まえがき

本書は Python によるデータ分析の書籍を読むための入門書です。話 題の Python を使ってデータを分析してみたい。ただ、プログラミング 経験がまったくないので不安だ、もしくは Python の入門書を読み始め たけど挫折したという読者が対象です。

本書はこれまでの Python 関連書籍を補完します。文系の読者が挫折 しやすいのは,理系向けの書籍は統計学・数学の説明が中心だからです。 一方で,Python の使い方を簡単に解説した書籍が物足りないのは,ど んな場面に応用できるかを踏み込んで示していないからです。

そこで、本書では細かいことから積み上げず、こんなことができると いう事例を概観して、イメージを持つことを重視します。このため、説 明は短く、数式や専門用語はほぼなし、各章ごとのテーマを体験しなが ら学ぶことが特徴になっています。

途中で挫折しない工夫もしています。各章は独立しており,経営学, 経済学,社会学に関連した事例を取り上げ,興味のある章から読めるように配慮しています。

Python や統計学・数学を1からすべて学ぶのは大変です。興味のあるトピックスに関連する部分から学べばよいのです。その際、Pythonの使い方だけではなく、分析手法の応用の仕方を学びましょう。

また, Python の関連書籍で初学者がわからなくなりやすい部分や警告・エラーが出る場合にどうしたらよいかや, さらにちょっと深く理解 したい場合の補足などを「よくある質問」「たまにある質問」というコ ラム形式で回答しています。

いろいろ詰め込むのではなく、本文は骨子にとどめ、補足を「質問」 にしたわけです。これは理解を助ける工夫です。コンパクトにすること で読みやすくしただけでなく、質問形式だと頭に入りやすいというメリ

i

ットがあります。とにかく手っ取り早くという要望に応え,できるだけ そぎ落として読みやすくしたのが特徴です。

■個人で使用される方へ

とりあえず,本書の内容を写経のようにとにかく打ち込んで慣れるこ とから始めましょう。はじめのうちはエラーが多く,思いのほか時間が かかるかもしれません。ただ,続けるのが大事です。だんだんと慣れて きます。

■講義で使用される方へ

1章1回分,半期を目安にされるとよいでしょう。本当に右も左もわ からない人向けの内容なので,社会科学にとどまらず,広くいろいろな 分野の初心者向けの講義(高校生の授業を含む)にも適していると思って います。

本書が、将来、次のステップの勉強に進むための役に立てれば幸いで す。その意味で、Python によるデータ分析の書籍を読むための入門書 としています。このため、厳密な説明や用語というよりは、わかりやす さを優先したところがあることをご理解いただけると幸いです。

最後になりましたが、本書のタイトルを含めいろいろご提案いただい たり、新しいバージョンのコードに修正していただいたりした有斐閣の 渡部一樹さん・渡辺晃さんに感謝申し上げます。本書がわかりやすい内 容になっているとすれば丁寧なコメントをくださったおふたりのおかげ です。

2024年9月

友原章典

動作環境

本書で示されるプログラムは、次の環境で動作確認しています(2024 年7月時点)。

OS: Windows 11 Home バージョン 22H2 Python: Python 3.11.8 Jupyter-notebook: 7.0.6 Web ブラウザ: Google Chrome

beautifulsoup44.12.2	pandas2.1.4
ipykernel6.28.0	pip23.3.1
ipython	PuLP2.7.0
matplotlib	scikit-learn1.3.0
mecab-python31.0.8	seaborn0.12.2
networkx	wasabi0.10.1
numpy1.24.3	watchdog2.1.6
openpyxl	wordcloud1.9.3

注意事項

さまざまな要因が影響しうるため、上記の条件を再現しても本書で示され るプログラムの実行結果がすべての環境で再現されるとは限りません。本書 の内容は 2024 年 3 月の原稿執筆時点のものであり、本書に掲載したソフト ウェアのバージョンや URL、プログラムの出力結果、操作方法・手順などは 変更される可能性があります。

