Python 3 エンジニア認定データ分析試験用模擬試験

一般社団法人Pythonエンジニア育成推進協会が運営する「Python 3 エンジニア認定データ分析試験」の試験範囲に準拠したオリジナル模擬試験です。 試験については、下記を参照してください。

- https://www.pythonic-exam.com/exam/analyist

本模擬試験は、Pythonプログラミング学習サービスPyQ（https://pyq.jp）が提供しています。

1. データ分析で使われるプログラミング言語やツールの説明として、誤っているものを選択してください（1つ選択）。

A. Pythonはデータ分析以外に、Webシステムの構築やIoTデバイスの操作でもよく使われる。一方で、Webアプリなどのフロントエンドや低レイヤー処理を行うことは苦手である

B. Javaなどの汎用プログラミング言語はソフトウェア開発が得意である。一方で、データ分析に関するライブラリが充実していなかったり、サンプルが少ないことがある

C. R言語はオープンソースのプログラミング言語であり、Pythonよりも特化した統計分野のツールを使える。一方で、機械学習の分野は苦手である

D. Microsoft Excelの利点はGUIでの操作が可能な点である。一方で、日々繰り返し発生するデータ取り込みを自動で行うには、VBAなどのプログラミングやツールが必要である

2. データサイエンティストとデータ分析エンジニアに関する説明として、正しいものを選択してください（1つ選択）。

A. データ分析エンジニアが付加的に持つべき技術のひとつに、高校から大学初等レベルの数学の知識がある

B. 実務の役割分担では、データサイエンティストはモデルやアルゴリズムの構築、新技術への取り組みが強く求められ、解決したい課題に向き合う実務は主にデータ分析エンジニアが担う

C. データハンドリングとは機械学習モデルの学習プロセスのことを指し、業務の8割とも9割とも言われている

D. データサイエンティストは、数学・情報工学・理学の3つの分野の知識を総合的に持つ職種である

3. Pythonの仮想環境に関する説明として、正しいものを選択してください（1つ選択）。

A. venvでは、1つの仮想環境内に同じパッケージの複数のバージョンをインストールできるため、状況に合わせてバージョンの切り替えができるようになる

B. venvやpipコマンドは開発者が1人だけの場合は不要だが、複数の開発者が共同作業を行うプロジェクトでは、パッケージのバージョンを統一するために必要になる

C. Anacondaで作った仮想環境でもpipコマンドが使えるため、Anaconda環境でも基本的にはpipコマンドを使うことが推奨される

D. 他の人とパッケージのバージョンを統一するには、`pip freeze` コマンドとrequirements.txtを利用するのが便利である

4. Pythonの標準コーディング規約に関係する説明として、正しいものを選択してください（1つ選択）。

A. Pythonの標準コーディング規約であるPEP 8では、複数のモジュールは1行にまとめてインポートすることが推奨されている

B. PEP 8違反をチェックするツールはPythonには同梱されていないため、pipコマンドでインストールする必要がある

C. flake8を使うとPEP 8違反は検出できるが、未使用の変数やモジュールなどの論理的なチェックはできない

D. PEP 8違反があるとプログラムの実行時にエラーが起きるため、実行前にツールでチェックすることが推奨される

5. 次のようなコードがあるとき、「期待する結果」のように出力する記述として【1】に当てはまるものを選択してください（1つ選択）。

コード:

```
data = ["python programming", "data science", "start python"]
print(  【1】  )
```

