

一般社団法人情報サービス産業協会主催



Rstudioによるデータハンドリング  
(データ集計・データ可視化)  
のご紹介

2017年11月27日



自己紹介

---

truestar

鈴木 貴士

学部時代は西洋近代史を学ぶ  
好きな歴史上の人物はユリウス・カエサル



2017年3月 筑波大学  
ビジネス科学研究科 経営システム科学専攻  
(修士課程・夜間) 修了  
現在, 同大学院の博士課程在籍中

現在, truestarにおいて, 広告投資と売  
上の関係を説明する新しい方法を検討中

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答

# Rとは何でしょう？

- 統計分析向けプログラミング言語



```
R Console

R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力$

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復帰します]

> print("Hello World")
[1] "Hello World"
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2   setosa
2           4.9           3.0           1.4           0.2   setosa
3           4.7           3.2           1.3           0.2   setosa
4           4.6           3.1           1.5           0.2   setosa
5           5.0           3.6           1.4           0.2   setosa
6           5.4           3.9           1.7           0.4   setosa
```

## 「R本体」

- Rの本体。これだけでも十分使用可能

## Rコマンドー

- プログラミング無しで**Rの一部機能**を使えるようにしたソフト

## Studio<sup>®</sup>

- Rの統合開発環境。Rを使いやすくする
- 通常はRstudioを用いる



```
R Console

R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力$

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復帰します]

> print("Hello World")
[1] "Hello World"
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2   setosa
2           4.9           3.0           1.4           0.2   setosa
3           4.7           3.2           1.3           0.2   setosa
4           4.6           3.1           1.5           0.2   setosa
5           5.0           3.6           1.4           0.2   setosa
6           5.4           3.9           1.7           0.4   setosa
```

➤ Rの本体。これだけでも十分使用可能

# Rコマンダー

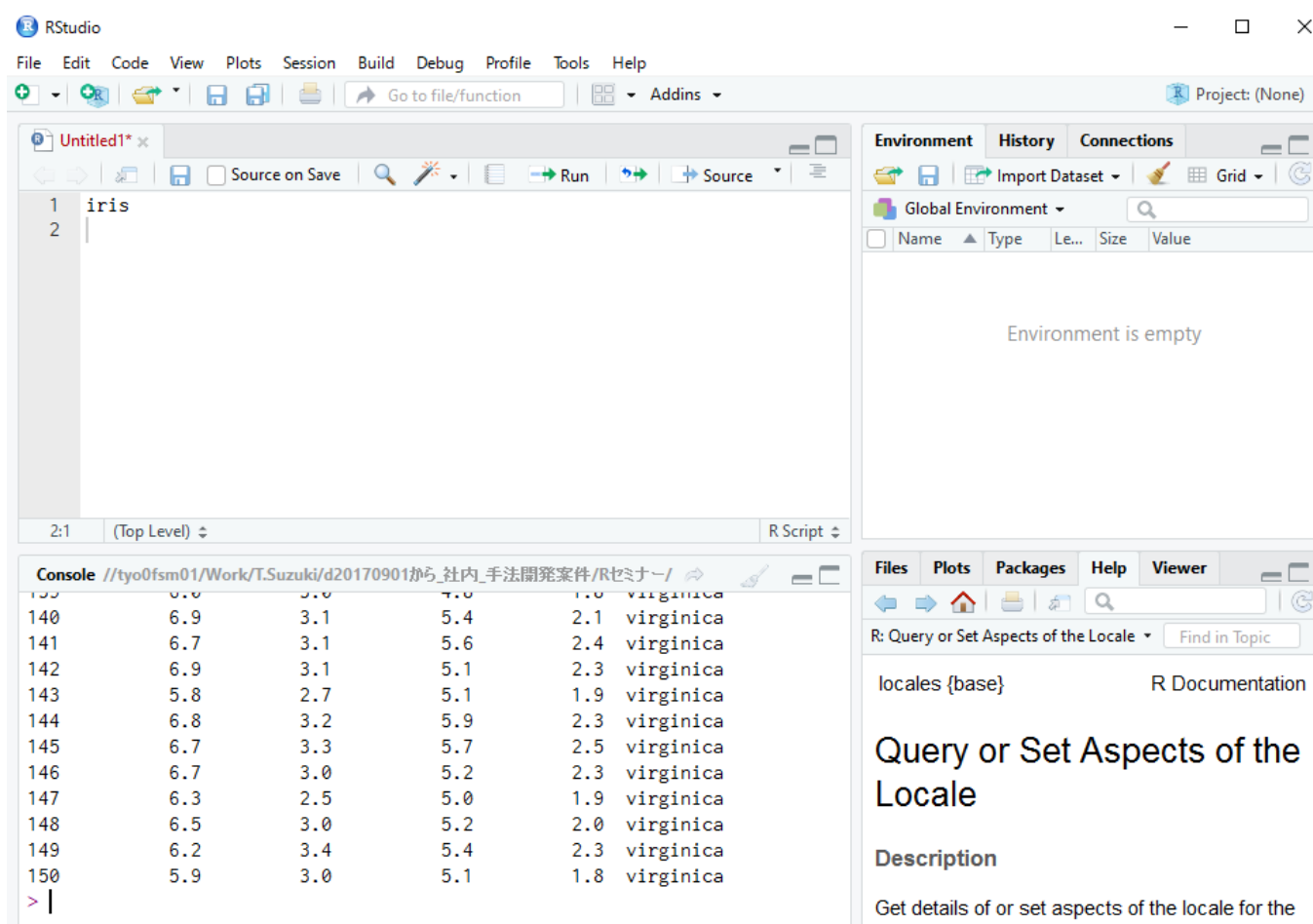


無料

➤ **プログラミング無し**で**Rの一部機能**を使えるようにした  
パッケージ



# RStudio



無料

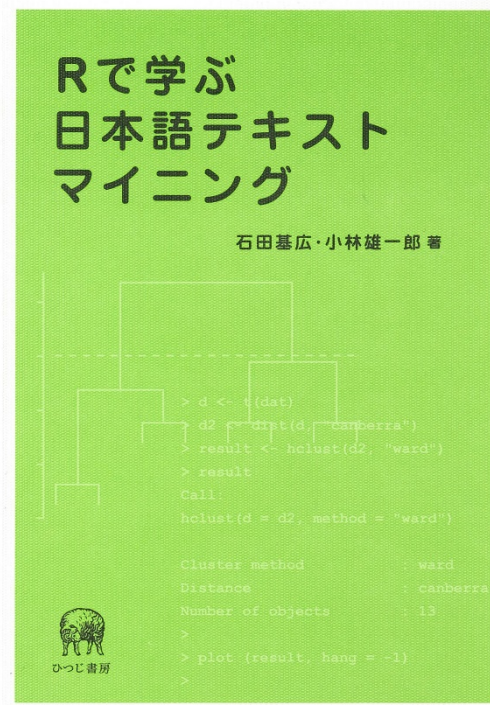
➤ Rの統合開発環境。Rを使いやすくする

➤ 通常はRstudioを用いる

**データ分析**に使いたいから

- **多様な統計分析**をおこなえる
- **集計・作図**のプロセスもおこなえる
- プログラムコードを残せるため、**再現性**がある
- 高機能にも関わらず**無料**で使うことができる
- 日本語で情報が取得できる

# なぜRを使うのか①：多様な統計分析をおこなえる



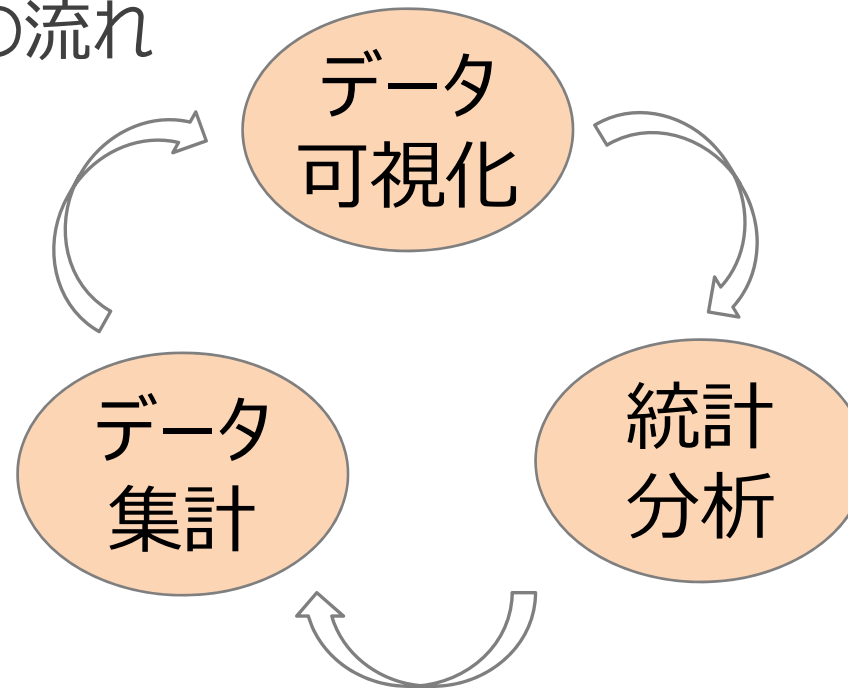
- 日々世界中で**新たなパッケージが開発**され、Rの機能は**進化し続けている**（Rは統計分析の**スタンダード**）

