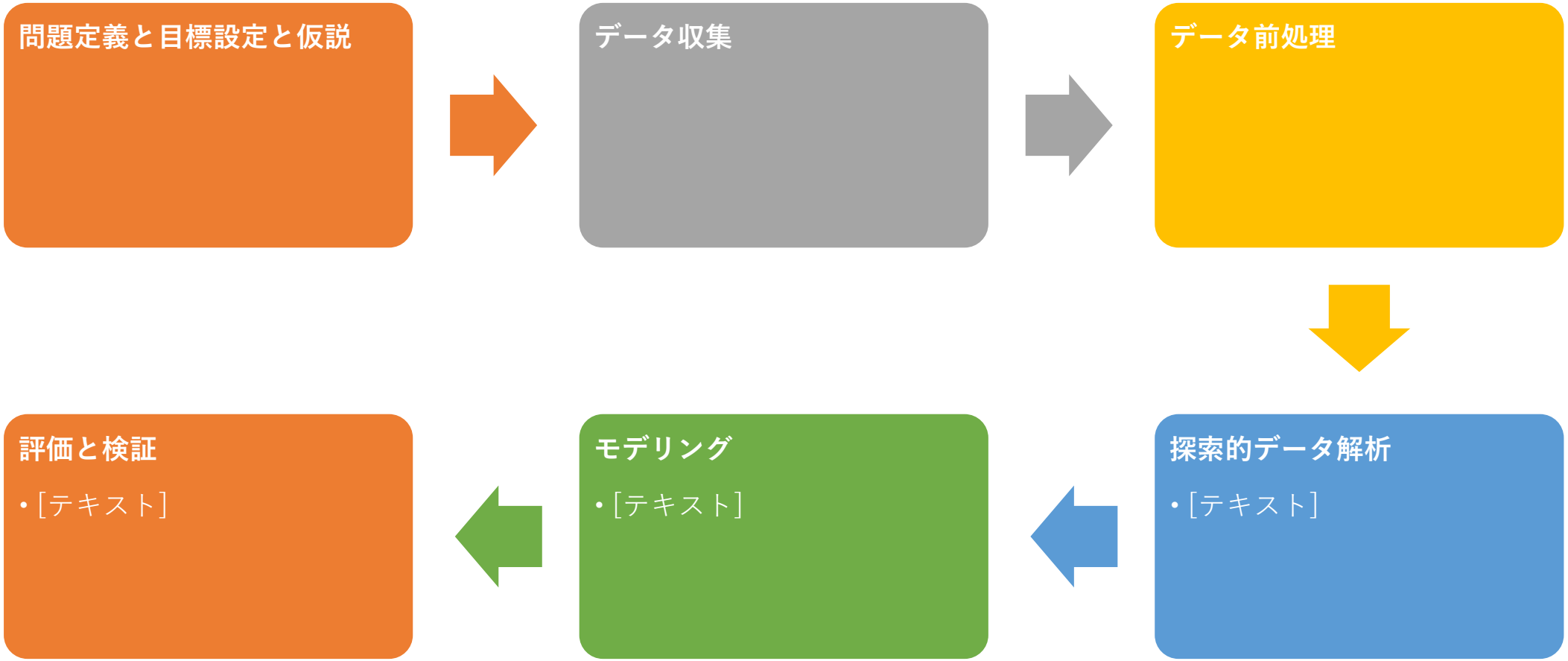
インテグレーションステップ
分析計画書

なぜ、成熟した民主国家は、投票率が下がるのか？？

データサイエンティスト育成講座
木曜クラス
佐藤 順


問題定義と目標設定と仮説
データ収集
データ前処理
探索的データ解析
• [テキスト]
モデリング
• [テキスト]
評価と検証
• [テキスト]

問題の定義と目標設定
仮説

問題の定義

- 日本の選挙の投票率が低いことが問題にされて、十数年経つがその問題は、現在でも未解決である。
- 国際的な統計でもＧ７の主要国でも選挙の投票率の低下は、指摘されている。

目標設定

- 世界の主要な民主国家の各種統計データと選挙の投票率の関連性を機械学習で分析する。
- モデルの精度や特徴量の重要度をもとに投票率と各種の統計データの関係性を考察する。
- 投票率の低下に影響する統計データについて考察する。

仮説

仮説

- 民主政治の経過年数が増えてくると国民の生活満足度が向上し政治に対する関心が希薄になり投票率が低下する。
- 各国の政治を取り巻く環境は変化するが、国民に対する教育の内容の更新が遅れ、教育レベル（教育の質）が低下し、政治への関心も低下する。

