初心者セッション：データ分析の進め方

BeginneR Session: Data Analytics 101

26th April 2025, Tokyo.R #117

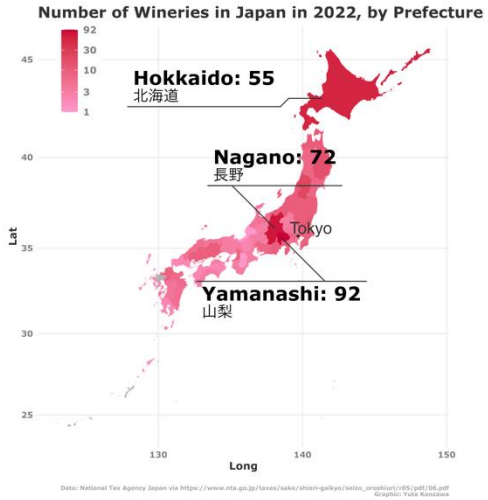
Yuta Kanzawa @yutakanzawa

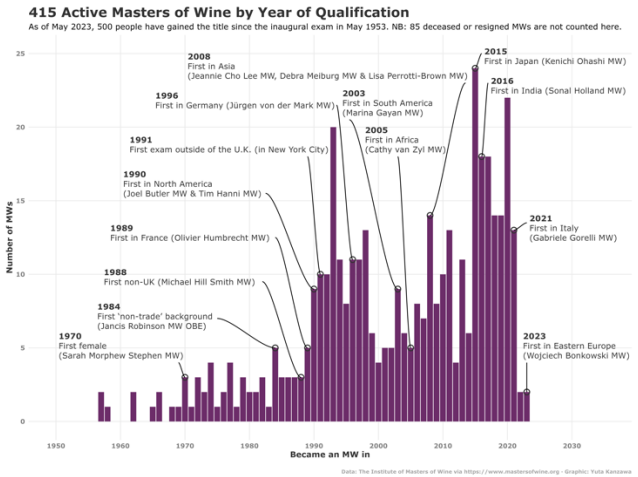
Senior Data Scientist at Zurich Insurance Company Limited, Japan Branch

