

データサイエンス・オンライン講座 特別セミナー

今日から始められる、 Pythonによるビジネスデータ解析プログラミング講座

2024/12/14

野口 怜



Agenda

■ Part 1 : 今日から始めるための Pythonプログラミングの基本 (30min)

- ✓ Google Colaboratoryの使い方 (10min)
 - [Google Colab使い方演習](#)
- ✓ Pythonプログラミングの基本 (20min)
 - [Python基本文法演習](#)

■ Part 2 : ビジネスデータ解析に 必要なスキル演習 (2hr00min)

- ✓ データ分析の進め方 (5min)
- ✓ 分析に必要なデータサイエンス力とデータエンジニアリング力
 - [Pythonでできる分析実体験 : 分析デモンストレーション](#) (5min)
 - 記述統計学の基礎 (15min)
 - ・ [Pythonによる記述統計量算出演習](#)
 - データ観察の基本 (25min)
 - ・ [Pythonによるグラフ描画演習](#)
 - データ加工の基本 (20min)
 - ・ [Pythonによるデータ加工演習 \(ダミー変数化、外れ値・欠損値の処理\)](#)
 - データ分析 (モデル構築) の基本 (50min)
 - ・ [Pythonによるモデル構築演習 \(単回帰分析、重回帰分析\)](#)

■ Part 3 : 実際のビジネスデータ解析 に向けた実践演習 (2.5hr)

- ✓ e-Stat, SSDSEの紹介と使い方 (15min)
- ✓ データ分析の実践演習 (2hr)
 - [決定木分析によるビジネス意思決定サポート](#)
 - [階層的クラスタリングによるデータ層別と戦略ポイントの検討](#)
- ✓ 今後の継続学習、実践活用のためのポイント (15min)

