数理・データサイエンス・AI入門

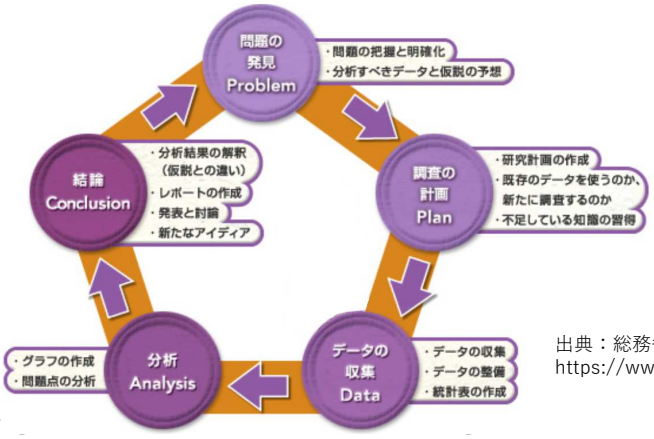
第6回
データサイエンス実践(1)

データサイエンスの基本知識

データを分析する際のステップ

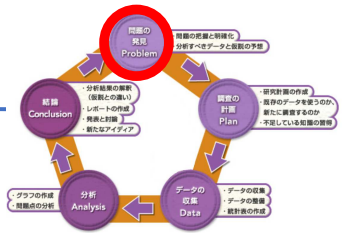
- PPDACサイクル
  - 「分析」テクニックを使いこなすためには「問題の発見（**仮説**構築）」と「計画」がしっかり出来上がっていることが条件
  - 「結果を解釈」し仮説の検証をして次の分析サイクルを回すことが不可欠

    ―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



出典：総務省統計局
https://www.stat.go.jp/naruhodo/12_ppdac/index.html

【Problem】課題の設定



分析で明らかにしたいことを定める

- **<u>問題を明示し、分析するべき課題（テーマ）を設定する</u>**
  - 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

