

データサイエンスのサイクルとは

- ・ データサイエンスを成功させるためのテンプレートを利用しよう
- ・ PPDACやCRISP-DMなど様々なフレームワークが存在

PPDAC

データを利用した問題解決の手法として知られるフレームワークであり、海外や統計教育でもよく使われるプロセス

PDCAサイクル

Plan(計画)

Do(実行)

Check(測定・評価)

Action(対策・改善)

P : Problem 問題・課題の発見

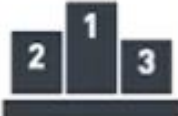
P : Plan 問題解決の計画を立てる

D : Data データを収集する

A : Analysis データを分析する

C : Conclusion 考察し活用をする



分類（水準）	解説と例	データの種類
名義尺度 同一性 	データは、単に記号としての意味だけを表す 大小関係や順序の概念はない 例：職業・性別・背番号・診断名	質的データ カテゴリカル データ
順序尺度 同一性 + 順序性 	データは、順番の概念があり、大小関係を表す 例：服サイズ（L・M・S）、薬効（悪化・無効・有効・著効）	
間隔尺度 同一性 + 順序性 + 加法性 	ある一定の単位を基にして測定された尺度 例：摂氏～華氏の温度・暦年	ここでの違いは、 データが等間隔で あるかどうか 量的データから質的 データへは変換可能○ 一方、逆はできない
比尺度（比例尺度） 同一性 + 順序性 + 加法性 + 等比性 	ある一定の単位を基にして測定された尺度に 加え原点が存在 原点=なし、0=なしとなる尺度です 例：長さ・重さ・濃度・人数・絶対温度	

小

情報量

大



データの型で適用できる分析手法がきまる

データサイエンス 力

外的基準	目的変数 y	説明変数 x	目的	手法
あり	量的	量的	数量予測と要因分析 目的変数に影響のある変数を測定したり、 モデル式を使って未知の目的変数を予測	重回帰分析 正準相関分析
		質的		数量化一類
	質的	量的	判別分析と要因分析 複数の説明変数から主に2値に分類される 目的変数を分類予測	判別分析
		質的		ロジスティック回帰分析 数量化二類
なし	—	量的	次元の縮約 複数の変数から少ない変数で情報を要約	主成分分析
		質的	パターン分類	数量化三類
		質的・量的	グループ化 関係が強いものをより近くに布置（対応 分析）、関係の強さを距離で表現するク ラスタ分析	対応分析 （双対尺度法）
		量的		クラスタ分析

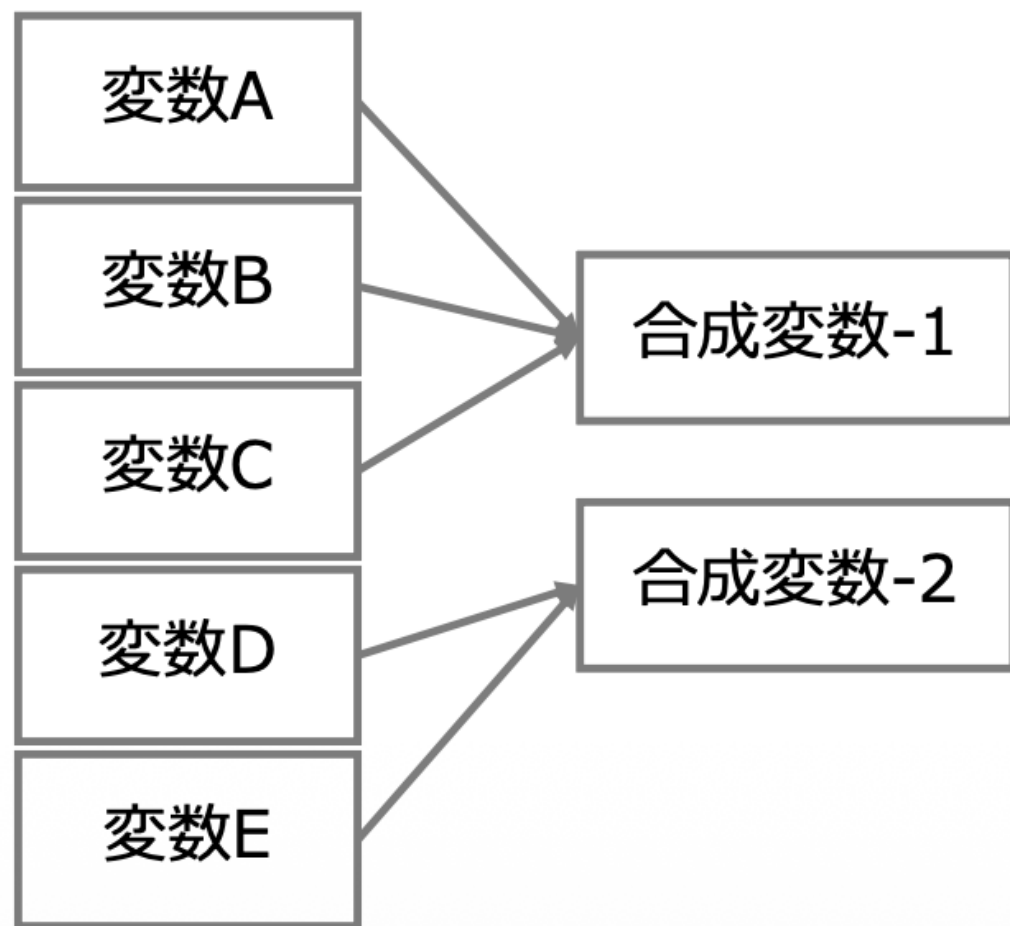
2変数の分析も変数の組み合わせによって手法とグラフが決定

パターン	主な分析手法	用いるグラフ
① 質的変数と質的変数	クロス集計表 カイ二乗検定	積み上げ棒グラフ
② 質的変数と量的変数	t検定や 一元配置分散分析など	エラーバー付の棒グラフ
③ 量的変数と量的変数	相関分析	散布図

データ間 対応の有無	グループの数	正規性	代表的な手法
独立(対応無) いずれか1つの グループに所属	2	正規分布	独立したサンプルの t 検定
		非正規分布	Mann-WhitneyのU検定
	3以上	正規分布	一元配置の分散分析
		非正規分布	Kruskal-Wallisの検定
対応あり	2	正規分布	対応のあるサンプルの t 検定
		非正規分布	Wilcoxonの符号付順位検定
	3以上	正規分布	反復測定
		非正規分布	Friedmanの検定

主成分分析

変数をまとめる：合成変数を作成



因子分析

変数を共通項目（因子）でわけける

