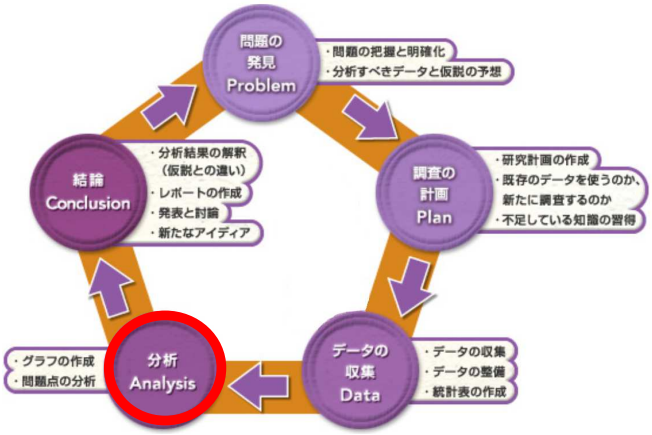
数理・データサイエンス・AI入門

第9回
データサイエンス実践(4)

クロス集計を使った分析と解釈



クロス集計の分析

- この表のどこに着目すると良いか一目で分かりますか？

	好き	普通	嫌い	合計
50代 以上	5	1	5	11
30代 ～40代	1	1	10	12
20代 以下	2	3	2	7
合計	8	5	17	30

クロス集計の分析手順

- まずは総数で全体を割ります

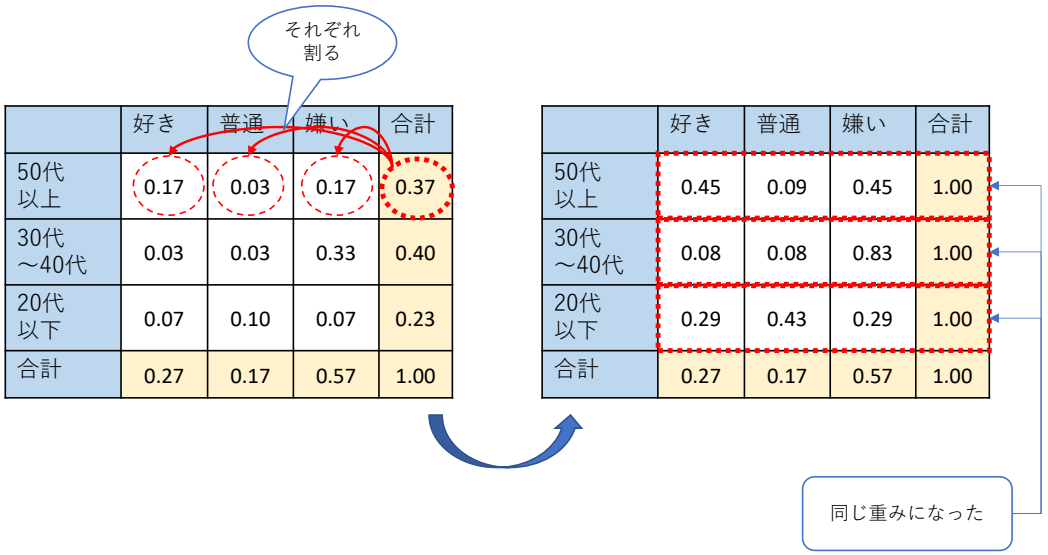
	好き	普通	嫌い	合計
50代 以上	5	1	5	11
30代 ～40代	1	1	10	12
20代 以下	2	3	2	7
合計	8	5	17	30

	好き	普通	嫌い	合計
50代 以上	0.17	0.03	0.17	0.37
30代 ～40代	0.03	0.03	0.33	0.40
20代 以下	0.07	0.10	0.07	0.23
合計	0.27	0.17	0.57	1.00