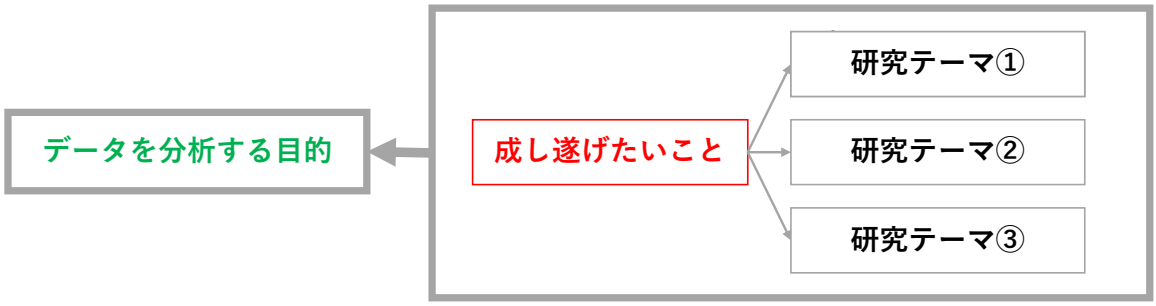
理想状態
↕
現状

工場の製造ラインの欠陥発生率が低い、ネット広告のクリック率が高い

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

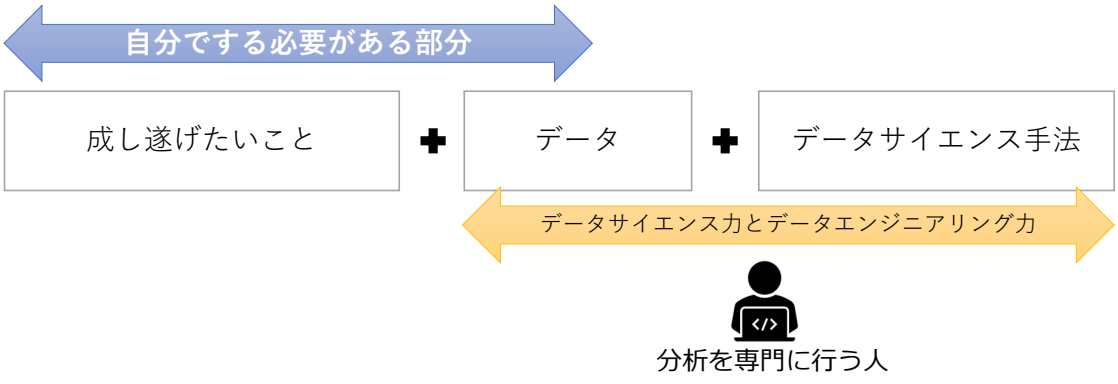
分析を行う目的とは

- 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



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

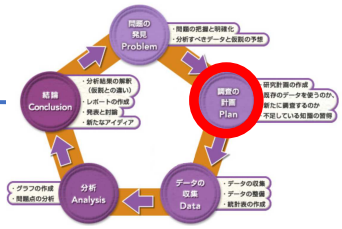


大事なポイント

- 「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



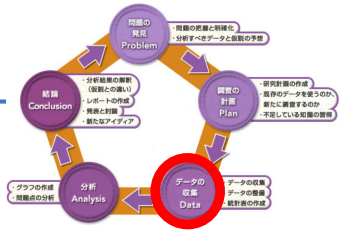
仮説を立て、分析アプローチをイメージする

- **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

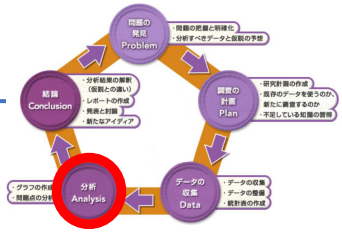


データを整形して分析ができるデータにする

- **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

【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

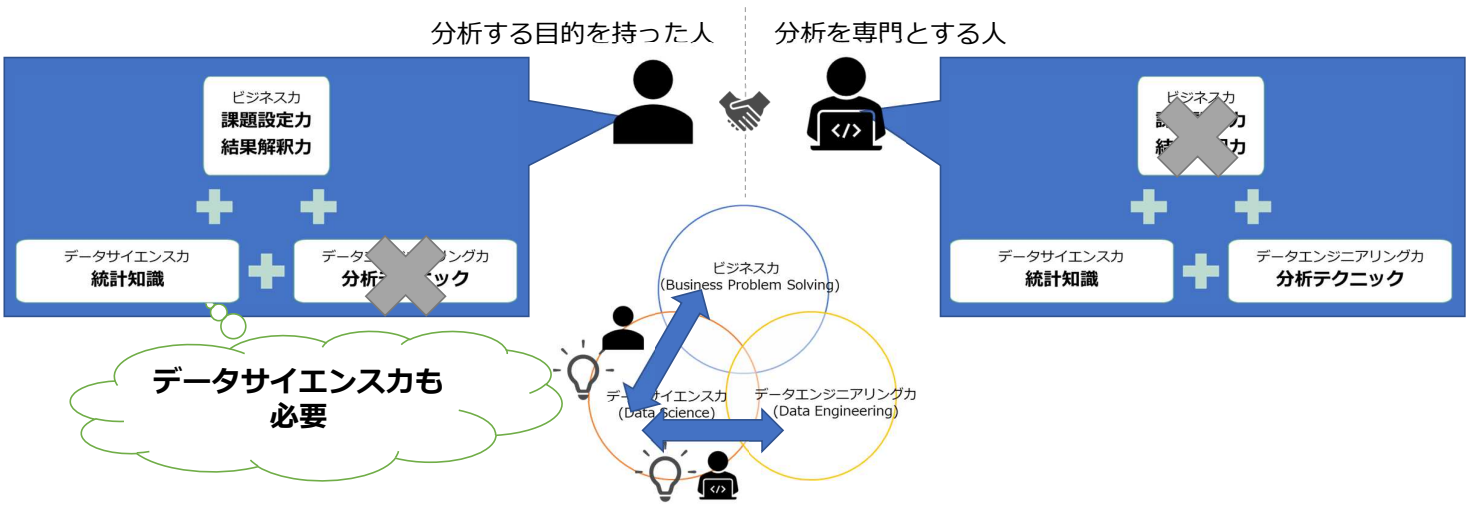
- さまざまな分析手法を使い仮説を検証する
  - １つの手法で期待した結果が出ない場合、手法を変えてみる
  - 手法を変えても良い結果が得られないときは、データを変えてみる