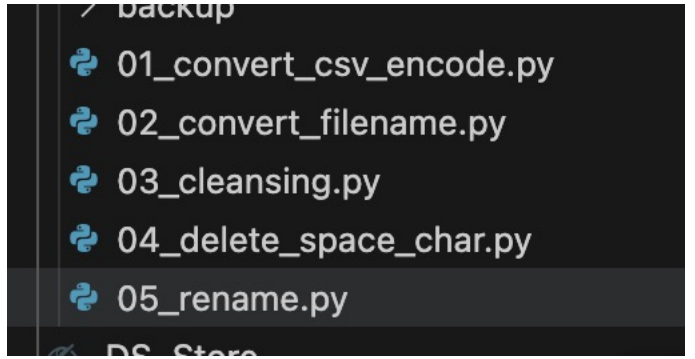
データ収集

- 各国の統計データは、Global Note を利用する。
  - https://www.globalnote.jp/

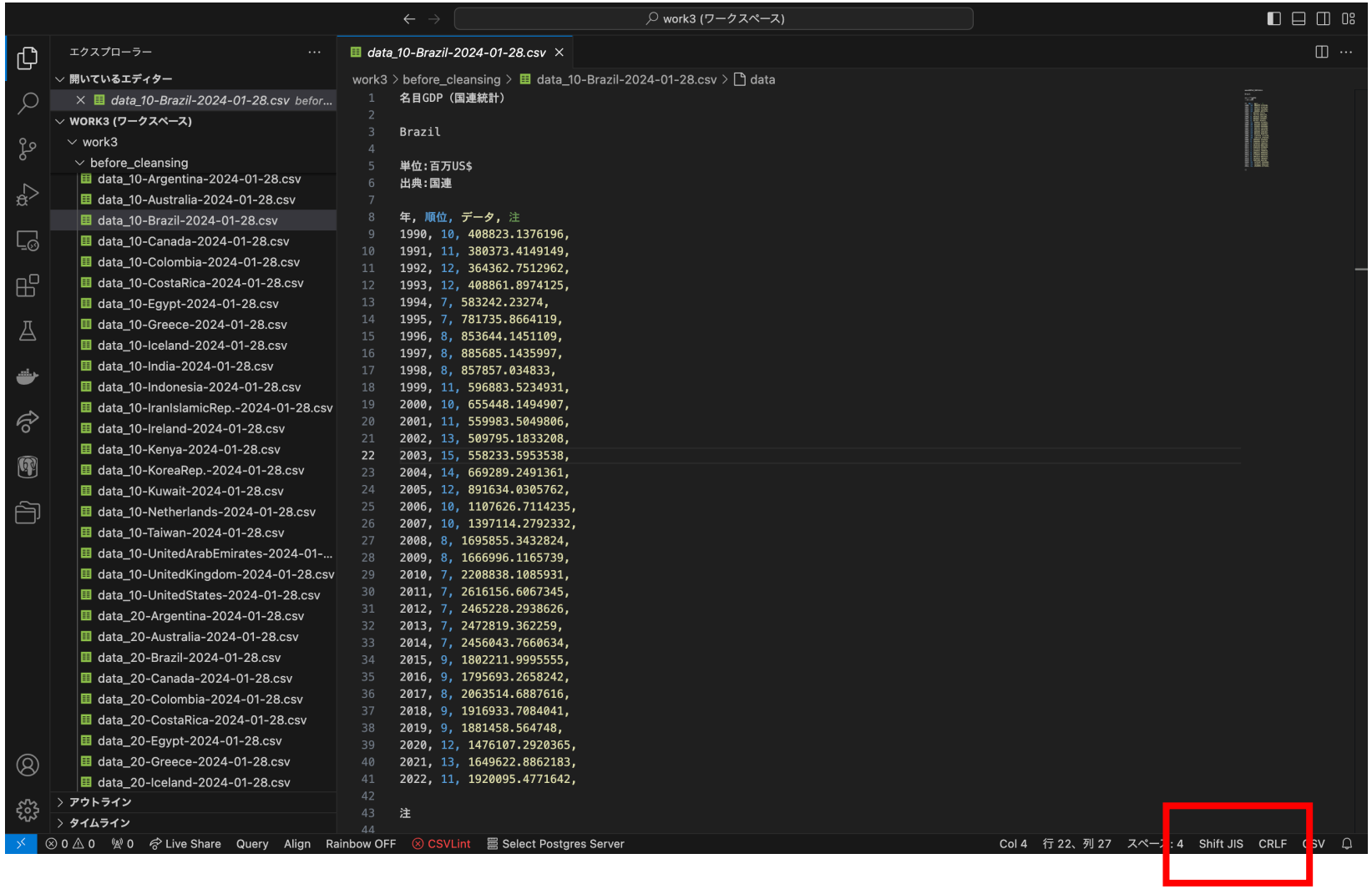
データ前処理

- Global Note のデータは、Shift JIS の CSV を含むテキストファイルとして、ダウンロードできる。

