

数理・データサイエンス・AI入門

第6回 データサイエンス実践(1)

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

0

データサイエンスの基本知識

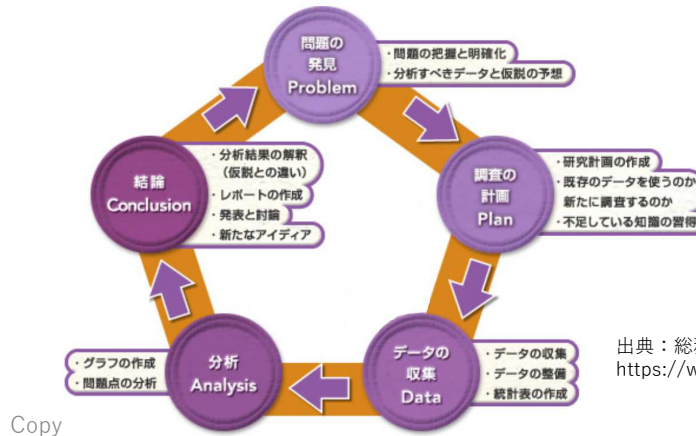
Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

1

データを分析する際のステップ

□ PPDACサイクル

- 「分析」テクニックを使いこなすためには「問題の発見（**仮説構築**）」と「計画」がしっかり出来上がっていることが条件
- 「結果を解釈」し仮説の検証をして次の分析サイクルを回すことが不可欠
—分析の細かなテクニックだけでなく、これら「問題の発見」「結果解釈」等の場面で必要な力が求められます

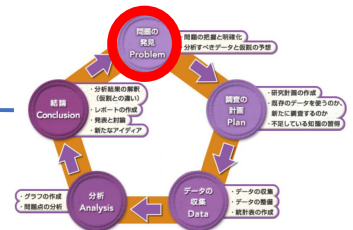


Copy

2

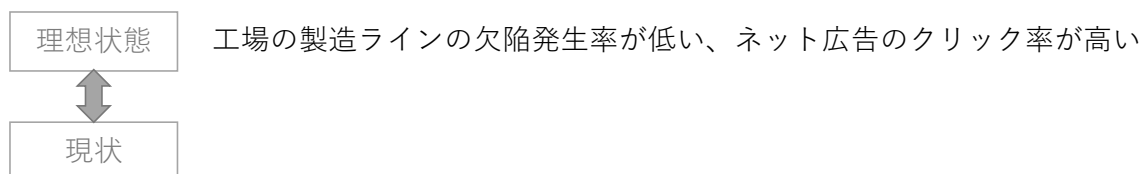
【Problem】課題の設定

分析で明らかにしたいことを定める



□ 問題を明示し、分析すべき課題（テーマ）を設定する

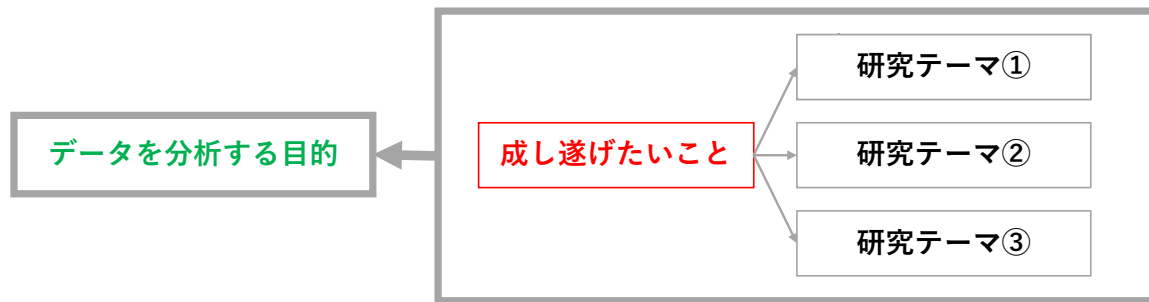
- 問題の明示とは、最終的に解決したい事項について、その理想状態に対して現状がどのような状態であるかを把握すること



- その問題となっている事象を捉えたデータ（可能性のあるもの）には何があるかを考えること

分析を行う目的とは

- 分析する動機として必ず何か成し遂げたいものがあるはず
 - ・ 但しデータがあるからという理由だけで着手するのはNG

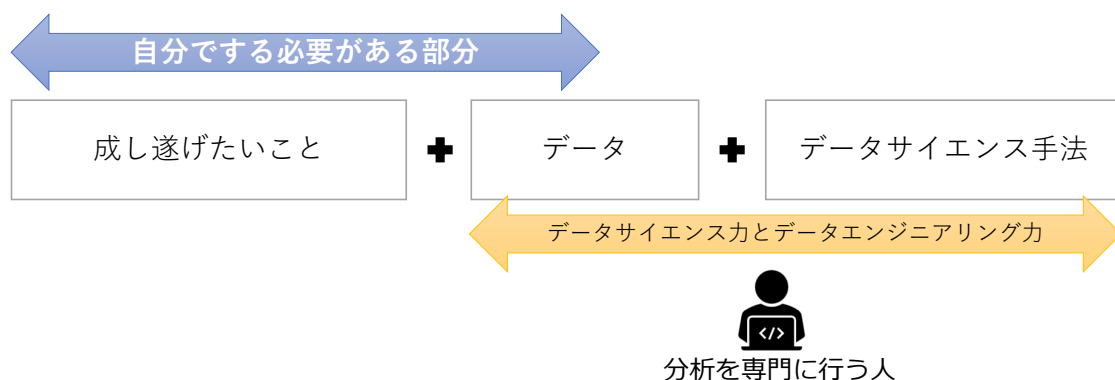


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

4

目的設定をするのは誰か？

- 分析の目的設定は分析をしたいと思った本人にしかできない



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

5

大事なポイント

□ 「当たり前」をスルーしない

- ・「なぜこう言えるのか／なぜわかるのか」と考えてみる（×経験・勘）
- ・経験則で疑問を持たないものにこそ「なぜ」と疑問を抱く姿勢を持つ

□ 各分野の問題点や先行研究等を事前に調査することが必要

- ・先人達が同じような課題に対してどのようなアプローチを採ってきたかも調べる

【Plan】計画

仮説を立て、分析アプローチをイメージする



□ どこからデータを集めるか計画する

- ・既にデータ化されているもの以外にも新規取得や代替候補を考える
- ・直接的に捉えられるデータ以外にも、間接的に事象を捉えられるデータは無いか考える

□ 分析を進めるための具体的な計画を立てる

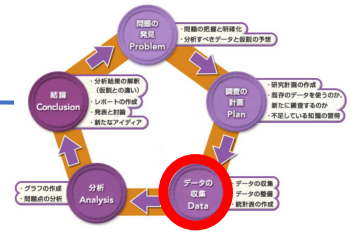
□ 仮説の検討を行う

- ・仮説なく闇雲に無計画な分析を進めると、やり直しが必要になったり、間違った分析を行ってしまう
- ・仮説を検討する事で分析の意義や目的もより明確になり、結果のレポートにも説得力が増す
- ・仮説を立証しようとするあまり、期待どおりの結果になるように調査の方法等を変更することのないように注意