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

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



【Conclusion】結論

結果を読み解く

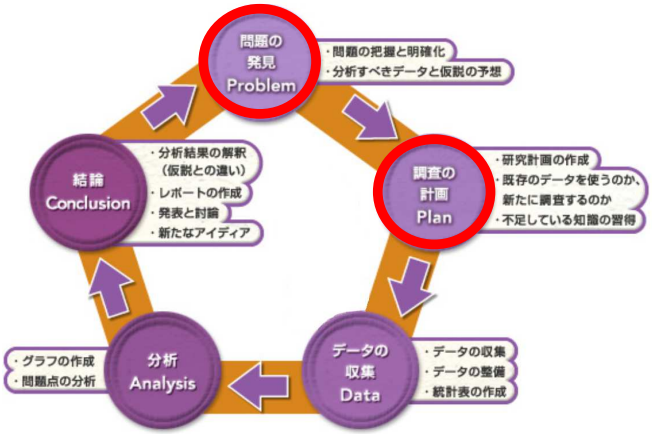
□ 結果の考察

- 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？

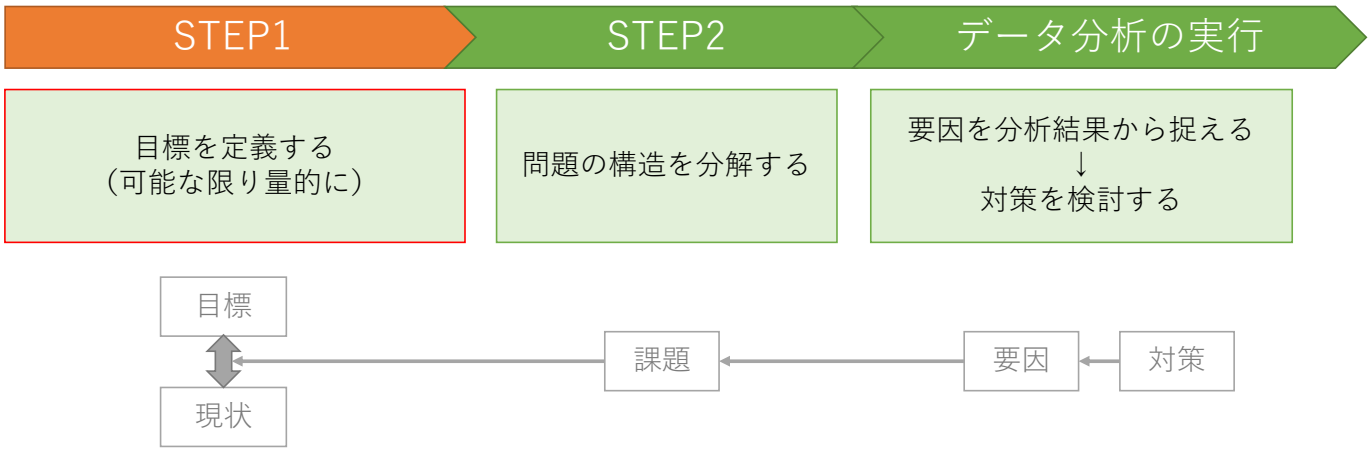
- 問題点に対して改善策が見つけられた場合は、改善に向けた目標を設定し、実践に移してみましょう

