

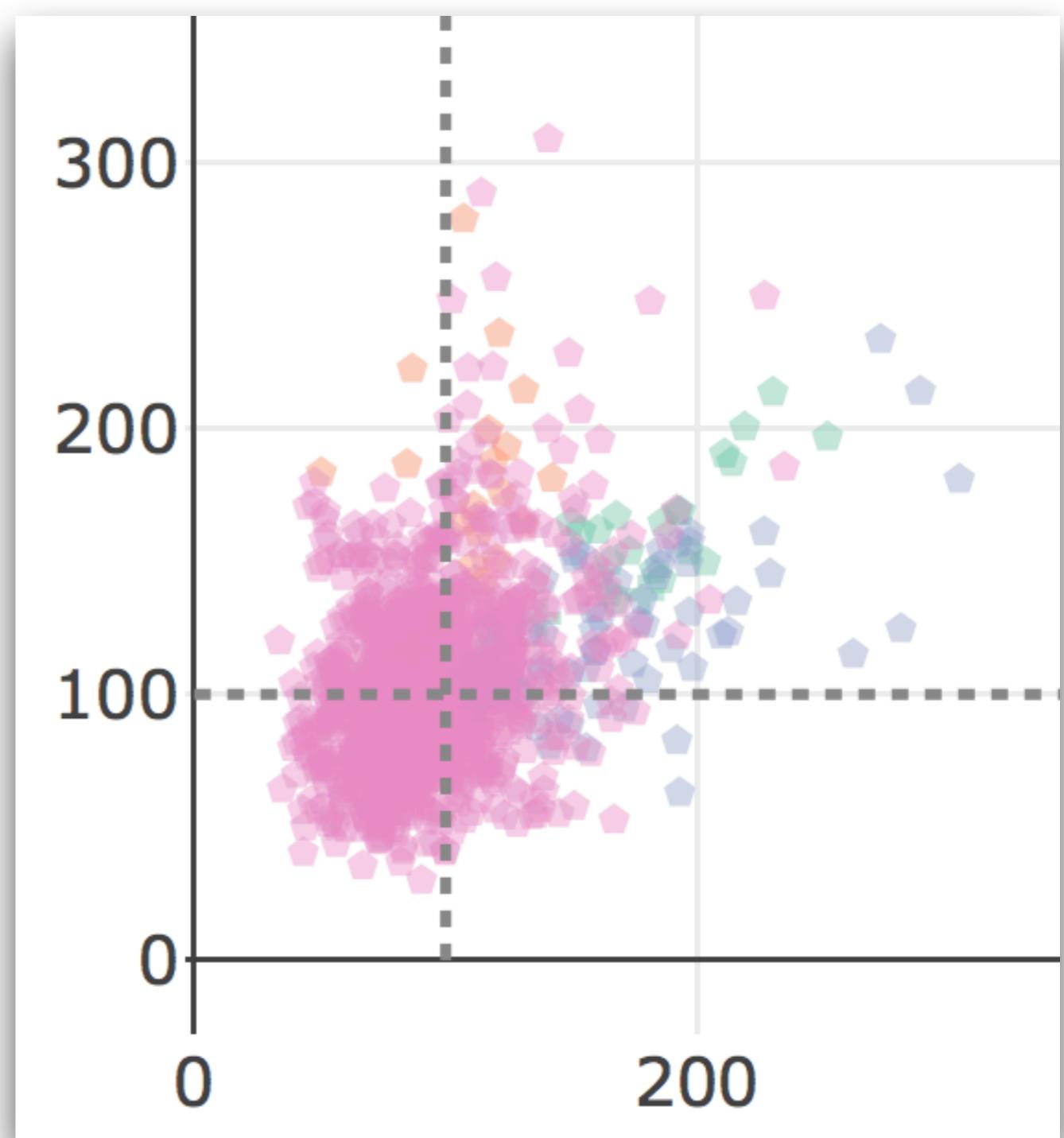
散布図から始めよう

ExcelとRの練習 データ分析の入り口まで

このスライドの最新版とデータは
github.com/nishioWU/JNPC
にあります

散布図とは

- ・ペアになっているデータを
 - ・XY平面に図示して
 - ・関係を探るのに使う
- ▶直線的な関係があるか？
- ▶関係は強いか弱いか？
- ▶トレンドやその変化を見る
- ▶外れ値に注目する



なぜ散布図か

- ① データ入手する
- ② 付き合わせたり表記の揺れを直したり加工する
- ③ 散布図描く
- ④ 相関係数計算したり、回帰直線を引いたりしてみる

という流れの真ん中に当たる。データの入手・加工の練習にもなるし、分析の糸口にもなる。さらに勉強したくなってくる（のでは）

データのペアの例

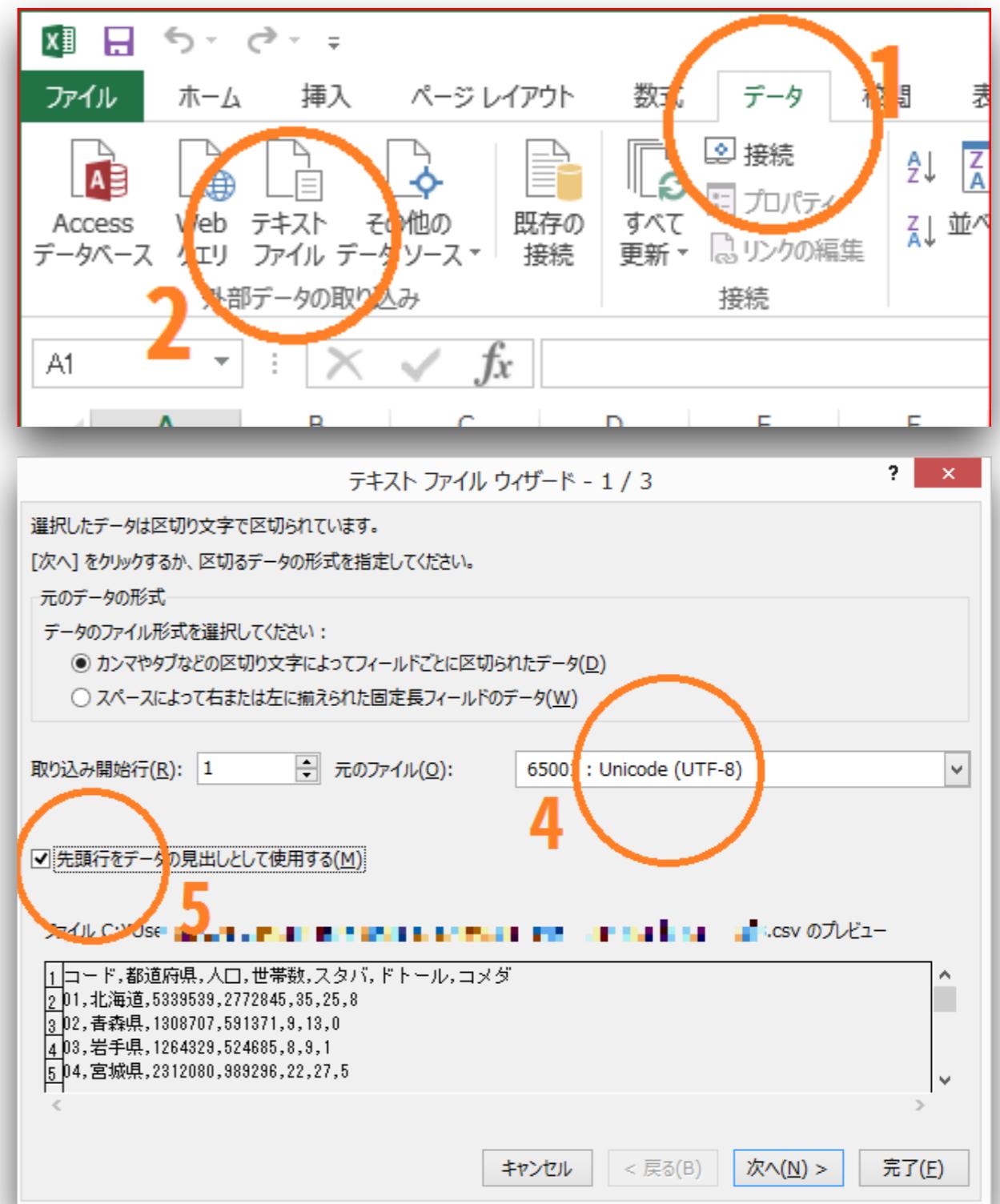
- 身長と、体重
- 親の身長と、子の身長
- 気温と、電力需要
- 気温と、ビール消費量
- 男性の合格率と、女性の合格率
- 自治体の人口と、ある疾患の死亡率
- 各県の人口と、喫茶店の店舗数

Excelに読み込む前に

- ・元データはコピーして別名で保存しておく（フォルダ丸ごとが楽）
- ・形式「CSV」は、コンマ区切りのテキストのこと
- ・テキストエディタで中身を確認しておくとよい
- ・文字コード違いで化ける。UTF-8か、それともS-JIS（CP932）か
- ・先頭の0を削るなど、Excelが気を利かせて勝手に変換するのが困りもの。日付も要注意。年がない場合は今年にされてしまう
なので、データファイルのダブルクリックでExcelを開くのはダメ
さきに空のExcelファイルを新規作成し、そこにデータを取り込む

CSVデータ読み込み 1

- ① Excel左上の「データ」タブで
- ② 「テキストファイル」を選択
- ③ CSVファイルを選ぶ。今回は、
dataフォルダにある
cafe_data.csv。すると「開く」
が「インポート」に変わるので、
そのボタンをクリック
- ④ 「Unicode (UTF-8)」を選ぶ
- ⑤ 「先頭行をデータの見出として使
用」に✓
- ⑥ 「次へ」



CSVデータ読み込み2

- ① 「区切り文字」を変更。「カンマ」に✓を
- ② プレビュー画面を確認して「次へ」
- ③ 列のデータ形式を指定。「コード」の列を選択してから、上のラジオボタン●で「文字列」にする
- ④ 0で始まるデータに注意。「標準」だと数値に変換され0が消えてしまうので、「文字列」にしておく
- ⑤ 年の入っていない日付も、勝手に今年にされてしまう。もしあれば、やはり「文字列」が安全
- ⑥ 今回は「コード」列以外はデフォルトで、「完了」
- ⑦ データの貼り付け先を適宜指定し、「OK」

CSVデータ読み込み2 続き

テキストファイル ウイザード - 2 / 3

フィールドの区切り文字を指定してください。[データのプレビュー] ボックスには区切り位置が表示されます。

区切り文字

タブ(I)
 セミコロン(M)
 カンマ(C)
 スペース(S)
 その他(O):

連続した区切り文字は 1 文字列の引用符(Q): "

1

テキストファイル ウイザード - 3 / 3

区切ったあとの列のデータ形式を選択してください。

列のデータ形式

G/標準(G)
 文字列(I)
 日付(D): YMD
 削除する(I)

[G/標準] を選択すると、数字は数値に、日付は日付形式の値に、その他の値は文字列に変換されます。

詳細(A)...