➤ **多種多様な統計分析**をおこなうことができる

## なぜRを使うのか②：集計・作図のプロセスもおこなえる

---

- データ分析の流れ



- データ分析の工程の6～8割はデータの「集計」「可視化」といった**データハンドリング**のプロセス
  - Rは「集計」「可視化」も得意（後ほどデモいたします）
  - Rを使えば、全工程含めた**一気通貫したデータ分析**が可能

## なぜRを使うのか③：再現性・繰り返しに強い

---

- Rはプログラミングコードを残せる

- 繰り返しに強い

- データを入れかえれば，他の分析に流用できる

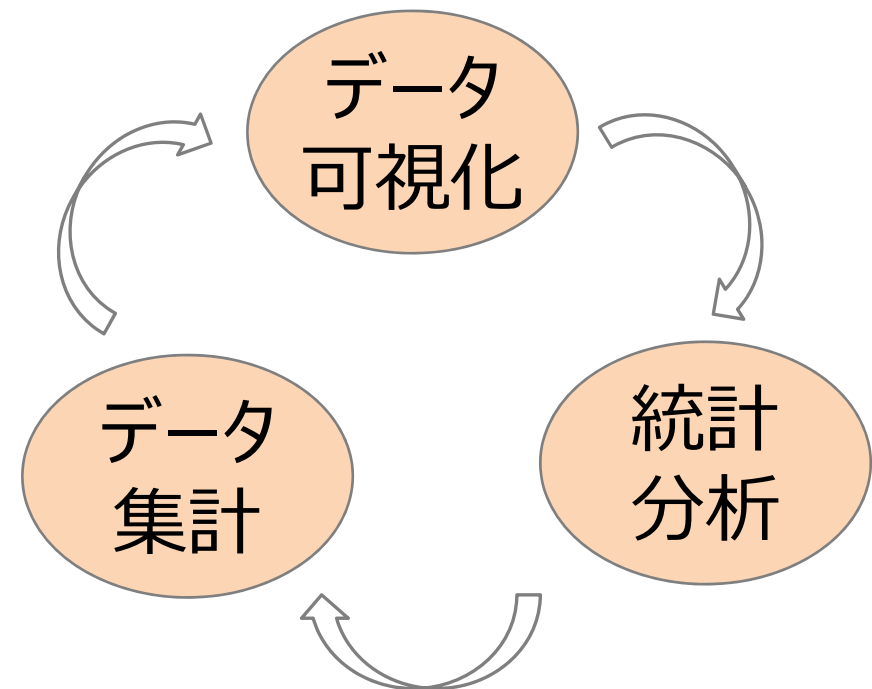
- 分析ミスがあった場合に追跡可能

- 実際のデータ分析は繰り返しが多数起こるため，上記は非常に大切

- ✓ データが間違っていた

- ✓ データ範囲を変えて再分析

- ✓ 同じ形式のデータが毎月来る

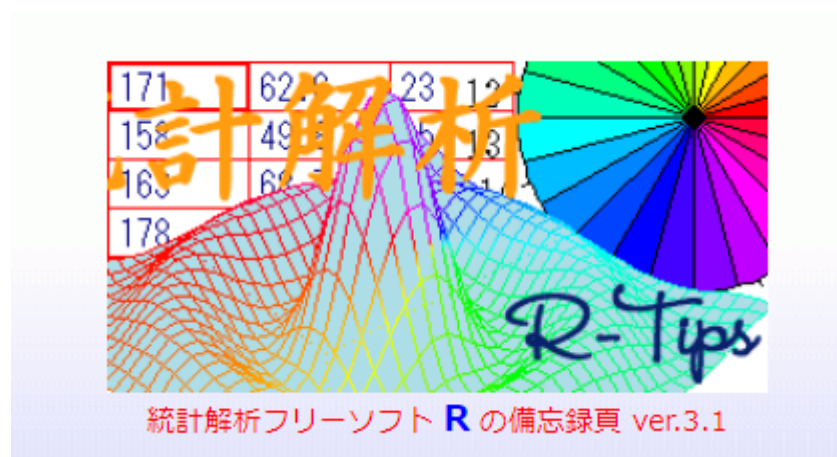
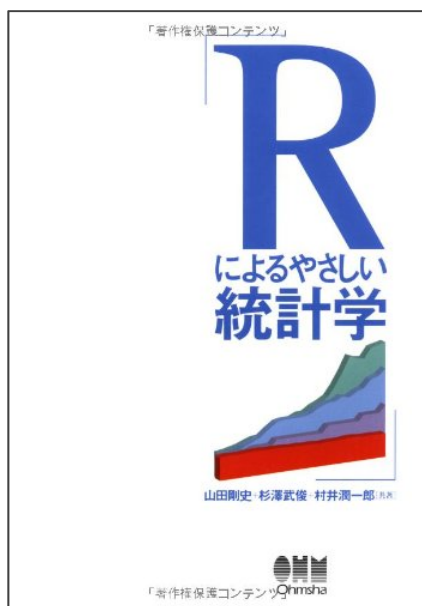


なぜRを使うのか④：高機能にも関わらず**無料**で使うことができる

---

- 統計分析ソフトで有名なIBM「**SPSS**」の価格は**数十万円**
- **Rは無料**なので、機能追加があっても問題なし
  - SPSSは更新に費用がかかる
- Rは無料だが、**信頼性は問題ない**
  - オープンゆえに、間違いがあったらすぐに修正される
    - 新パッケージの場合は検証が不十分という面も…
  - 学術論文にも多数使用されている

## なぜRを使うのか⑤：日本語で情報が入手できる






<http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>

- Rは**統計分析**における**スタンダード**な存在なので、**日本語**書籍・ウェブ情報が**豊富**



## なぜRを使うのか：R以外のソフトウェアは？

			
価格	無料	数十万円～	1万円
操作方法	<u>プログラミング</u>	<u>マウス操作</u>	
学習コスト	△	○	◎
データ前処理・図示化	◎	△	△
統計分析	◎	○	×～△

- 複雑な分析や繰り返し処理をしないなら「**Excel**」
- ある程度複雑な分析をおこないたいが、プログラミングは避けたいなら「**SPSS**」
- 分析に留まらずシステム開発も含めるなら「**python**」

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

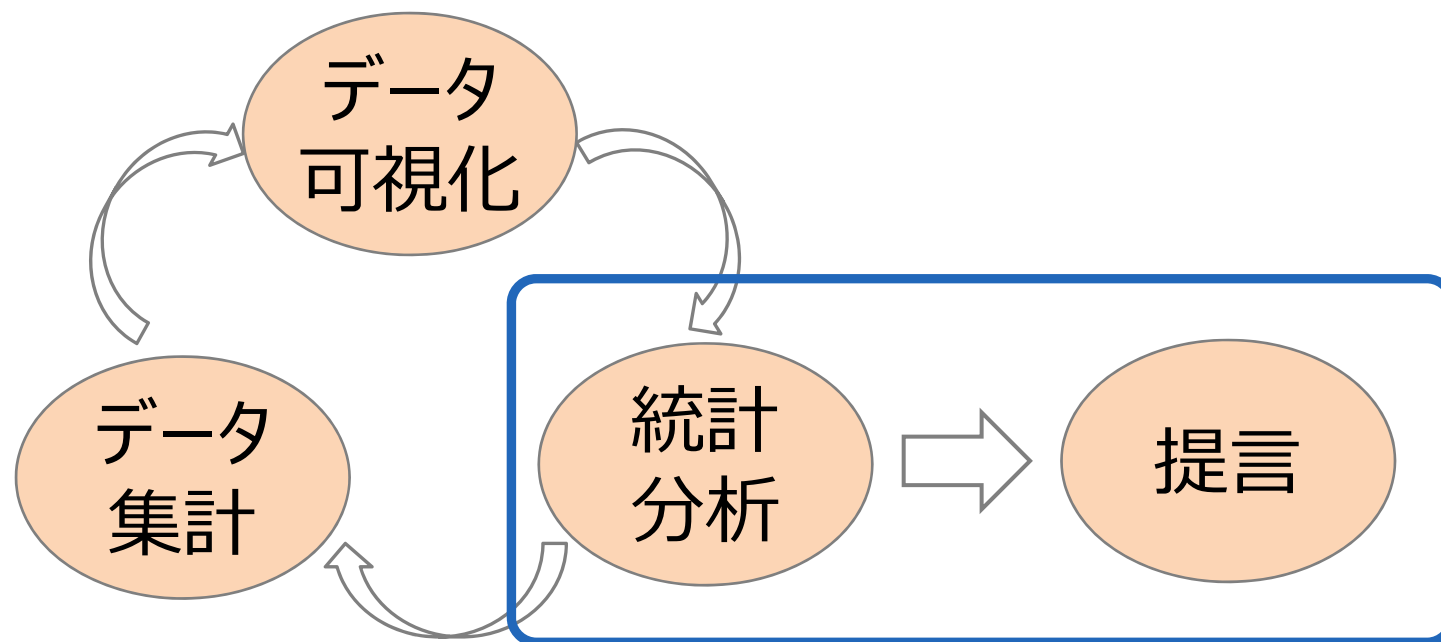
総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答