分析の設計と計画



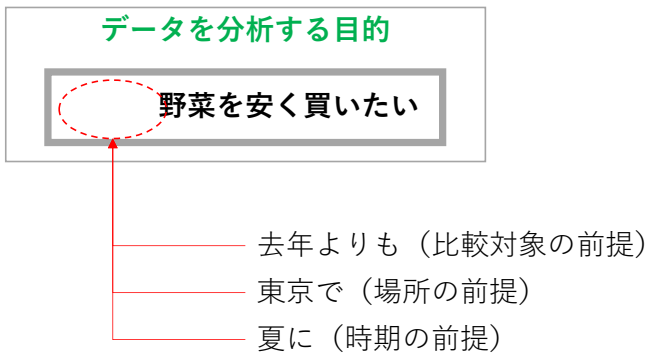
分析設計のステップ

- まずは目標を定めましょう



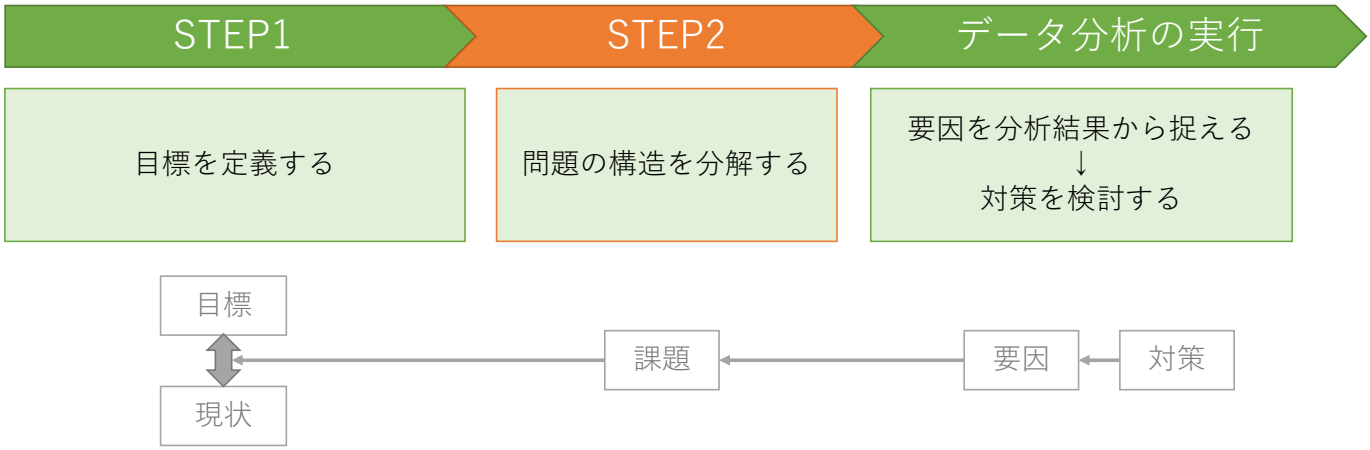
まずは前提を確認する

- 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



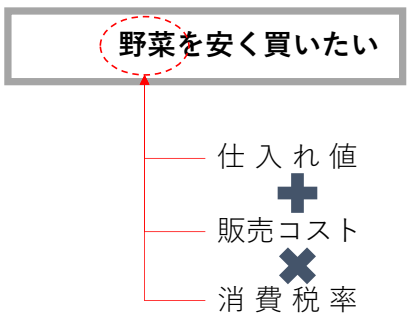
分析設計のステップ

- 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



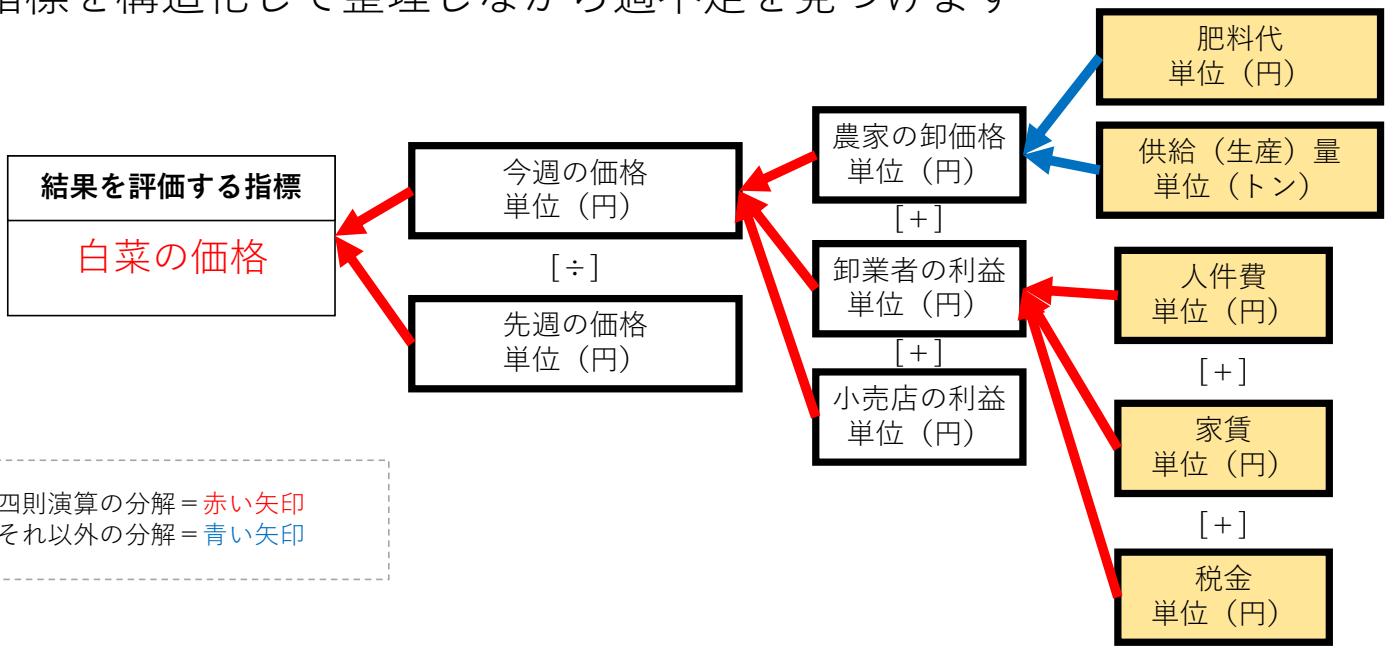
変数を構造分解して考える