重みがバラバラで
比較できない

クロス集計の分析手順

- 横方向に割り算して縦項目同士の重みを揃える

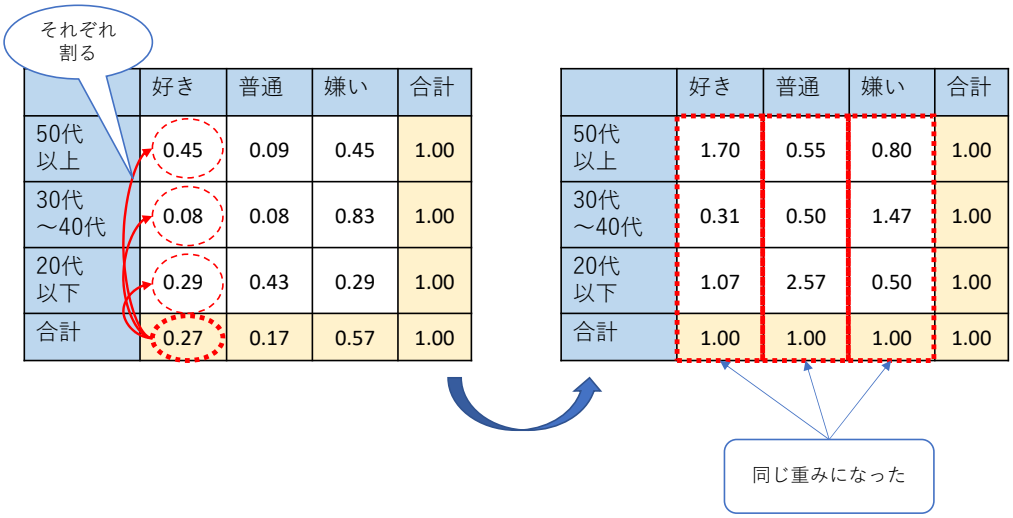


	好き	普通	嫌い	合計
50代以上	0.17	0.03	0.17	0.37
30代～40代	0.03	0.03	0.33	0.40
20代以下	0.07	0.10	0.07	0.23
合計	0.27	0.17	0.57	1.00

	好き	普通	嫌い	合計
50代以上	0.45	0.09	0.45	1.00
30代～40代	0.08	0.08	0.83	1.00
20代以下	0.29	0.43	0.29	1.00
合計	0.27	0.17	0.57	1.00

クロス集計の分析手順

- 同様に縦方向に割り算して横項目同士の重みを揃える



	好き	普通	嫌い	合計
50代以上	0.45	0.09	0.45	1.00
30代～40代	0.08	0.08	0.83	1.00
20代以下	0.29	0.43	0.29	1.00
合計	0.27	0.17	0.57	1.00

	好き	普通	嫌い	合計
50代以上	1.70	0.55	0.80	1.00
30代～40代	0.31	0.50	1.47	1.00
20代以下	1.07	2.57	0.50	1.00
合計	1.00	1.00	1.00	1.00

クロス集計の分析手順

- 最終的に計算された値を見て
  数値が特に大きい／小さい箇所にフォーカスして考察する
  - 実数では見えにくい課題を抽出することができる

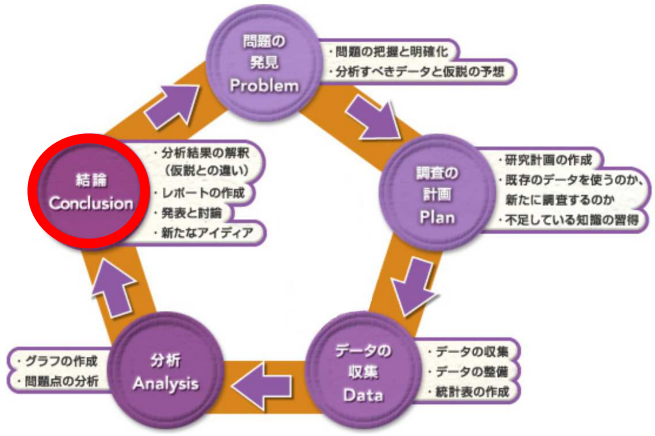
	好き	普通	嫌い	合計
50代以上	5	1	5	11
30代～40代	1	1	10	12
20代以下	2	3	2	7
合計	8	5	17	30

実数では特に多い／少ないと
言えるのかどうか
判断できない

	好き	普通	嫌い	合計
50代以上	1.70	0.55	0.80	1.00
30代～40代	0.31	0.50	1.47	1.00
20代以下	1.07	2.57	0.50	1.00
合計	1.00	1.00	1.00	1.00