## データ分析の流れ

---



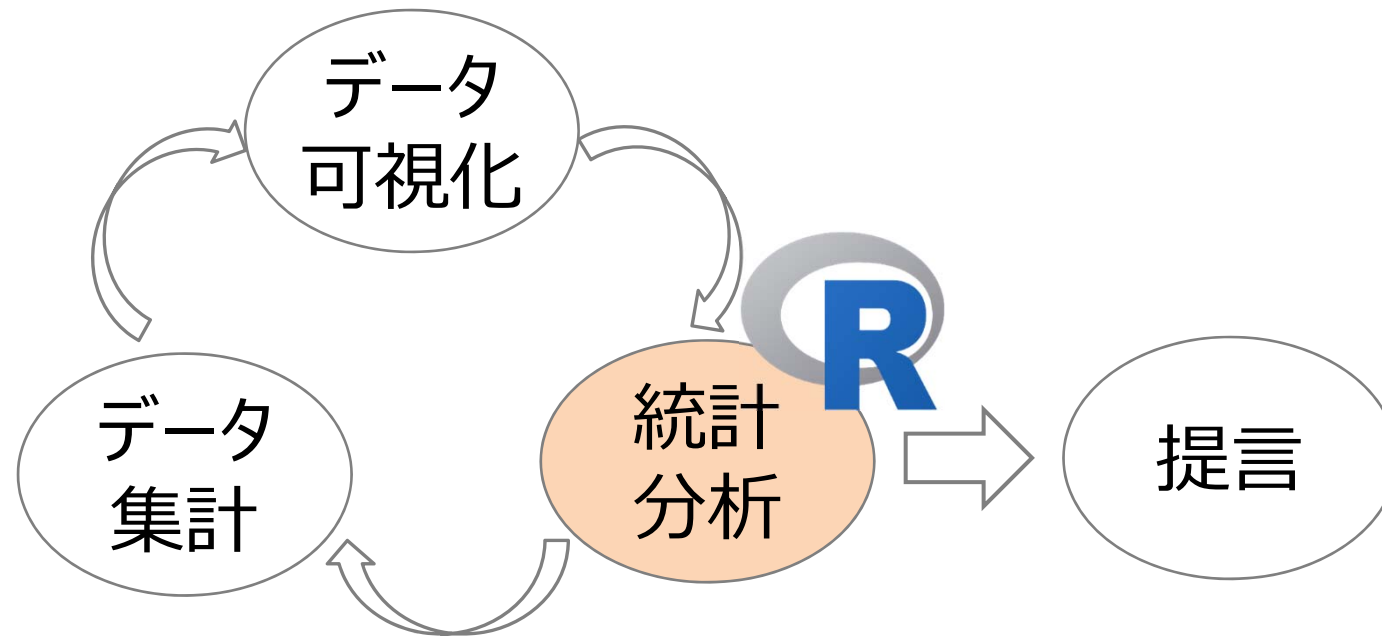
- データ分析の工程の**6～8割**はデータの「**集計**」「**可視化**」といった**データハンドリング**のプロセス