データのプレビュー(P)

コード	都道府県	人口	世帯数	スタバ	ドトール	コメダ
01	北海道	5339539	2772845	35	25	8
02	青森県	1308707	591371	9	13	0
03	岩手県	1264329	524685	8	9	1
04	宮城県	2312080	989296	22	27	5

3

データのプレビュー(P)

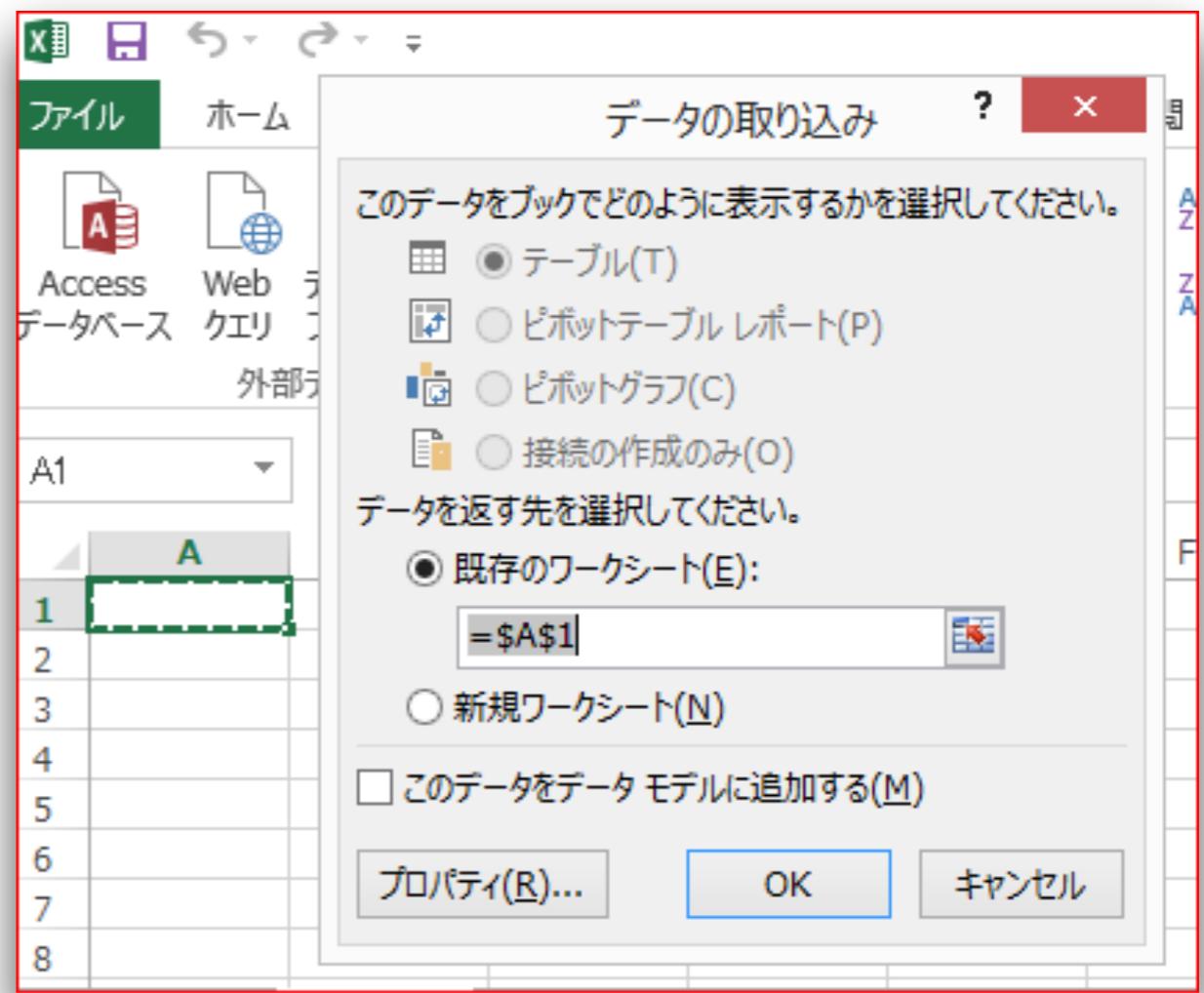
文字列	G/標準	G/標準	G/標準	G/標準	G/標準	G/標準
コード	都道府県	人口	世帯数	スタバ	ドトール	コメダ
01	北海道	5339539	2772845	35	25	8
02	青森県	1308707	591371	9	13	0
03	岩手県	1264329	524685	8	9	1
04	宮城県	2312080	989296	22	27	5

3

キャンセル < 戻る(B) 次へ(N) > 完了(E)

データ読み込み完了

- OKを押せば、シートの左上隅を起点に貼りつく
- 左上隅以外や、新しくシートを増やして貼る場合は、その旨指定を



ここまで済んだ状態：[cafe excel 01.xlsx](#)

絞り込み可能な表にする

- ① Excelのシートを複写しておく。失敗した場合のリカバリー用。
簡単なやり方は、Ctrl + ドラッグ
- ② データに通し番号を打つ。今回は「コード」で代用して省略
- ③ 絞り込み・ソート可能な表になると便利。Ctrl + A または Ctrl + * で範囲指定してCtrl + Tが早い

- ▶ 人口順とかコメダが多い順とか、並べ替え可能になった
- ▶ 元に戻すときは、通し番号の列で「昇順」に
- ▶ フィルターが散布図と連動するので便利

A	B	C	D	E	F	G
1 コード	都道府県	人口	世帯数	スタバ	ドトール	コメダ
2 01	北海道	5339539	2772845	35	25	8
3 02	青森県	1308707	591371	9	13	0
4 03	岩手県	1264329	524685	8	9	1
5 04	宮城県	2312080	989296	22	27	5
6 05	秋田県	1015057	425933	7	2	1
7 06	山形県	11069				
8 07	福島県	19196				
9 08	茨城県	29510				
10 09	栃木県	19857				
11 10	群馬県	19905				
12 11	埼玉県	73630				
13 12	千葉県	62989				
14 13	東京都	136373				
15 14	神奈川県	91712				
16 15	新潟県	89810				

テーブルの作成 ? ×

テーブルに変換するデータ範囲を指定してください(W)
=\$A\$1:\$G\$49

先頭行をテーブルの見出しとして使用する(M)

OK キャンセル

表にする2

- ④ 「都道府県」列の「合計」、または「コード列」の「NULL」を除外しておくと、散布図を描くときに困らない。「都道府県」の列でフィルターを使い、「合計」だけ✓を外しておく。これで、全国計が表示されなくなる

表にしておくと、後が楽

The screenshot shows a Microsoft Excel spreadsheet titled "cafe_excel_02.xlsx". The active cell is B1. A filter menu is open over column B, showing options like '昇順(S)', '降順(O)', '色で並べ替え(I)', and '検索'. Below these are checkboxes for various prefectures: 京都府, 熊本県, 群馬県, 広島県, 香川県, 高知県, 合計, 佐賀県, 埼玉県, 三重県, 山形県, and 山口県. The '合計' checkbox is highlighted with a red circle and a large red number 4. The rest of the column B data is visible, showing numerical values. The Excel ribbon at the top includes tabs for ファイル, ホーム, 挿入, ページレイアウト, 数式, データ, and 校閲.

	A	B	C	D	E	F
1	コード	都道府県	人口	世帯数	スタバ	ドト
	昇順(S)			2772845	35	
	降順(O)			591371	9	
	色で並べ替え(I)			524685	8	
	"都道府県" からフィルターをクリア(C)			989296	22	
	色フィルター(I)			425933	7	
	テキストフィルター(E)			413685	7	
	検索			781157	9	
	京都府			1235665	31	
	熊本県			826672	24	
	群馬県			841085	15	
	広島県			3259736	66	
	香川県			2851491	66	
	高知県			7096622	320	
	合計			4280874	104	
	佐賀県			895463	13	
	埼玉県			418653	9	
	三重県			482491	10	
	山形県			292518	6	
	山口県			358393	11	
				866562	20	
				816077	13	
				1571636	31	
				3257903	95	
				789961	16	
				1834269		

