

第10回

前回演習とデータ処理入門 (Pandas)

出席認証コード: 1699

本日の流れ

1. **前回の演習課題 (続き)** (約40分)
 - 他の人のアプリを見てみよう
2. **イントロダクション** (約5分)
 - データ分析パートの紹介
3. **講義：PandasとDataFrameの基本** (約20分)
 - データをプログラムで扱う準備
4. **講義&演習：データの絞り込み** (約20分)
 - 必要なデータだけを取り出す
5. **まとめと次回予告** (約5分)

1. 前回の演習課題 (続き) (40分)

- 前回の資料を参照

データ分析の世界へようこそ

ここからは、Webアプリケーションに「データを扱う力」を加えていきます。

2. イントロダクション (5分)

データ分析パート (第10～12回)

- **第10回 (本日):** データ処理入門
 - まずはデータをプログラムに読み込む
- **第11回:** 実践データ分析
 - データを集計し、ランキングを作る
- **第12回:** データ可視化
 - データをグラフにして、わかりやすく表現する

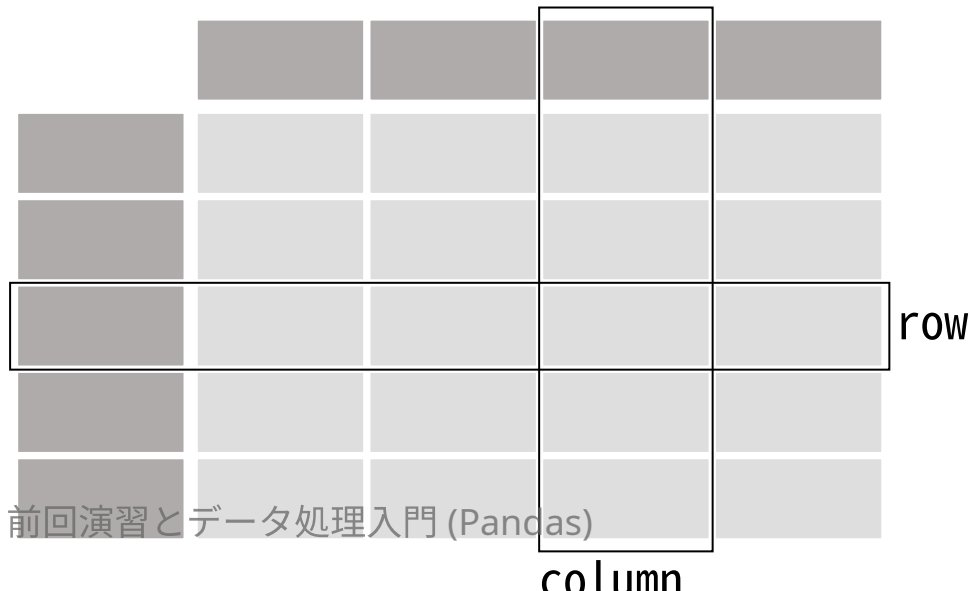
本日のゴール:

外部データを読み込み、中身を覗き、簡単な条件で絞り込めるようになる

3. 講義：PandasとDataFrameの基本 (20分)

- **Pandasとは？**
 - Pythonでデータを簡単に扱うための「工具箱 (ライブラリ)」。
 - Excelの表のようなデータを扱うのが得意。
- **DataFrameとは？**
 - Pandasが扱う、行と列からなる表形式データのこと。

DataFrame



row

column

データセットの紹介

Kaggle: Visa Issuance by Nationality and Region in Japan

- <https://www.kaggle.com/datasets/yutodennou/visa-issuance-by-nationality-and-region-in-japan>
- 日本政府が発行したビザの統計データ。
- `Year`, `Country`, `Number of issued_numerical` などの列が含まれる。

アップロード手順

1. 上記リンクからデータをダウンロード
2. ZIPファイルを解凍し、 `visa_number_in_japan.csv` を取得
3. Colabの左側のファイルパネルにドラッグ＆ドロップでアップロード

Colabで演習：データの読み込みと観察

`10_lecture.ipynb` を開いてください。

Step 1: Pandasをインポート

```
import pandas as pd
```

Step 2: CSVファイルを読み込む

```
# 'visa_number_in_japan.csv'をアップロードしておく  
df = pd.read_csv('visa_number_in_japan.csv')
```

Step 3: データの中身を覗いてみる

```
# 最初の5行を表示  
df.head()  
  
# データの基本情報を表示（行数、列数、データ型など）  
df.info()
```

4. 講義 & 演習：データの絞り込み (20分)

- 大量のデータから、必要な部分だけを取り出す操作です。

列の選択 (縦方向に絞る)

`DataFrame['列名']`

```
# 'Country' (国籍) の列だけを取り出す  
countries = df['Country']  
print(countries)
```

行の選択 (横方向に絞る)

DataFrame[条件式]

```
# 2023年のデータだけを取り出す  
df_2023 = df[df['Year'] == 2023]  
df_2023.head()
```

- `df['Year'] == 2023` の部分が「Year 列が 2023 である」という条件を表す。

Colabで演習：データの絞り込み

10_lecture.ipynb の続きです。

演習1: 「中国 (China)」のデータだけを抽出してみよう

```
# df_china に結果を代入してみよう
df_china = df[df['Country'] == 'China']
df_china.head()
```

演習2: 「Number of issued_numerical」が10000以上のデータだけを抽出してみよう

```
# df_large_count に結果を代入してみよう
df_large_count = df[df['Number of issued_numerical'] >= 10000]
df_large_count.head()
```

5. まとめと次回予告

本日のまとめ

- 前回の演習課題の時間を設け、Streamlitの復習と実践を行った。
- Pandasライブラリの基本を学んだ。
 - `pd.read_csv()` で外部データを読み込める。
 - `df.head()`, `df.info()` でデータの概要を掴める。
 - `df[条件式]` で特定のデータだけを絞り込める。

次回予告: 第11回 実践データ分析

- **複数条件での絞り込み:** 「2023年」かつ「韓国」のような、より複雑な条件を扱う。
- **集計 (`groupby`):** 国別の合計発行数などを計算する。
- **ランキング作成:** 発行数が多い順に並び替える。
- **Streamlit連携:** `st.selectbox` で選んだ国のデータを表示するアプリを作成する。

Q & A

質疑応答

お疲れ様でした！