プログラムファイルの作成にあたっては,内容に誤りのないようできる限 リ注意を払いましたが,結果生じたこと(損害等)には,著者,出版社とも 責任を負うことはできません。

本書に記載されている会社名,製品名ならびにサービス名等は,各社の商標および登録商標です。

次

* ……Python の文法やデータ分析について, やや発展的な話題を扱っていることを意 味します。はじめは読み飛ばしてもかま いません。

- 🕄 Python を学ぶメリット 🛛 4
- 小書の概要と各章のモチベーション 5
 - 1 数値データの分析 7
 - 2 数理モデルによるシミュレーション 9
 - 3 テキストデータの分析 10
- ふ要なものをインストールしよう
 Jupyter Notebook 10
 - **1** Anaconda のインストール 11
 - 2 Jupyter Notebook の手始め 11
 - 3 ライブラリをダウンロードする 13

数値データ 16
 データフレーム 16
 データフレーム 16
 データの行数(個数)の確認 19
 要素の抽出 20
 要素の削除 21
 データフレーム同士の連結 23
 欠損値 24
 条件を満たす要素の抽出 25
 ループ(繰り返し) 27
 文字データ 28

1 データの種類 28

- 2 分割 29
- 3 結合 29
- 4 出力する文字に改行を含める 30
- 🕄 データの取り込み 31
-] 平均 36
 - 1 データの入力と住民数の表示 36
 - 2 平均の計算 37
- 22 ヒストグラムで分布を把握する 38
 - **1** ヒストグラムの描き方 **39**
 - **2** データの抽出* 42
- 3 分布の特徴を表す指標 44
 - 1 極端な値の影響 44
 - 2 老年人口指数――高齢化を議論するための指標* 46
 - 3 中央値 47
 - 4 最頻値 48
 - 5 標準偏差 49

- 1. 相関係数 52
- 名 散布図 53
 - **1** データの取り込み 53
 - 2 関係性を散布図にする 54

3 相関の数値化 57

- 1 相関係数を求める 57
- 2 各組み合わせの相関係数を一括で出力する 60
- 3 相関係数を視覚的に表す――ヒートマップ 61

4 相関関係と因果関係* 63

- 1 相関関係と因果関係の違い① 63
- 2 相関関係と因果関係の違い② 64

3 相関関係と因果関係の違い③ 65

-] 回帰分析 70
- 23 傾向線 71
 - 1 広告費と売上高の傾向線 71
 - 距離と売上高の傾向線
 72
 - 3 広告費と距離の傾向線 72
- 🕄 単回帰 74
 - 広告費と売上高の回帰式 74
 - **2** 距離と売上高の回帰式 76
- ④ 重回帰 77
 - 1 広告費・距離・売上高の3次元のグラフ 77
 - 2 広告費・距離と売上高の関係式 78

- 1 クラスタリング 82
- 2 グループに分ける 83
 - **1** データの取り込み 83
 - **2** グループ分け 84
 - 3 KMeans()では何をしているのか* 85
 - 4 グループの平均値 88
 - 5 図の描き方 88

3 グループに分ける 92

- **1** グループ分け 92
- 2 グループの平均値 93
- 3 図の描き方 93
- 4 グループ分けの結果 94
- 4 グループに分ける 97
 - **1** グループ分け 97
 - 2 グループの平均値 98

- 3 図の描き方 99
- 4 グループ分けの結果 100

-] 決定木 106
- 2 決定木による分類の準備 107
 - **1** データの取り込み 107
 - 2 データの確認と図の表示 108
 - **3** データの整理 110
 - 4 学習用データと評価用データの確認 112
- 3 決定木による分類と予測・評価 113
 - 1 学習用データによる分類 113
 - 2 モデルの予測・評価 114
 - 3 決定木のイメージ* 115
 - 4 分類のやり直しと決定木の図* 117
- ④ 決定木とクラスタリング* 122
 - **1** 決定木の活用例 122
 - 2 クラスタリングの活用例 124
 - 3 いずれを使うかは目的次第 125