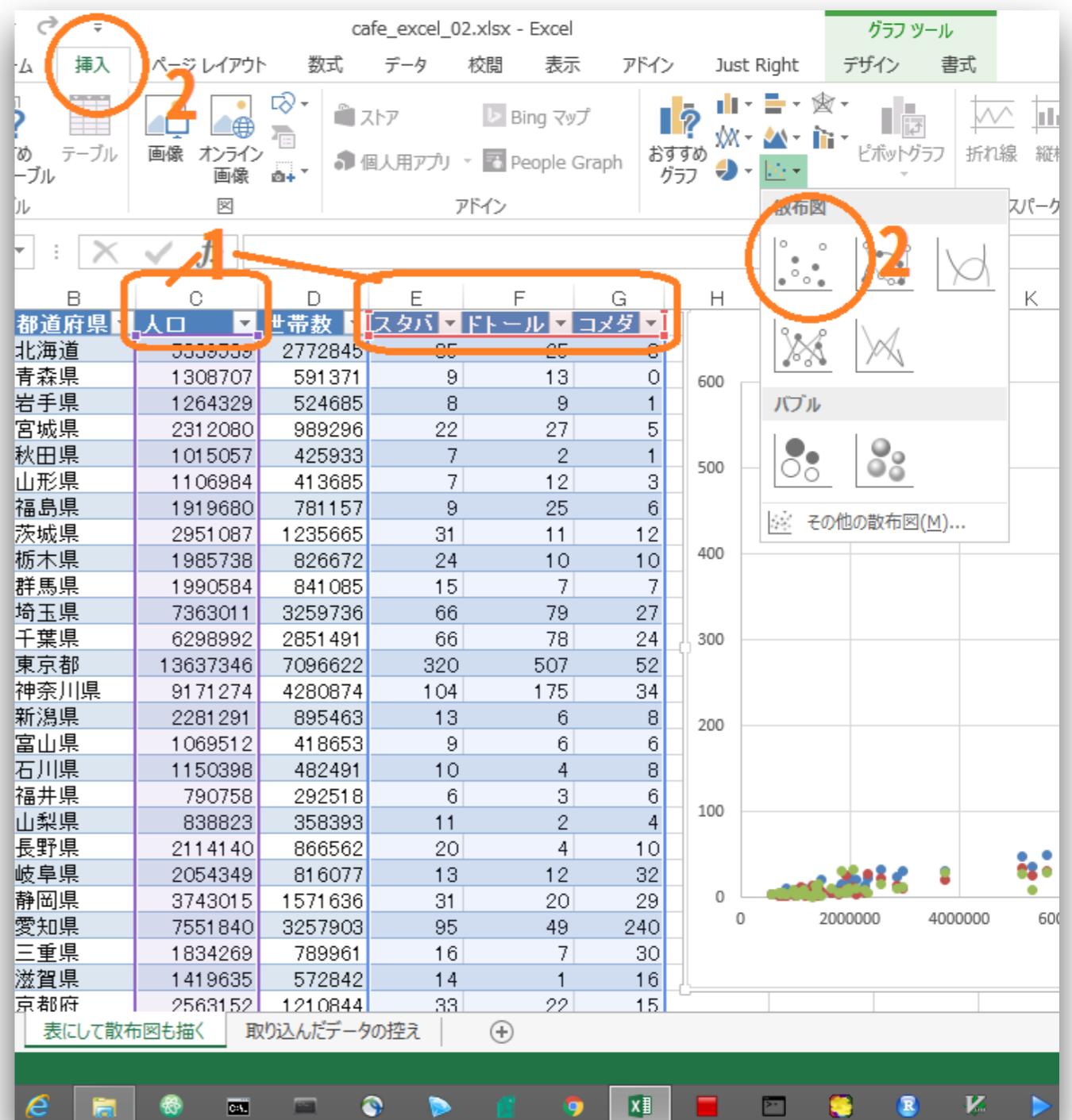
散布図を描く 簡単な方法

① 表にある列を2つ選ぶ。2列目が離れているなら、1列目を選んだ後、Ctrl + クリックで追加

② 「挿入」タブから「グラフ」の「散布図」を選ぶ

③ もし3列以上選んだ場合は、左端の列がX座標、残りはY座標になる

では、トライ！ まず人口と世帯数で練習してみてもいい



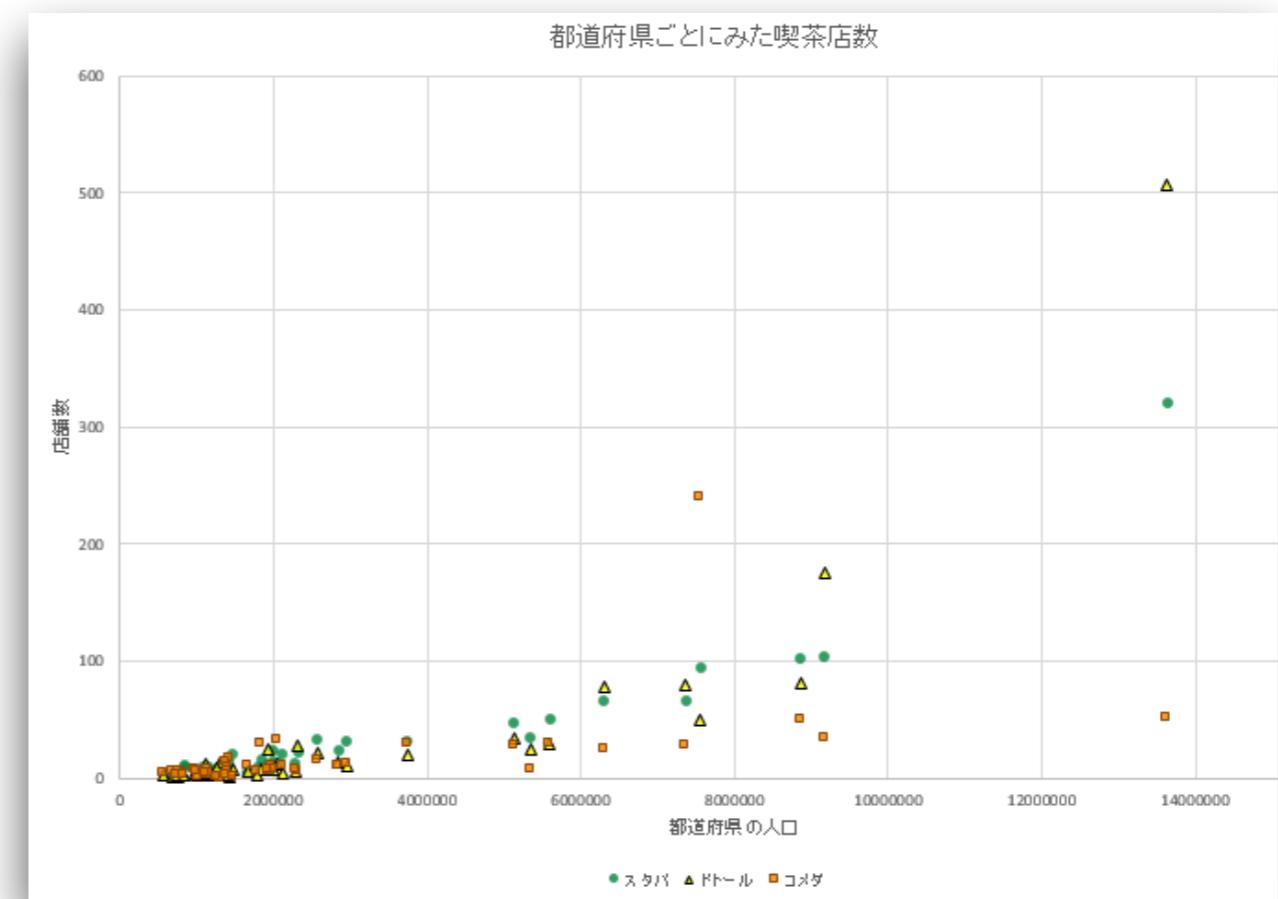
散布図を描く 手動でやるなら

- ① 今あるグラフを手直しするか、列を選ばずに空っぽのままの散布図を挿入。グラフエリア右の「漏斗」のアイコンから、「データの選択」に進み、左側の「凡例項目」の窓で指示する
- ② 「追加」で開く「系列の編集」パネルで、上から順に項目名（じか打ちしても、入っているセルを指定しても可）、X座標のデータ範囲、Y座標のデータ範囲、を指定。XとYは先頭のイコールを残しておく。「OK」で完了
- ③ やり直す場合は「編集」。不要な列があれば「削除」

実は…体裁変更が面倒

- ・色や形の変更は、点を右クリックして「データ系列の書式設定」「マーカー」と進み、「マーカーのオプション」から
- ・都道府県名を添えたければ、グラフエリア右の「+」アイコン →「データラベルの書式設定」から「セルの値」で都道府県名の列を指定。「Y値」の✓は外す

体裁を直してみた例：
[cafe_excel_02.xlsx](#)



表にデータ列を追加

グラフの元データの表には、データ列をさらに追加できる。入力済みのテーブルの隣の列に新たに何か入力すると、表の範囲が自動的に広がる仕組み。現有データをもとに計算して追加することも可能。

たとえば、人口の多い県に店が多くても不思議はないので、人口10万人あたりに直したスタバ店舗数を使おう、と考えたら……

- ① 「H2」セルをクリックして「=」だけ打つ
- ② 「E2 (スタバ店舗数)」セルをクリックして、「/」を打つ。「C2 (人口)」セルをクリックして、「*100000」と打つ
- ③ 列見出しの「H1」を適宜付け直す

これを使って新たに散布図を描けばよい

ここまでを反映：
[cafe_excel_03.xlsx](#)

参考になるサイト

- ・「エクセル2016 散布図グラフの作り方」

2016以外のバージョンでも参考になる。マーカー（点）の色の変更、軸の目盛りなど、細かな設定の解説あり

<https://www.tipsfound.com/excel/05036>

Excelのグラフ全般については

- ・「エクセル2016 グラフの作り方」

<https://www.tipsfound.com/excel/05001>

文字化けしたとき IEの小技

CSVかTXT形式のデータが文字化けしているときは、エンコードの違いが原因。UTF-8かSHIFT-JISを試してみる。実はInternet Exploreでコードを変換して保存し直す手がある。割と役立つ。

- ① 拡張子が「.csv」の場合は「.txt」に変えて保存。ピリオドまで消さないように注意
- ② IEでファイルを開く。化けていれば、画面を右クリックしてエンコードを直す。たいてい、自動認識してくれる
- ③ 保存したい形式（上記2通りのどちらか）を選び、別名で保存。別名にしないと、元が消えてしまう
- ④ 必要なら拡張子を「.csv」に戻す。戻さなくても、表計算ソフトに読み込むことは可能

Excelの勘どころ

▶ 行と列

横を行、縦を列とかカラムと呼んで区別している。

▶ 式は小文字で

Excelは（Rと違って）大文字でも小文字でも命令を聞いてくれる。なので、関数は小文字で入力するとよい。正しく認識されれば大文字に変換される。小文字のまま残ったら、打ち間違いだと分かる。

▶ 絶対参照

式をコピーすると、気を利かして、計算対象の行や列をずらしてくれる。それが便利だからだが、困る場合もある。そのときは、ずらされては困るものに「\$」マークをつけると、コピー先でもずれない。これが絶対参照。式の入力中に「F4」キーを押すと、行と列の両方またはどちらかに、\$がついたり消えたりして切り替えられる

能率が上がるショートカット

- ▶ Ctrl + ドラッグ
シート名のタブをつかみながらだと、そのシートのコピーを作成
- ▶ Shift + ドラッグ
行や列を選択し、その境目をつかみながらだと、並び替え
- ▶ Ctrl + 1
セルの書式設定。エルではなくて数字の一（テンキーの1はダメ）
- ▶ Ctrl + Z
直前の変更を元に戻す
- ▶ Ctrl + A
シート全体を選択
- ▶ Ctrl + C
コピー

ショートカットその2

▶ Ctrl + V

通常の貼り付け。セル幅以外すべて引き継ぐ。もう一度押すと、貼り付けの形式を選べる

▶ Alt + Ctrl + V

形式を選択して貼り付け。Ctrl + Vでは困るときに使う。関数を使って整形をした後、貼り直して「値だけ」にするのに便利（Vを選ぶ）

▶ Alt + ;

絞り込み時に表示されているセルだけをコピー元にする。重宝する

▶ Ctrl + S

ファイルを上書き保存

▶ 「F12」

ファイルを別名で保存

ショートカットその3

- ▶ 「F2」
セルの編集
- ▶ Shift + Ctrl + @
セルの表示を「処理の結果」か「数式そのもの」か切り替える
- ▶ Ctrl + F
検索
- ▶ Ctrl + H
置換
- ▶ Ctrl + カーソルキー
空白セルは飛ばし、その次にデータの入っているセルにジャンプ
- ▶ Ctrl + Home
A1セル（左上）にジャンプ

ショートカットその4（完）

- ▶ Ctrl + End
データの入っている最終セルにジャンプ
- ▶ Ctrl + ;
きょうの日付を入力。便利
- ▶ Ctrl + :
現在の時刻を入力。便利
- ▶ Ctrl + *
データが入っている範囲を選択。離れ小島は選択されない
- ▶ Ctrl + T
テーブルにする
- ▶ Ctrl + Enter
複数のセルに同じデータを入れる。一括して修正するときに便利
- ▶ Alt + 下矢印
そのカラムに入力済みのデータのリストから選ぶ

計算モデルを当てはめる

Y軸の喫茶店数を、X軸の人口を使った数式で計算・説明できないだろうか？ 準備として、人口を万人単位に直したものを作る（やらなくててもよいのだが、数式の係数の桁数をそこそこ確保するため）。

直線的な分布なら、 $y = a + bx$ の一次関数で表せるのでは？

- ① グラフの元表に1列追加。都道府県の人口を10000で割った数が入るようにする。15ページの方法の応用
- ② 都道府県人口と3チェーンの店舗数をプロットしたシートを複写。Ctrl + ドラッグで。名前を適宜付け替える
- ③ 13ページの方法で各チェーンの系列データを編集。X軸の値だけ、①で作った万人単位に差し替える。あとは触らず

軸を対数に変更する

人口と喫茶店数の関係は、直線的なものではなさそう。対数軸にしてみる。つまり、 $\log y = \log a + b \log x$ を試してみることにする

- ① X軸をクリック。軸の書式設定パネルが開く
- ② 「対数目盛を表示する」に✓。最小値・最大値も変えると見やすい
- ③ 最小50万人、最大2000万人にしてみた



近似曲線を引く①

Y軸も対数目盛に変更する。「負の値やゼロは……」という警告が出るかもしれない。これは、青森にコメダがないため。そのまま続行。だいたい直線的になったので、近似曲線を追加してみる

- ④ スタバのマーカーのどれかを右クリック。「近似曲線の追加」を選ぶ。いきなり線が引かれてしまったら、右クリックして書式変更。線が引かれても、引かれなくても、次ページへ