30・40代に好まれていない
また、20代は普通評価が
特に多いと言える

分析結果の
読み取りと解釈



結果を読み解く際のポイント

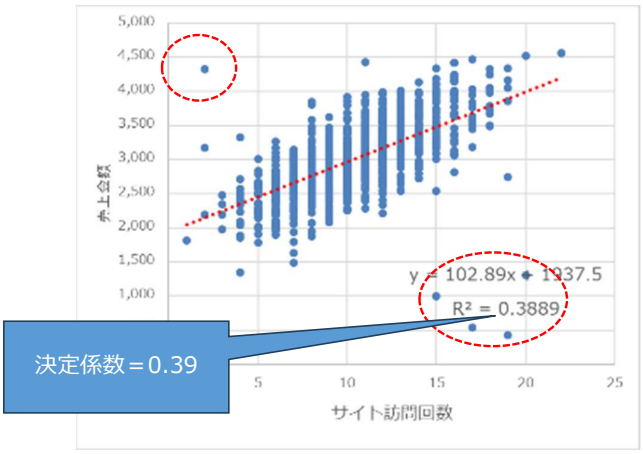
- 結果を考察する
  - 結果は他の人にも共有し、他者の視点を入れるようにしましょう
  - 新たな観点や見逃していた課題が出て来たら、次のテーマにしましょう

- 結果を当初の目的と接合させる
  - 結果（当初の仮説）ありきで解釈を歪めることが無いようにしましょう
  - 仮説と異なれば理由をしっかりと考えましょう

外れ値による影響

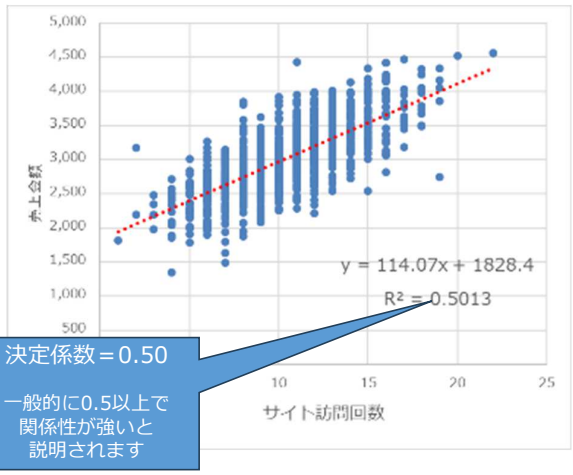
- 関係性が強く出ない場合は外れ値の存在を疑ってみる

外れ値がある結果



決定係数＝0.39

外れ値を取り除いた結果

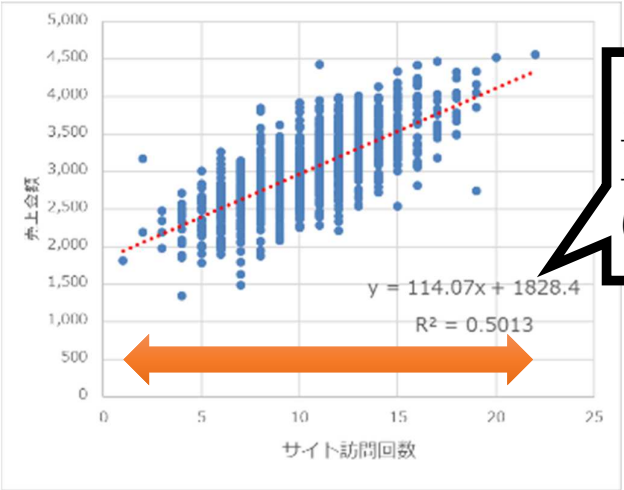


決定係数＝0.50

一般的に0.5以上で
関係性が強いと
説明されます

データのカバー範囲

- 予測対象が元データの範囲外になっていないか確認が必要

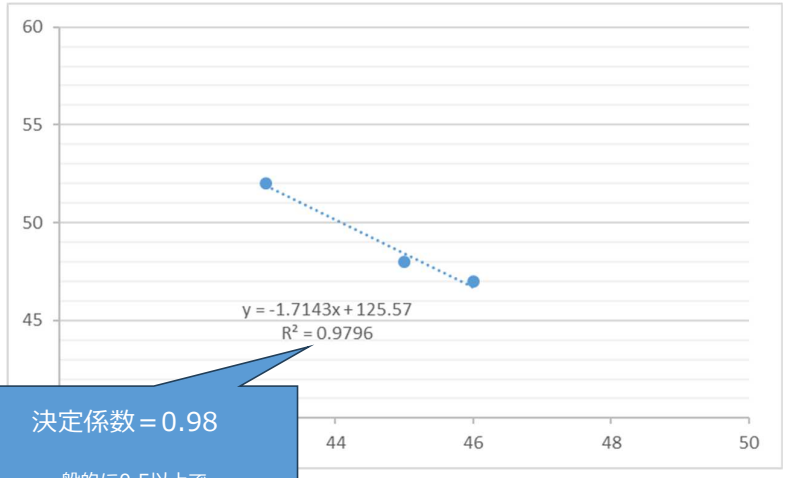


このモデル式なら
サイト訪問回数(x=)５回なら売り上げ(y=)約2,500円
サイト訪問回数(x=)２０回なら売り上げ(y=)約4,000円
になると予測ができる