↓ けれども…

- 普通, 興味があるのは「**提言**」と「**統計分析**」

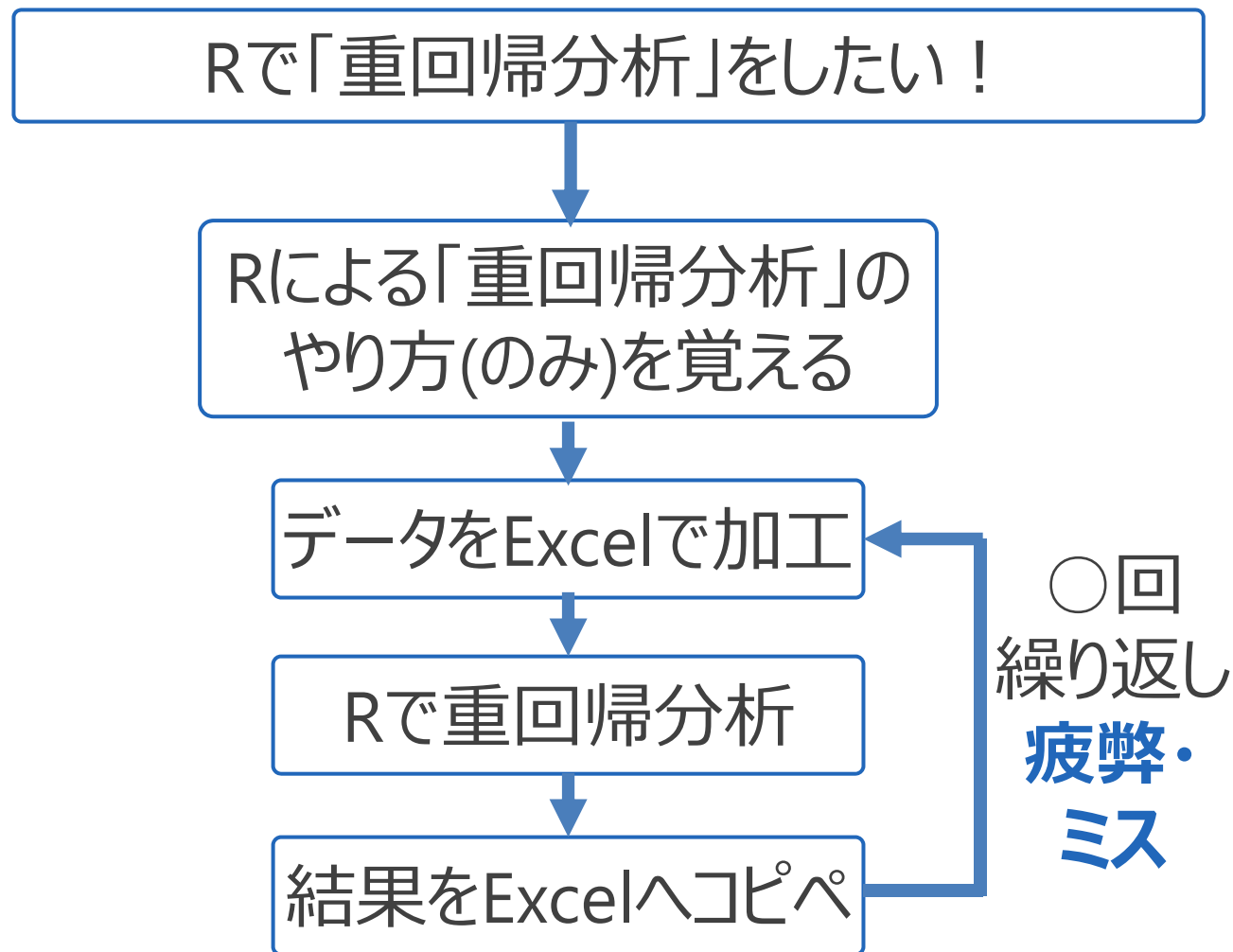
## Rで学ぶ〇〇に書いてあること

---



- Rで学ぶ〇〇という本はデータハンドリングができることが前提

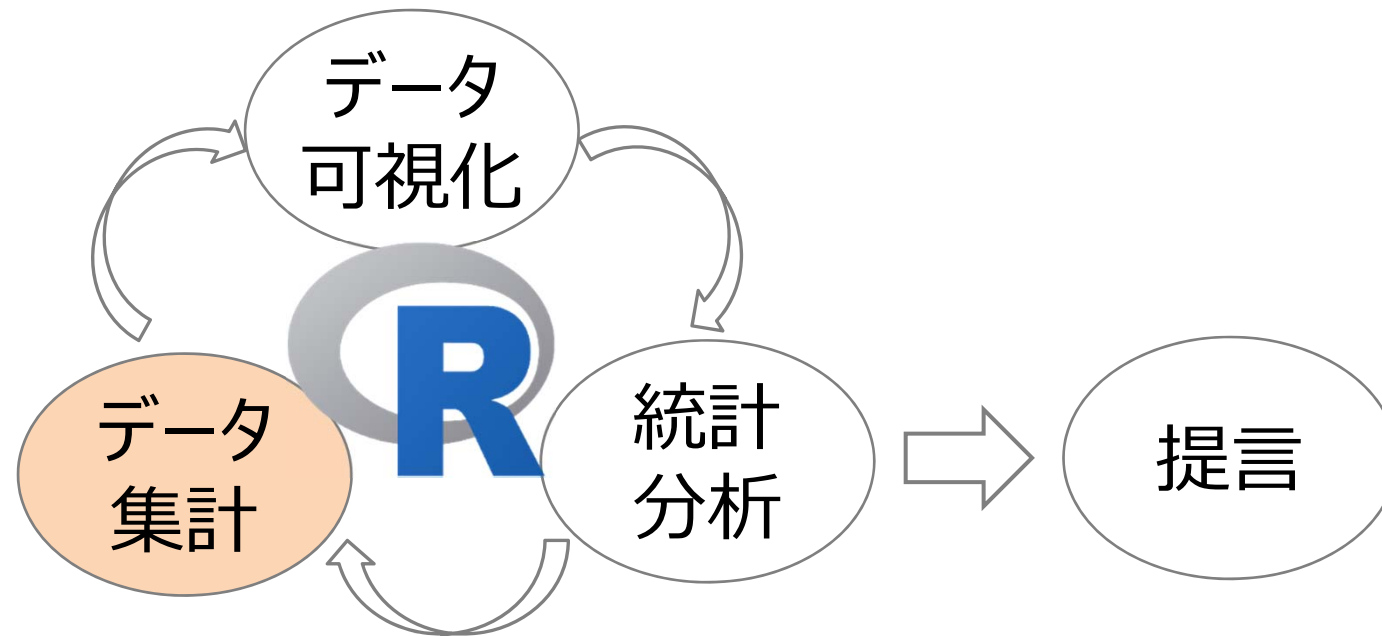
## 統計分析と提言だけに興味がある場合に**ありがちなプロセス**



➤ Rによるデータハンドリングを身に着ければ上記は避けられる

## 本来あるべきデータ分析の流れ①

---

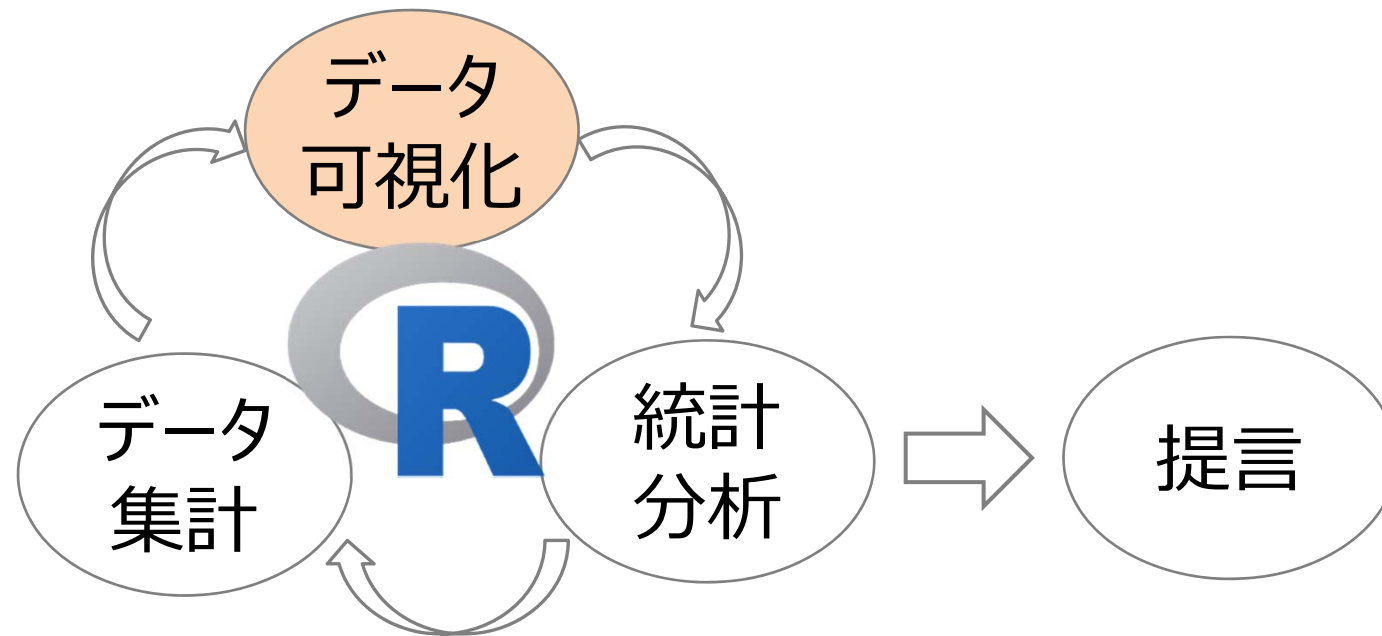


- 手元に集まったデータは「整理されていない」「複数ファイルに分散している」場合が普通なので、**データ分析ができる形式に整える必要**がある

➤ 本に載っているサンプルデータの場合は初めからデータが整っており、省かれるプロセス

## 本来あるべきデータ分析の流れ②

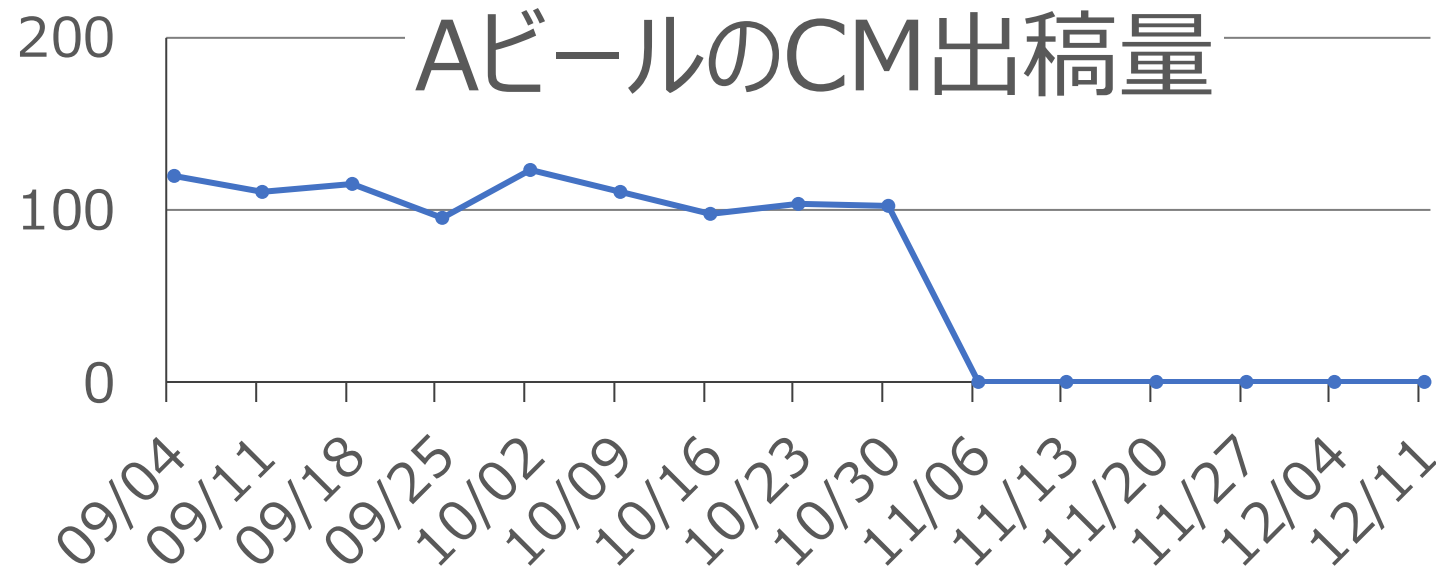
---



- 手元に集まったデータの関係性を把握するには数字を見るだけでは分からないので、**グラフにして可視化する必要**が
- この段階でデータの異常に気づくことが多いので、欠かせないプロセス
  - 本に載っているサンプルデータの場合は省かれるプロセス