期待する結果:

```
{'python programming', 'start python'}
```

A. `[x for x in data if "python" in x]`

B. `[x for x in data if x in "python"]`

C. `{x for x in data if "python" in x}`

D. `{x for x in data if x in "python"}`

6. Pythonの標準ライブラリに関する説明として、正しいものを選択してください（1つ選択）。

A. reモジュールの`search`メソッドで、文字列が正規表現にマッチしなかった場合、戻り値は`None`になる

B. loggingモジュールで出力できるログレベルのうち、重要度が最も低いのは`INFO`である

C. 任意の書式を指定して日時を文字列に変換するには、datetimeモジュールの`isoformat()`メソッドを使う

D. pathlibモジュールの`Path`クラスを使うと、`Path("/root") + "sub_dir" + "sample.txt"`のように`+`演算子でパスを作れるようになる

7. JupyterLabに関する説明として、誤っているものを選択してください（1つ選択）。

A. Notebookのファイル本体はJSON形式で記述されているが、MarkdownやHTMLなどの形式でダウンロードすることも可能である

B. JupyterLabを使うと、プログラム・実行結果・ドキュメントを1つのファイルにまとめられる

C. JupyterLabでは`!`で始まるマジックコマンドがあり 、実行時間の計測など便利な機能を利用できる

D. JupyterLabにおけるプログラムの記述時の動作はIPythonをベースとしており、Tabキーによる補完機能も拡張機能なしで利用できる

8. x_0からx_nをすべて掛け合わせる数式として、正しいものを選択してください（1つ選択）。

A.

$$x_0!$$

B.

$$x_n!$$

C.

$$\prod_{i=0}^{n-1} x_i$$

D.

$$\prod_{i=0}^{n} x_i$$

9. ベクトルや行列の操作の説明として、誤っているものを選択してください（1つ選択）。

A. ベクトルの内積は、ベクトルになる

B. A が2×3行列、 B が3×2行列のとき、 AB は2x2行列になる

C. 行列とベクトルの掛け算ができるとき、その結果は行列になる

D. ベクトルにスカラーを掛けた結果は、ベクトルである

10. ベクトルや行列の操作として、計算可能なものを選択してください（1つ選択）。

A. 3×2行列に、要素数3のベクトルを掛ける

B. 3×2行列を転置した行列に、3×2行列を掛ける

C. 2×3行列に、2次の正方行列を掛ける

D. 要素数3のベクトルから、要素数が2のベクトルを引く

11. 微分と積分に関する説明として、誤っているものを選択してください（1つ選択）。

A. 不定積分には積分定数がつくが、定積分にはつかない

B. 微分係数を計算することで、関数が増加しているか減少しているかがわかる

C. 関数F(x)を微分してf(x)になるとき、Fをfの原始関数、fをFの導関数と呼ぶ

D. 積分は下記の記号を使って表現できる

$$\frac{dy}{dx}$$

12. データ間の関係性を見る指標の説明として、正しいものを選択してください（1つ選択）。

A. すべてのデータが同じ値のとき、分散は1になる

B. 相関係数は、共分散を2つの変数の分散で割った値であり、-1以上1以下の値になる

C. 分散と共分散は必ず非負の値になる

D. スピアマンの順位相関係数はデータの順番だけに着目するため、実際のデータの値は必要ない

13. 以下のような8面体のサイコロの目の確率変数と確率分布があるとき、誤っているものを選択してください（1つ選択）。

X	1	2	3	4	5	6	7	8	計
$P(X)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	1

A. この確率分布は、離散一様分布である

B. 期待値は4になる

C. サイコロを1回ふって「偶数が出た」という情報を知らされたとき、この条件のもとで出た目が2である確率は0.25である

D. サイコロを1回ふって「2以下の目が出た」という情報を知らされたとき、その情報量は2ビットである

14. 次のコードを実行したとき、`a`の中身として正しいものを選択してください（1つ選択）。

コード:

```
import numpy as np
```

```
a = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

a[1, 1:] = 10
```

A.

```
array([[10, 10, 10],
       [ 4,  5,  6],
       [ 7,  8,  9]])
```

B.

```
array([[ 1,  2,  3],
       [ 4, 10, 10],
       [ 7,  8,  9]])
```

C.

```
array([[10,  2,  3],
       [10,  5,  6],
       [10,  8,  9]])
```

D.

```
array([[ 1,  2,  3],
       [ 4, 10,  6],
       [ 7, 10,  9]])
```

15. `np.linspace(2, 11, 3)` の結果として、正しいものを選択してください（1つ選択）。
なお、事前に `import numpy as np` を実行しているものとします。

A. `array([2., 5., 8.])`

B. `array([2., 5., 8., 11.])`

C. `array([ 2., 6., 10.])`

D. `array([ 2. , 6.5, 11. ])`

16. 次のようなデータがあるとき、NumPy配列を操作するコードと実行結果の組み合わせとして、正しいものを選択してください（1つ選択）。

コード:

```
import numpy as np

a = np.array([0, 1, 2])
b = np.array([[0, 1, 2], [3, 4, 5]])
```