- \bigcirc ランダム・フォレスト 130
- 2 ランダム・フォレストの準備 131
 - **1** データの取り込み 131
 - 2 行数と列数の表示 133
 - 3 欠損値の確認 133
- 3 ランダム・フォレストによる分類と予測 134
 - 1 学習用データによる分類 134
 - 2 学習用データによる分類の評価(評価用データ) 135
 - 3 学習用データによる分類の評価(学習用データ) 135
- 4 特徴量と予測の精度 136

- 1 予測の確認 136
- 2 特徴量の重要性
 138
- **3** 手順のイメージ* 138

第	♪ 章 データの規則性を探って将来を予測しよう③
1	ランダム・フォレストによる回帰分析の準備 147 1 データの取り込み 147 2 学習用データによる分析 148 3 分析結果の評価の準備 149 4 決定係数* 149 5 分析結果の評価 150
\mathcal{D}	ランダム・フォレストによる回帰分析 152
	 犯罪件数と予測した犯罪件数の図 152 45 度線の追加 153 特徴量の寄与 154
3	データの再現* 156 1 乱数のシード 156 2 疑似乱数 158
第	 施策の効果を調べよう
1	傾向スコア・マッチング 162
2	 傾向スコアの導出 164 1 データの取り込み 164 2 データの確認 165 3 傾向スコアの導出 166 4 重なり具合のチェック 168
B	傾向スコア・マッチングによる分析 169
_	1 グループ分け 170
	2 グループごとの平均の比較 171
	 2 グループごとの平均の比較 171 3 グループごとの平均の差 172
	 グループごとの平均の比較 171 グループごとの平均の差 172 図の出力 173

- 1 ダイクストラ法 181
- 2 ダイクストラ法のアルゴリズムのイメージ* 183
- 3 ダイクストラ法の図の導出* 187

- SIR モデル ロコミの分析 193
 分析の準備 193
 分析結果の図示 194
- 2 数値による SIR モデルの確認 196

- 1 線形計画法 204
- 2 線形計画法の実践 205
 - 1 計算のための準備 205
 - 2 式の入力 206
- 3 線形計画法の数理モデル* 207
 - 1 数理モデルと図 207
 - 2 図で確かめる① 209
 - 3 図で確かめる② 210
- 4 線形計画法の適用* 212

- 1 文章の分析 216
- ⑦ 形態素解析 217
 - 1 データの取り込み 217
 - 2 分析の準備 217
 - 3 リストの作成 218

- 4 分析結果の確認① 220
- 5 分析結果の確認② 220
- 3 配列と文字列* 224
- 読書案内 227
- 索引 229

ウェブサポート

本書で使用するデータや練習問題の解答例など,各種サポート情報を下記 のページから提供していきます。ぜひご利用ください。



https://www.yuhikaku.co.jp/books/detail/9784641166363



特徴を踏まえて適切な計画を立てよう



市町村の住民の特徴を比べて,地域住民の違いを調べます。住民の特徴を 1つの数値で示して比べることで,適切な政策を考えてみましょう。たとえ ば,住民の平均年齢が50歳と同じであるA市とB市を例にとります。平 均年齢で見た住民の特徴が同じなので,同じ政策(例:年金生活者に配慮す る政策)を行えば,いずれの市の住民も満足するでしょうか。