## 私がデータを触っていた時の実例

【コンビニ商材の**CM出稿量**と**売上**の関係を分析していた時】



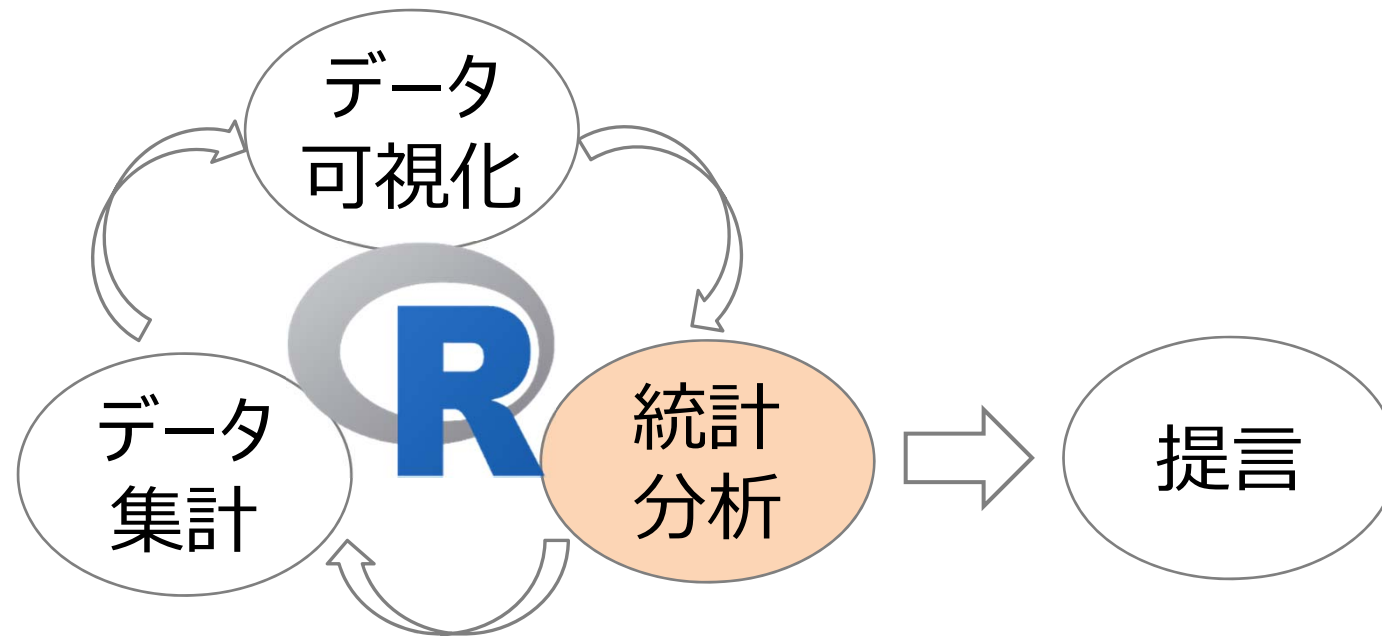
- ある時期からCM出稿量がゼロになっていたが、商品リニューアルによって、商品コードが変わっただけで、AビールのCM出稿がなくなった訳ではなかった

➤クライアントにデータを貰い直し、**データ集計からやり直し**



## 本来あるべきデータ分析の流れ③

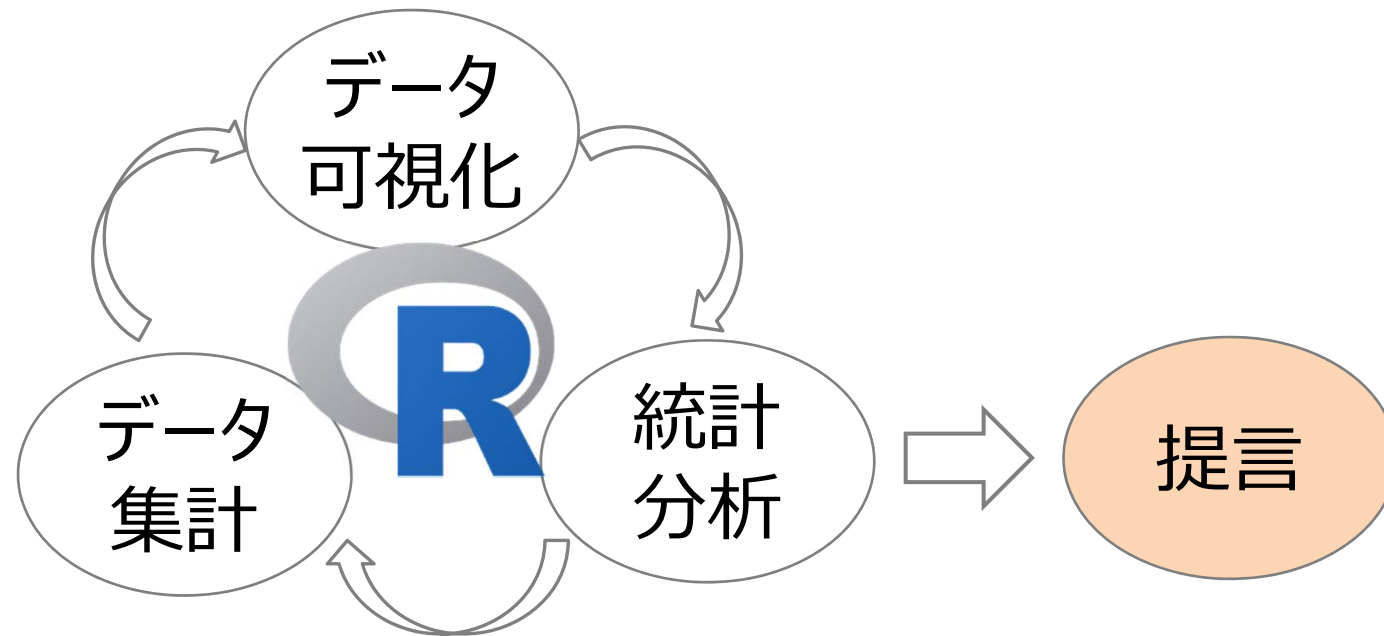
---



- データ集計・可視化をおこなない，データについて理解を深めてから統計分析へ移る
- 本に載っているサンプルデータの場合はデータが正しいことが前提なので，初めからここから

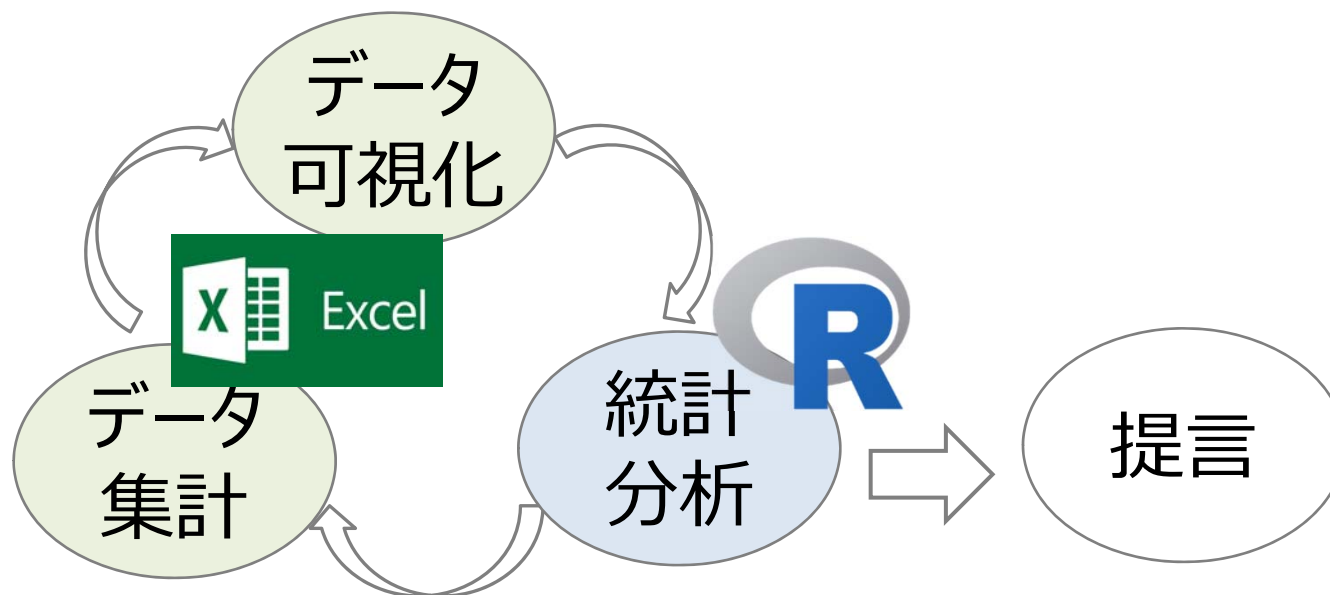
## 本来あるべきデータ分析の流れ④

---



- 統計分析を元に、「提言」をおこなう

## Excelでデータハンドリングをおこなって、Rで分析するのは？



- 簡単な処理が**数回程度**なら**Excel**でも構わない
  - それが何十回・何百回となると…?!
- **Excel**で扱えるデータは**100万行**まで
  - Rは**100万行以上のデータ**を扱える
- Rで一気通貫**したデータ分析をおこないましょう！

Rによるデータハンドリングを身につければ

---

- データ集計・データ可視化をRプログラミングによっておこなうことで、**正確性が増す**と共に**生産性が向上**
  - Rプログラミングスキルが向上するので、**新しい統計分析の実装が早くなる**
    - 統計分析は手法ごとに必要なデータ構造や方法が異なる
- **Rによるデータハンドリング**を身につけましょう