※青字はGoogle Colaboratory による演習

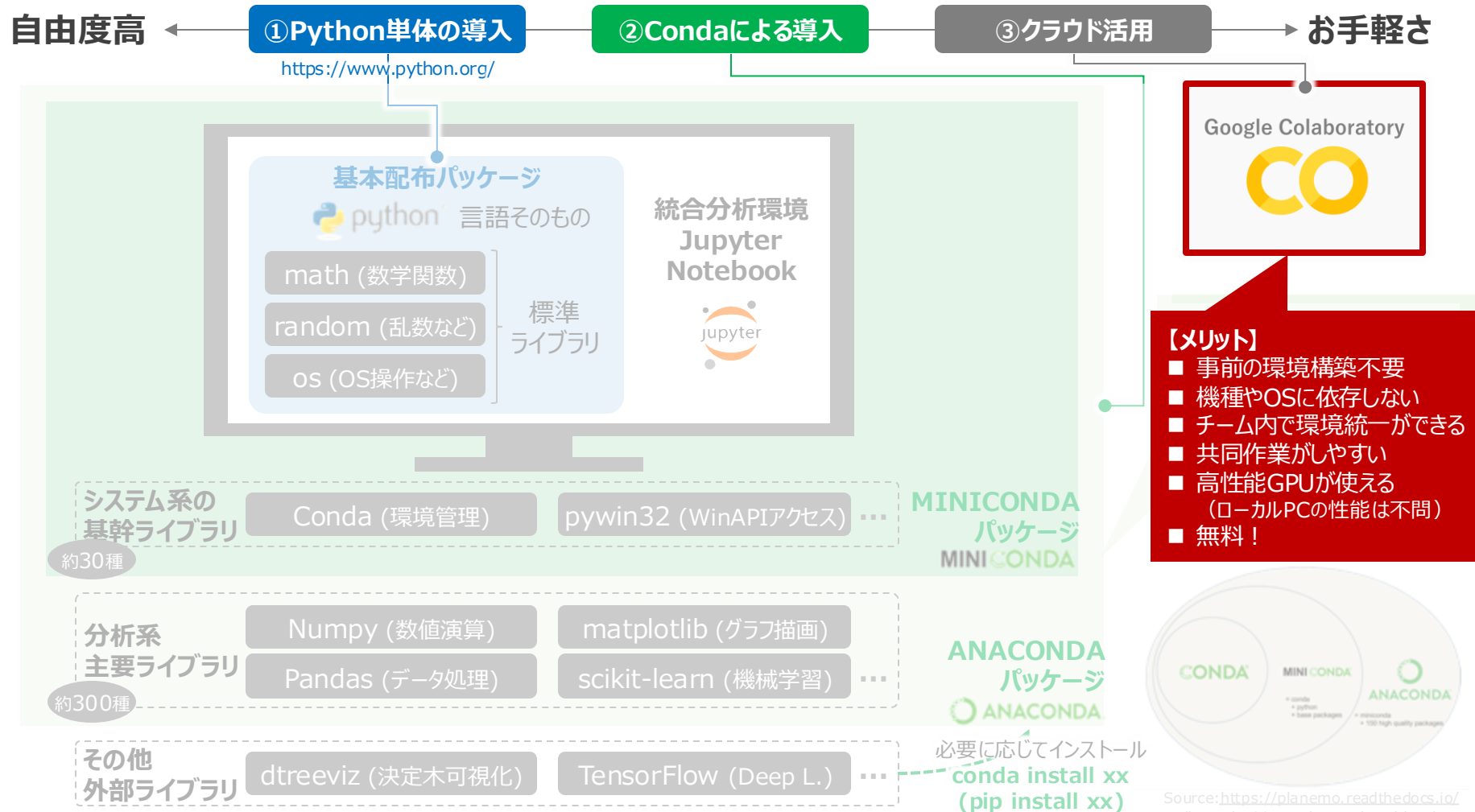
Part 1

今日から始めるための Pythonプログラミングの基本

- ✓ Pythonプログラミング環境：Google Colaboratoryの使い方
- ✓ Pythonプログラミングの基本

Python導入方法の3パターンとGoogle Colaboratory

- Pythonの初期環境構築には、①Python単体の導入、②Condaによる導入、③構築済のクラウド活用（Google Colaboratory）の大きく3パターンが存在。環境統一や共同作業を行う場合は、③が適している



※今のところ、miniconda+conda-forgeの組合せは無償！

Google Colaboratoryへのアクセス方法とノートブックの作成

- Google Colaboratory はインストール不要のPython開発環境であり、Googleのサーバーにインターネット接続（クラウド環境）することで、ブラウザ上でコーディング、実行ができる
- 必要なデータも随時、Googleのサーバーにアップする必要がある

1

Google Colaboratory へアクセス
<https://colab.research.google.com>

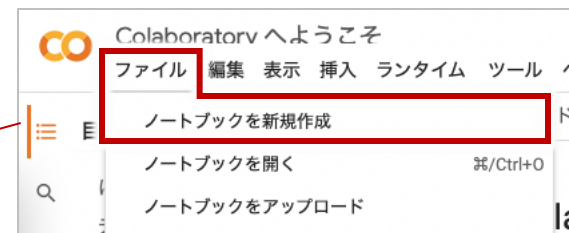
2

右上の [ログイン] ボタンより、Googleアカウントでログイン



3

左上の [ファイル] → [ノートブックを新規作成]



Google Colaboratory画面の概要

画面構成

このスクリーンショットはGoogle Colaboratoryのインターフェースを示しています。左側のサイドバーには「コード目次」のアイコンと「ファイル一覧」のアイコンがあります。上部のメニューバーには「ファイル」「編集」「表示」「挿入」「ランタイム」「ツール」「ヘルプ」があり、その右側には「コメント」「共有」「設定」のボタンがあります。中央の作業領域には「+ コード」と「+ テキスト」のボタンがあり、それぞれ「コードセルを追加」「テキストセルを追加（見出しなどを作りたい場合）」と説明されています。右側の「設定」ウィンドウには「フォントサイズ」が14pxに設定されており、「行番号」「インデントガイド」がチェックされています。また、「コードセルを閉じかっこと閉じ引用符を自動的に追加する」もチェックされています。