ならば、訪問回数100回(x=100)のときに
売り上げ(y)は約13,000円になるといえるか？
→データが存在する範囲についての説明しかできません

データ量不足による影響

- 決定係数だけ見ると非常に良い分析結果に見える



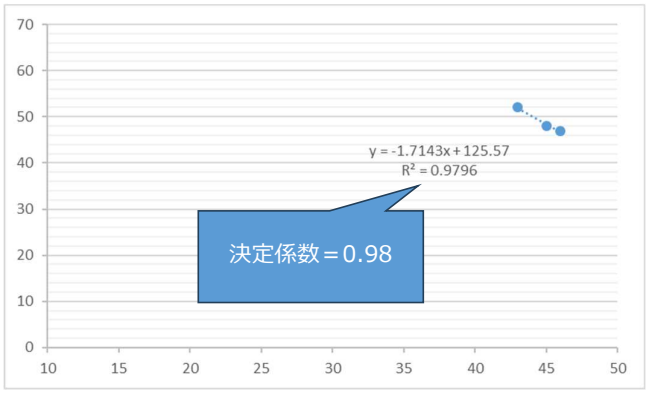
決定係数＝0.98

一般的に0.5以上で
関係性が強いと説明されます

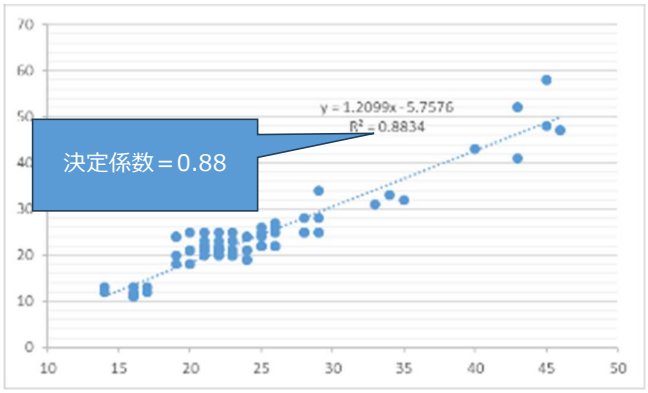
データ量不足による影響

- 元データを恣意的に選択することで意図的な結果にもできる

データが少ないときの分析結果

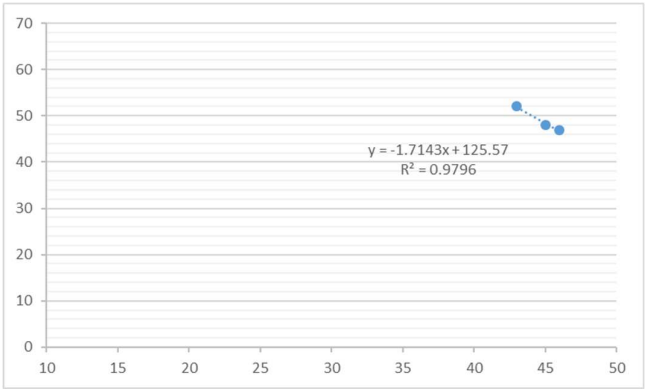


データ量を十分に増やした分析結果



データ量不足による影響

- データ量が十分でない場合、モデルの信頼性の指標が著しく悪化



概要

回帰統計	
重相関 R	0.989743319
重決定 R2	0.979591837
補正 R2	0.959183673
標準誤差	0.3086067
観測数	3

Excelで回帰分析した場合は
有意Fがモデルの信頼性の指標

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	4.571428571	4.571428571	48	0.091257897
残差	1	0.095238095	0.095238095		
合計	2	4.666666667			