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答

## 総務省・家計調査データとは？

---

- 皆様に事前に配布したデータをExcelで開いてください
  - 開いた後に、**Excel形式で保存**し直してください。
  
- 総務省・家計調査とは？
  - 全国の家計の収入・支出、貯蓄・負債などを毎月調査している
  
  - **県庁所在地 + 政令指定都市**の物品の**年間支出金額**がわかる

# 総務省・家計調査データとは？

	A	B	C	D	E	F	G	H	I	J						
1	統計名：	家計調査 家計収支編 二人以上の世帯														
2	表番号：	10														
3	表題：	[品目分類] 品目分類（平成27年改定）（総数：金額）														
4	実施年月：	- -														
5	市区町村別：	-														
6	表章項目：	1 金額														
7	世帯区分：	3 二人以上の世帯														
8	時間軸（年####）	2016年														
9																
10																
11					0	1003	2003	3003	4003	5003	6003					
12			全国	01100	札幌市	02201	青森市	03201	盛岡市	04100	仙台市	05201	秋田市	06201	山形市	072
12	10000000	1	食料【円】	947,618	904,155	894,438	918,234	955,619	838,891	1,018,483						
13	10100000	1.1	穀類【円】	76,988	78,737	71,566	70,959	73,794	62,015	76,057						
14	10110001	1.1.1	米【円】	23,522	28,911	20,690	22,342	20,776	17,210	23,236						
15	10111001	1.1.1.2	パン【円】	30,294	26,329	25,299	25,031	29,426	21,200	26,733						
16	10112001	120	食パン【円】	8,904	7,772	6,796	6,922	7,782	5,048	7,222						
17	10112901	129	他のパン【円】	21,390	18,557	18,504	18,108	21,643	16,152	19,511						
18	10113001	1.1.3	麺類【円】	17,606	17,398	21,471	19,425	18,840	20,139	21,004						
19	10130010	130	生うどん・そば【円】	3,410	3,013	3,480	2,907	2,973	3,073	3,999						
20	10130020	131	乾うどん・そば【円】	2,462	2,471	2,119	2,353	2,384	3,738	3,276						
21	10130030	134	スパゲッティ【円】	1,202	1,296	1,166	1,103	1,405	935	1,307						
22	10130040	133	中華麺【円】	3,926	3,750	5,637	5,274	4,833	4,897	4,691						
23	10130050	135	カップ麺【円】	4,061	4,597	6,027	5,012	4,622	4,905	5,058						

都市別

品目

まずはExcelでデータハンドリングを体験

---

## 【お題①】

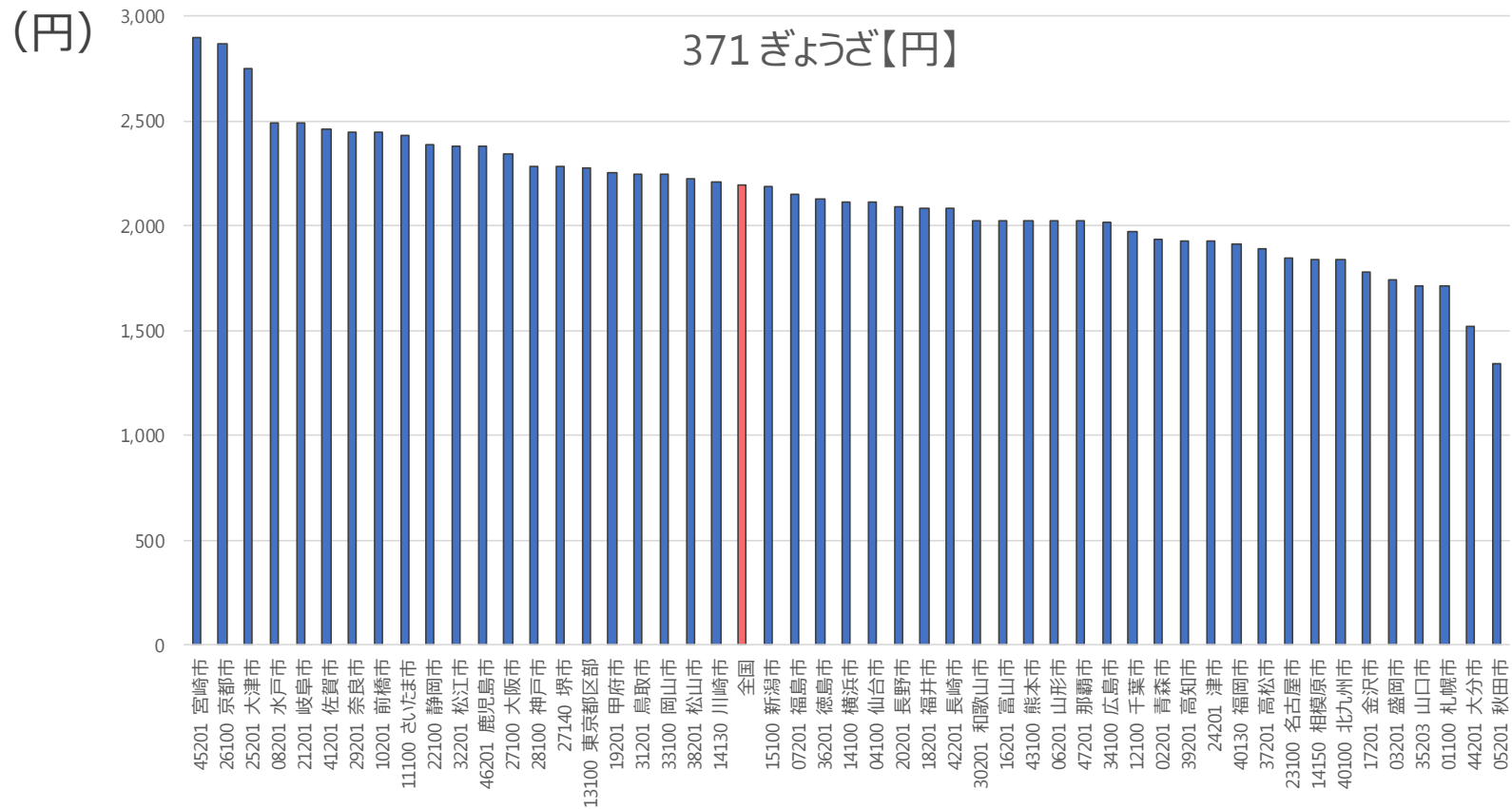
- あなたは「**宇都宮市**」の観光課の職員です
- 「**ぎょうざ**」を観光に活かしたいと考えています
- そこで、「**家計調査**」掲載都市の「**ぎょうざ**」購入金額を**グラフ化**し、「宇都宮市」の立ち位置を把握してください