コード目次

ファイル一覧

テキストセルを追加
(見出しなどを作りたい場合)

コードセルを追加

コードは基本的に自動保存！

クリックでファイル名変更可能

ノートブックを共有したい場合

設定画面（以下オススメ設定）

設定

サイト

フォント... 14 px フォントサイズ

編集者

Colab Pro

インデント幅 (スペース)

2

GitHub

他の詳細

80

☒ コード入力時の候補を自動的に表示する

☒ 行番号を表示

☒ インデントガイドの表示

☐ エディタでコードの折りたたみを可能にする

☒ コードセルで閉じかっこと閉じ引用符を自動的に追加する

行番号

インデントガイド

キャンセル 保存

コードセルの編集

このスクリーンショットはGoogle Colaboratoryのインターフェースを示しています。上部の「編集」メニューが開かれ、様々なセル操作が可能であることが示されています。また、下部の「実行ボタン」や「コード記載領域」も示されています。右側の「セル操作」メニューには「セルを上へ」「セルを下へ」「リンクをコピー」「コメントを追加」「設定を開く」「別タブで開く」「セルを削除」などの操作がリストアップされています。また、「+ コード」と「+ テキスト」のボタンも表示されています。

上部の編集メニューからも
様々なセル操作が可能

編集 表示 挿入 ランタイム ツール ヘルプ

スクラッチセルの削除を元に戻す ⌘/Ctrl+M Z

やり直す ⌘/Ctrl+Shift+Y

すべてのセルを選択 ⌘/Ctrl+Shift+A

セルまたは選択範囲を切り取り

セルまたは選択範囲をコピー

貼り付け

選択したセルを削除 ⌘/Ctrl+M D

実行ボタン

コード記載領域

セル操作

セルを上へ

セルを下へ

リンクをコピー

コメントを追加

設定を開く

別タブで開く

セルを削除

その他操作

+ コード

+ テキスト

コードセル／テキストセルを追加
(マウスオーバーで表示される)



Google Colaboratory上での レクチャー&演習

「変数」とは

- 変数とは、「データを入れておく箱」のことであり、数値に限らず、日付型や文字列型のデータも格納しておくことができる。変数同士の演算も可能で、実際の分析では変数を用いて行う
- 変数名は自由に設定できるが、使用可能な文字には制限があるため、注意が必要