A. `np.vstack([a, b]).shape` の結果は、`(3, 3)` である

B. `np.diff(a).shape` の結果は、`(3,)` である

C. `np.ravel(b).shape` の結果は、`(6, 1)` である

D. `b.reshape((3, 1)).shape` の結果は、`(3, 1)` である

17. 次のようなNumPy配列を作成するコードとして、正しいものを選択してください（1つ選択）。

```
array([[1., 1.],
       [1., 1.]])
```

A. `np.ones(2)`

B. `np.full((2, 2), 1.0)`

C. `np.all((2, 2), 1)`

D. `np.eye(2)`

18. 2×2のNumPy配列 `a` 、 `b` があるとき、 `a` と `b` の要素同士の掛け算をするコードとして正しいものを選択してください（1つ選択）。

A. `a @ b`

B. `a * b`

C. `np.matmul(a, b)`

D. `np.dot(a, b)`

19. NumPyの機能の説明として、正しいものを選択してください（1つ選択）。
なお、選択肢内の `np` は `import numpy as np` でインポートしたNumPyモジュールを指します。

A. NumPyのユニバーサルファンクションとは、形状の異なるデータ同士の演算を可能にする機能のことである

B. ユニバーサルファンクションには、`np.hstack()` や `np.hsplit()` がある

C. ユニバーサルファンクションを使うと、`for` 文が必要となるようなコードが1行で書けるようになるため、コードが簡潔になる

D. NumPy配列の `a` があるとき、`np.sum(a)` と `a.sum()` は異なる結果になる

20. 次のようなDataFrame `df` があるときに、`6` が格納されている要素を参照する方法として正しいものを選択してください（1つ選択）。

コード:

```
import pandas as pd

df = pd.DataFrame([[1, 2, 3], [4, 5, 6]], index=["a", "b"])
```

A. `df.loc[2, "b"]`

B. `df.loc["b", -1]`

C. `df.iloc[-1, -1]`

D. `df.iloc[2, 1]`

21. pandasのファイル入出力機能の説明として、正しいものを選択してください（1つ選択）。
なお、選択肢内の `pd` は `import pandas as pd` でインポートしたpandasモジュールを指します。

A. `pd.read_html()` を使うと、HTML内に複数の表があった場合でも読み込みができる

B. `pd.read_csv()` は、ファイルの文字コードを自動で判別するため、エンコーディングの指定は不要である

C. `df.serialize()` と `pd.deserialize()` を使うと、DataFrameを直列化したファイルの読み書きができる

D. `pd.read_xlsx()` と `df.to_xlsx()` を使うと、Excelファイルの読み書きができる

22. 次のDataFrame `df` について、 `日付` が小さい順で並べ替えた後、`日付` カラムをインデックスに設定したいです。コードとして、正しいものを選択してください（1つ選択）。

DataFrame `df`:

	日付	売上
0	2024/01/04	98
1	2024/01/05	102
2	2024/01/01	105
3	2024/01/02	140
4	2024/01/03	200

A.

```
df = df.sort_values(by='日付', ascending=True)
df.index = '日付'
```

B.

```
df = df.sort_values(by='日付', ascending=True)
df = df.set_index('日付')
```

C.

```
df = df.sort_values(by='日付', ascending=False)
df.index = '日付'
```

D.

```
df = df.sort_values(by='日付', ascending=False)
df = df.set_index('日付')
```

23. 2024年1月1日から2024年1月31日までの、31日分の日付データを作るコードとして、正しいものを選択してください（1つ選択）。
なお、`import pandas as pd` が既に実行されているものとします。

A. `pd.date_range(start="2024-01-01", freq="1M")`

B. `pd.date_range(start="2024-01-01", periods=31)`

C. `pd.date_range(start="2024-01-01", days=31)`

D. `pd.date_range(start="2024-01-01", end="2024-02-01")`

24. 次のデータがあるとき、欠損値に関する説明として正しいものを選択してください（1つ選択）。

コード:

```
import numpy as np
import pandas as pd

df = pd.DataFrame({"A": [0, "Python", "nan", "nan"],
                   "B": [np.nan, "", np.nan, "<nan>"]})
```