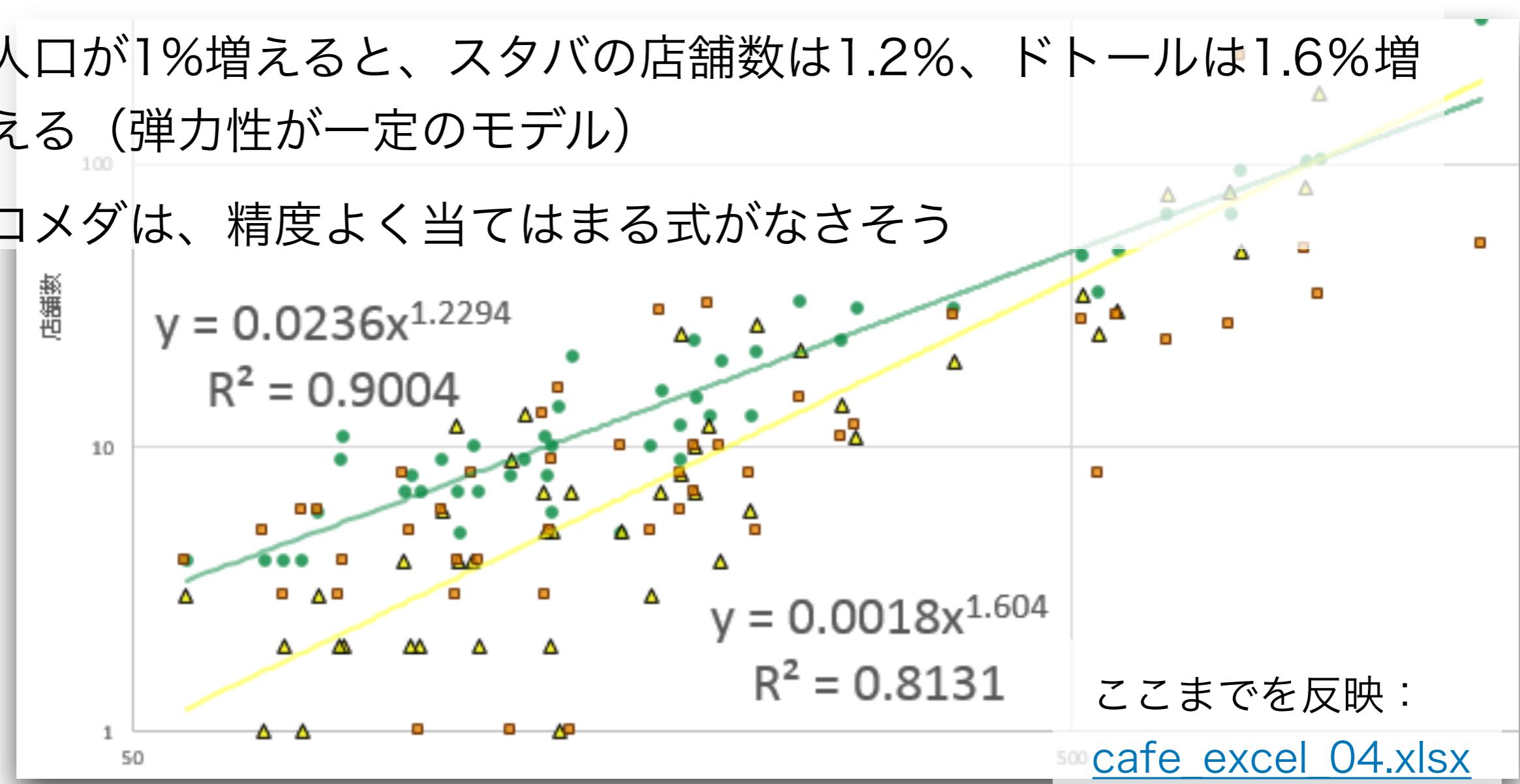
近似曲線を引く2

- ④ 続き。ボタンで「累乗近似」を選ぶ。対数軸なので直線状になる。
なるべく多くのマーカーの近くを通る線が引かれる
- ⑤ 「グラフに数式を表示」「R2乗値を表示」の両方に✓
- ⑥ ペンキ缶をこぼしたようなアイコンで、線の色など適宜変更
- ⑦ 数式の文字も大きくした。人口が1%増えたら店は1.2%増える
- ⑧ ドトールもやってみて



近似曲線を引く3（完）

- ・「R²乗値」とは、Xの式でYの値をどれぐらい説明できているかの目安。1に近いほど、よく近似できている
- ・人口が1%増えると、スタバの店舗数は1.2%、ドトールは1.6%増える（弾力性が一定のモデル）
- ・コメダは、精度よく当てはまる式がなさそう



相関係数を計算する

散布図を眺めて、X軸とY軸のデータに直線的な関係がありそうなら、相関係数を計算してみる。-1～1の値になる。

ExcelではCORREL関数、Rならcor関数を使う。Excelではアドインの分析ツールを有効にして、それを使ってもいい。

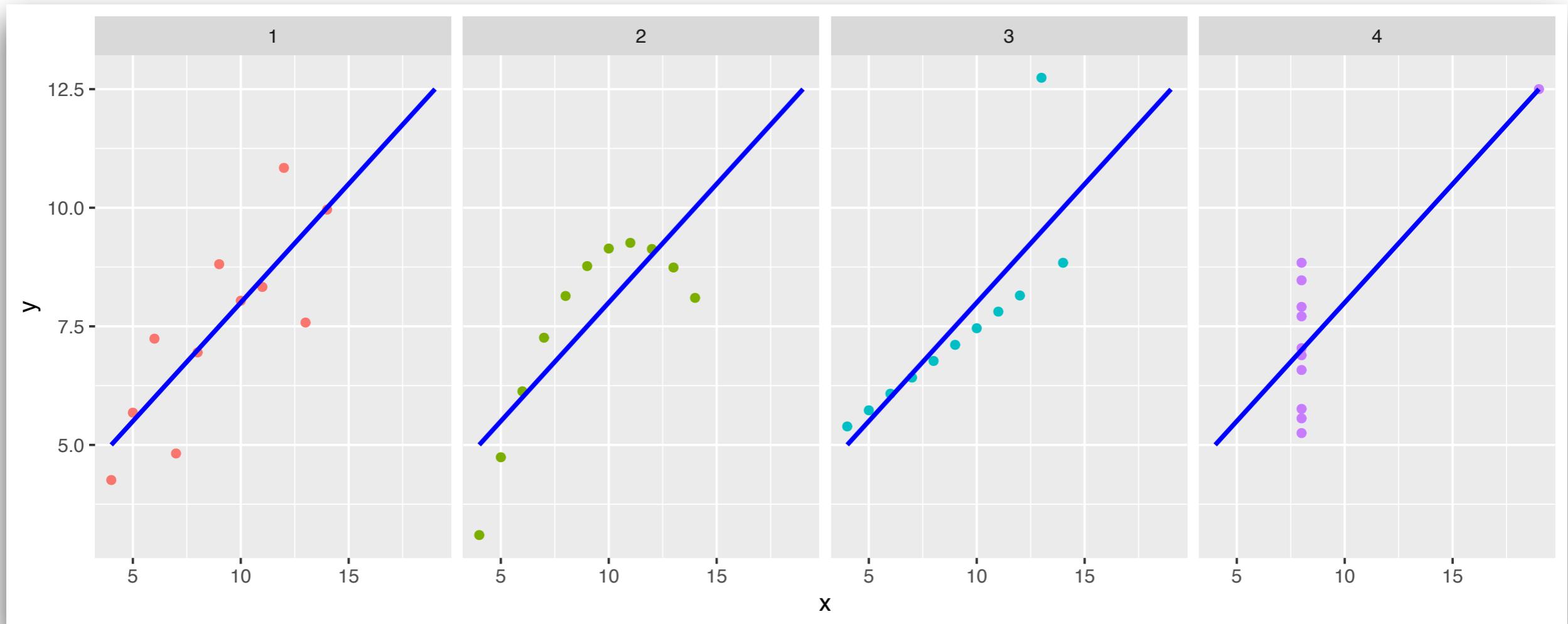
▶絶対値が1に近ければ、強い関係

▶0なら相関なし

で、関係の強弱の目安になる。散布図を描きもしないで、相関係数を求めるのは、実は危ない。

ただし散布図を眺めてから

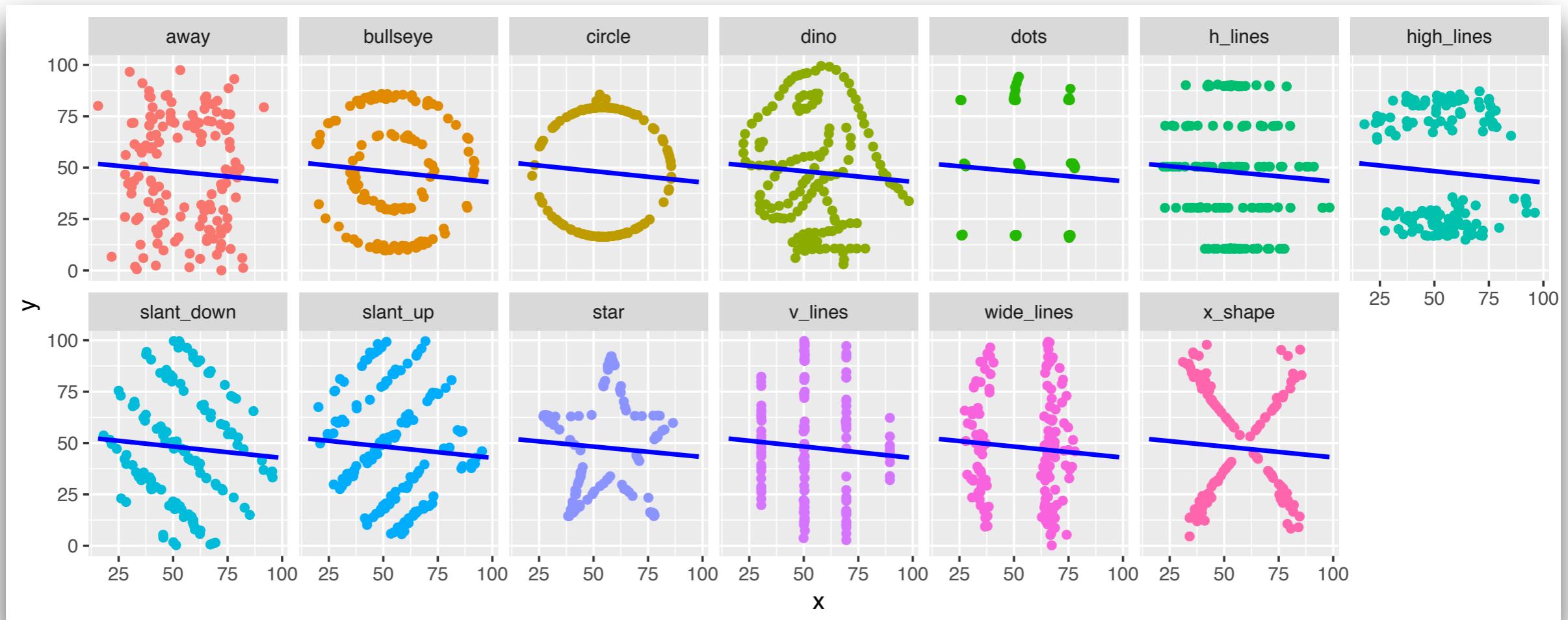
直線的な関係でなさそうなら、相関係数を求めるのはナンセンス。
機械的に計算したなら、下の図は4つとも、Xの平均が9.0、標準偏差3.3、Yの平均7.5、標準偏差2.0、相関係数は0.82。それでいいの？



一見に如かず

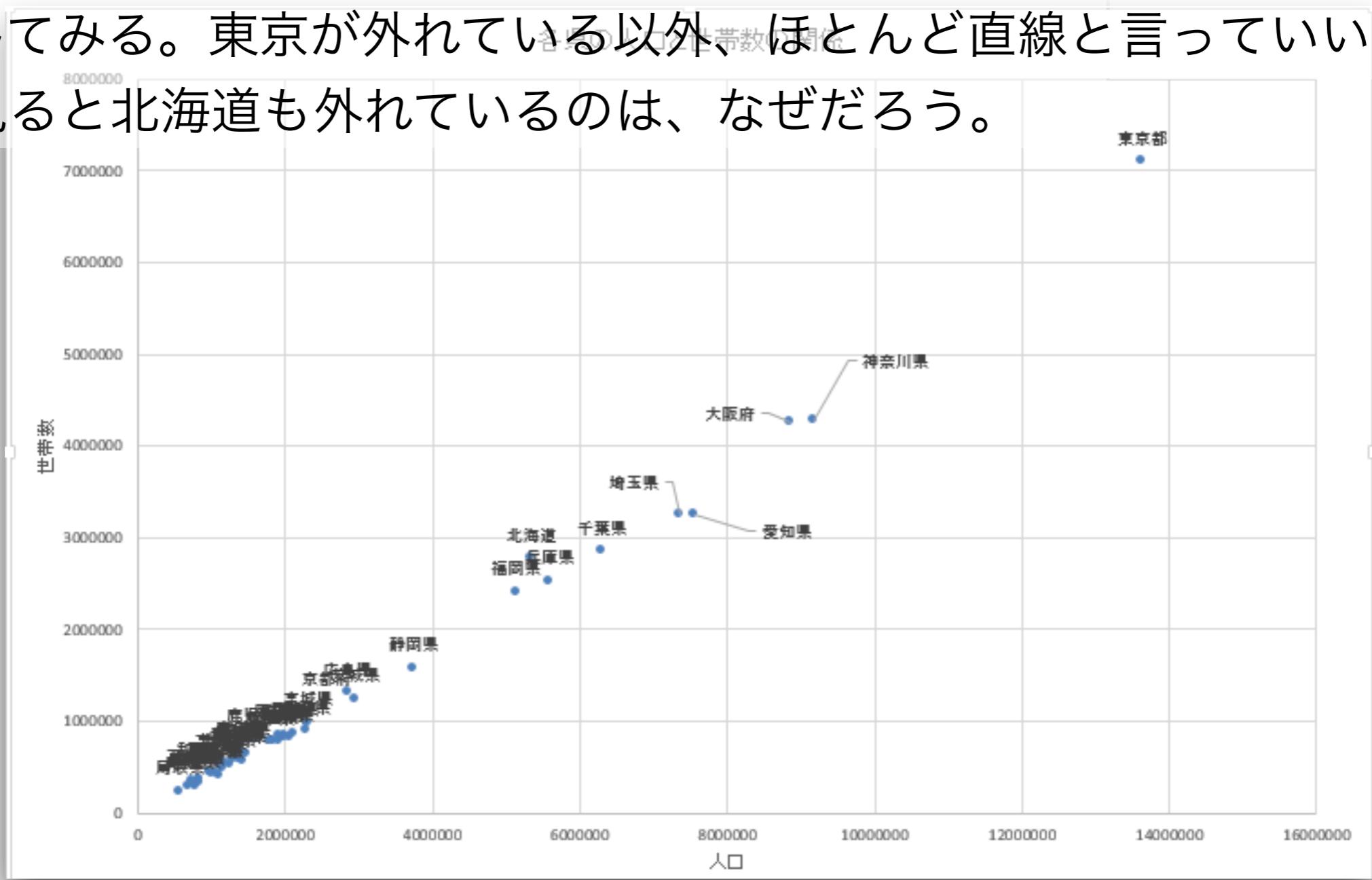
“Never trust summary statistics alone; always visualize your data”