- 「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



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

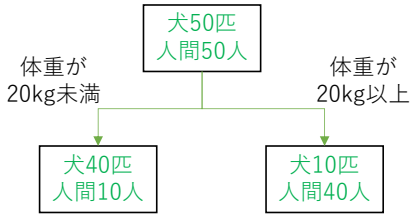
ある飲食店の15時台の年代別来店人数

	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

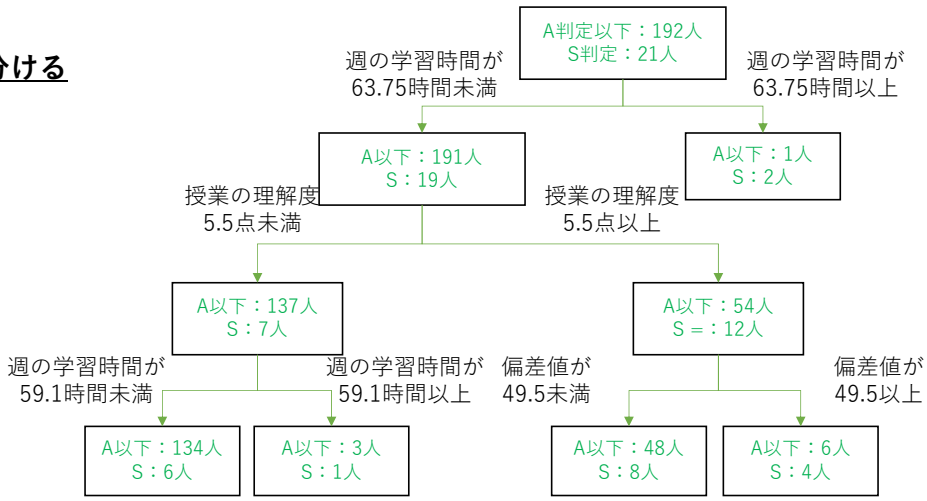
決定木分析とは

- 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