] _{平 均}

1 データの入力と住民数の表示

A市, B市, C市という3つの市の住民の年齢について架空データを 作りましょう(C市には179歳以上の年齢が入っていますが,データ分析の練 習のためのデータなので気にせず入力してください)。最初の4行については 後ほど説明します。

```
コード 3-1
import pandas as pd
import matplotlib.pyplot as plt
import japanize_matplotlib
import seaborn as sns
data = {
    "A市":「
         42, 43, 45, 46, 47, 48, 49, 50, 51,
         52, 52, 52, 52, 53, 55, 56, 57,
     ٦,
     "B市":「
         10, 14, 20, 23, 28, 33, 40, 51, 53,
         54, 62, 62, 70, 75, 80, 85, 90,
     ],
     "C市":「
         8, 13, 18, 26, 27, 29, 33, 35, 37,
         38, 39, 45, 45, 48, 179, 186, 198,
     ],
}
df = pd.DataFrame(data)
df.head()
```

い出力結果

	A市	B市	C 市
Θ	42	10	8

1	43	14	13
2	45	20	18
3	46	23	26
4	47	28	27

df.head() で最初の5行(それぞれの市に住む5人の年齢)が示されま す。前章で述べたように、Python では最初の行は「0」から始まります。 次にデータの「長さ」を調べる len()を使って、住民数を示します。

1	コード 3-2
ten(u)	

└⇒出力結果

17

それぞれの市に住民が17人いることがわかります。

ここではすべての市で住民数が同じですが、A市の住民数だけを示 したいときには次のように入力します。

1 on (df["A ="])	コード 3-3

い出力結果

17

2 平均の計算

mean()を使って,それぞれの市の年齢の平均値(mean)を計算します。 これまではセルの最終的な処理結果をそのまま出力することが多くあり ましたが,出力をより見やすくするために,ここでは print()を使い ます。出力したい内容をカンマ区切りで指定すると,それぞれをスペー

1] 平均 37

print("A市", df["A市"].mean())
print("B市", df["B市"].mean())
print("C市", df["C市"].mean())

し、出力結果

A市 50.0 B市 50.0 C市 59.05882352941177

なお, df.mean() と入力すると, 市ごとに計算するのではなく, すべ ての市の平均年齢を一度に計算できます。

	コード 3-5
dr.mean()	

い出力結果

A市 50.000000 B市 50.000000 C市 59.058824

A 市の平均年齢は 50 歳。B 市の平均年齢も 50 歳です。C 市の平均年 齢は約 59 歳です。

2 ヒストグラムで分布を把握する

A市とB市の平均年齢は同じですが、データを見るかぎり、2つの市 の特徴が同じようには見えません。そこで、年齢構成の特徴を可視化す るために、グラフを描いて、詳しく見てみましょう。

コード 3-4

1 ヒストグラムの描き方

新たなライブラリを使って、A市の住民の年齢の分布を図で描いて みましょう。分布図のことをヒストグラム(histogram)といいます。

```
import pandas as pd
import matplotlib.pyplot as plt
import japanize_matplotlib
import seaborn as sns
age_spans = [
    10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110,
    120, 130, 140, 150, 160, 170, 180, 190, 200,
]
df["A市"].plot(
    kind="hist",
    bins=age_spans,
    title="A市の平均年齢",
    grid=True
)
```

最初の4行の import で, 計算したり, 図を描いたりする際に必要な ものをインポートします。

pandas は第2章でも使いましたが、データの取り込みや分析などに 使うものです。matplotlib.pyplot は図を作成するときに使うもので、 japanese_matplotlib は日本語で表記されている変数名(売上高など) が図にきちんと表示されるようにするためのものです。seaborn は図を より見やすくするために使います(本章中のコードでは使用しませんが、練 習問題で使用します。これら4つはセットで使用することが多いので、「おまじ ない」のようにノートブックの先頭に書くクセをつけておくのもよいでしょう)。

ここでは, as を使うことで, pandas を pd, matplotlib.pyplot を plt などと省略して使えるようにしています。これによってコードを書 く負担が減ります。

import はまとめて記述する必要があるのですか?