➤ **Excel**でやってみましょう！

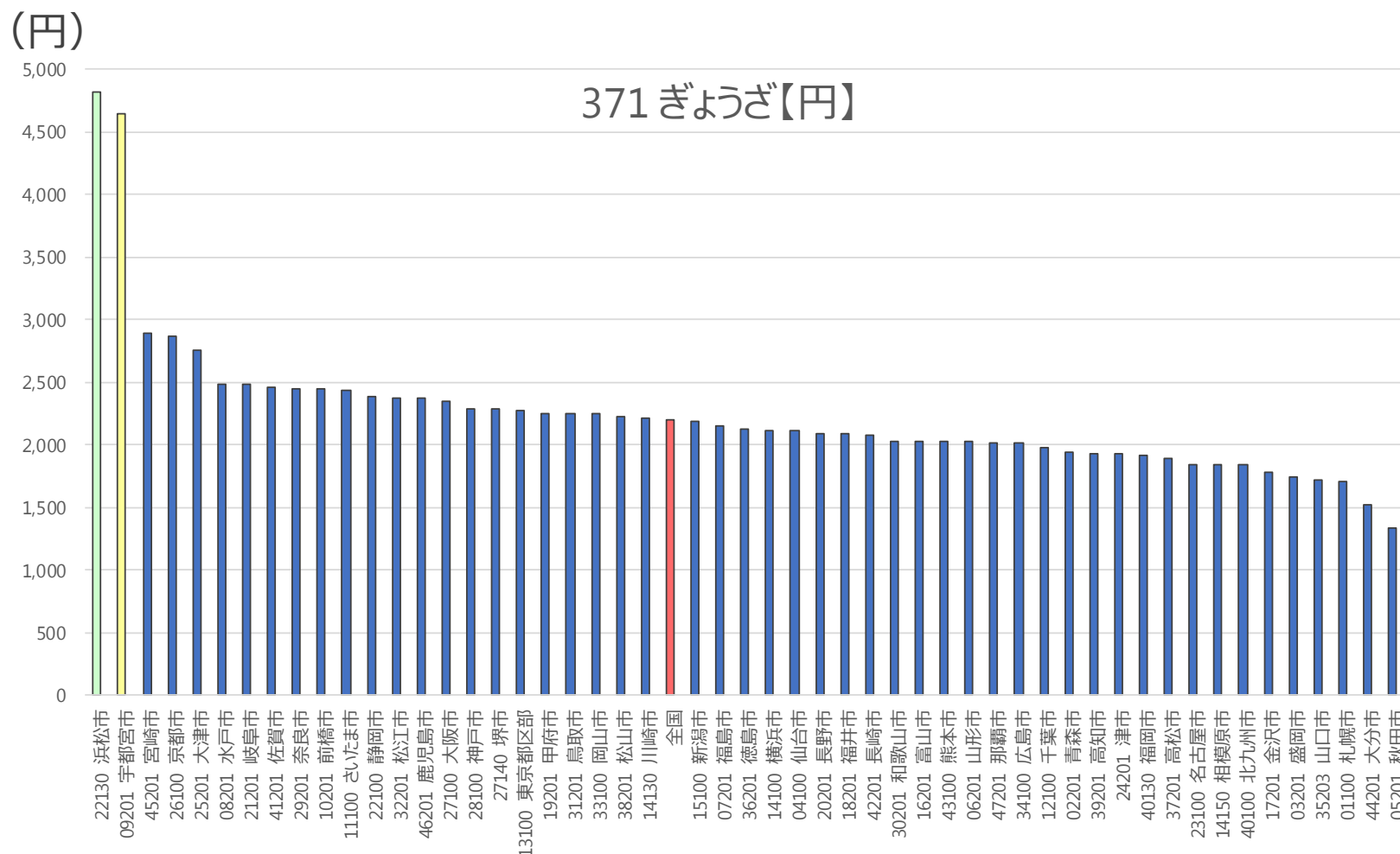


## グラフのイメージ

**Q:「家計調査」掲載都市の「ぎょうざ」購入金額をグラフ化し、「宇都宮市」の立ち位置を把握してください**

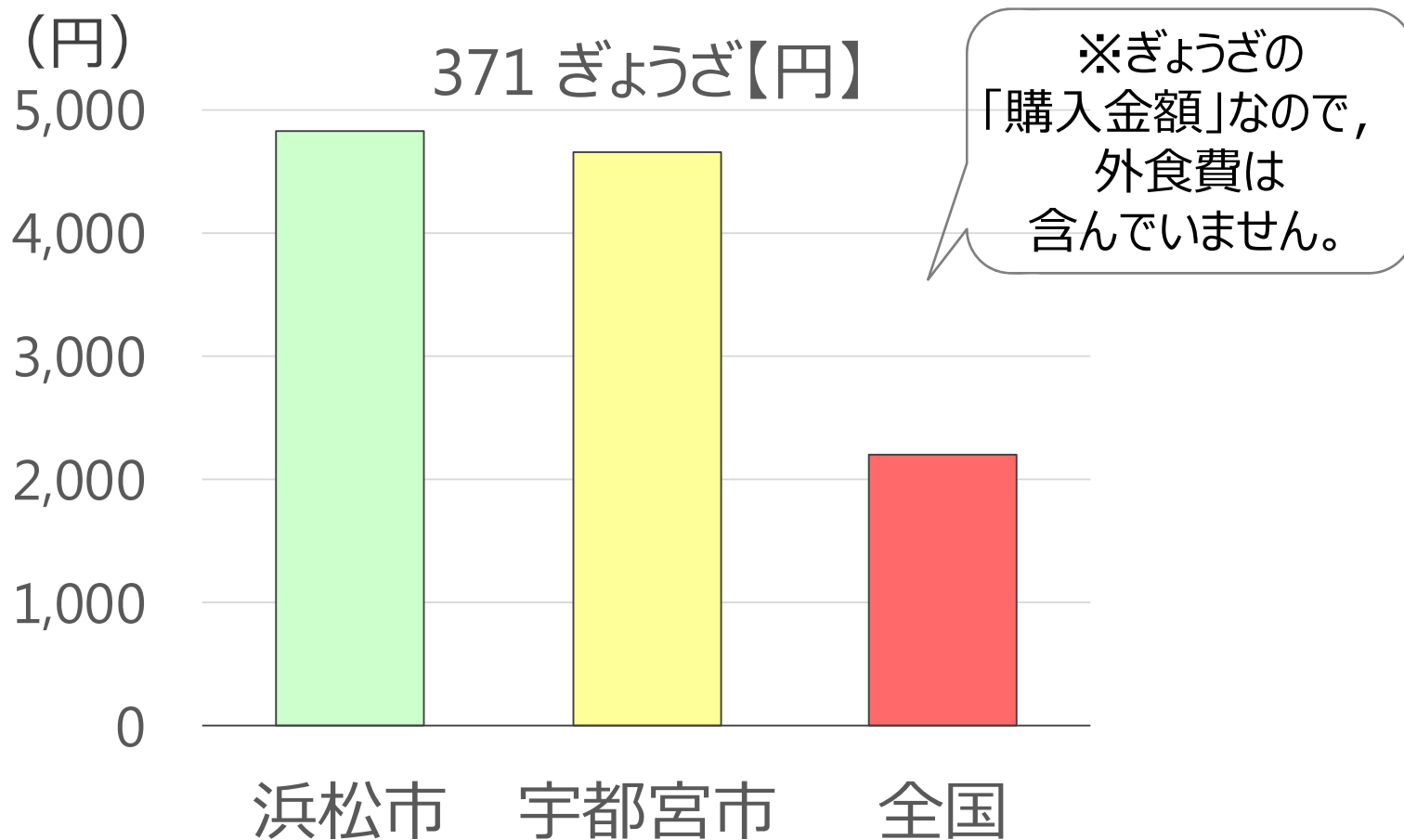


# 「ぎょうざ」の購入金額1位の都市は・・・！？



●「宇都宮市」はごくわずかに及ばず「浜松市」が1位

## 「ぎょうざ」の購入金額1位の都市は・・・！？



- 「**宇都宮市**」はごくわずかに及ばず「**浜松市**」が1位

まずはExcelでデータハンドリングを体験

---

## 【お題②】

- 「**ぎょうざ**」を観光に活かしたいと考えていましたが、**購入額1位は「浜松市」**でした
- そこで、観光に活かせる品目を探さねばなりません
- 各都市の中で「**宇都宮市**」が**トップテンに入る品目**を**グラフ化**してください
  
- **Excel**でやってみましょう！

お疲れ様でした

---

- ○分間の時間を取りましたが、いくつできたでしょうか？
- **宇都宮市がベスト10**に入る品目は692品目中**121品目**あるので、全て終わる時間を計算してください
- **同じ作業を繰り返す**ことは**非常に大変**
- もし、作業後にデータ差し替えなどがあったら？

Rで同じことをやるには

---

- 「dplyr」と「ggplot2」というパッケージを使用
- コードを書くのに合計**30分**使用

```
9 d00 <- fread("FEH_00200561_171106202630.csv", stringsAsFactors = F, skip = 10) %>% as_tibble()
0 # d00 <- read.csv("FEH_00200561_171106202630.csv", skip = 10)
1 # d00 <- read_csv("FEH_00200561_171106202630.csv", skip = 10, local = locale(encoding = "SJIS"))
2
3 d01 <- d00 %>%
4   as_tibble %>%
5   select(- V1, - V2) %>%
6   rename("goods" = V3) %>%
7   gather(area, yen, - goods) %>%
8   mutate(area = if_else(area != "全国", substring(area, 7, 99), area)) %>%
9   mutate(mycity = if_else(area == "宇都宮市", "宇都宮市",
0     if_else(area == "浜松市", "浜松市", "その他")))
1
2 # カテゴリ
3 cate <- unique(d01$goods)
4 for(i in 1:length(cate))
5 {
6   d02 <- d01 %>%
7     filter(goods == cate[i]) %>%
8     arrange(yen)
9
0 }
```