—仮説が間違っていたという結果も、重要な発見である

【Data】データの収集

データを整形して分析ができるデータにする



□ 内容を検査する

- 集めてきたデータがそのまま分析に適した形であるとは限りません
- この後の集計作業を行いやすくするためデータの状態を正しく把握する

□ 加工する

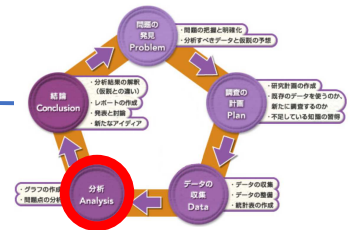
- 使用する分析手法に合わせてデータの形を整える
- 質的データの場合符号付けを行う
- 例えば性別の場合、男性は“0”、女性は“1”のように符号を付けます

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

8

【Analysis】分析

データを分析する



- データの分析を行なっていく上では、その特徴を統計量などの具体的な数値で把握することが重要
- さまざまな分析手法を使い仮説を検証する
 - 1つの手法で期待した結果が出ない場合、手法を変えてみる
 - 手法を変えても良い結果が得られないときは、データを変えてみる

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

9

データ分析をするのは誰か？

- ケース 1：全部誰かにやってもらう
 - ・ 文脈や意図を汲んで結果解釈してもらうことは難しい
- ケース 2：全部自分で行う
 - ・ 分析テクニックを全て習得するハードルは高い
- ケース 3：分担する
 - ・ 分析したい人は最低限の統計知識は持ち合わせることは必須
 - ・ 高度なテクニック部分は分析を専門とする人に頼ることができる

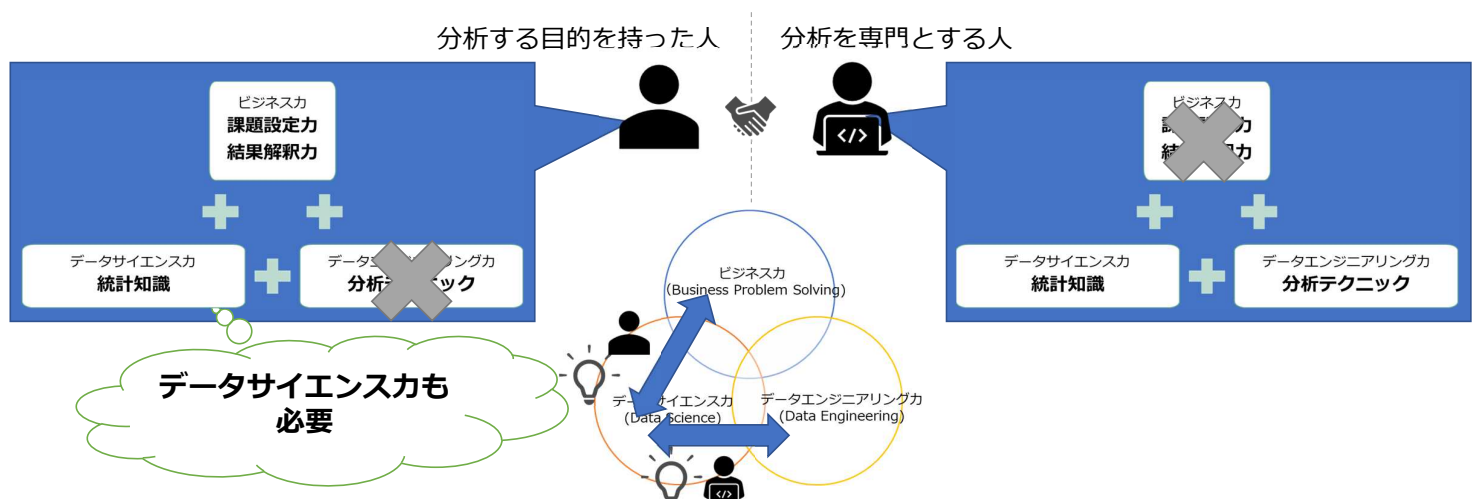


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

10

理想的な役割分担

- ・ データサイエンスの最低限の知識を持つことによって分担した場合でも共通の理解ができ正しく意思疎通が行えます

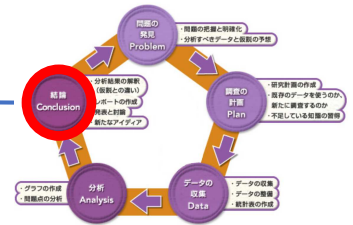


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

11

【Conclusion】結論

結果を読み解く



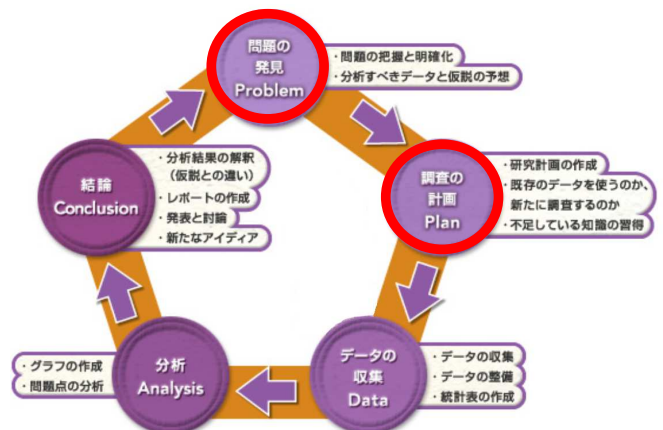
□ 結果の考察

- 分析を行った本人では思い込み等で視野が狭くなってしまいがち
— 結果は第三者にも共有し、出てきた質問等から気づけなかった観点や課題をあぶり出しましょう
- 計画通り分析ができなかった場合は、どのようにすればその問題が解決するのか考えてから次の分析へ