▼プログラミングでは「変数」と呼ばれる「箱」にデータを入れる

例：数値を入れる場合

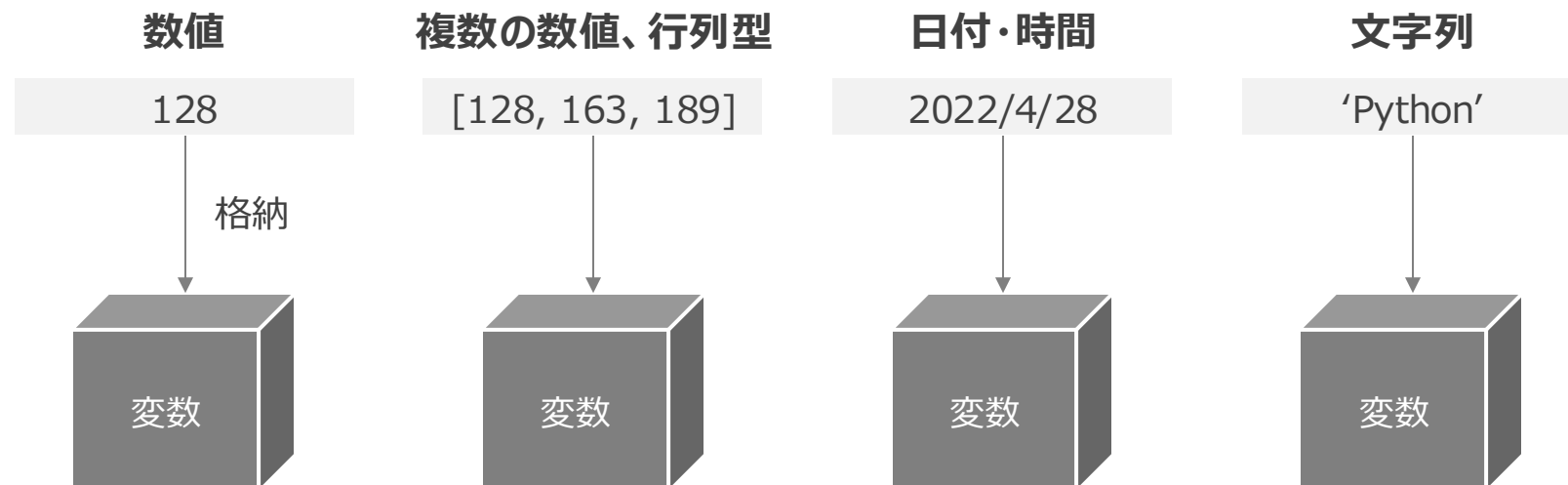
```
num = 10  
print(num)  
>> 10
```

例：文字を入れる場合

```
str = "Hello"  
print(str)  
>> Hello
```

例：演算（足し算）の処理

```
num1 = 1  
num2 = 2  
num1 + num2  
>> 3
```



▼変数名の制約

OK

- アルファベット（大文字・小文字は区別）
- ひらがな／カタカナ／漢字 ※非推奨
- 数字（ただし頭文字は不可）

NG

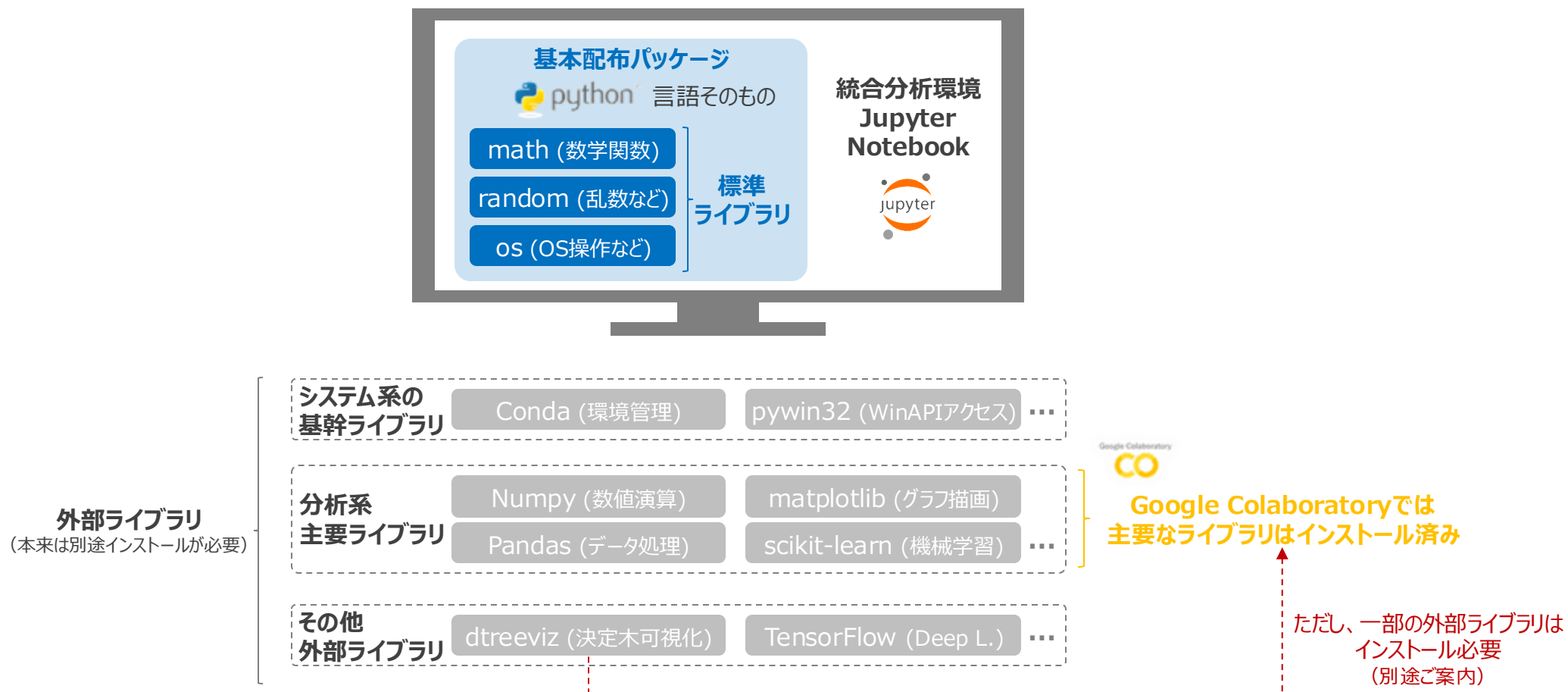
- 記号（@ ! ? - など。アンダーバー _ はOK）
- 環境依存文字（① * などの記号）
- 予約語や関数名