➤ R実行

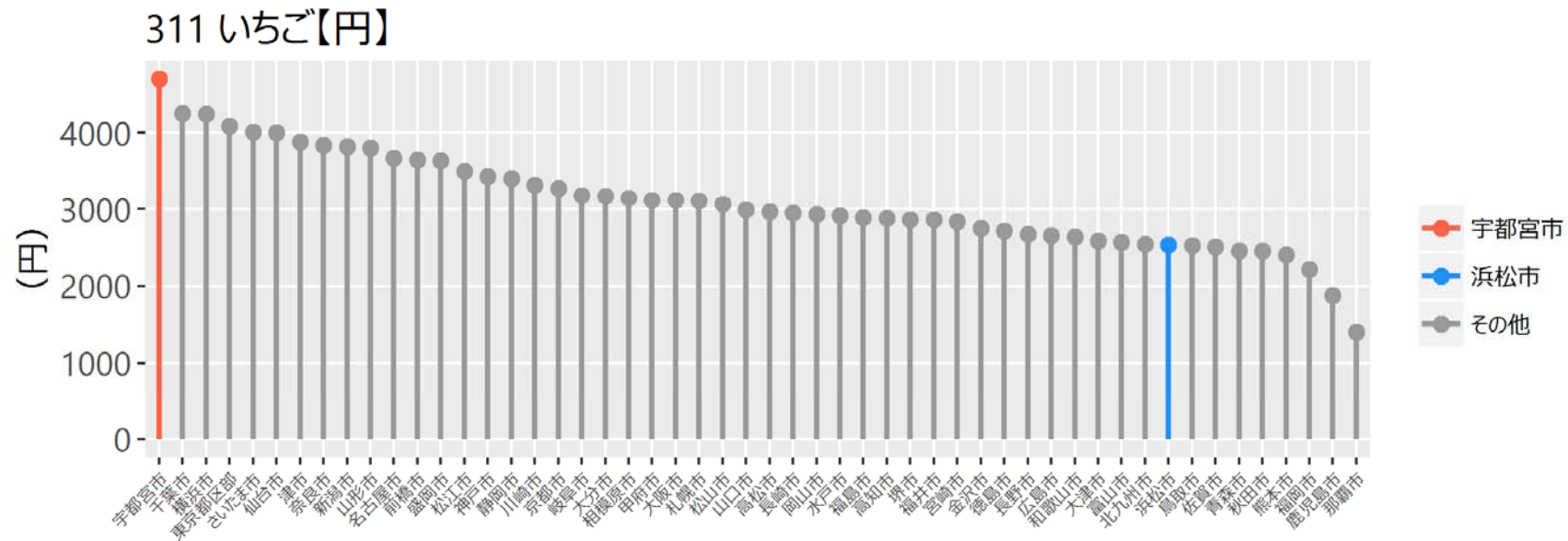
## Rでデータ整形およびグラフによる可視化

---

- 「dplyr」というパッケージでデータ整形をおこなった後  
「ggplot2」というパッケージでグラフによる可視化

➤ できたグラフを見ていきましょう

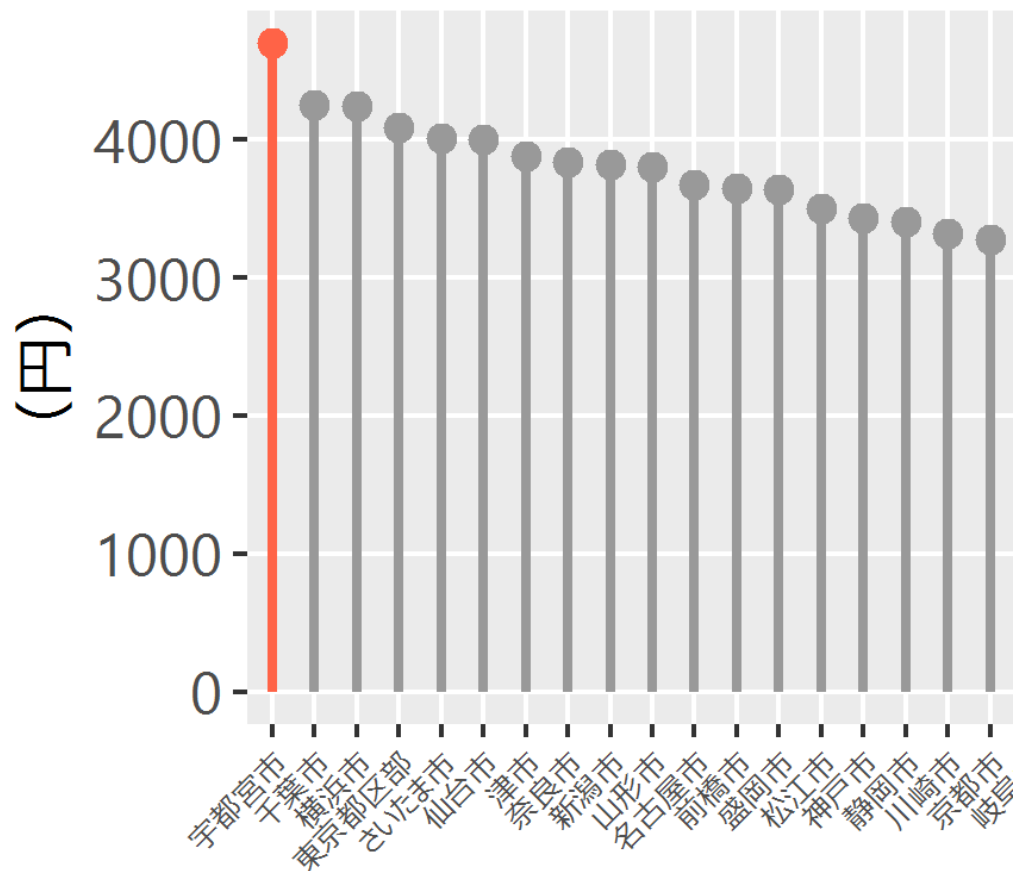
# Rでデータ整形およびグラフによる可視化：「いちご」





# Rでデータ整形およびグラフによる可視化：「いちご」

## 311 いちご【円】

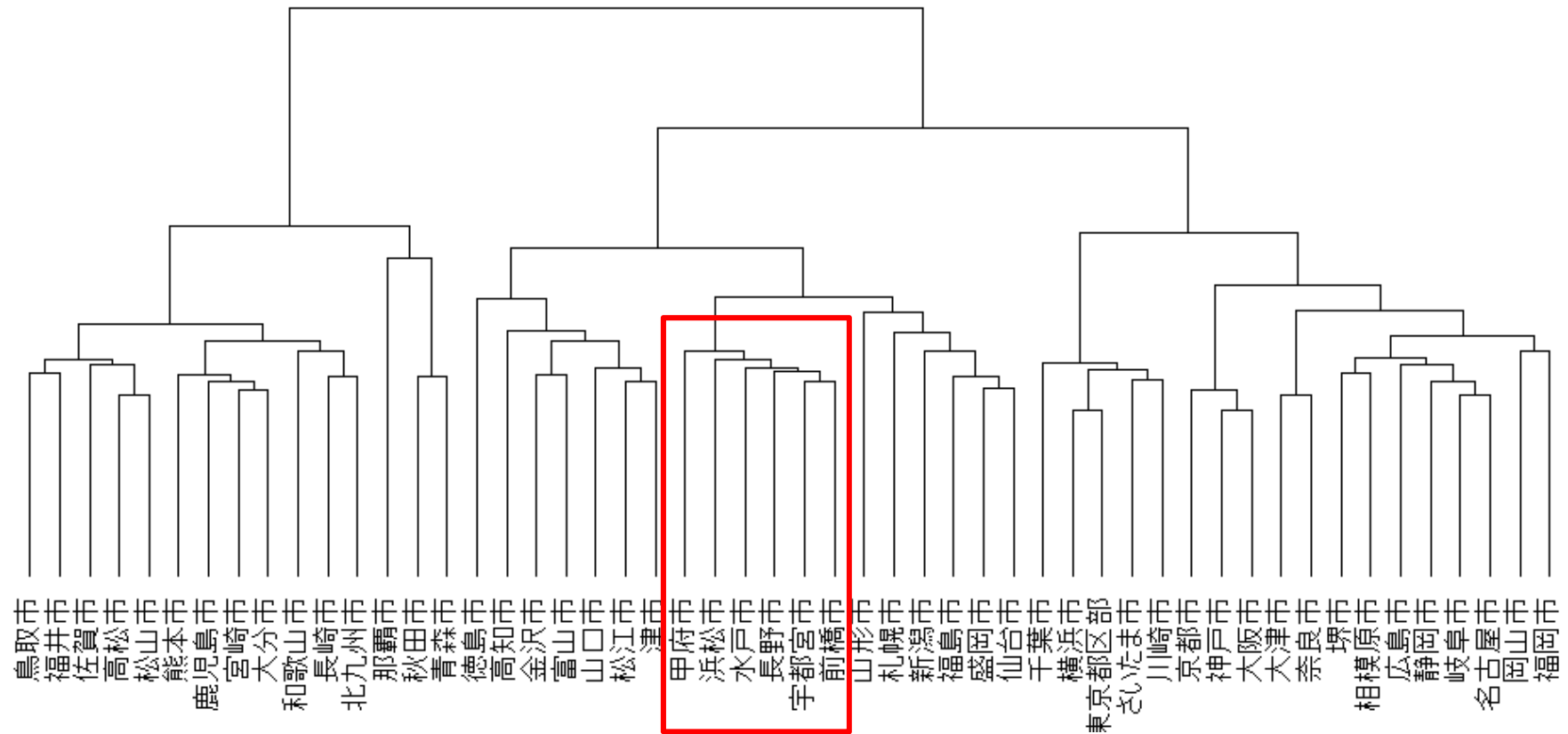


## VBAでプログラムを書けば良いのでは？

---

- **R**でデータ集計・可視化をおこなえば、「統計分析」まで**一気通貫**して処理をおこなえる
  
- わざわざ**VBA**でデータハンドリングをおこなった後、**R**で統計分析をおこなうのではなく、初めから**R**でやった方が**効率**が良い

## 例えば、**クラスター分析**を実行



➤ クラスター分析で各都市をグルーピングしたところ、  
**宇都宮市**に最も近いのは**前橋市**という結果に

## 100万行を超えるビッグデータもRで処理可能

---

- **Excel**では**100万行**までしか開けないですが、**R**では**100万行を超える**データも扱えます
- 実際に**500万行**のファイルを開いてみると…

## デモを通じてわかったこと

---

- データ集計・データ可視化をRプログラミングによっておこなうことで、**正確性が増す**と共に**生産性が向上**
  - Rプログラミングスキルが向上するので、**新しい統計分析の実装が早くなる**
    - 統計分析は手法ごとに必要なデータ構造や方法が異なる
- **Rによるデータハンドリング**を身につけましょう

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答

# 年明け以降のセミナーの概要

---

セミナー当日は説明

## 本日の流れ

---

なぜRを使うのか

なぜデータハンドリングが必要か

総務省 家計調査を用いたRによる  
データハンドリングのデモ

年明け以降のセミナーの概要

アンケート記入と質疑応答



アンケートのご協力をお願い申し上げます

本日はありがとうございました