□ 分析を終えて次に何をするか？

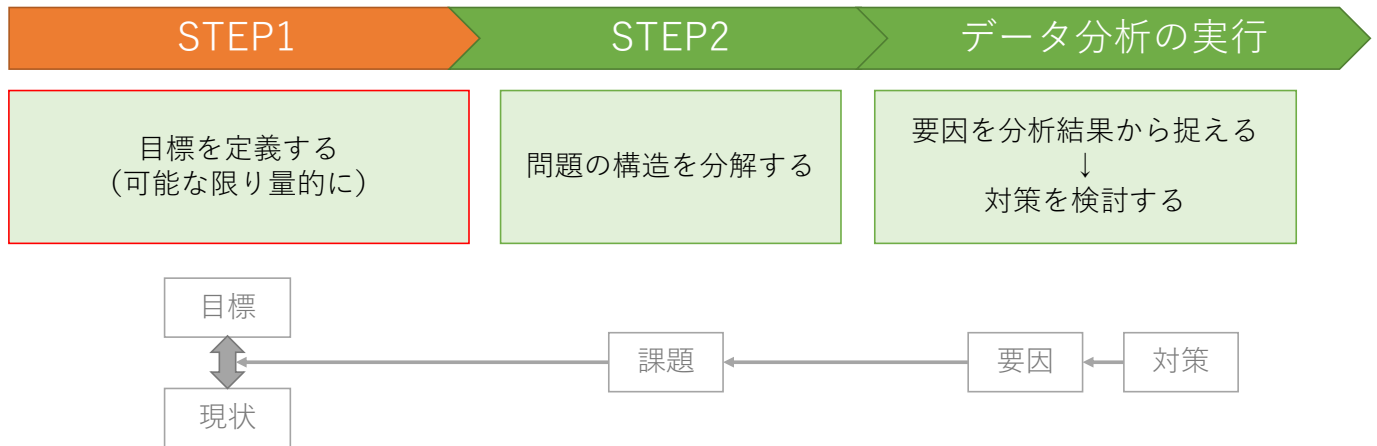
- 問題点に対して改善策が見つけられた場合は、改善に向けた目標を設定し、実践に移してみましょう

分析の設計と計画



分析設計のステップ

□ まずは目標を定めましょう



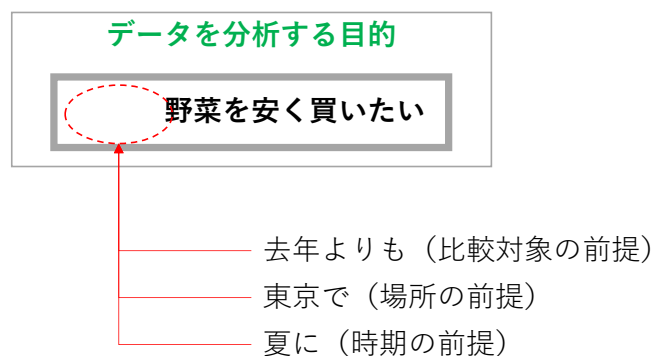
Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

14

まずは前提を確認する

□ 分析の前提として設定されているものを明らかに

- 前提のあいまいさはやり直しにつながる恐れ
- 前提を明確にしていないと、必要のない無駄なデータを収集してしまったり、必要以上に分析作業を続けてしまったりする

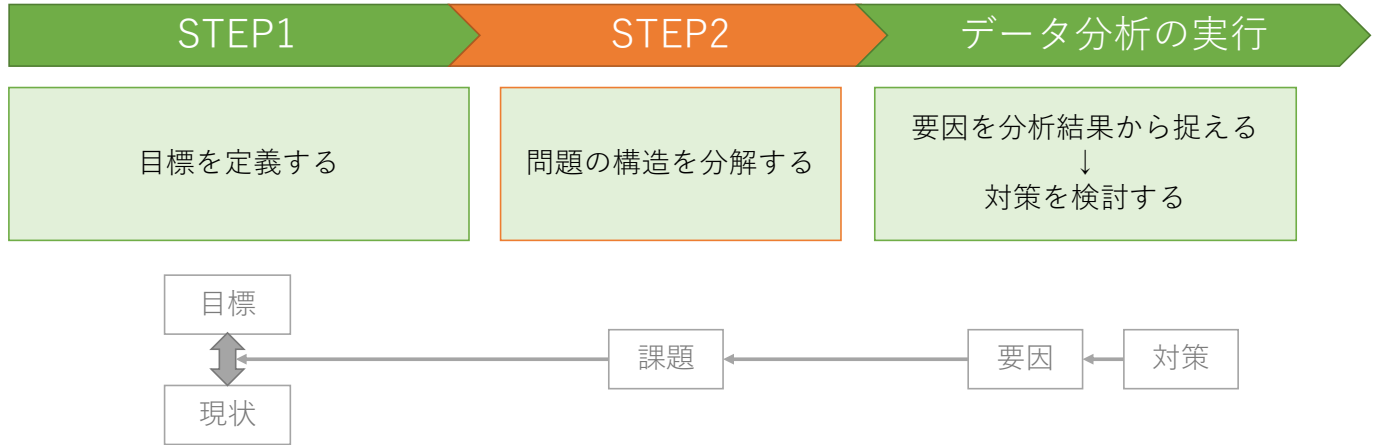


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

15

分析設計のステップ

- 指標を構造分解して課題の**仮説**を立てられるようにします

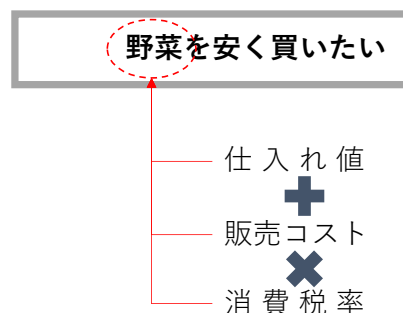


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

16

変数を構造分解して考える

- 「売上 = 客数 × 単価」のような構造分解を行います
- 変数を分解して細かく構造を捉えておくことで、分析の結果から問題発生箇所が明確になり、施策として手を加える対象がより明確になります