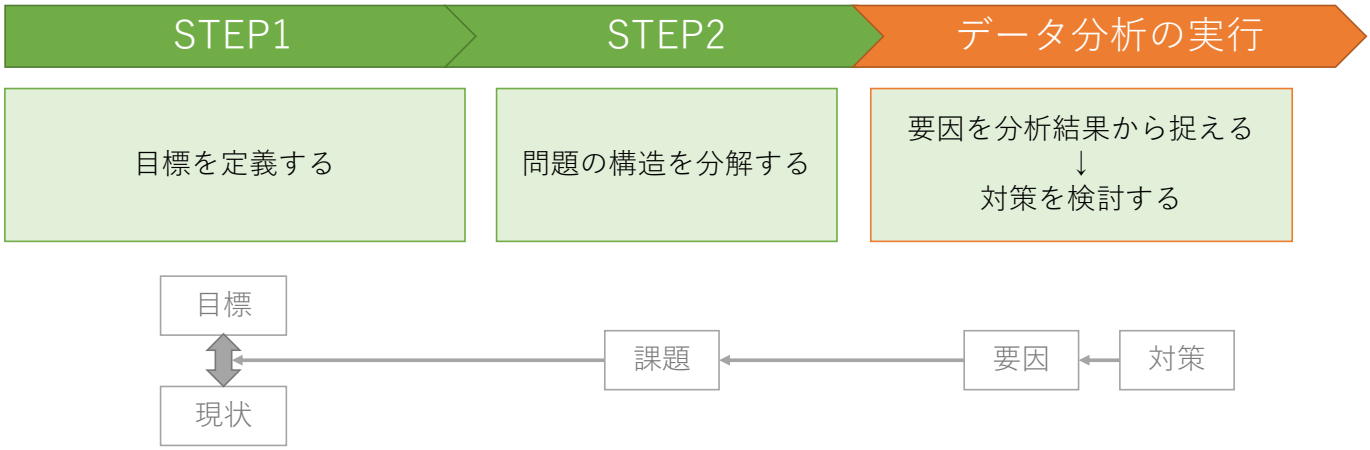


ある大学の合格判定



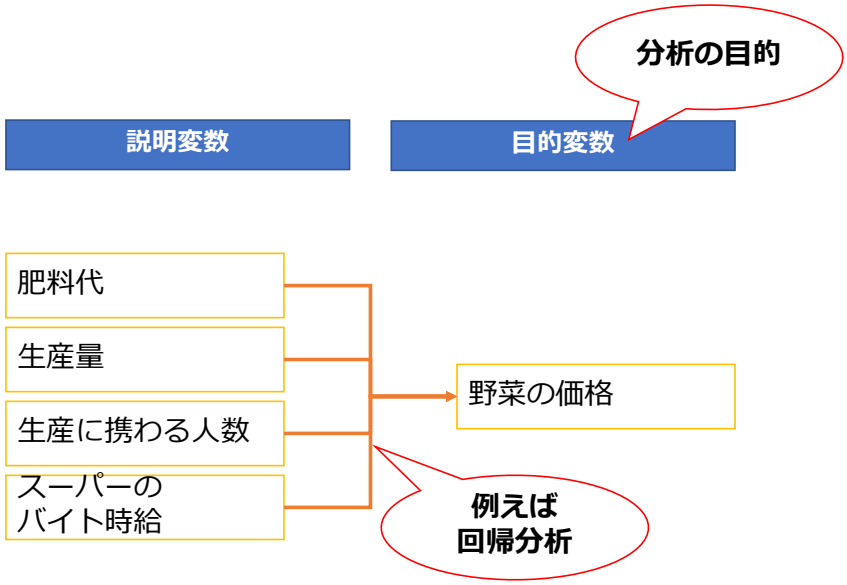
分析設計のステップ

- データを実際に集め、分析し、結果を解釈します



分析の実行へ

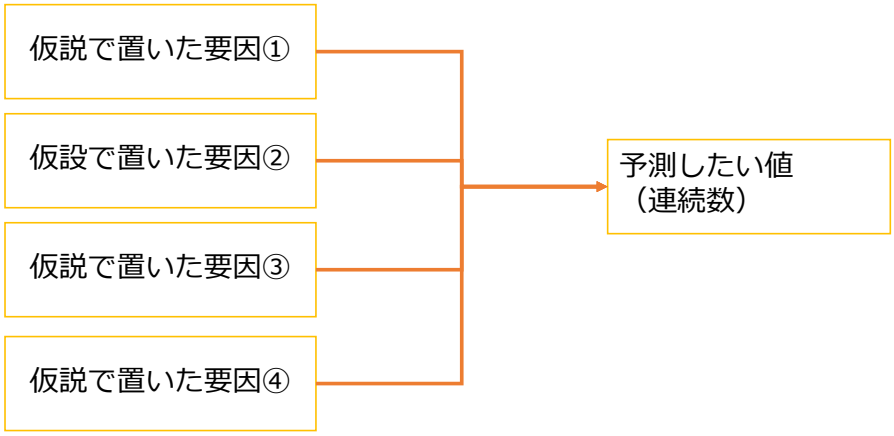
- 指標を回帰分析する場合のイメージ



目的変数を考える

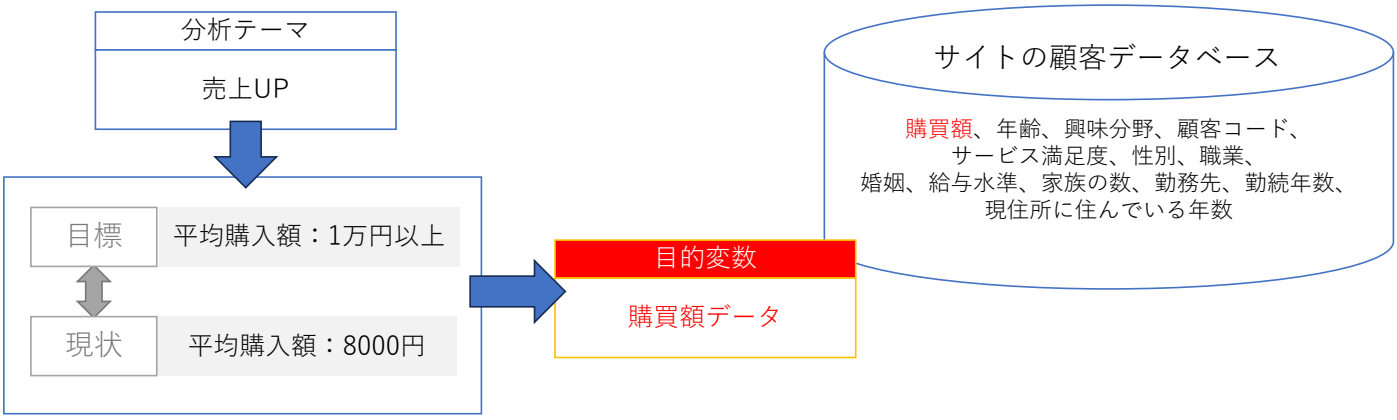
回帰分析による予測イメージ　パターン１

- 重回帰分析



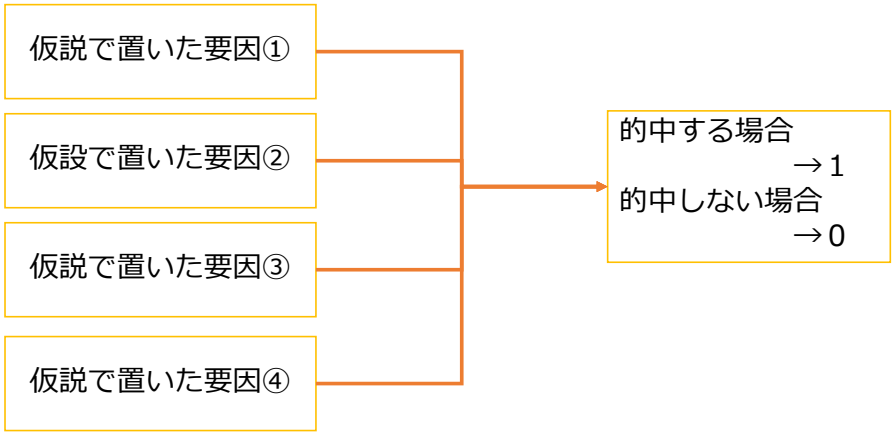
目的変数の決定　パターン１