A. `df.dropna()` とすると、欠損値を含むカラムを削除できる

B. `df.fill(10)` とすると、欠損値を10で置換できる

C. `df.isnull()` とすると、DataFrame内に1つ以上の欠損値が含まれるかどうか判定できる

D. このDataFrame内に、欠損値の要素は2個ある

25. 2つのDataFrame（`df_1` と `df_2`）があります。これらを行方向に連結するコードとして、正しいものを選択してください（1つ選択）。
なお、`import pandas as pd` が既に実行されているものとします。

A. `pd.concat(df_1, df_2, axis=0)`

B. `pd.concat([df_1, df_2], axis=0)`

C. `pd.concat(df_1, df_2, axis=1)`

D. `pd.concat([df_1, df_2], axis=1)`

26. データの傾向を見るpandasの機能の説明として、正しいものを選択してください（1つ選択）。

A. `df.loc[:, "部活"].common()` で、`部活` カラムの最頻値を確認できる

B. `df.corr()` を使うと、カラム間のコサイン類似度を計算できる

C. `pd.scatter_matrix()` を使うと、各カラムのヒストグラムとカラム間の散布図のグラフが表示される

D. `df.describe()` を使うと、データ個数や平均値、最小値、最大値、四分位数などの統計量をまとめて確認できる

27. Matplotlibを使ったグラフ描画の説明として、正しいものを選択してください（1つ選択）。

A. pyplotインタフェースでは、サブプロットを作成するために `plt.subplots()` を使用する

B. オブジェクト指向インタフェースでは、描画オブジェクトの `plot()` メソッドを実行して折れ線グラフを描画する

C. オブジェクト指向インタフェースでは、1つの描画オブジェクトに複数のサブプロットを指定できる

D. 描画オブジェクトにタイトルを指定するには、描画オブジェクトの `set_title()` メソッドを使う

28. Matplotlibのサブプロットの説明として、正しいものを選択してください（1つ選択）。
なお、事前に `import matplotlib.pyplot as plt` が実行されているものとします。

A. `plt.subplots(2)` とすると、2つのサブプロットが1行2列で配置される

B. `plt.subplots(2, 3)` とすると、6つのサブプロットが2行3列で配置される

C. `plt.subplots(size=(2, 2))` とすると、4つのサブプロットが2行2列で配置される

D. `plt.subplots(6, ncols=2)` とすると、6つのサブロットが3行2列で配置される

29. 次のような描画オブジェクトとサブプロット、データがあるとき、ヒストグラムを描くコードの説明として正しいものを選択してください（1つ選択）。

コード:

```
import numpy as np
import matplotlib.pyplot as plt

# 正規分布に従う乱数1000個のデータを2組作成
rng = np.random.default_rng(123)
x0 = rng.normal(100, 25, 1000)
x1 = rng.normal(100, 10, 1000)

fig, ax = plt.subplots()
```

A. `ax.hist(x0, bins=10)` で、ビンの幅が10のヒストグラムを描画できる

B. `ax.histh(x0)` で、横向きのヒストグラムを描画できる

C. `ax.hist((x0, x1), stacked=True)` で、`x0` と `x1` を積み上げたヒストグラムを描画できる

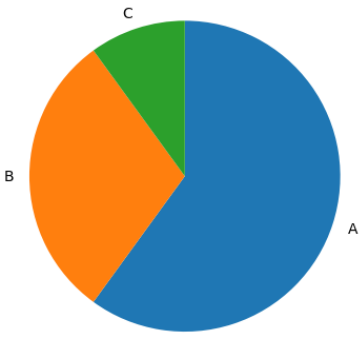
D. `ax.hist(x0)` 実行後に `ax.hist(x1)` を実行すると、1つのサブプロット内に2つのヒストグラムを重ねずに並べて表示できる

30. 次のコードの実行結果として、正しいものを選択してください（1つ選択）。

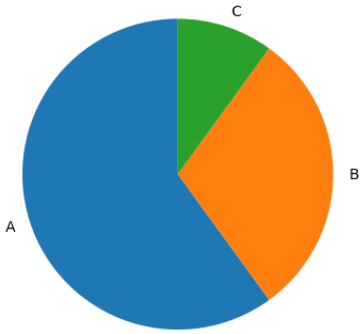
コード:

```
import numpy as np
import matplotlib.pyplot as plt

labels = ["A", "B", "C"]
x =  [20, 10, 3]

fig, ax = plt.subplots()
ax.pie(x, labels=labels, startangle=0)
plt.show()
```

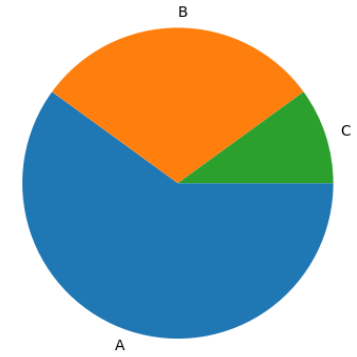
A.