Jupyter Notebook では、一度ライブラリをインポートすれば、ノートブッ クのそれ以降の箇所でいつでも使えるので、まとめて書く必要はありません。 ただ、使用するライブラリをまとめて書くことで、そのノートブックで行う分 析の概要を示せてわかりやすいので、まとめて書く慣習があります。

仕事(=分析)をするにあたり,能力(=機能)に応じて,一緒に働く仲間 を招集するイメージです。今回は,4行なので4人のパーティメンバーを集め たといったところでしょうか。

age_spansの部分でたくさんの数字を書いているのは、年齢幅を10歳区切りにするためです。

plot()の各引数を指定することでヒストグラムを生成できます。 kind は図の種類をヒストグラム(hist)として指定するもので, bins ではデータの区切りを指定しています。titleとgridの引数はその名 のとおり、タイトルとグリッド(目盛り線)を指定するオプションです。

図 3-1 のようなグラフが出力されます。



図 3-1 コード 3-6 の出力結果

横軸が年齢,縦軸は住民数です。40~49歳の住民が7人,50~59歳 の住民が10人いるのがわかります。

A市では、住民の年齢が50歳付近に集中しています。

たまにある質問

図の日本語が文字化けします。どうすればいいですか?

japanize_matplotlib が正しくインストールされていない可能性があり ます。エラーが表示されていたらその文面でネット検索することで解決のヒン トが得られます。

どうしてもエラーが解消しないときには、sns.set(font="Yu Gothic") や sns.set(font="Meiryo")のように個別にフォントを指定しても大丈夫 です。それぞれ游ゴシックやメイリオのことです(Jupyter Notebook を実行 しているパソコンにインストールされているフォントでないものを指定しても 反映されないのでご注意ください)。

import japanize_matplotlibの1行を入れると, sns.set()を使う必要はありません。どちらか1つで構いません。

B市の住民の年齢の分布を図で描いてみましょう。

df["B市"].plot(kind="hist", bins=age_spans,	コード 3-7
title="B市の平均年齢", grid= True	
)	

図 3-2 のようなグラフが出力されます。

B市では,住民の年齢が10~100歳の間に散らばっているのがわかります。

著者紹介

青山学院大学国際政治経済学部教授 早稲田大学政治経済学部卒。ジョンズ・ホプキンス大学大学院 Ph.D.(経済学)。米州開発銀行,世界銀行コンサルタントから, ニューヨーク市立大学助教授,UCLA経営大学院エコノミスト などを経て現職。『移民の経済学――雇用,経済成長から治安 まで,日本は変わるか』(2020年,中央公論新社),『会社では ネガティブな人を活かしなさい』(2021年,集英社)など。

文系のための Python データ分析 — 最短で基本をマスター

友原章典(ともはら・あきのり)

Introduction to Data Analysis with Python

2024年10月30日初版第1刷発行

- 著 者 友原章典
- 発行者 江草貞治
- 発行所 株式会社有斐閣
 〒101-0051 東京都千代田区神田神保町 2-17
 https://www.yuhikaku.co.jp/
- 装 丁 嶋田典彦 (PAPER)
- 印 刷 大日本法令印刷株式会社
- 製 本 牧製本印刷株式会社
- 装丁印刷 株式会社亨有堂印刷所

落丁・乱丁本はお取替えいたします。定価はカバーに表示してあります。 ©2024, Akinori Tomohara Printed in Japan. ISBN 978-4-641-16636-3

本書のコピー、スキャン、デジタル化等の無断複製は著作権法上での例外を除き禁じられています。本書を代行 業者等の第三者に依頼してスキャンやデジタル化することは、たとえ個人や家庭内の利用でも著作権法違反です。

【JCOPY】 本書の無断複写 (コピー)は、著作権法上での例外を除き、禁じられています。複写される場合は、そのつど事前に、(− 社)出版者著作権管理機構(電話03-5244-5088、FAX 03-5244-5089, e-mail:info@jcopy.or.jp)の許諾を得てください。