Alberto Cairoの「Datasaurus」。このデータセットの13種はどれも、Xの平均54.3、標準偏差16.8、Yの平均47.8、標準偏差26.9。XとYの相関係数も、わずかに差はあるが、ほぼ-0.06で揃っている。でも、分布は全然違う。散布図を描く手間を惜しまないこと。



強い相関の例

喫茶店データにある、都道府県の人口の列と、世帯数の列を散布図にしてみる。東京が外れている以外、ほとんど直線と言っていい。よく見ると北海道も外れているのは、なぜだろう。



相関をCORREL(X, Y)関数で

- ① 適当なセルに「=correl(」と入力
- ② X要素の範囲、この場合は「c2:c48」を指定。打ち込んで、マウスやカーソルで範囲指定してもよい。C列全部を意味する「c:c」にはないこと。散布図で除外した全国計が含まれてしまう
- ③ コンマで区切り、Y要素の範囲「d2:d48」を指定する
- ④ 「)」を閉じて改行。相関係数は0.99586と計算された

cafe_excel_05.xlsx - Excel

ホーム

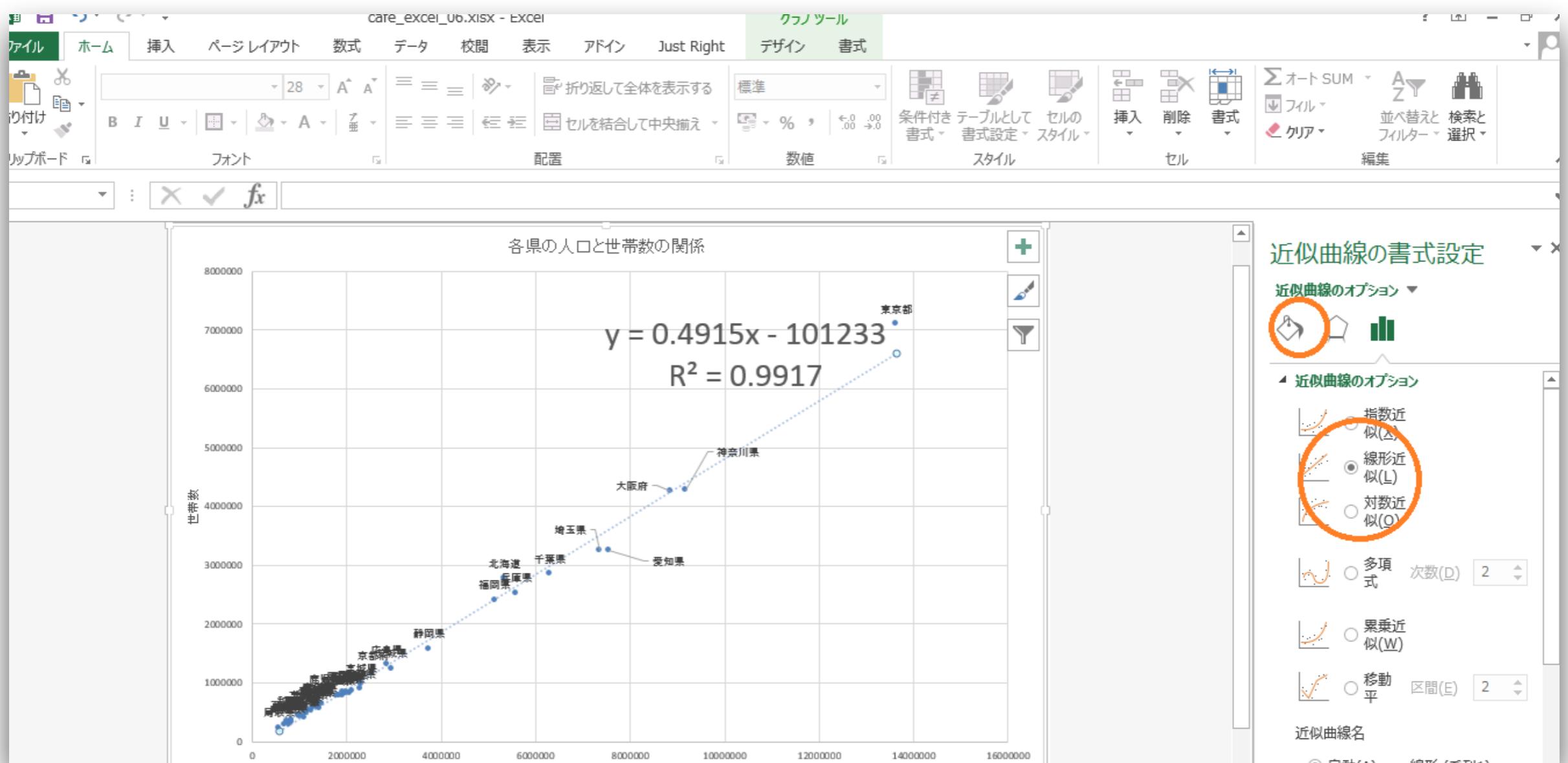
1,2,3,4

=CORREL(C2:C48,D2:D48)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	コード	都道府県	人口	世帯数	タバコ	ドトール	コメダ	S人口	D人口	K人口	人口(ア)		人口(C列)と世帯数(D列)の相関係数を求める。PEARSON				
2	01	北海道	5339539	2772845	35	25	8	0.655487	0.468205	0.149826	533.9539		0.995819	列すべてを指定した			
3	02	青森県	1308707	591371	9	13	0	0.687702	0.993347	0	130.8707		相関係数 = CORREL	8行目まで全国計を含まない			
4	03	岩手県	1264329	524685	8	9	1	0.632747	0.71184	0.079093	126.4329		0.995819	49行目までにしたら、列すべてと同じ値に			
5	04	宮城県	2312080	989296	22	27	5	0.951524	1.16778	0.216255	231.208		この場合、適切なのは全国計を除いたもの				
6	05	秋田県	1015057	425933	7	2	1	0.689616	0.197033	0.098517	101.5057		決定係数 = 0.99175	算で囲ったセルの値を2乗した			
7	06	山形県	1106984	413685	7	12	3	0.632349	1.084027	0.271007	110.6984		近似曲線のR2乗値(決定係数)と同じ				
8	07	福島県	1919680	781157	9	25	6	0.468828	1.3023	0.312552	191.968		こちらは「データ」タブの「データ分析」から「相関」を選んで計算				
9	08	茨城県	2951087	1235665	31	11	12	1.05046	0.372744	0.40663	295.1087		アドイン「分析ツール」を有効にしておく必要がある				
10	09	栃木県	1985738	826672	24	10	10	1.208619	0.503591	0.503591	198.5738		人口	世帯数			
11	10	群馬県	1990584	841085	15	7	7	0.753548	0.351656	0.351656	199.0584						
12	11	埼玉県	7363011	3259736	66	79	27	0.896372	1.072931	0.366698	736.3011						

回帰直線を引く

散布図に近似曲線を追加。軸を触らない今まで直線的関係があるので「線形近似」にする。これが回帰直線。線種や文字サイズを調整。



ところで回帰直線とは

- ・Y列の値をX列の1次式で説明・予測しようとするもの。直線で近似するのが妥当なときしか、やる意味がない。名前はいかめしいし、回帰ってどういうことよ、と思うが、そういうこと
- ・直線的な関係があるとき、という以外にも制約はあるが、省略
- ・X列とY列の平均値を通り、かつ、実際の値とのY軸方向のズレを最小にする直線を引く。あくまで、X列でY列を説明するために使う。
- ・相関係数はX列とY列を入れ替えても値が変わらない。が、回帰直線の式でX列とY列を入れ替えると、別の直線になってしまう。X列からY列を説明するのと、Y列からX列を説明するのとでは、異なる直線になる

分析ツールを使うと

「相関」

- CORREL関数と同様に、2つでペアになったデータ列の相関係数を求めることができる
- CORREL関数と違い、3つ以上のデータ列の相関係数を求めることができる。どのペアに強い相関があるか、を探しやすい