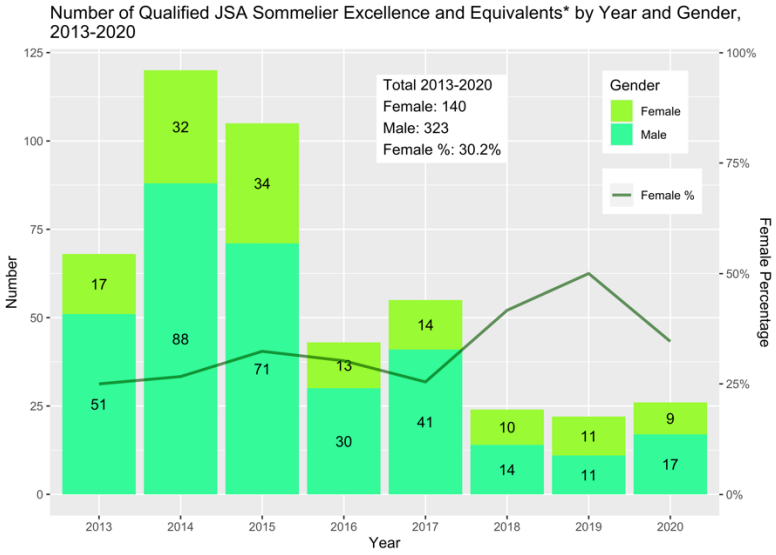
神沢雄大 Yuta Kanzawa

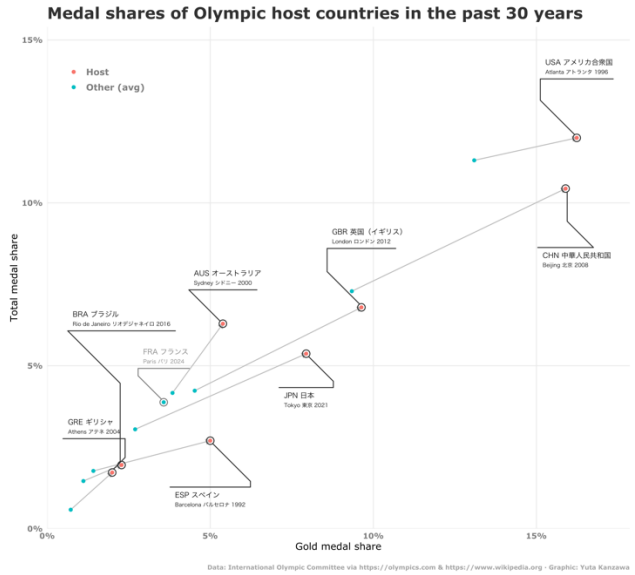
- データサイエンティスト@チューリッヒ保険会社
  - 日本支店
- Twitter: @yutakanzawa
- 好きなもの：オペラとワイン
  - ワーグナー
  - ブルゴーニュ (WSET Lv 3→?)
- 使用可能言語：7
  - 人間：日本語、英語、ドイツ語
  - コンピューター：R, Python, SAS, SQL

