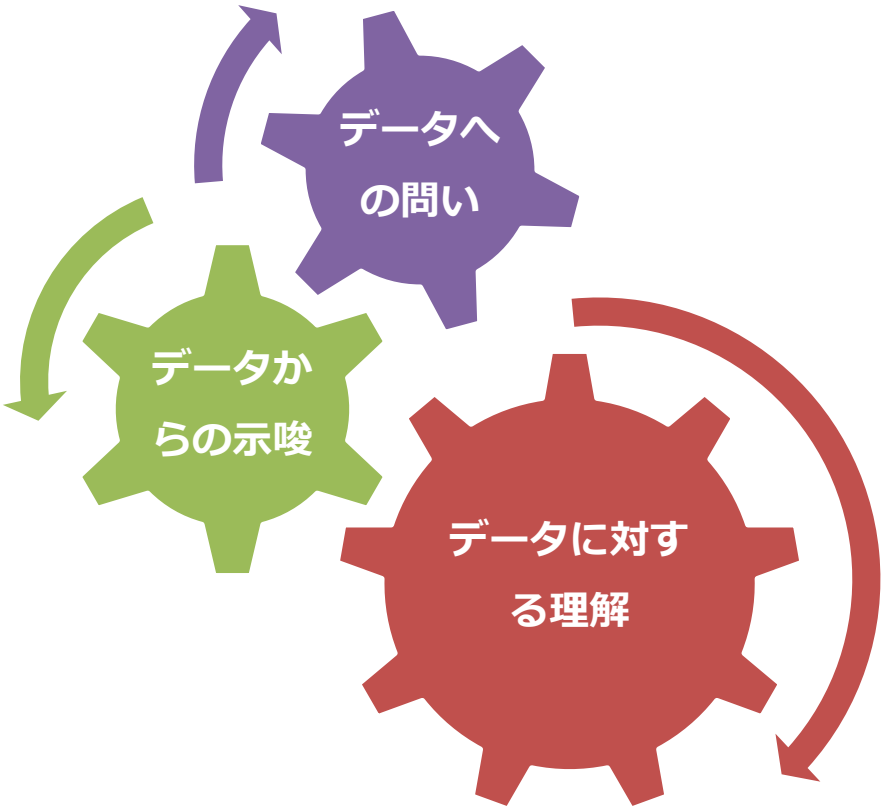
- **例：連続欠席をした学生に関する分析（vs. 成績）**

アイディア出し（データへの問い）

a. 連続欠席をした学生を特定したい
b. 学期前半に連続欠席をした学生を特定したい
c. 頑張って出席していたけど、
   ラストで息切れした学生を特定したい
d. x回以上連続欠席した学生を特定したい
e. 連続欠席を複数回行った学生を特定したい, etc.

【ちょっと脱線】Quality in, quality out

■ データ分析の質を上げるには？
→ 扱うデータを熟知する → データ探索が重要



データに投げかける問いの多さ

＝データから得られる示唆の数

＝データに対する理解度の深さ

＝ 「RQ（問い）」の質の高さ

＝提供する情報の質の高さ

気をつけて欲しいこと
データを理解していないまま
「RQ」を立てて分析をしても、
質の高い情報提供は期待できません

各学生群の抽出：a, b & c

a. 連続欠席をした学生

```
# A tibble: 380 x 1
      ID
   <dbl>
 1     2
 2     3
 3     5
 4     7
 5     9
 6    10
 7    13
 8    14
 9    16
10    20
# ... with 370 more rows
```

b. 連続欠席（学期前半）

```
# A tibble: 219 x 1
      ID
   <dbl>
 1     3
 2     5
 3     9
 4    22
 5    33
 6    40
 7    44
 8    46
 9    49
10    55
# ... with 209 more rows
```

- **学期前半:**
  第８回目までと定義

c. ラストで息切れ

```
# A tibble: 16 x 1
      ID
   <dbl>
 1    92
 2   198
 3   256
 4   409
 5   458
 6   474
 7   519
 8   521
 9   624
10   626
# ... with 6 more rows
```

- **ラストで息切れ：**
  C10まで連続出席 +
  C11-C15で連続欠席アリ

各学生群の抽出：d & e

d. 3回以上連続欠席をした学生

```
# A tibble: 95 x 4
      ID 授業回 出席状況   連続
   <dbl> <chr>  <chr>      <int>
 1     9 C07    欠             3
 2    16 C10    欠             3
 3    33 C10    欠             4
 4    48 C15    欠             3
 5    89 C04    欠             3
 6    94 C04    欠             3
 7   109 C09    欠             3
 8   118 C08    欠             3
 9   119 C14    欠             3
10   122 C07    欠             3
# ... with 85 more rows
```

■ **重複処理の条件：**
３回以上の連続欠席を複数回した場合、
最初の連続欠席情報を反映

e. 連続欠席を複数回行った学生

```
# A tibble: 55 x 2
      ID 連続欠席_複数回
   <dbl>           <int>
 1     9               2
 2    89               2
 3   106               2
 4   109               2
 5   116               2
 6   119               2
 7   122               2
 8   133               2
 9   134               2
10   143               2
# ... with 45 more rows
```