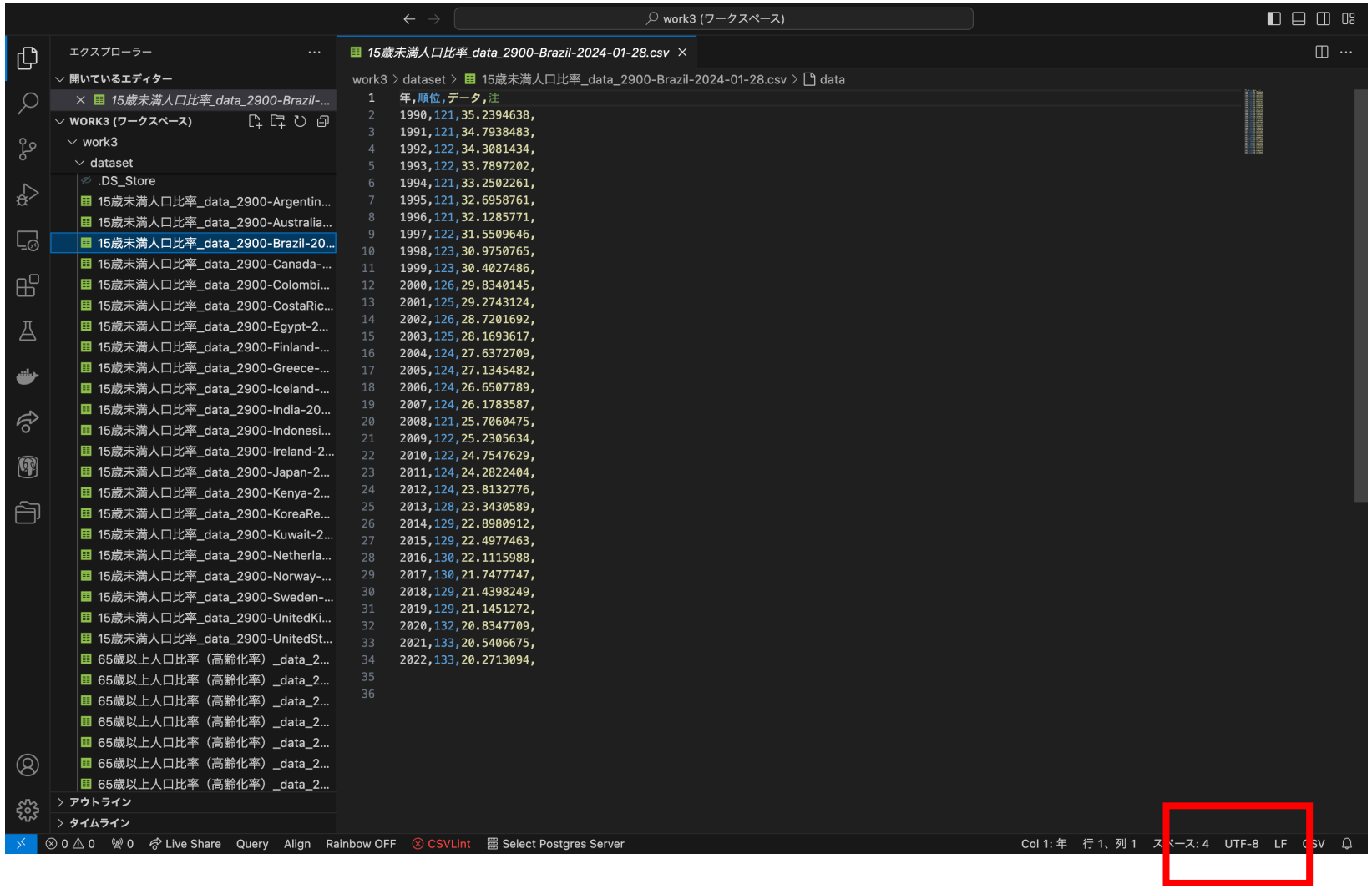
- 以下の処理を行うPythonコードを作成
  - 文字コードの変換（Shift JIS → UTF-8）
  - 不要な行の削除
  - 国名をファイルないから削除し、ファイル名に付与する。
  - 不要な半角スペースを削除する。
  - ファイル名の変更。