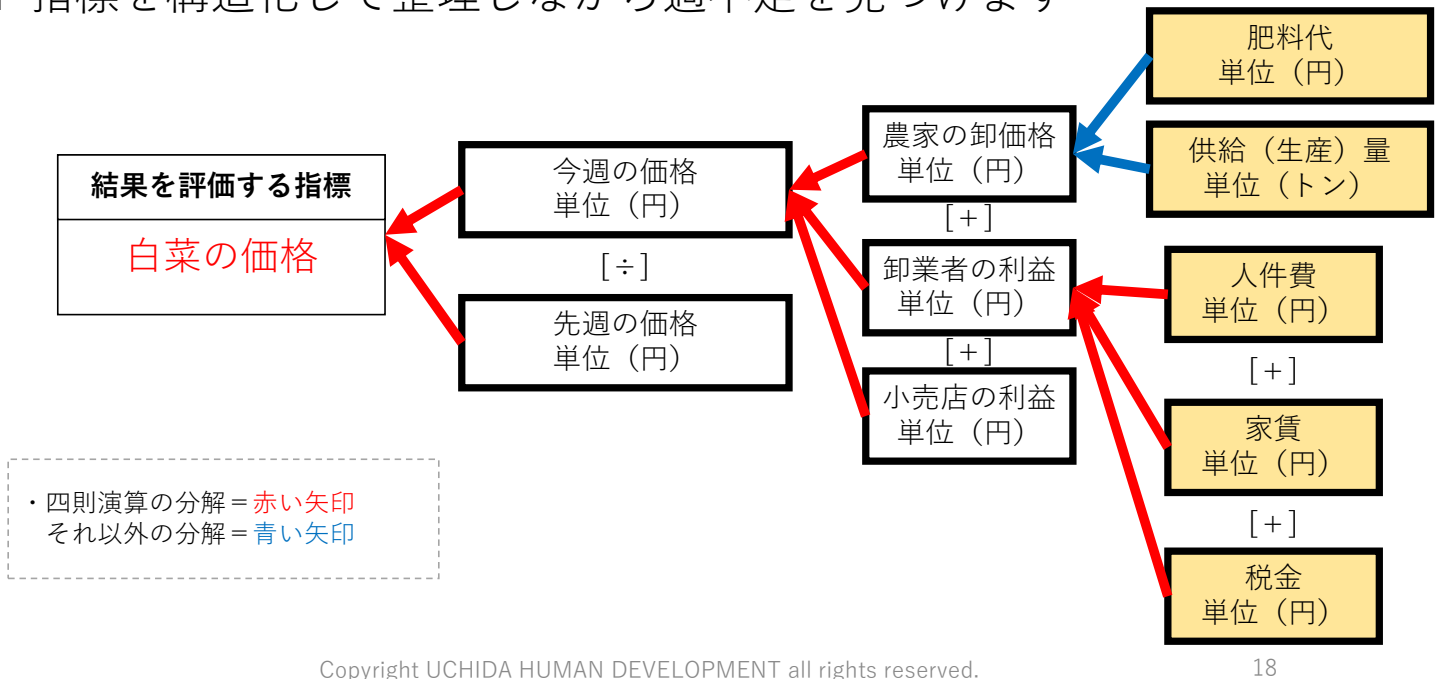


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

17

構造化の例

- 指標を構造化して整理しながら過不足を見つけます



分析手法を決める

- 関係性
 - 単回帰分析
 - 重回帰分析
 - クロス表
- 分類
 - 決定木分析
 - クラスター分析
- 上記を説明変数を変化させながら繰り返して精度を高める
- 過去の実績データで関係性のモデルを作ると未来の予測ができる