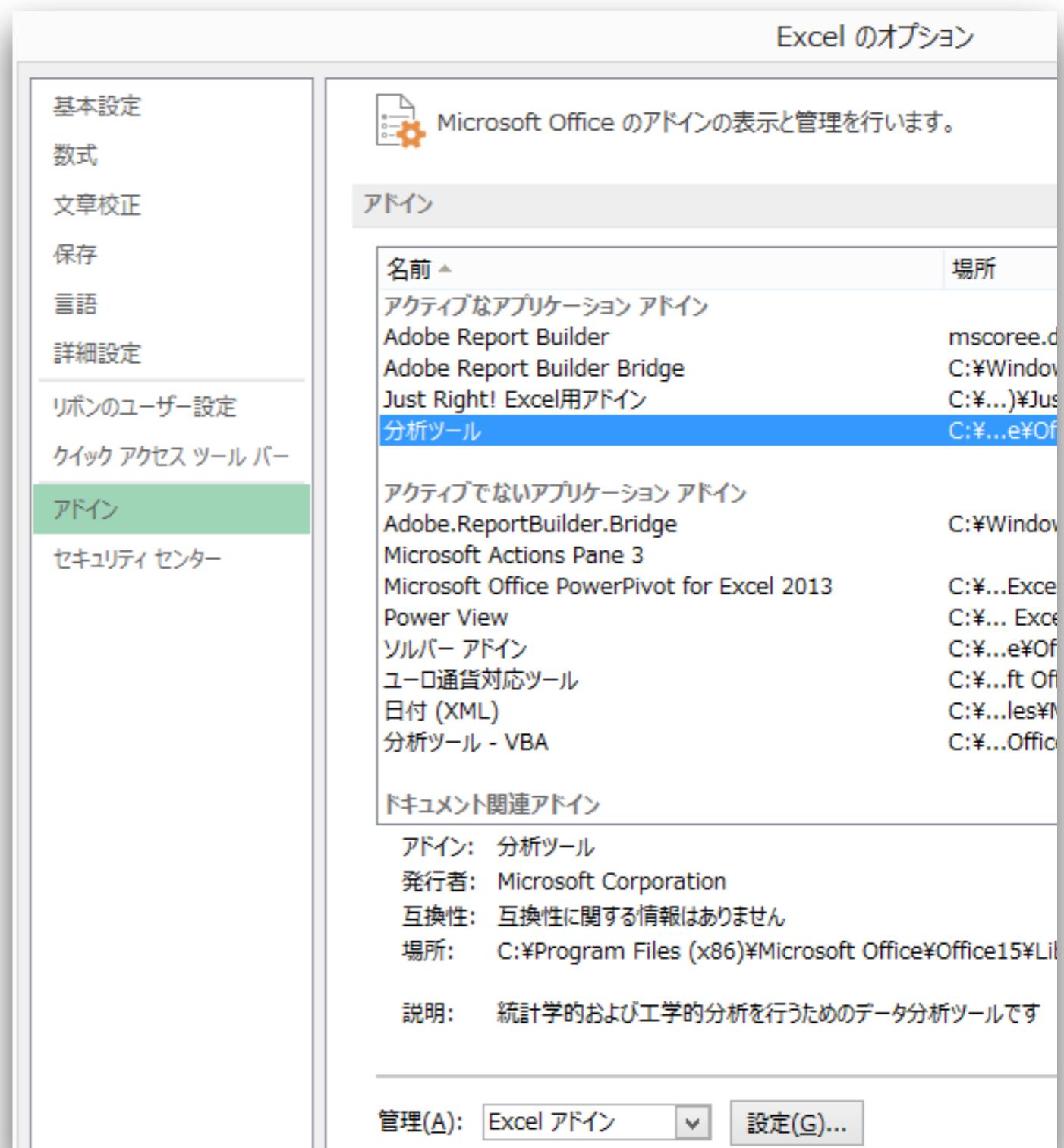
「回帰分析」

- 回帰直線の式の係数を求めることができる。散布図に「近似曲線を追加」→「線形近似」で追加するのと同じもの。説明に複数の変数を使う「重回帰分析」もできる

分析ツールのアドイン

デフォルトではツールが使えない
ので、使えるようにする
(一度だけやればOK)。

左上の「ファイル」タブから「オプション」→「アドイン」→
「Excel アドイン」と進み、「分析ツール」を有効にする
設定



相関を分析ツールで1

- ① 「データ」タブから「データ分析」（「分析」の中にある）を選ぶ
- ② 開いたパネルで「相関」を選ぶ

The screenshot shows the Microsoft Excel ribbon with the 'Data' tab selected (marked with a red circle labeled 1). In the bottom right corner of the ribbon, there is a 'Data Analysis' button (also marked with a red circle labeled 1). The main worksheet area displays a table of data with columns labeled C through R. A formula bar at the top shows '=CORREL(C2:C48,D2:D48)'. To the right of the table, there is a text box containing the results of the correlation analysis. A 'Data Analysis' dialog box is open in the foreground, with the 'Correlation' option highlighted (marked with a red circle labeled 2). The dialog box also lists other statistical analysis options like Descriptive Statistics and Histogram.

人口(C列)と世帯数(D列)の相関係数を求める。PEARSON
0.999819 列すべてを指定した
相関係数 = 0.99586 48行目まで全国計を含まない
0.999819 49行目までにしたら、列すべてと同じ値に
この場合、適切なのは全国計を除いたもの
決定係数 = 0.99175 署で囲ったセルの値を2乗した
近似曲線のR2乗値(決定係数)と同じ

データ分析

分散分析: 繰り返しのない二元配置
相関
共分散
基本統計量
指標平均
F検定: 2 標本を使った分散の検定
フーリエ解析
ヒストグラム
移動平均
乱数発生

相関を分析ツールで2

- ③ 「入力範囲」はXとYをまとめて指定。項目名の行（1行目）も含めるとよい。含めない場合は「列1」「列2」と表示されるので分かりにくい
- ④ 「先頭行をラベルとして使用」に✓
- ⑤ 「出力先」を指定して、「OK」

CORREL関数を使ったときと同じ結果になる。対角線に1が入っているのは、自分自身との相関係数を求めているから。

相関係数の2乗は、近似曲線で出てきたR2乗値に等しい。

相関を分析ツールで3 (完)

お疲れさまでした！

さて。各チェーンの店舗数データはどこから持ってきたか？

① ドトール

サイト (<https://sasp.mapion.co.jp/b/doutor/attr/>) からデータを貼り付けた。リンクを有効にしておくと点検・校閲が楽

② コメダ

いい一覧ページがない。サイト (<http://www.komeda.co.jp/search/>) で検索する。Googleスプレッドシートにやらせた

③ スターバックス

サイト (<http://www.starbucks.co.jp/store/search/>) で検索。同上

※「経済センサス」には、個人経営も含めた喫茶店数のデータがある。

喫茶店は小分類の767

ドトールの店舗数 1

- ① 店舗一覧サイト (<https://sasp.mapion.co.jp/b/doutor/attr/>) をブラウザで開く
- ② 左の「エリア絞込み」の列のデータをExcelのA列にコピペする。ChromeならCopyTablesを使う（Altを押して範囲指定）とよい。これはドトール系列の全ブランドの店舗数。もし「黄色いドトール」のみに絞るなら、店舗検索トップで指定してから、同様に

店舗検索トップ > 全国の店舗一覧

全国の店舗一覧

エリア絞込み

北海道 (26) 2
青森県 (13)
岩手県 (9)
宮城県 (27)
秋田県 (2)

○ 条件を指定して絞り込み 開く

1334件見つかりました。
1 ~ 20件を表示しています。
次の20件 >

ドトールの店舗数 2

- ③ 都道府県名を取り出す。まず、全角開きかっこが何文字目かを「SEARCH」関数で調べる。「B4」セルに「=search(" (",a4)」と入力する。引用符の中は全角かっこ。タイプミスがなければ、大文字になるはず。A列の中身のうち、全角開きかっこは何文字目か、という意味。きちんと動いたら、下までコピーで増やす。セルの右下角にマウスを当てるとき十字型になるので、ダブルクリックすると早い（フラッシュフィルという）。苦手ならドラッグで
- ④ 「LEFT」関数を使う。「C4」セルに、「=left(a4, b4-1)」と入れる。A列の中身のうち、B列に入っている数より 1 文字少ない文字数を、左端から取ってくる、という意味になる。つまり、かっここの手前まで。きちんと動いたら、下までコピーで増やす

ドトールの店舗数 3

- ⑤ 店舗数を取り出す。いろいろ方法は考えられるが、「MID」関数を使うことにする。LEFTに似ているが、左端ではなく、途中から指定文字数分を抜き出してくる関数。つまり、全角開きかっここの次から、××文字持ってくる、という指定にする。××に幅があるので困るが、10桁ということはないので、仮に10にする
- ⑥ 「D4」セルに「=mid(a4, b4+1, 10)」と入れる。開きかっこ次の数字から始まって最後の閉じかっこまでが抜き出されるはず。下までコピーして増やす
- ⑦ 閉じかっこを捨てる。「SUBSTITUTE」関数を使って、空文字に置き換える。「E4」セルに「=substitute(d4, " ", "")」と入れる。最初の引用符の中は全角閉じかっこ。後の引用符の中は何もなしで、引用符を2つ続けて打つ。コピーして下まで増やす

ドトルの店舗数 4

- ⑧ ちゃんと数字になっているか、余計な文字が紛れ込んでいないかを確かめる。「F4」セルに「=value(e4)」と入力し、下までコピー

3から8まで

A	B	C	D	E	F	G	H	I	J	K
1 サイトから貼り付けた	開き括弧は何文字目？	その1字前まで取得	店舗数の数字以降	閉じ括弧を捨てる	ちゃんと数値になるか					
2 加工に使う関数	SEARCH	LEFT	MID	SUBSTITUTE	VALUE					
3 式の書き方	=SEARCH("(",A4)	=LEFT(A4,B4-1)	=MID(A4,B4+1,10)	=SUBSTITUTE(D4,"")	=VALUE(E4)					
4 北海道(26)		4 北海道	26)	26		26				
5 青森県(13)		4 青森県	13)	13		13				
6 岩手県(9)		4 岩手県	9)	9		9				
7 宮城県(27)		4 宮城県	27)	27		27				
8 秋田県(2)		4 秋田県	2)	2		2				
9 山形県(12)		4 山形県	12)	12		12				
10 福島県(25)		4 福島県	25)	25		25				
11 茨城県(11)		4 茨城県	11)	11		11				
12 栃木県(10)		4 栃木県	10)	10		10				
13 群馬県(7)		4 群馬県	7)	7		7				
14 埼玉県(79)		4 埼玉県	79)	79		79				
15 千葉県(77)		4 千葉県	77)	77		77				
16 東京都(506)		4 東京都	506)	506		506				
17 神奈川県(175)		5 神奈川県	175)	175		175				
18 新潟県(6)		4 新潟県	6)	6		6				
19 富山県(6)		4 富山県	6)	6		6				
20 石川県(4)		4 石川県	4)	4		4				
21 福井県(3)		4 福井県	3)	3		3				
22 山梨県(2)		4 山梨県	2)	2		2				
23 長野県(4)		4 長野県	4)	4		4				
24 岐阜県(11)		4 岐阜県	11)	11		11				
25 静岡県(19)		4 静岡県	19)	19		19				
26 愛知県(48)		4 愛知県	48)	48		48				
27 三重県(7)		4 三重県	7)	7		7				
28 滋賀県(1)		4 滋賀県	1)	1		1				
29 京都府(22)		4 京都府	22)	22		22				
30 大阪府(81)		4 大阪府	81)	81		81				

ドトルの店舗数5（完）

- ⑨ シート全体をCtrl + ドラッグでコピー
- ⑩ そのシートをCtrl + Aで全範囲指定。Ctrl + Cでコピー。Alt + Ctrl + V、今度は単独でV。式を消して値だけにした

The screenshot shows a Microsoft Excel spreadsheet with data in columns A through F. Column A lists locations with their counts, and column E shows the formula =SUBSTITUTE(E4," ","") to remove spaces. A 'Paste' dialog box is open over the spreadsheet, specifically the 'Paste Options' section. The 'Values' option (radio button) is highlighted with a red circle. The number '10' is displayed prominently in orange in the center of the dialog box. The dialog box has buttons for 'OK' and 'Cancel' at the bottom.

A	B	C	D	E	F
1 サイトから貼り付けた	開き括弧は何文字目？	その1字前まで取得	店舗数の数字以降	閉じ括弧を捨てる	ちゃんと数値になるか
2 加工に使う関数	SEARCH	LEFT	MID	SUBSTITUTE	VALUE
3 式の書き方	=SEARCH(E(D4,"")	=VALUE(E4)
4 北海道(26)					26
5 青森県(13)					13
6 岩手県(9)					9
7 宮城県(27)					27
8 秋田県(2)					2
9 山形県(12)					12
10 福島県(25)					25
11 茨城県(11)					11
12 栃木県(10)					10
13 群馬県(7)					7
14 埼玉県(79)					79
15 千葉県(77)					77
16 東京都(506)					506
17 神奈川県(175)					175
18 新潟県(6)					6
19 富山県(6)					6
20 石川県(4)					4
21 福井県(3)					3
22 山梨県(2)					2

ここまでを反映：
[data_preparation_01.xlsx](#)