仮説検定

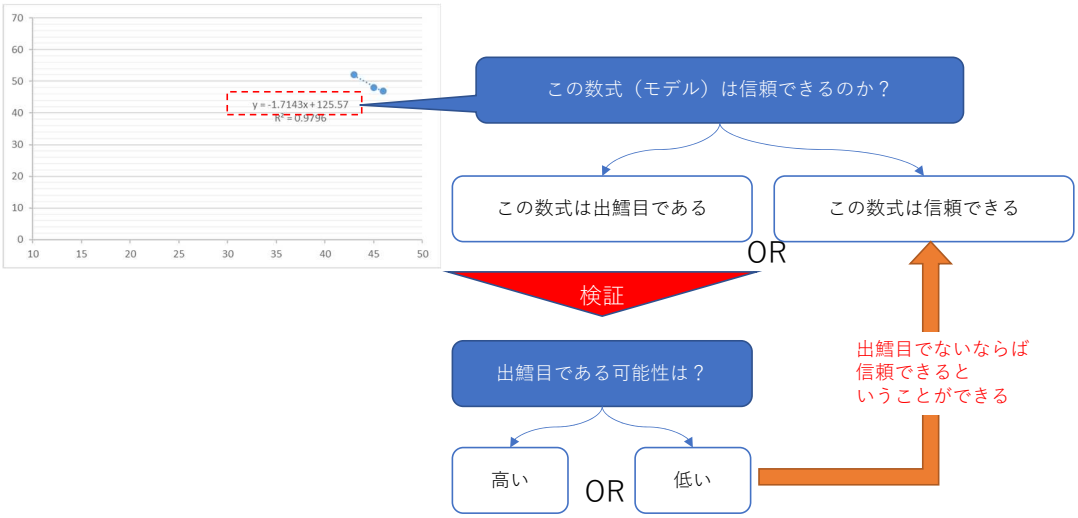
仮説検定

- 仮説検定とは、各種の統計分析を行う上で算出した統計量がどの程度信頼できるのかを判定する手法
  - 分析結果の信頼性を検証するためには仮説検定の考え方を使います

- 用語の理解
  - 対立仮説＝本来主張したい内容
  - 帰無仮説＝主張の逆の内容

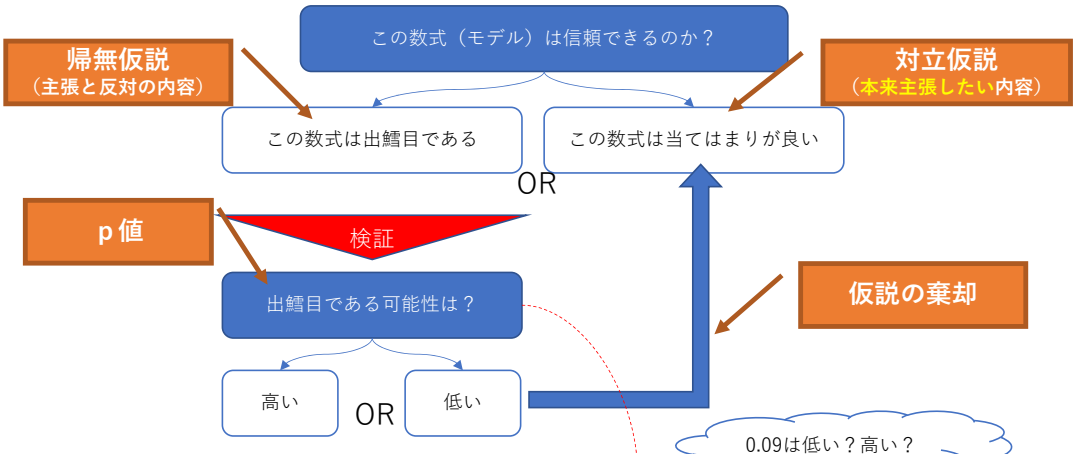
仮説検定

- モデルの信頼性を仮説検定で検証
  - 検証したい回帰式が「出鱈目である」という仮説（帰無仮説）を立てます
  - その可能性が低い場合「当てはまりが良い」と判断します



回帰分析における仮説検定

- モデル全体の信頼度は「有意Ｆ」の数値で判断



分散分析表

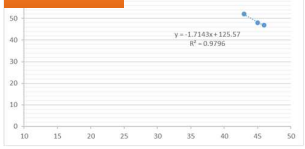
	自由度	変動	分散	観測された分散比	有意 F
回帰	1	4.571428571	4.571428571	48	0.091257897
残差	1	0.095238095	0.095238095		
合計	2	4.666666667			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	72.66666667	4.045377531	17.96288878	0.03540429	21.26527153	124.0680618	21.26527153	124.0680618
X 値 1	−0.571428571	0.08247861	−6.92820323	0.091257897	−1.619418675	0.476561532	−1.619418675	0.476561532