クロス表とは

- 2つのカテゴリ変数を縦横にとって件数を集計したもの

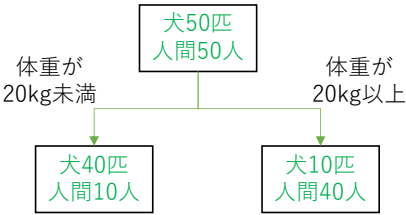
ある飲食店の15時台の年代別来店人数

	金曜	土曜	日曜	合計
50代以上	1	5	5	11
30代～40代	1	1	10	12
20代以下	3	2	2	7
合計	5	8	17	30

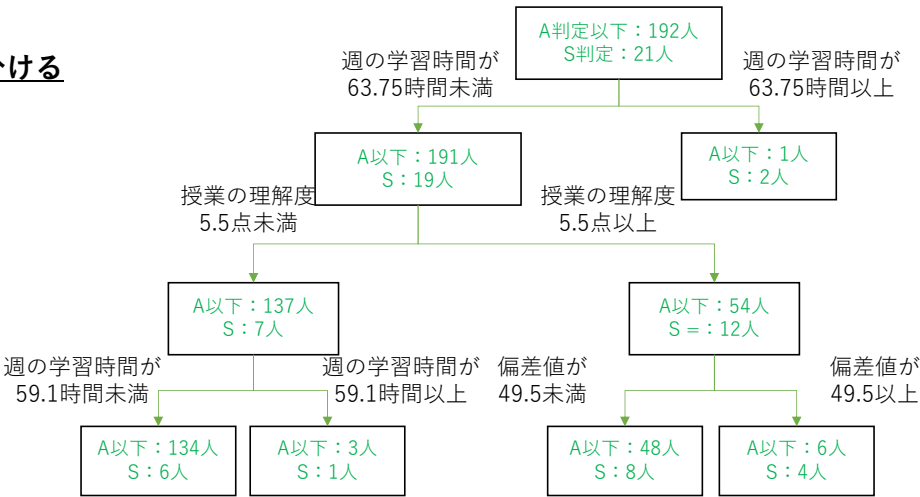
決定木分析とは

- 分けた後に最も純度が高くなる条件を見つけながら分割を繰り返す手法で、解釈が容易

体重のデータだけで犬と人間を仕分ける

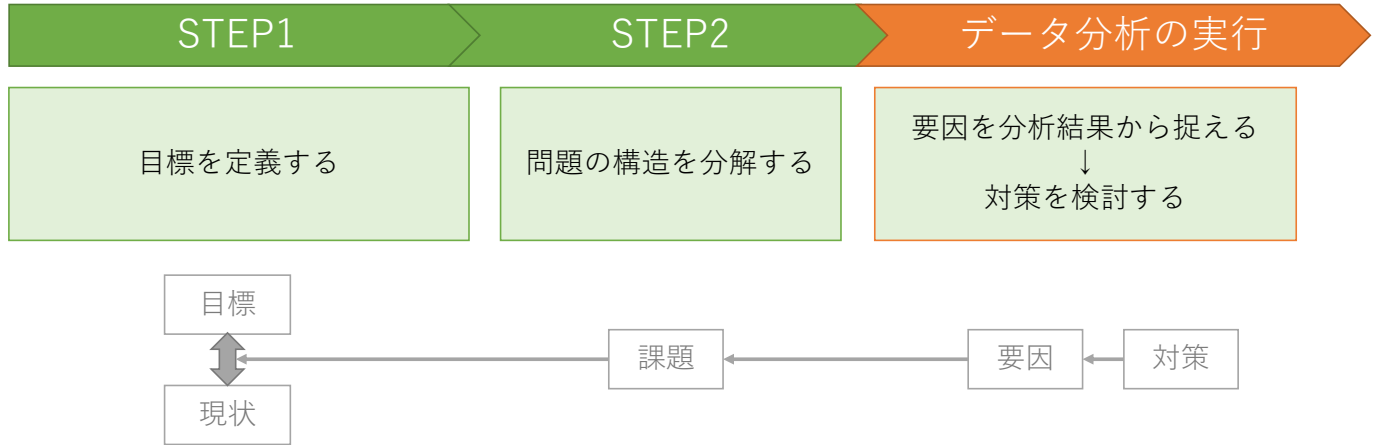


ある大学の合格判定



分析設計のステップ

- データを実際に集め、分析し、結果を解釈します

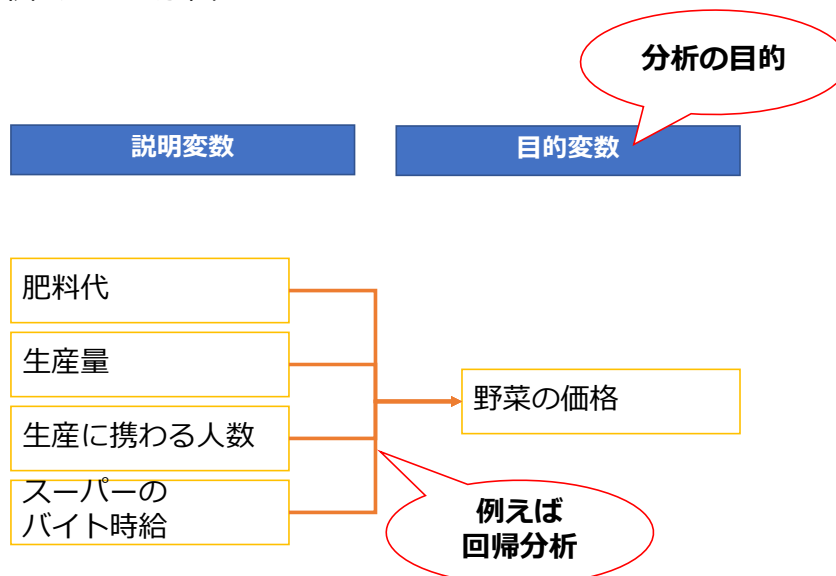


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

22

分析の実行へ

- 指標を回帰分析する場合のイメージ



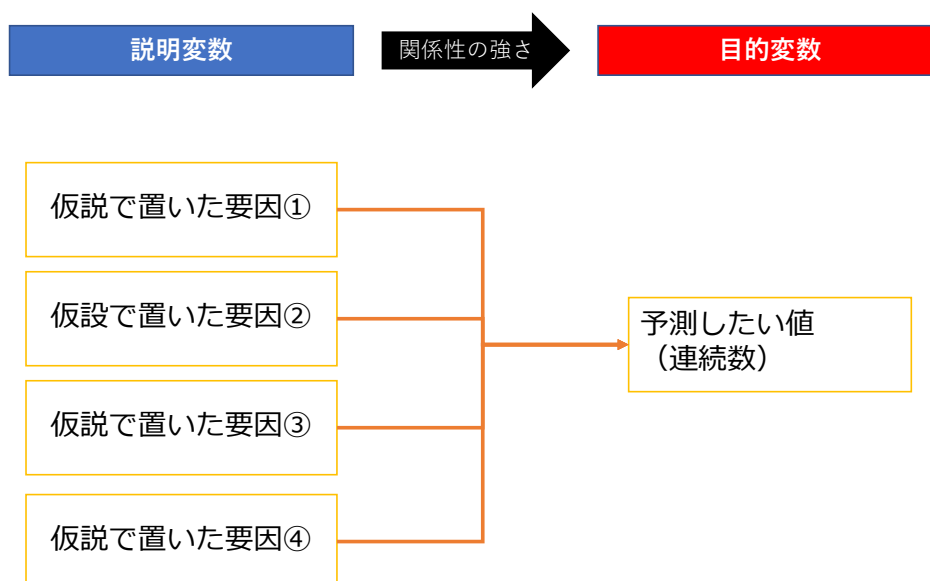
Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