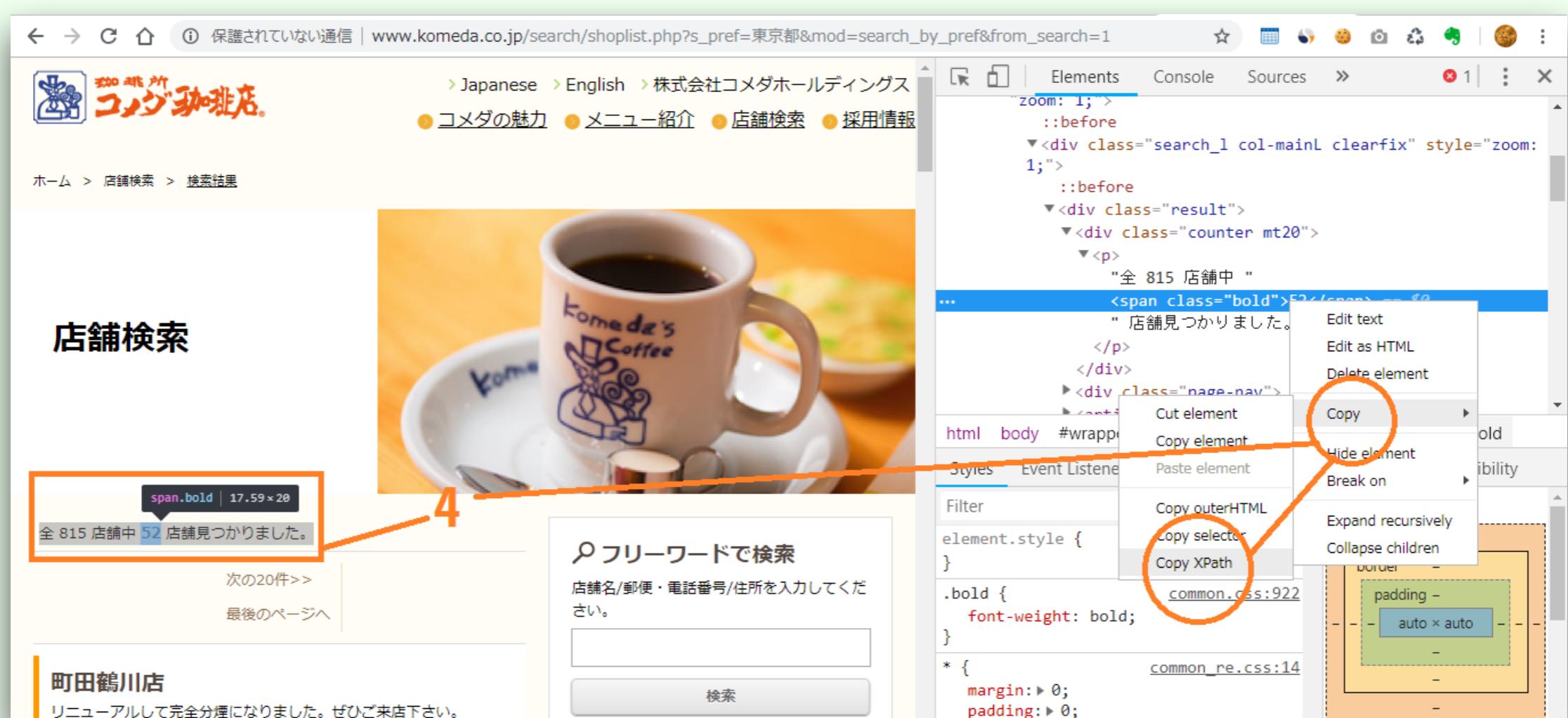
- ⑪ 不要な列や行を削除。列はCとFだけ残せばよい

コメダの店舗数 1

- ① サイト(<http://www.komeda.co.jp/search/>) で検索する。地図からの検索を試してみると、「http://www.komeda.co.jp/search/shoplist.php?s_pref=<<ここに都道府県名>>&mod=search_by_pref&from_search=1」で、その都道府県の店舗一覧を返してくれるようだ
- ② 47回繰り返したくない。Googleスプレッドシートを使って、自動取得することにする。シートを新規作成し、名前をつける
- ③ コードや都道府県名が必要になるので、A列とB列に入れておく
- ④ B列に都道府県名が入っているとして、C列に「IMPORTXML」という命令を入れる。取得先のURLと、XPathというものを指定する必要がある。XPATHは、Chromeなら「F12」（検証）→要素を右クリック→「Copy」→「Copy XPath」で調べる

コメダの店舗数 2

- ⑤ 「全815店舗中××店舗見つかりました」の太字××部分のXPathを調べる。④でクリップボードにコピーされるので、テキストエディタなどに貼り付ける。「//*[@id="search"]/div[1]/div/div[1]/p/span」だと分かる



コメダの店舗数3 (完)

- ⑥ Googleスプレッドシートの「C1」セルに以下を入力。URLの都道府県名部分をB列から持ってくるように指定している。XPathの中、searchを囲む引用符は一重に変更している

```
=importxml("http://www.komeda.co.jp/search/shoplist.php?  
s_pref=&B1&"&mod=search_by_pref&from_search=1", "//  
*[@id='search']/div[1]/div/div[1]/p/span")
```

The screenshot shows a Google Sheets interface with the following data:

	A	B	C	D	E	F	G	H	I	J	K
1		1 北海道	8								
2		2 青森県	0								
3		3 岩手県	1								
4		4 宮城県	5								
5		5 秋田県	1								

- ⑦ うまく動いたら、下までコピーする

スタバの店舗数 1

- ① サイト(<http://www.starbucks.co.jp/store/search/>) で検索して試す。「http://www.starbucks.co.jp/store/search/result.php?search_type=1&pref_code=<<ここに都道府県コード>>」で、都道府県の一覧を返してくれる
- ② XPathは「/html/body/div[2]/article/header/h2」と出たが、単に「//h2」に直さないと動作しなかった。数字だけでなく、見出し部分を丸ごと取ってくるしかなさそう
- ③ コメダの隣、D列にスタバのデータを入れる。コメダと区別がつくよう、1行挿入して項目名を入れた
- ④ 「D2」セルに以下を入力

```
=importxml("http://www.starbucks.co.jp/store/search/
result.php?search_type=1&pref_code=&A1, "//h2")
```

スタバの店舗数 2

⑤ うまく動いたら、下までコピー

JNPC_cafe_data

ファイル 編集 表示 挿入 表示形式 データ ツール アドオン ヘルプ 変更内容をすべてドライブに保存しました

fx =importxml("http://www.starbucks.co.jp/store/search/result.php?search_type=1&pref_code=""&A2, "//h2")

	A	B	C	D	E	F	G	H	I	J	K	L
1	都道府県コード	都道府県名	コメダ店舗数	スタバ店舗数（加工前）								
2	1 北海道		8 店舗検索：北海道で検索した結果（35 件）									
3	2 青森県		0 店舗検索：青森県で検索した結果（9 件）									
4	3 岩手県		1 店舗検索：岩手県で検索した結果（8 件）									
5	4 宮城県		5 店舗検索：宮城県で検索した結果（22 件）									
6	5 秋田県		1 店舗検索：秋田県で検索した結果（7 件）									
7	6 山形県		3 店舗検索：山形県で検索した結果（7 件）									
8	7 福島県		6 店舗検索：福島県で検索した結果（9 件）									
9	8 茨城県		12 店舗検索：茨城県で検索した結果（31 件）									
10	9 栃木県		10 店舗検索：栃木県で検索した結果（24 件）									
11	10 群馬県		7 店舗検索：群馬県で検索した結果（15 件）									
12	11 埼玉県		28 店舗検索：埼玉県で検索した結果（66 件）									
13	12 千葉県		24 店舗検索：千葉県で検索した結果（67 件）									
14	13 東京都		52 店舗検索：東京都で検索した結果（321 件）									
15	14 神奈川県		35 店舗検索：神奈川県で検索した結果（105 件）									
16	15 新潟県		8 店舗検索：新潟県で検索した結果（13 件）									
17	16 富山県		6 店舗検索：富山県で検索した結果（9 件）									
18	17 石川県		8 店舗検索：石川県で検索した結果（10 件）									
19	18 福井県		6 店舗検索：福井県で検索した結果（6 件）									
20	19 山梨県		4 店舗検索：山梨県で検索した結果（11 件）									
21	20 長野県		10 店舗検索：長野県で検索した結果（20 件）									
22	21 岐阜県		31 店舗検索：岐阜県で検索した結果（13 件）									
23	22 静岡県		29 店舗検索：静岡県で検索した結果（31 件）									
24	23 愛知県		239 店舗検索：愛知県で検索した結果（97 件）									
25	24 三重県		30 店舗検索：三重県で検索した結果（16 件）									

スタバの店舗数 3

- ⑤ Excelに戻って、スタバの店舗数だけを切り出す。CSVでダウンロードしても、Excelにコピペしてもいい。が、Excel形式でダウンロードすることはおすすめしない。スタバで都道府県コードを順々に差し込んでいく部分が、すべて北海道に化けてしまった
- ⑥ CSV形式のデータを読み込み、42～43ページの方法を応用してスタバの店舗数の数字だけを抜き出す。D列 4 行目からスタバのデータが入っているなら、このように。引用符中のかっこは全角
- E4← =search("結果 (", D4)
 - F4← =mid(d4, e4+3, 10)
 - G4← =substitute(f4, " 件) ","")
 - H4← =value(G4)

スタバの店舗数4（完）

⑦ シートを複写し、値だけに。不要な列と行を削って完成

The screenshot shows an Excel spreadsheet titled "data_preparation_02.xlsx". The spreadsheet has a header row with formulas and data rows below it. A note in cell A1 says: "A~D列はGoogleスプレッドシートを読み込んだもの。E列以降でスタバを数値のみに加工". The columns are labeled A through H. Column A contains row numbers from 1 to 33. Columns B and C contain data from Google Sheets. Column D contains the formula: "=SEARCH("結果(",D4)&MID(D4,E4+3,10)". Column E contains the result of the formula, which is the number of stores. Column F contains the formula: "=SUBSTITUTE(E4,"件","",")". Column G contains the result of the formula, which is the value without the character "件". Column H contains the formula: "=VALUE(G4)". The last row (row 33) shows the formula for the last data point: "=SUBSTITUTE(E32,"件","",")". The text "ここまでを反映" and "data preparation 02.xlsx" is overlaid on the right side of the screenshot.

A	B	C	D	E	F	G	H
1	A~D列はGoogleスプレッドシートを読み込んだもの。E列以降でスタバを数値のみに加工			「結果(」は何文字目から? それ以降を持ってきて		アキ+件)を捨てる	数値になるか
2				SEARCH	MID	SUBSTITUTE	VALUE
3	都道府県コード	都道府県名	コメダ店舗数	=SEARCH("結果(",D4)	=MID(D4,E4+3,10)	=SUBSTITUTE(F4,"件","",")")	=VALUE(G4)
4	1	北海道	8 店舗検索:北海道で検索した結果(35 件)	14 35 件)	35		35
5	2	青森県	0 店舗検索:青森県で検索した結果(9 件)	14 9 件)	9		9
6	3	岩手県	1 店舗検索:岩手県で検索した結果(8 件)	14 8 件)	8		8
7	4	宮城県	5 店舗検索:宮城県で検索した結果(22 件)	14 22 件)	22		22
8	5	秋田県	1 店舗検索:秋田県で検索した結果(7 件)	14 7 件)	7		7
9	6	山形県	3 店舗検索:山形県で検索した結果(7 件)	14 7 件)	7		7
10	7	福島県	6 店舗検索:福島県で検索した結果(9 件)	14 9 件)	9		9
11	8	茨城県	12 店舗検索:茨城県で検索した結果(31 件)	14 31 件)	31		31
12	9	栃木県	10 店舗検索:栃木県で検索した結果(24 件)	14 24 件)	24		24
13	10	群馬県	7 店舗検索:群馬県で検索した結果(15 件)	14 15 件)	15		15
14	11	埼玉県	28 店舗検索:埼玉県で検索した結果(66 件)	14 66 件)	66		66
15	12	千葉県	24 店舗検索:千葉県で検索した結果(67 件)	14 67 件)	67		67
16	13	東京都	52 店舗検索:東京都で検索した結果(321 件)	14 321 件)	321		321
17	14	神奈川県	35 店舗検索:神奈川県で検索した結果(105 件)	15 105 件)	105		105
18	15	新潟県	8 店舗検索:新潟県で検索した結果(13 件)	14 13 件)	13		13
19	16	富山県	6 店舗検索:富山県で検索した結果(9 件)	14 9 件)	9		9
20	17	石川県	8 店舗検索:石川県で検索した結果(10 件)	14 10 件)	10		10
21	18	福井県	6 店舗検索:福井県で検索した結果(6 件)	14 6 件)	6		6
22	19	山梨県	4 店舗検索:山梨県で検索した結果(11 件)	14 11 件)	11		11
23	20	長野県	10 店舗検索:長野県で検索した結果(20 件)	14 20 件)	20		20
24	21	岐阜県	31 店舗検索:岐阜県で検索した結果(13 件)	14 13 件)	13		13
25	22	静岡県	29 店舗検索:静岡県で検索した結果(31 件)	14 31 件)	31		31
26	23	愛知県	239 店舗検索:愛知県で検索した結果(97 件)	14 97 件)	97		97
27	24	三重県	30 店舗検索:三重県で検索した結果(16 件)	14 16 件)	16		16
28	25	滋賀県	16 店舗検索:滋賀県で検索した結果(15 件)	14 15 件)	15		15
29	26	京都府	15 店舗検索:京都府で検索した結果(33 件)	14 33 件)	33		33
30	27	大阪府	50 店舗検索:大阪府で検索した結果(102 件)	14 102 件)	102		102
31	28	兵庫県	29 店舗検索:兵庫県で検索した結果(49 件)	14 49 件)	49		49
32	29	奈良県	13 店舗検索:奈良県で検索した結果(11 件)	14 11 件)	11		11
33	30	和歌山県	8 店舗検索:和歌山県で検索した結果(7 件)	15 7 件)	7		7