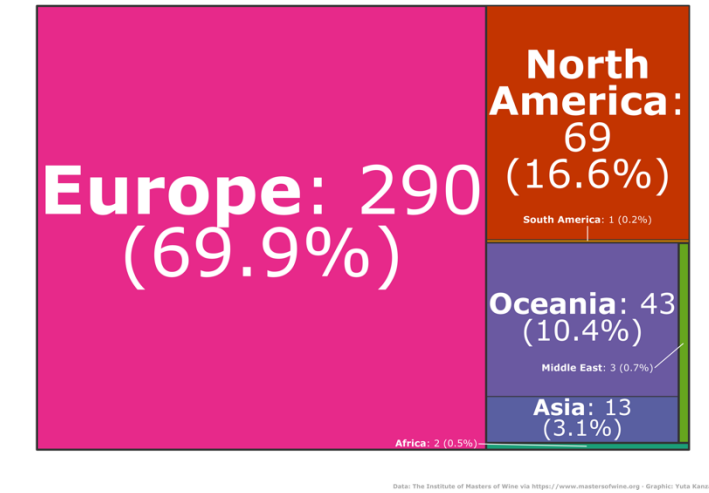
ポートフォリオ

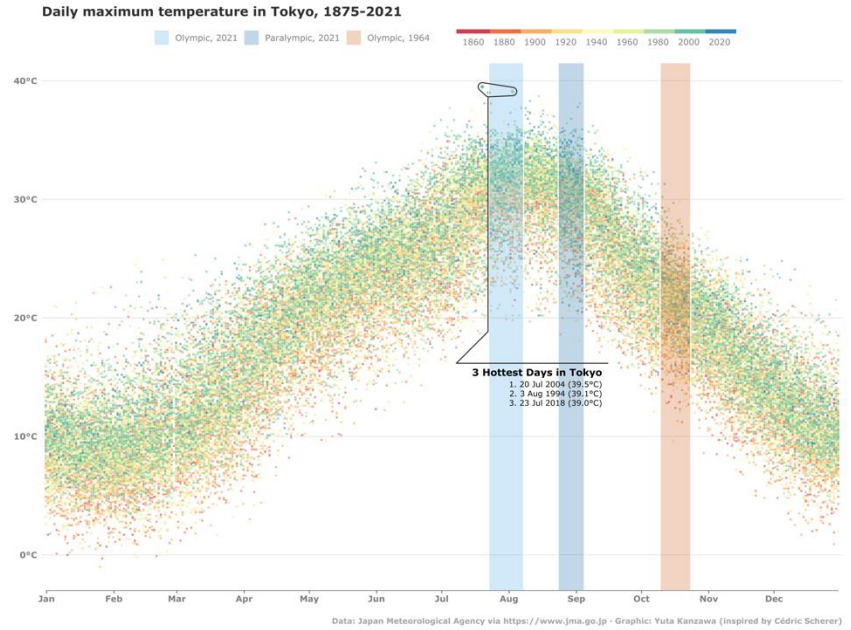




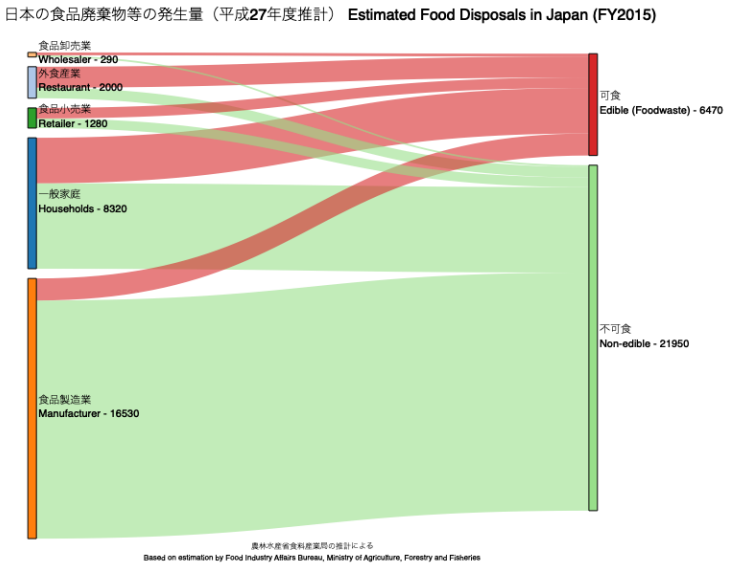


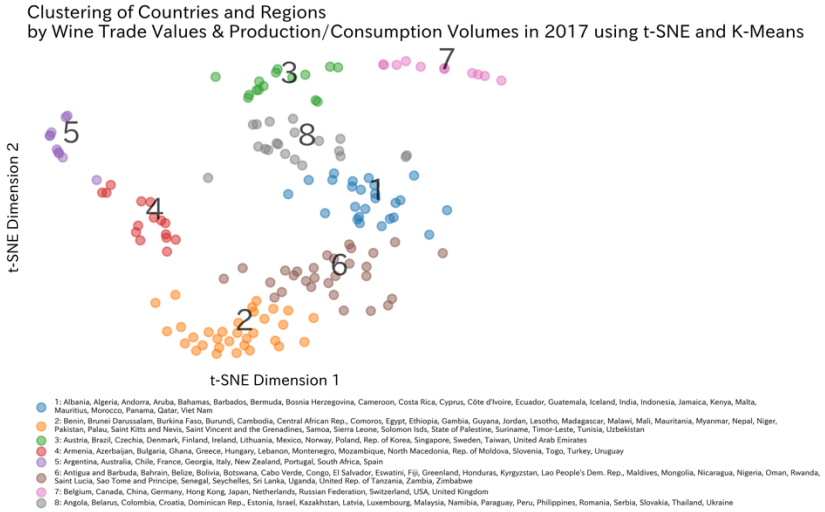


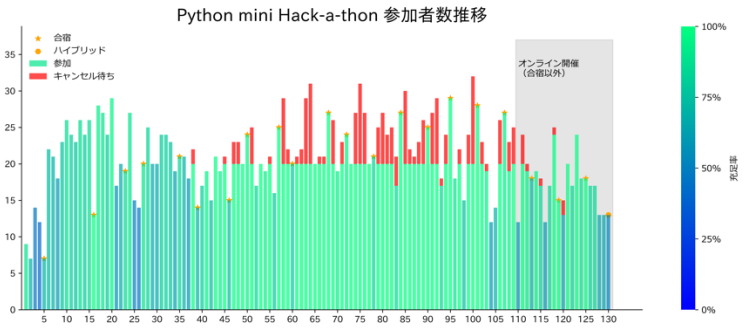


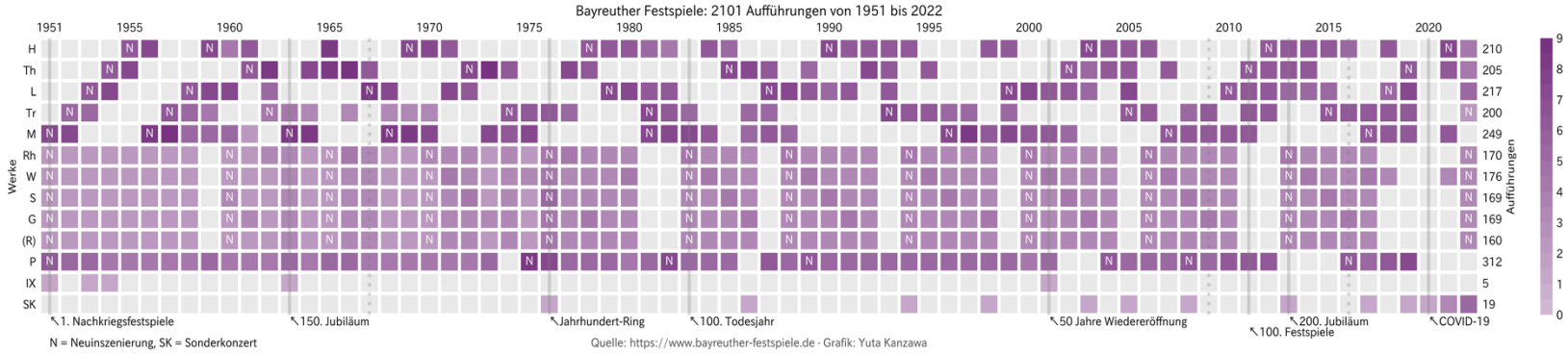


ポートフォリオ（参考までにR以外も）









アジェンダ

- 今日話すこと
  - ビジネスにおけるデータ分析フロー
  - 各ステップの詳細
- 今日話さないこと
  - R

- 対象（以下のいずれか）
  - データ分析を初めてする人
  - なんとなくデータ分析をしている人

TL;DR

- データ分析はフローであり、ループであり、スパイラル（螺旋）
  - 問題設定 → 結果の解釈 を繰り返す。
    - 期待値調整、ドメイン知識
  - 他部署との協力
- ビジネスを意識する。
  - 技術的な映え要素は不要。