23

目的変数を考える

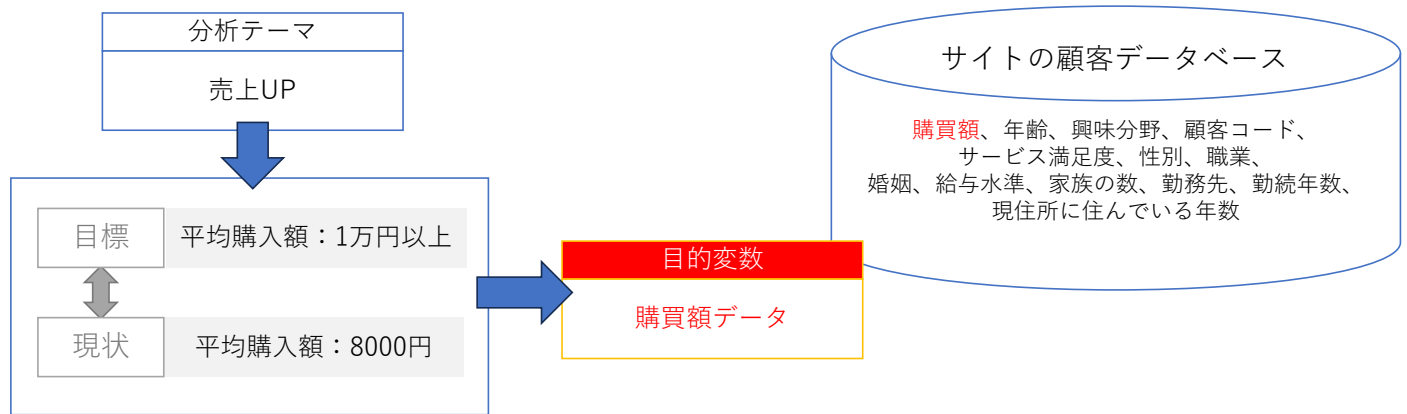
回帰分析による予測イメージ パターン1

□ 重回帰分析



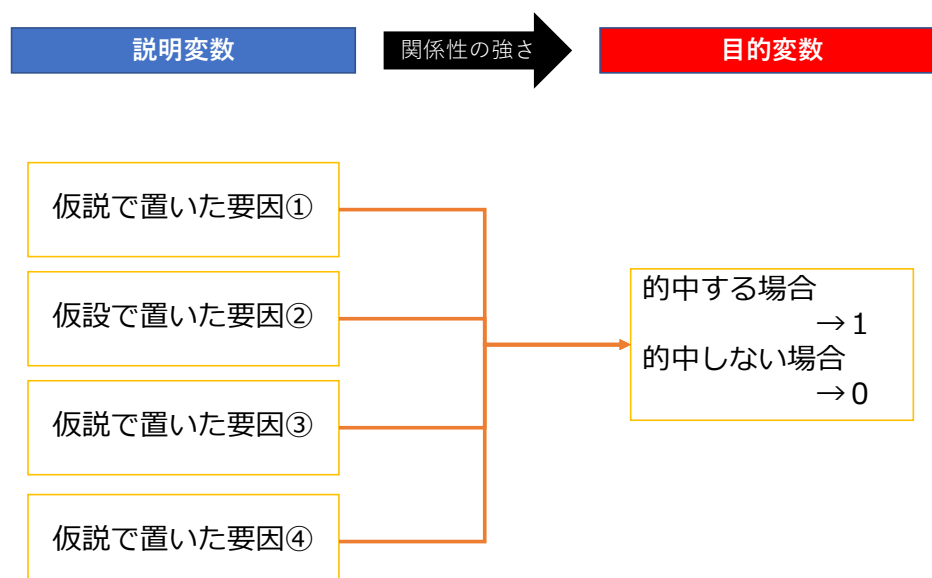
目的変数の決定 パターン1

- ECサイトでの顧客の購買額を予測する場合、サイトの顧客データベース内の購買額データがそのまま目的変数として使えます



回帰分析による予測イメージ パターン2

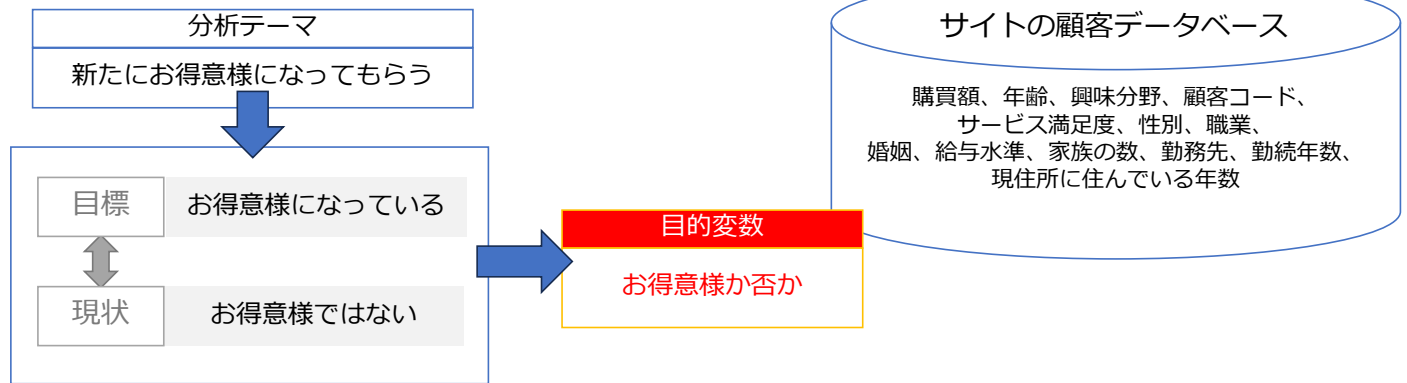
- 重回帰分析



目的変数の決定 パターン2

□ 「お得意様」を増やすための分析

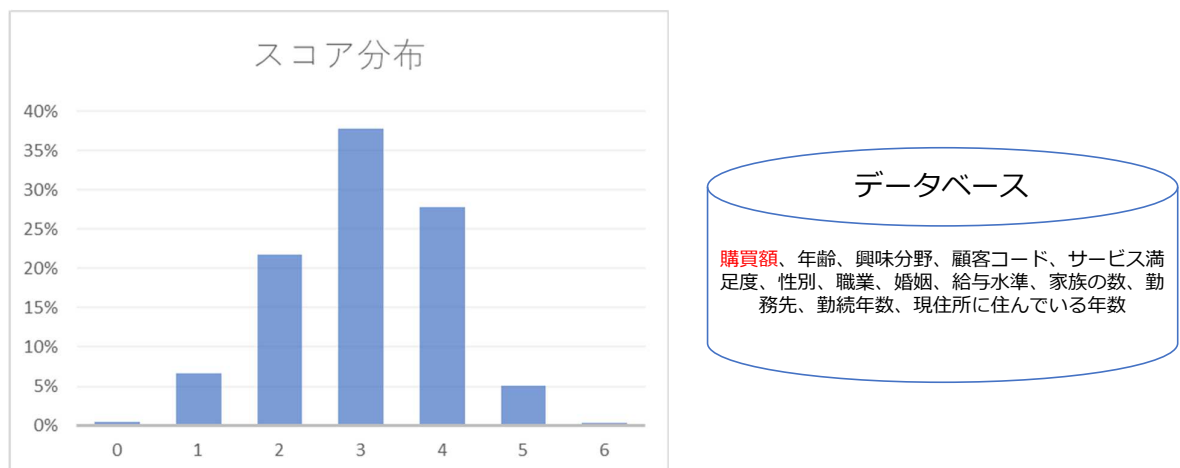
- 今お得意様かどうかを判定する必要がありますが、データにはそれを判断する情報が存在しません
- このような場合は、**お得意様とはどのような属性なのか**仮説をまず立ててから目的変数を定義することがポイントになります