■ **max(連続欠席_複数回) = 3**

```
# A tibble: 3 x 2
     ID 連続欠席_複数回
  <dbl>           <int>
1   350               3
2   494               3
3   997               3
```

データマネジメント：何をしたのか？

ID	授業回	出席状況	連続
5	C01	欠	1
5	C02	出	1
5	C03	出	2
5	C04	欠	1
5	C05	出	1
5	C06	出	2
5	C07	欠	1
5	C08	欠	2
5	C09	出	1
5	C10	出	2
5	C11	出	3
5	C12	出	4
5	C13	出	5
5	C14	出	6
5	C15	出	7
9	C01	出	1
9	C02	欠	1
9	C03	欠	2
9	C04	出	1
9	C05	欠	1
9	C06	欠	2
9	C07	欠	3
9	C08	出	1
9	C09	出	2
9	C10	出	3
9	C11	出	4
9	C12	出	5
9	C13	出	6
9	C14	出	7
9	C15	出	8

① **ロング型**のデータを準備する。（**データ変形**）

② **「連続」列の追加**

a. 「ID」と「出席状況（**そのまま**）」の組み合わせごとに、**連番を振る**。

■ **フィルター処理**
条件：出席状況 = “**欠**” & 連続 > **1**

ID	授業回	出席状況	連続
5	C08	欠	2
9	C03	欠	2
9	C06	欠	2
9	C07	欠	3

■ **「ID」の重複処理** → 「a」等のデータ
■ 「授業回」のフィルター処理 → 「b」
■ 「連続 > **2**」のフィルター処理 → 「d」

・**「c」および「e」のデータ作成**に関しては、もう少し複雑なデータ処理が必要となります。よって、講演時間の関係上、割愛します。

データ処理後：分析用データの例

ID	授業回数	出席回数	欠席回数	出席率	欠席率	初回欠席回	通算欠席5到達回	連続欠席	前半連続欠席	息切れ	連続3欠席以上	連続欠席_複数回
1	15	13	2	0.87	0.13	C04						
2	15	10	5	0.67	0.33	C03	C14	1				
3	15	12	3	0.8	0.2	C06		1	1			
4	15	13	2	0.87	0.13	C09						
5	15	11	4	0.73	0.27	C01		1	1			
6	15	14	1	0.93	0.07	C13						
7	15	11	4	0.73	0.27	C10		1				
8	15	13	2	0.87	0.13	C11						
9	15	10	5	0.67	0.33	C02	C07	1	1		1	2
10	15	11	4	0.73	0.27	C04		1				
11	15	11	4	0.73	0.27	C04						
12	15	13	2	0.87	0.13	C08						
13	15	11	4	0.73	0.27	C01		1				
14	15	12	3	0.8	0.2	C03		1				
15	15	12	3	0.8	0.2	C02						
⋮							⋮			⋮		
990	15	14	1	0.93	0.07	C03						
991	15	13	2	0.87	0.13	C03						
992	15	12	3	0.8	0.2	C01						
993	15	11	4	0.73	0.27	C03						
994	15	13	2	0.87	0.13	C01						
995	15	15	0	1	0							
996	15	10	5	0.67	0.33	C01	C15	1	1			
997	15	8	7	0.53	0.47	C06	C11	1	1		1	3
998	15	11	4	0.73	0.27	C02		1	1			
999	15	13	2	0.87	0.13	C01						
1000	15	10	5	0.67	0.33	C01	C13					

まとめ

まとめ（1）

- **IR担当者の役割: 意思決定者のSupport**

  → **エラーが無いデータを手際よく準備**して、
  効率的にデータ分析と情報提供を行う必要があります

- データマネジメントとは、分析に必要なデータを
  a. **分析可能な形で「速やかに」準備**すること
  b. **将来に備えて整理**すること

- なぜ、データマネジメントが重要か？
  - データ分析の**成否**に影響するから
  - データ分析および情報提供の**質**にも影響するから

  データ探索も重要
  データを理解していないまま、分析を止めていませんか？
  （「RQの設定 → 分析」の前提が「データの理解」）

まとめ（2）

■ 幸せになるために

- **IR業務は地味**な物です
  （日本では、誤解している人が本当に多い）

- 「当たり前」と思われることを、
  「当たり前だったね」と確認する作業が大多数となります

- なので、

  「情報提供をした」 ＝ 「業務を遂行した」

  と考えるようにしましょう

  （そう考えないと、徒労に終わってばかりになります）

まとめ（3）

- 幸せになるために（contd.）
  - データマネジメントは、（地味なIR機能を支える）
    さらに地味な作業です
  - コピペ等の「苦行（手作業）」は止めて、
    「R」等の「ツール」を使って**楽をしましょう**
- How to improve data management skills

Thank You!

Any Questions, Comments or Suggestions?

藤原 宏司｜Koji Fujiwara, Ph.D.

kfujiwara@cc.yamagata-u.ac.jp

山形大学
Yamagata University