処理<**前**>のCSVファイル



処理<**後**>のCSVファイル



探索的データ解析（EDA）

- Jupyter Notebook での<**前**>処理
  -

モデリング

- 以下のモデルを作成して分析した。
  - LinearRegression（線形回帰分析）
  - RandomForestRegressor（ランダムフォレスト）
  - LGBMRegressor（Light GBM）

- 国については、以下の 22 カ国を用いた。
  - アルゼンチン、オーストラリア、ブラジル、カナダ、コロンビア
  - コスタリカ、エジプト、フィンランド、ギリシャ、アイスランド
  - インド、インドネシア、アイルランド、日本、ケニヤ、韓国
  - クウェート、ニュージーランド、ノルウェー、スウェーデン
  - イギリス、アメリカ
- 主に民主的に政治が行われていると思える国を選んだ。
  - 中国、北朝鮮、ロシアは、含めなかった。

- 特徴量として、以下の統計量を用いた。
  - 15歳未満人口比率、65歳以上人口比率（高齢化率）
  - ひとり当たり名目GDP（国連統計）、科学論文数（全分野合計）
  - 公的教育費の対GDP比、政治の民主化度
  - 大学進学率（短期大学含む）、名目GDP（国連統計）

- 以下の特徴量は、国によって、統計が存在しないものがあり、使用できなかった。
  - 経常収支、輸入総額、輸出総額、トマトの生産量
  - 教育制度への市民満足度、医療制度への市民満足度
  - 大卒人口比率（25歳-64歳）、15歳-64歳人口比率
  - 高齢者扶養率、年間労働時間（全就業者）

- 以下の特徴量については、対数変換を行なった。
  - 科学論文数（全分野合計）
  - 名目GDP（国連統計）
  - ひとり当たり名目GDP（国連統計）

- 最終的に 96 件のデータを準備することができた。

評価と検証

	LinearRegression	RandomForestRegressor	LightGBMRegressor
RMSE	115.64	73.54	8.89

- LinearRegression

順位	特徴量	Coefficient（影響度）
1位	政治の民主化度	18.27
2位	名目GDP（国連統計）	4.94
3位	公的教育費の対GDP比	0.62
4位	15歳未満人口比率	0.44
5位	大学進学率（短期大学含む）	0.13
-3位	65歳以上人口比率（高齢化率）	-0.46
-2位	ひとり当たり名目GDP（国連統計）	-2.30
-1位	科学論文数（全分野合計）	-5.50

- RandomForestRegressor

順位	特徴量	Importance（重要度）
1 位	政治の民主化度	0.42
2 位	公的教育費の対GDP比	0.14
3 位	大学進学率（短期大学含む）	0.098
4 位	ひとり当たり名目GDP（国連統計）	0.095
5 位	科学論文数（全分野合計）	0.08
6 位	15歳未満人口比率	0.06
7 位	65歳以上人口比率（高齢化率）	0.04
8 位	名目GDP（国連統計）	0.04

- LightGBMRegressor

順位	特徴量	Importance（重要度）
1 位	名目GDP（国連統計）	47
2 位	公的教育費の対GDP比	27
3 位	ひとり当たり名目GDP（国連統計）	20
4 位	政治の民主化度	17
5 位	科学論文数（全分野合計）	16
6 位	65歳以上人口比率（高齢化率）	15
7 位	15歳未満人口比率	10
8 位	大学進学率（短期大学含む）	5

結論

- 三つのモデルで共に「公的教育費の対GDP比」が比較的に重要度が高いことがわかった。
- 今回の分析では、公的教育費を高めることで、日本の選挙の投票率を上げることができる。という結論とする。

今後の展望と課題

- より深度を深めるためになぜ公的教育費を高めることが選挙の投票率に影響するのか？？という研究が今後の課題かと思う。
- 過去の関連研究に対する文献調査を実施したい。
- もともとの主題と仮説に対する調査が不十分だった。

データミックス教官の評価

- 用いたデータの件数が少なく、対象の国の範囲も狭いため学術的に価値を認めるには、まだ、不十分である。