目的変数を考える際のポイント

□ まずは購買額をヒストグラムを使って可視化してみます

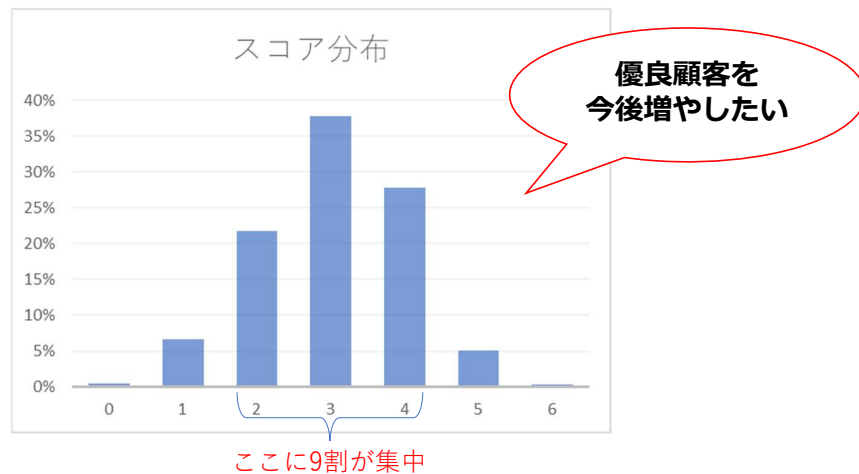
- 例えば購買額に応じて等間隔に区分を設け、上位2つの区分に含まれる場合に有料顧客と定義すると該当するのは全体の5%程度だけだった
- 区分を上位3つまでにすると30%超で3分の1が該当・・・



目的変数を考える際のポイント

□ この上位 5 %とは特殊要因ではないかどうか注意する

- ・ もしも一部の特殊な人たちがばかりなのであれば、上から 3 つ目の中から、さらに別の条件で絞って優良顧客を見つける方法を考えてみましょう



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

30

目的変数を考える際のポイント

□ 現在使えるデータの中から組合せで条件設定を作る

分析テーマ
新たにお得意様になってもらう

目的変数
お得意様 = 1
そうではない = 0

これらの項目を使って条件を設定して「優良顧客」の定義をつくる

「目的変数」という列を作ります

既婚者	職業：正社員	職業：パート	最大の連続利用月数	直近1年間の購入回数	サイトの滞在時間	サイトの利用回数	ユーザーアンケートで「不満がある」と答えた回数	年齢	性別	月間利用額「10万円」以上になった回数
0	0	0	40	0	101	3	20	2	女性	0
0	1	0	35	0	248	1	22	1	女性	0
0	0	0	53	0	248.1	1	22	1	女性	0
0	1	0	50	0	32	0	25	0	女性	0

優良顧客	既婚者	職業：正社員	職業：パート	最大の連続利用月数	直近1年間の購入回数	サイトの滞在時間	ユーザーアンケートで「不満がある」と答えた回数	年齢	性別	月間利用額「10万円」以上になった回数
1	0	0	0	40	0	101	3	20	2	女性
1	0	1	0	35	0	248	1	22	1	女性
1	0	0	0	53	0	248.1	1	22	1	女性
1	0	1	0	50	0	32	0	25	0	女性

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

31

データの準備

□ e-Stat (https://www.e-stat.go.jp/)

行財政

主な調査

地方公務員給与実態調査

すべて見る (45 調査)

政府統計一覧

政府統計コード	政府統計名	概要
00000002	一般職国家公務員在職状況統計表 (人事統計報告)	詳細
00000003	国家公務員退職手当実態調査	詳細
00020112	国家公務員死因調査	詳細
00020131	国家公務員災害補償統計	詳細
00020151	退職公務員生活状況調査	詳細
00020211	一般職の国家公務員の任用状況調査	詳細
00020312	国家公務員給与等実態調査	詳細
00100502	SDGsに関する全国アンケート調査	詳細
00160003	地方消費者行政の現況調査	詳細

データの準備

□ 令和5年度のデータを使用します

SDGsに関する全国アンケート調査

一覧形式で表示

本調査は、自治体のSDGs達成に向けた取組の実施状況を調査することを目的に、貴自治体におけるSDGsの認知度や取組度合いについてお伺いさせて頂くものです。また、令和4年12月に閣議決定された「デジタル田園都市国家構想総合戦略」に「地方公共団体によるSDGs達成に向けた取組割合の把握を行う」ことが明記されています。

SDGsに関する全国アンケート調査	データベース	件数 更新日	ファイル	件数 更新日	概要
令和2年度 SDGsに関する全国アンケート調査				3件 2021-04-01	
令和3年度SDGsに関する全国アンケート調査				3件 2021-12-22	
令和4年度SDGsに関する全国アンケート調査				3件 2022-12-06	
令和5年度SDGsに関する全国アンケート調査				3件 2024-01-19	

データの準備

SDGsに関する全国アンケート調査 [詳細](#)

本調査は、自治体のSDGs達成に向けた取組の実施状況を調査することを目的に、貴自治体におけるSDGsの認知度や取組度合いについて伺いさせて頂くものです。また、令和4年12月に閣議決定された「デジタル田園都市国家構想総合戦略」に「地方公共団体によるSDGs達成に向けた取組割合の把握を行う」ことが明記されています。

令和5年度SDGsに関する全国アンケート調査

公開（更新）日

- [3件]

2024-01-19

[一覧形式で表示](#)

データセット一覧

[戻る](#)

[URLをコピー](#)[一覧形式で表示](#)