データ分析フロー

Data analytics flow

一般的な流れ

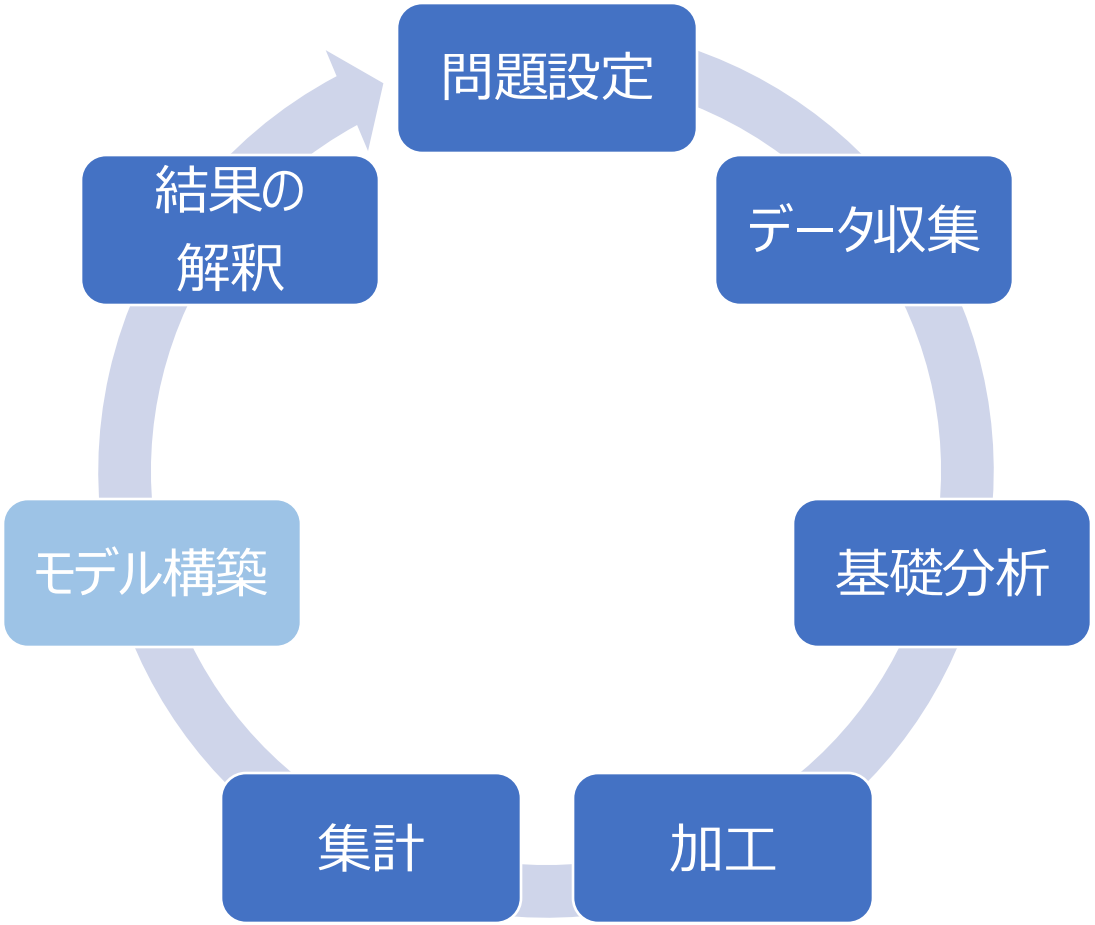
問題設定	データ収集	基礎分析	加工	集計	モデル構築	結果の解釈

• 分析の大枠を決める
  • 目的
  • 期間
• 偉い人と一緒の会議
	• ITや関連部署を拝み倒す
	• Excelがいっぱい
	  • xls👎
		• 読込→バイナリー化
		• クレンジング
		  • GIGO
		• 1マス戻る
			• 結合
			• 集約
				• いわゆる「切り口」
				• クロス集計
					• 線形回帰？機械学習？AI？
					• 動いてるからヨシ！
						• 悪魔と取引
						• 振り出しに戻る
						• 偉い人と一緒の会議