回帰分析における仮説検定

- ２つのモデルを比較
  - 下のほうが相関係数はやや下がるものの信頼性は非常に高いといえます

決定係数 $R^2 = 0.9796$



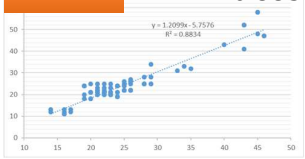
分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	4.571428571	4.571428571	48	0.091257897
残差	1	0.095238095	0.095238095		
合計	2	4.666666667			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	72.66666667	4.045377531	17.96288878	0.03540429	21.26527153	124.0680618	21.26527153	124.0680618
X 値 1	−0.571428571	0.08247861	−6.92820323	0.091257897	−1.619418675	0.476561532	−1.619418675	0.476561532

小数以下47桁までゼロ
＝限りなくゼロ

決定係数 $R^2 = 0.8834$



分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	4551.447932	4551.447932	742.4230791	1.58945E−47
残差	98	600.7920684	6.13053131		
合計	99	5152.24			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	7.090995549	0.694126643	10.21570865	4.12022E−17	5.713523817	8.468467282	5.713523817	8.468467282
X 値 1	0.730124151	0.026796063	27.2474417	1.58945E−47	0.676948237	0.783300065	0.676948237	0.783300065

演習

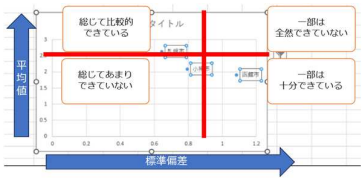
演習内容

- SDGsの取り組みに課題がある自治体が、なぜ推進されないのかデータ分析を通じて原因を調査したいと思います

- 今日のゴール
  - SDGsの取り組みに対して
    課題がない自治体を「0」
    課題がある自治体を「1」
    として、各自が考えた条件で分類した結果を
    「目的変数」の列に記入していきましょう

作業内容の確認

- ここまでの作業のおさらい



② 平均と標準偏差を使って散布図を描き自治体を２つのグループに分けた

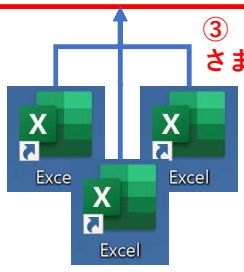
① 自治体ごとのアンケートのスコアの平均と標準偏差を計算した

	自治体名	目的変数	説明変数1	説明変数2	説明変数3	説明変数4	説明変数5	説明変数6	平均値	標準偏差	SDGｓについて理解している自	自治体内における意見発表会	2030年のあるべき姿を設定した将来のビジョンづくり	将来展望の設定等によるバックキャスト思考による価値づくり	SDGs推進のために、自治体で作成している各種方針との計画との	SDGs推進のために、自治体で作成している各種方針を総合的	SDGsに関する取組の現場視察、視察の	国内外を問わないステークホルダー（関係者）との連携	ローカル指標（自治体独自の評価指標）の設定	その他ＳＤＧｓ推進に向けた取組	「地方創生SDGs金融」に関する取組内容
12																					
13	北海道	1							3	0.4264014	4	3	3	3	3	3	3	3	3	3	2
14	札幌市	1							2.83838384	0.84282439	4	3	3	2	3	2	3	2	3	2	2
15	函館市	0							1.90909091	1.08330884	4	1	1	1	3	3	1	3	1	2	1
16	小樽市	0							2.09090909	0.79292708	4	2	2	1	2	3	2	2	2	2	1
17	旭川市	0							2.38383838	0.77138922	4	2	3	2	3	3	2	2	2	2	1
18	室蘭市	0							1.54545455	0.4979296	2	2	1	1	2	2	2	2	1	1	1

作業内容の確認

- ここまでの作業のおさらい

③ さまざまなデータを収集



こちらはもう分析には使わない

但し、削除してしまうと
関数の計算結果が変わるので残しておく

作業内容の確認

◘ これから行うこと

0 ＝SDG s に取り組めている
1 ＝SDG s に取り組めていない

目的変数 ← 関係性の強さ → **説明変数**

●●（要因①）が多いほど「SDG s に取り組めている」のではないか？　等の仮説

結果

仮説で置いた要因①
仮設で置いた要因②
仮説で置いた要因③

作業内容の確認

◘ 結果を記録

この「列名」を記入します

分析結果の記録												
	説明変数											
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩		
列名を記入→	婚姻有無	正社員か否か	パートか否か	最大連続利用月数	年間購入回数	サイト閲覧時間	サービスへの満足度	年齢	性別			補正R2
1回目	○	○	○	○	○	○	○	○	○		→	0.438
2回目			○	○	○	○	○		○		→	0.441
3回目						○	○				→	0.443
4回目											→	
5回目											→