Google Colaboratory上での レクチャー&演習

「ライブラリ」について

- Pythonでは、よく使う関数や便利な機能をひとまとめでした「**ライブラリ**」が豊富に提供されており、コード作成の際は、随時必要なライブラリを import して用いるのが基本
- インストール不要な標準ライブラリその他、別途インストールが必要なライブラリもある



ビッグデータ分析／機械学習でよく用いられるライブラリ

ライブラリ名 (パッケージ名)	用途
matplotlib	基本グラフ描画
Seaborn	拡張グラフ描画
Numpy	数値演算、ベクトル/行列演算
Pandas	テーブル形式データのデータ処理
scikit-learn	機械学習モデル



表データの処理ライブラリ：pandas

- pandasは、**表形式のデータ**を処理する上で必須のライブラリである
- 表データをそのまま読み込んだ**DataFrame型**と、一部の列（行）のみを抽出した**Series型**とが存在する

表形式のデータ（テーブルデータ）

Excel, csvファイル など

顧客番号	性別	...	残高	契約有無
000011	男性	...	2,520,000	契約
000012	女性	...	12,353,000	契約
000013	女性	...	4,531,000	未契約
⋮	⋮	⋮	⋮	⋮

Pandas

- Excel
- CSV
- テキスト

Pandasのデータフレーム型

インデックス

カラム名／列名

	顧客番号	性別	残高	契約有無
0	000011	男性	2520000	契約
1	000012	女性	12353000	契約
2	000013	男性	4531000	未契約

DataFrame型

行
(row)

列
(column)

Series型

💡 余談：pandasの由来

諸説ありますが・・・下記が有力です
**Python and data analysis
panel data**



画像出典：Adobe Stock
<https://stock.adobe.com/jp>

標準ライブラリのみで
読み込むと・・・

- ×Excel
- CSV
- テキスト

ネストされた2重のリスト
(2次元配列とも呼ばれる)

```
[['000011', '男性', 2520000, '契約'],  
 ['000012', '女性', 12353000, '契約'],  
 ['000013', '男性', 4531000, '未契約']]
```



Google Colaboratory上での レクチャー&演習

(参考) 用語の整理 : 「ライブラリ」「パッケージ」「モジュール」「クラス」