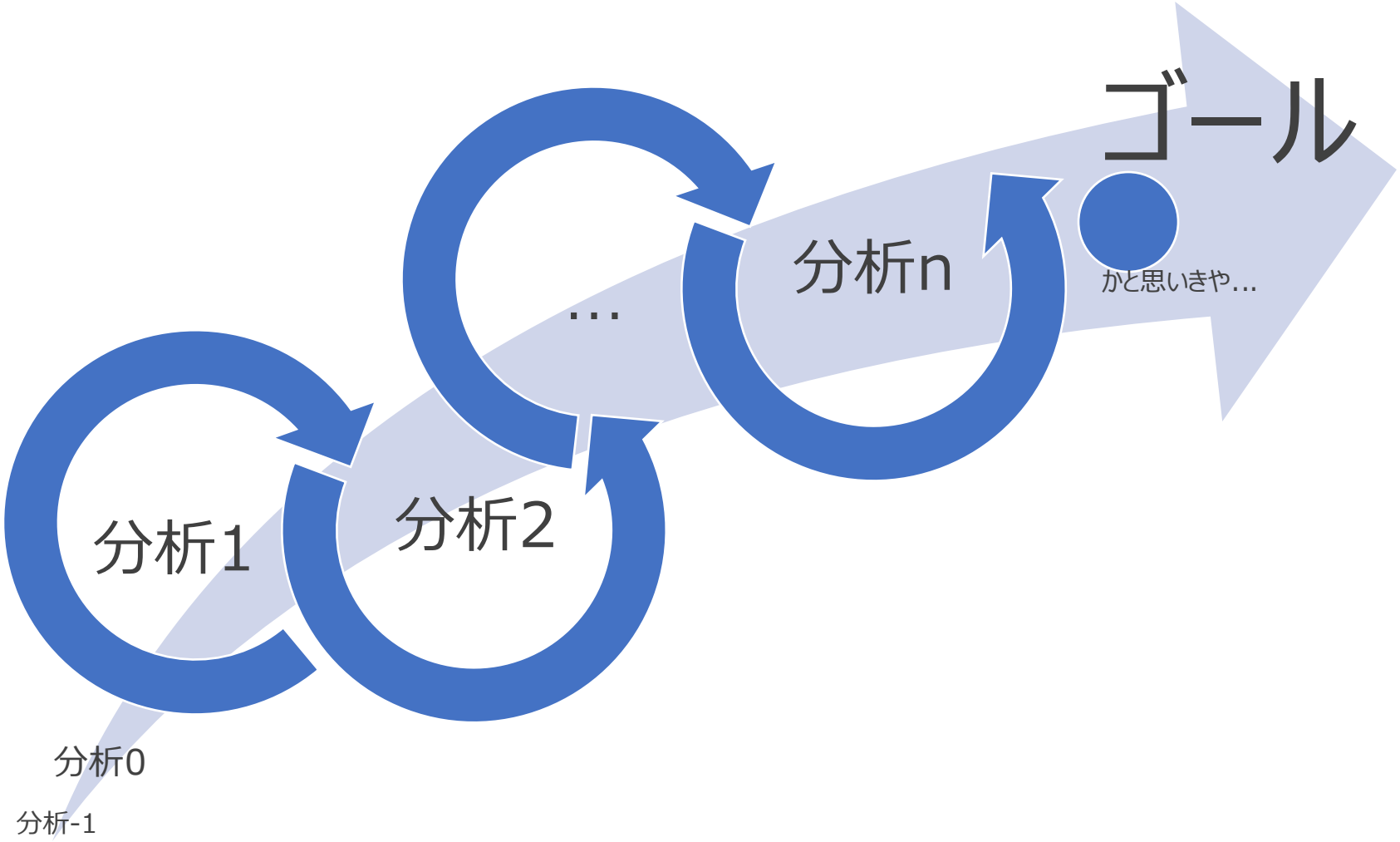
■会議
■作業

* https://twitter.com/karaage_rutsubo/status/1853190311645974904
* https://twitter.com/genbaneko_bot/status/1265783114959212544

分析ループ



分析スパイラル（螺旋）



各ステップの詳細

Details of each step

問題設定

- 分析の大枠を決める（要件定義）。
  - 目的
  - 期間
  - 工数
- **期待値調整！**
  - 失敗すると後で苦労することに…
- **自社や顧客のビジネスのことを分かって**いないと、作業するだけになってしまう。（ドメイン知識）

データ収集

- ITや関連部署を拝み倒す。
  - **データベース**、テーブル、定義書
  - データソースとなるソフトウェア
  - **コンプライアンス**
- Excelがいっぱい。
  - 打倒xls形式！

基礎分析とデータクレンジング（1/2）

- データ読込→バイナリー化
  - **データ型を保存**する。
    → 他の言語でも読み書きできる**parquet形式**がおすすめ（**arrow パッケージ**）。
- 定量的評価
  - **要約統計量**、相関係数
- 定性的評価
  - **分布の確認**
    - 統計量、ヒストグラム、散布図

基礎分析とデータクレンジング（2/2）

- 外れ値、異常値
  - 箱ひげ図、分位点、正規性
  - 対処法：削除、上下限値置換
- 欠測値（欠損値）
  - 原因に応じた対処法：削除、**層化**平均値代入、多変量補完、多重代入法
    - 時系列データ
- 値の重複
  - 対処法：削除、集約

加工

- データエンジニアが担当することも。
- 結合
  - **結合キー**を基礎分析で確認しておく。
  - サイズが大きい場合は、**データベース上で実行**した方がよい。
- 集約
  - 日次 → 週次
  - 明細 → 属性単位

集計

- いわゆる「切り口」
  - 属性（年代、居住地、職業、など）
  - 時系列（年、月、週、曜日）
- 組み合わせ → 「クロス集計」
  - Excelの**ピボットテーブル**

モデル構築

- 目的
  - 説明？予測？
- 多重共線性？
  - 予測が目的であればそのままでもよいかも？
    - ただし、**モデルの解釈性**は低下する。
- 線形回帰？機械学習？AI？
  - 映え要素は不要。→ 目的に実直に。

* https://twitter.com/genbaneko_bot/status/1265783114959212544

結果の解釈

- 「悪魔」と取引
  - **科学的な解が正解とは限らない。**
  - ただし、良心を失ってはいけない。
- 振り出しに戻る
- モデルやインサイトが**社会的バイアスを助長**していないか？
- 偉い人との会議
  - 技術的な話は不要。→ **ビジネスの話**をする。
  - 'So, what?'

まとめ

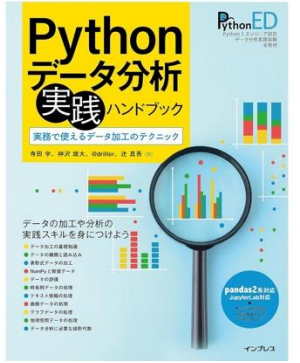
Long story short

Long story short

- データ分析はフローであり、ループであり、**スパイラル**（螺旋）
  - 問題設定 → 結果の解釈 を繰り返す。
    - 期待値調整、ドメイン知識
  - 他部署との協力
- **ビジネスを意識**する。
  - 技術的な**映え要素は不要**。

参考書

- 『イシューからはじめよ』改訂版（安宅、2024年）
- 『ビジネスデータサイエンスの教科書』（Taddy、2020年）
- 『Pythonデータ分析 実践ハンドブック』（寺田、神沢、@driller、辻、2023年）



Enjoy!