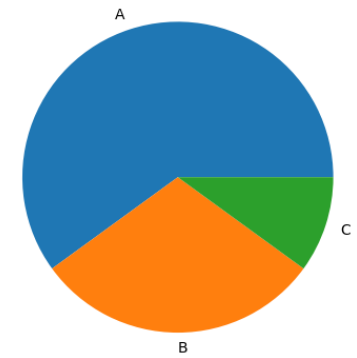
B.



C.



D.



31. Matplotlibにおけるスタイル設定の説明として、正しいものを選択してください（1つ選択）。

A. サブプロットの `set_xlabel` メソッドなどで、`fontdict` 引数と個別の引数（`size`引数など）両方でスタイルを指定した場合、個別の引数で指定した設定が優先される

B. `matplotlib.style.set` 関数を使うと、グラフの表示スタイル全体（線の色、太さ、背景色など）にあらかじめ用意されている設定を適用できる

C. サブプロットの `set_text` メソッドを使うと、サブプロット内の任意の位置にテキストを設定できる

D. 折れ線グラフの線の色を設定する方法のひとつとして、`color` 引数に0-255の範囲でRGBAの値を指定できる（`(255, 255, 0, 50)`など）

32. 次のようなDataFrame `df` があるとき、「期待する結果」のようなグラフを描画するコードとして正しいものを選択してください（1つ選択）。

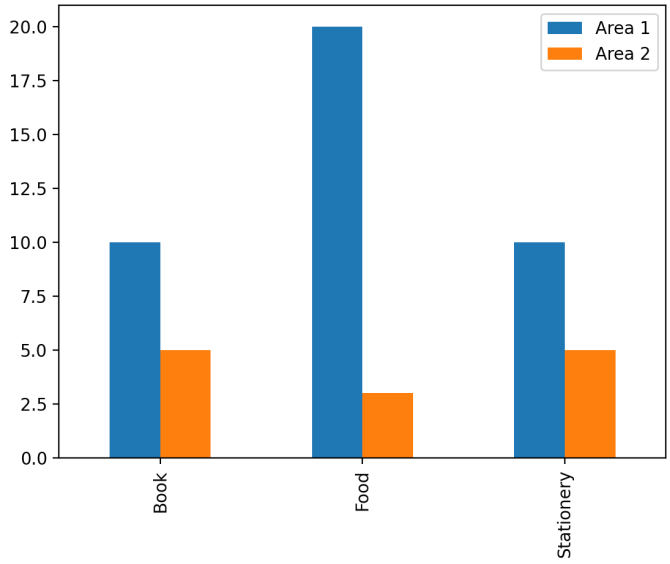
DataFrame `df`:

	Area 1	Area 2
Book	10	5
Food	20	3
Stationery	10	5

コード:

```
import pandas as pd

df = pd.DataFrame({"Area 1": [10, 20, 10], "Area 2": [5, 3, 5]}, index=
["Book", "Food", "Stationery"])
```

期待する結果:



A. `df.bar()`

B. `df.plot.bar()`

C. `df.T.bar()`

D. `df.T.plot.bar()`

33. 機械学習の前処理の説明として、正しいものを選択してください（1つ選択）。

A. One-hotエンコーディングは、a→0、b→1、c→2のようにカテゴリ変数を整数に変換する

B. ダミー変数化を行うと、K種類の値が入力された列がK+1個の列に展開される

C. 最小最大正規化とは、特徴量の最小値が0、最大値が1をとるようにデータの分布を変換する処理のことで、標準化とも呼ばれる

D. 分散正規化とは、特徴量の平均が0、標準偏差が1となるように特徴量を変換する処理のことで、z変換とも呼ばれる

34. 分類モデルに関するデータセット分割の説明として、正しいものを選択してください（1つ選択）。

A. scikit-learnのインタフェースでは、学習は `train` メソッド、予測は `predict` メソッドを用いて行う

B. 学習データセットとテストデータセットの分割を1回行った後、同じデータセットでモデルの構築と評価を複数回行う方法のことを、交差検証と呼ぶ

C. 学習に使ったデータセットでは、汎化能力を評価できない

D. 層化k分割交差検証では、すべての目的変数（クラスラベル）が同じ件数になるようデータが分割される

35. 分類に関するアルゴリズムの説明として、正しいものを選択してください（1つ選択）。

A. サポートベクタマシンは、分類だけではなく回帰や外れ値検出にも使えるアルゴリズムである

B. サポートベクタマシンは線形分離できるデータの分類が得意な一方で、線形分離できないデータの分類は苦手である

C. 複数の学習器を用いた学習方法のことをアンサンブル学習と呼び、該当する手法には決定木などがある

D. データセットを高次元に変換したものを、ブートストラップデータと呼ぶ

36. 回帰に関する説明として、正しいものを選択してください（1つ選択）。

A. 回帰は教師あり学習の一種であり、目的変数が離散値になるようなタスクで使われる

B. 目的変数が、各説明変数の1次式の和で表せる回帰のことを単回帰、2次式の和で表せる回帰のことを重回帰と呼ぶ

C. 回帰モデルを定量的に評価する指標として、F値がある

D. 横軸を予測値、縦軸を実測値とする散布図をプロットしたとき、y=xの直線に近づくほど回帰モデルの性能が良いと言える

37. 次元削減に関する説明として、正しいものを選択してください（1つ選択）。

A. 特徴量の種類が数千を超えるような教師あり学習において、計算量がネックで学習が進まないとき、次元削減が効果的なことがある

B. 次元削減とは教師なし学習のひとつであり、ブラックボックス的な環境の中で行動するエージェントが、得られる報酬を最大化するように次元を削減する手法である

C. 次元削減では、元のデータセットのすべての情報を保持したまま数個〜数十個の新しい特徴量を生成する

D. 主成分分析は、データのばらつきを最小に保ちながら（高次元のデータに対して分散が小さくなる方向を探して）、元の次元と同じかそれよりも低次元にデータを変換する手法である

38. 「ある病気に関する検査データから、陽性か陰性か判断する」という分類タスクについて考えます。

「陽性である」を正例とした場合、評価指標の説明として正しいものを選択してください（1つ選択）。

A. ROC曲線が `(0, 0)` と `(1, 1)` を結ぶ対角線に近くなるほど、分類精度が悪い

B. 正解率は適合率と再現率の調和平均であり、両方の指標のバランスの良さを考慮して評価を行う

C. 真に陽性であるケースを見逃さないことを重視する場合は、適合率に着目して評価を行う

D. AUCの値が1に近づくほど、モデルが正例と負例を区別する能力がないことを意味する

39. 機械学習のハイパーパラメータの最適化の説明として、正しいものを選択してください（1つ選択）。

A. グリッドサーチではパラメータの探索処理が進むごとに精度が改善するため、最初に作られたモデルより最後に作られたモデルの方が精度が良い傾向にある

B. ハイパーパラメータには欠損値を補完する方法やデータセットの分割比率などがあり、学習とは別にユーザが値を指定する必要がある

C. scikit-learnの `GridSearchCV` クラスの `param_grid` 引数には、ハイパーパラメータ名とその候補値のリストを対応づけた辞書を指定する

D. ハイパーパラメータの最適化を行うことで、モデルの学習時間が短くなり、計算コストが削減される

40. クラスタリングに関する説明として、誤っているものを選択してください（1つ選択）。

A.「音楽のストリーミングサービスで、ユーザーの利用履歴、利用時間、評価、音楽ジャンルなどのデータを使って、ユーザーをグループ分けし、どのような利用パターンがあるのかを分析する」といったタスクは、クラスタリングになる

B. 凝集型クラスタリングとは、小さなクラスタの状態から開始して、似ているデータ同士を順次まとめていくアプローチであり、k-meansもこの手法のひとつである

C. 階層型クラスタリングの結果はデンドログラムで可視化できるが、k-meansの結果ではできない

D. scikit-learnの `AgglomerativeClustering` クラスでは、クラスタ数を指定可能である