- 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



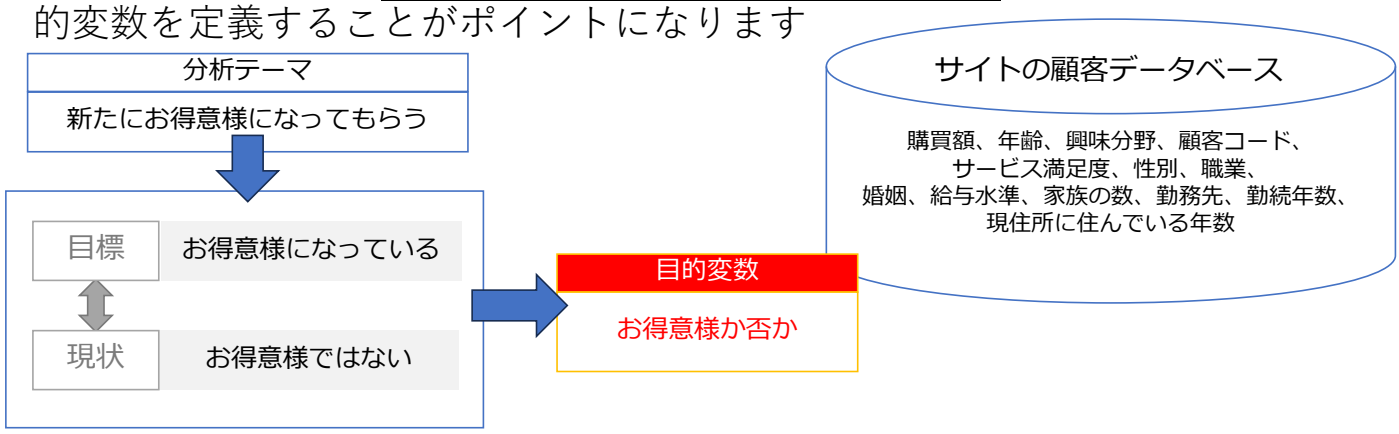
回帰分析による予測イメージ　パターン２

- 重回帰分析



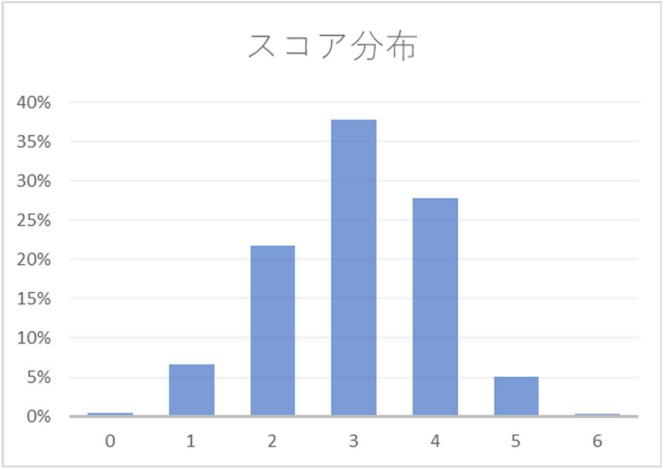
目的変数の決定　パターン２

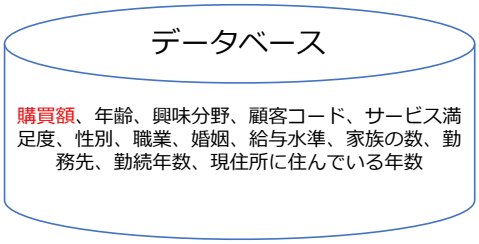
- 「お得意様」を増やすための分析
  - 今お得意様かどうかを判定する必要がありますが、データにはそれを判断する情報が存在しません
  - このような場合は、**<u>お得意様とはどのような属性なのか</u>**仮説をまず立ててから目的変数を定義することがポイントになります