ドトールの分と併せて1枚のシートにし、CSVで書き出したものが、今回の教材用データ

データの出所について

- ・喫茶店数の人口と世帯数は、総務省の住民基本台帳から

http://www.soumu.go.jp/main_sosiki/jichi_gyousei/daityo/jinkou_jinkoudoutai-setaisuu.html

- ・東京の天候データは気象庁と、電力需要は東京電力の以下のページから。天候は、ここだと1ヶ月分ぐらいまとめてダウンロード可能。欠測や、降水の有無（降水ありなのに量は0ミリ）に注意

<https://www.data.jma.go.jp/risk/obssl/index.php>

<http://www.tepco.co.jp/forecast/html/download-j.html>

- ・肝疾患のデータは厚労省の人口動態統計特殊報告「平成20～24年 人口動態保健所・市区町村別統計」から第3表と5表。人口は住民基本台帳の平成24年分。自治体の並び順の違いや、熊本市の政令指定市移行による自治体コード変更に注意

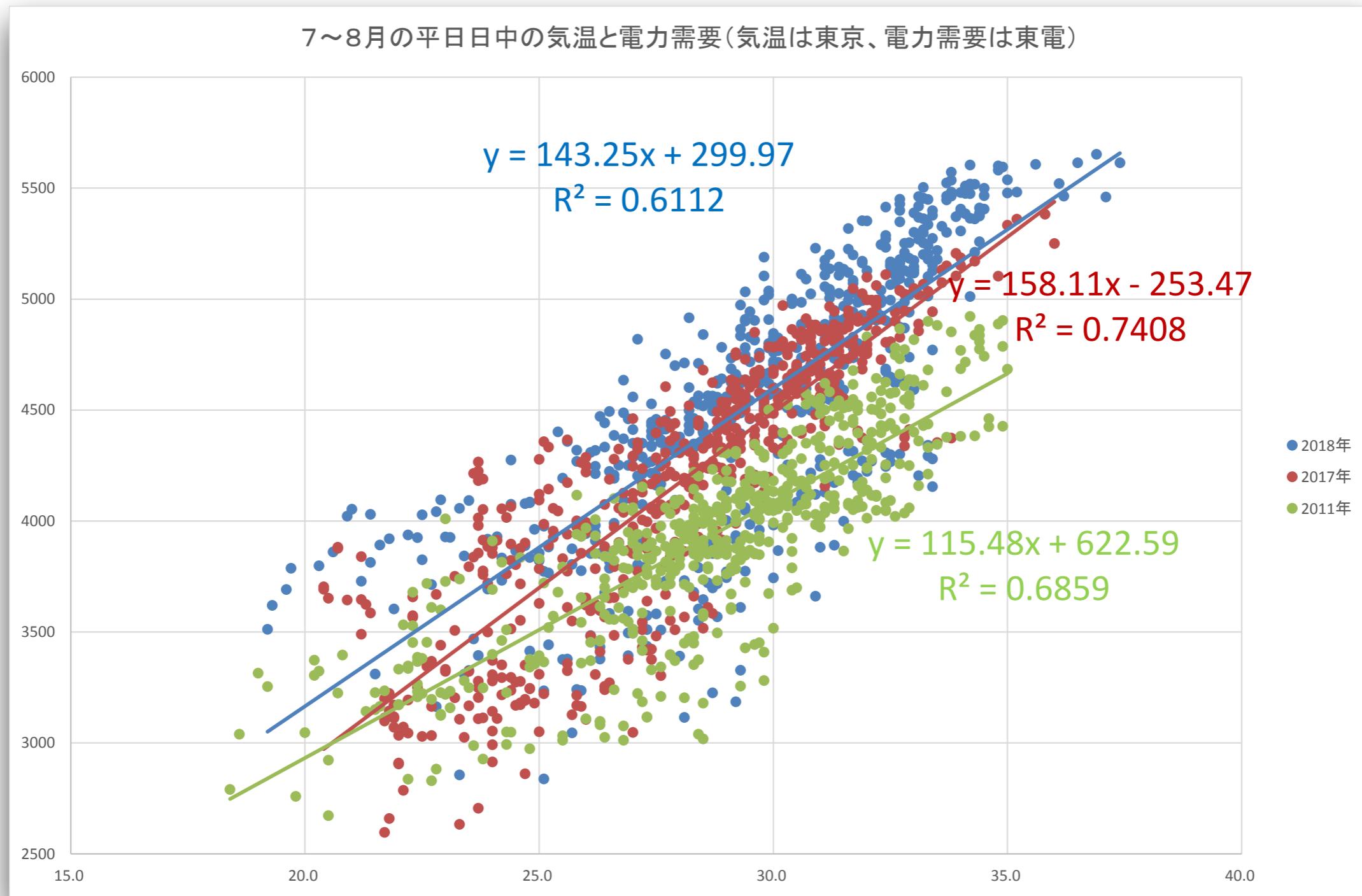
<https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450013&tstat=000001063680&cycle=0>

トレンドに着目した例

猛暑の今夏、電力需要は例年に比べてどうだったのだろう。必要なときには冷房を使わないと命に関わる、ということが広く認識されて、昨年と変わった点はないか。節電意識の高かった2011年とも比較。

- ① データは、electricity18/17/11.csvとclimate18/17/11.csv。読み込んで成型する過程が、electricity_excel01/02.xlsx。結合の手順をざっと見れば結構。東京だけの気温で、東京電力全体の電力需要を説明するのは、かなり大ざっぱではある。エアコンの効率向上や電気料金の変化も無視している
- ② electricity_03.xlsx の1枚目のシートをスタート台にして、気温と電力需要の関係を散布図にして下さい。今年と昨年、2011年それぞれに色分けし、回帰直線も年ごとに引く

気温×電力需要



トレンドが変化している

気温と電力需要には直線的な関係があることが分かる。その直線が、昨年と今年では別物になっている。今年は、より積極的に冷房を使うようになったのではないだろうか。

2011年の回帰直線は昨年よりもさらに低い位置にある。東日本大震災後の節電を反映したものだろう。さらに、今年や昨年ほどは気温が上がっていない。涼しい年だったといえるのではないか。

Excelでの作業例：
[electricity_excel_03.xlsx](#)

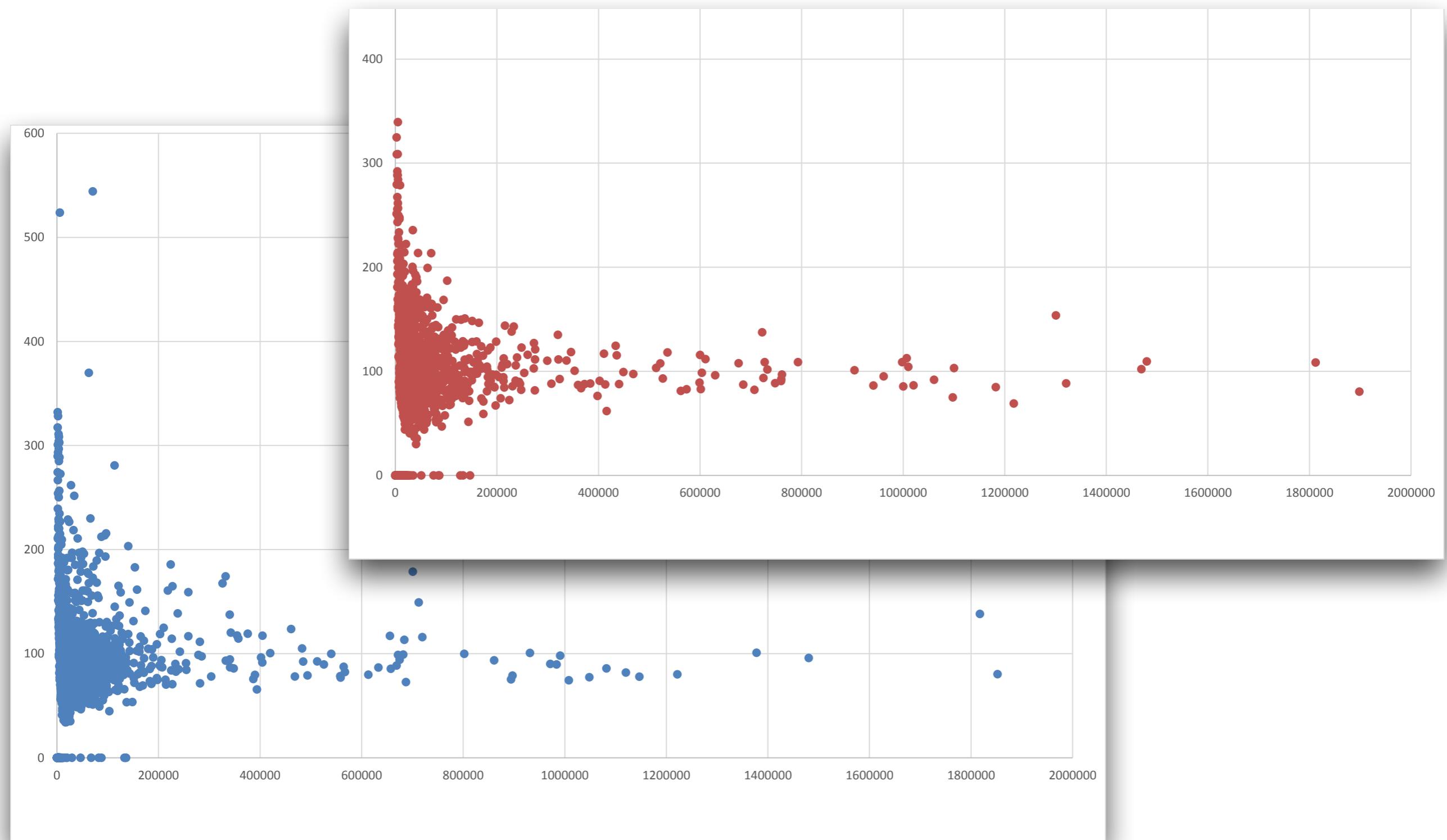
散布図と回帰直線：
[electricity1\)excel.pdf](#)

外れ値にニュースがある例

肝疾患による自治体ごとの死亡率を縦軸に、自治体の人口を横軸にしてプロットしてみる。統計学を持ち出さなくても、見ただけで分かることがある。死亡率が極端に高いのは、小規模な自治体。人口の多い自治体は、全国平均からさほどぶれない。なお、死亡率は全国平均を100として標準化・年齢調整ずみ。

- ① データは、liver_pop.csv。53ページの出所のデータをすでに結合すみ。やってみれば分かるが、突き合わせには難渋した
- ② [liver_excel_01.xlsx](#)にデータを読み込み、男女別の散布図にしてある。死亡率が高いのはどこの自治体？ この図の各点に自治体名のラベルをつけたらどうなる？

肝疾患の死亡率×自治体人口



自治体名を見やすくしてみた

飛び抜けて高い小規模自治体もニュースだろうが、それ以上に気になるのは、人口の多い自治体の外れ値。どんな理由があるのか。改善する方法はあるのか。ニュースの種、取材のきっかけになりそう。

確率算定にあたって年齢調整はしていないので、参考程度に。HTML版は自治体名をポップアップ式にして見やすくしてみたもの。

Excelでの作業例：

[liver_excel_01.xlsx](#)

各種散布図：

[liver0\)Female.pdf](#)

[liver0\)Male.pdf](#)

[liver1-1\)女性の死亡率と自治体規模.html](#)

[liver1-2\)女性の死亡率に等確率線を重ねた.pdf](#)

[liver2-1\) 男性の死亡率と自治体規模.html](#)

[liver2-2\) 男性の死亡率に等確率線を重ねた.pdf](#)

[liver3\) 男性と女性の死亡率.html](#)

ご参加 ありがとうございます

- ・質問はみんなの宝物です。ご遠慮なく質問を。分からるのは、説明が悪いから。たぶん、ほかの受講者も困っています
- ・内容改善につなげるために、ぜひフィードバックをお願いします
- ・そのほか、お問い合わせは

yoshito.nishio+JNPC@gmail.com へ

川上さん、田中さん、山本さんはじめ、資料作成やヘルプにご協力いただいたアドバイザリーのみなさま、いろいろと助けていただいている事務局のみなさまに深謝します。