政府統計名	SDGsに関する全国アンケート調査			詳細
提供統計名	令和5年度SDGsに関する全国アンケート調査			

表番号	統計表	調査年月	公開（更新）日	データダウンロード
	自治体別 地方創生SDGs達成に向けた取組状況	2023年度	2024-01-19	EXCEL
	都道府県別 地方創生SDGs達成の取組を推進している自治体割合	2023年度	2024-01-19	EXCEL
	地方創生SDGsを推進している自治体一覧	2023年度	2024-01-19	EXCEL

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

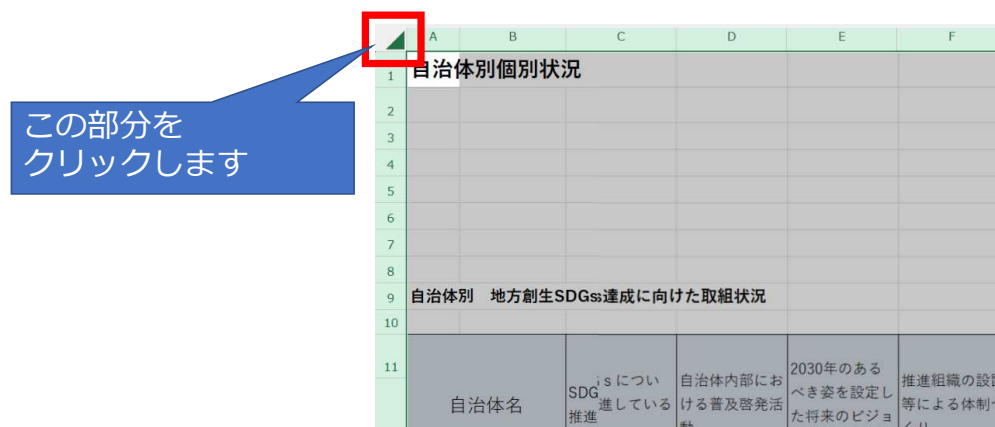
34

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

演習手順

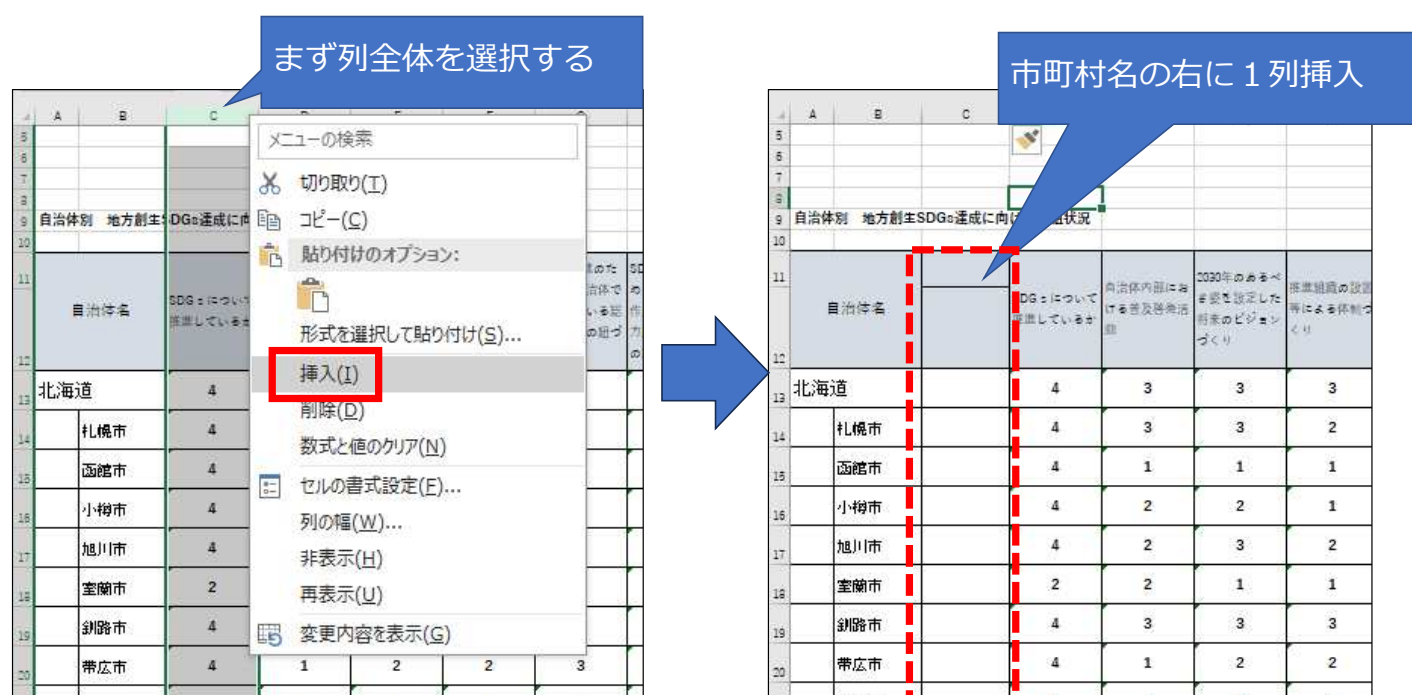
- 元のシートのデータを全コピーして新しいシートに貼り付け
 - ・ [Ctrl+a]キーを押して全部選択できます
 - ・ シートの左上をクリックしても全部選択できます



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

36

演習手順



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

37

演習手順

□ 完成イメージ

- IF関数を使いましょう
- 関数の計算結果がエラーになったものが一部あった場合、今日はそのままで構いません

	A	B	C	D
10				
11				
12	自治体名	目的家数	80G以上の	
13	北海道	1	4	3
14	札幌市	0	4	3
15	函館市	1	4	1
16	小樽市	1	4	2
17	旭川市	0	4	2
18	室蘭市	0	2	2
19	釧路市	1	4	3

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

38

Excel関数の復習

□ 「IF」

- 条件を設定して、値をその条件に合わせて変化させる関数です。

基本形

=IF(【条件】 , 【条件に合致の場合】 , 【条件に合致しない場合】)

例

=IF(A2>=A3, “A2が大きいです”, “A3が大きいです”)

応用例

=IF(A2>=A3, IF(A2=A3, “A2とA3は等しいです”, “A2が大きいです”), “A3が大きいです”)

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

39