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

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

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	サイト利用者アンケートで「不満がある」と答えた項目数	年齢	月間利用額「10万円」以上になった回数	性別
0	0	0	40	0	101	3	20	2	女性
0	1	0	35	0	248	1	22	1	女性
0	0	0	53	0	248.1	1	22	1	女性
0	1	0	50	0	32	0	25	0	女性

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

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

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	サイト利用者アンケートで「不満がある」と答えた項目数	年齢	性別	月間利用額「10万円」以上になった回数	
1								20	0	2	
1								22	0	1	
1	0						1	22	0	1	
0	0	1				32	0	25	0	0	

「目的変数」
という列を作ります

データの準備

- e-Stat（https://www.e-stat.go.jp/）

データの準備

- 令和5年度のデータを使用します

データの準備

演習内容

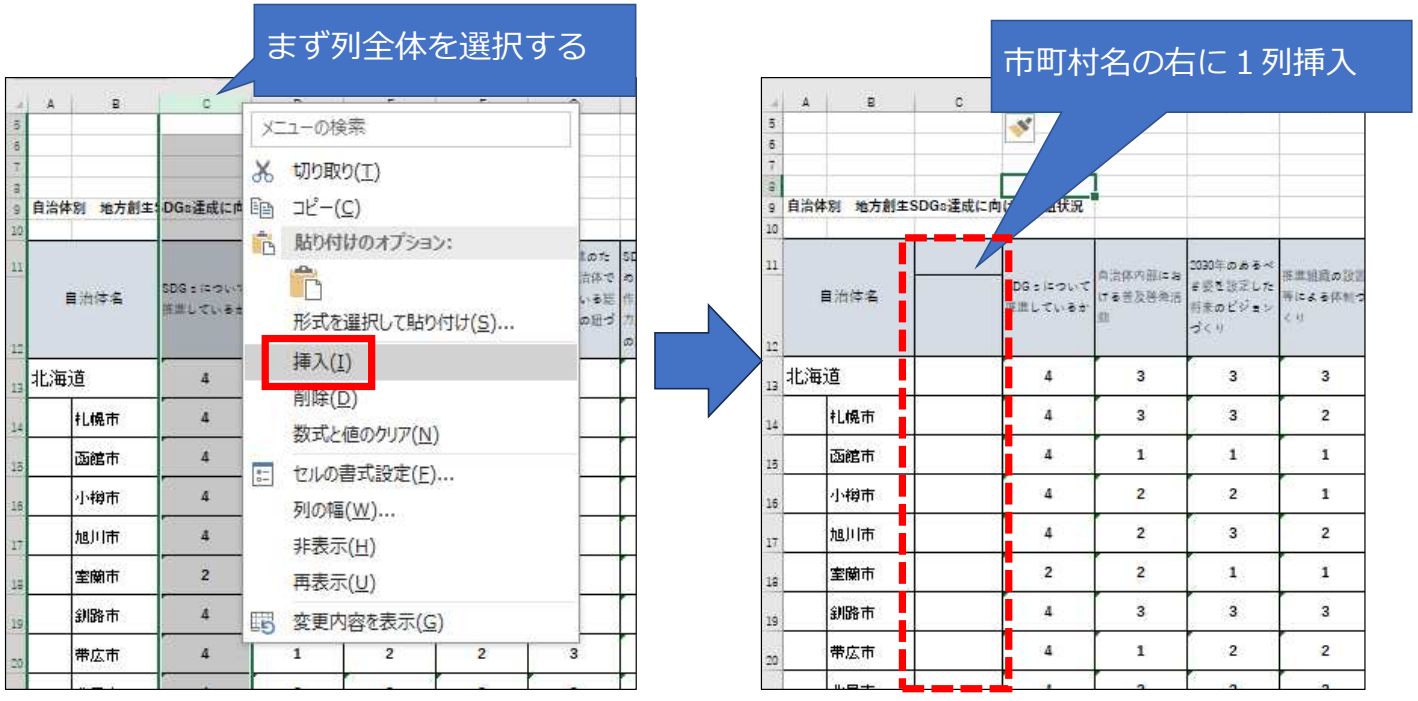
- SDGsの取り組みに課題がある自治体が、なぜ推進されないのかデータ分析を通じて原因を調査したいと思います

- 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

演習手順



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	E
10					
11	自治体名		目的変数	SDGsに…	
12					
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

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”)