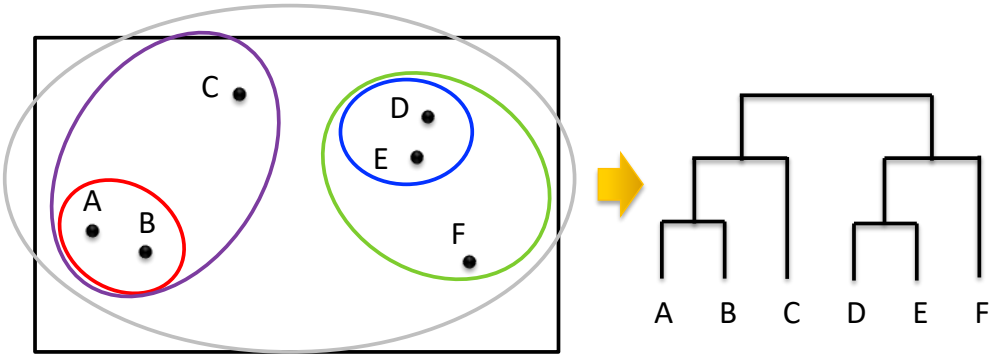
第10回データの集計と可視化3

SciPyを用いたクラスタリング

目次

事例:クラスタリング

- データの集合を、クラスタと呼ぶ部分集合に分ける処理をクラスタリングやクラスター分析と呼びます。
- クラスタは、データ同士がどれだけ似ているかを距離という指標をもとに評価して分割します。
- クラスタリングには、階層的手法と非階層的手法があります。
- 階層的手法
  - もっとも似ているデータから順番にクラスタにまとめていく方法で、途中経過が階層的に表すことができ、最終的に樹形図(デンドログラム)というトーナメント表のような図を得ることができます。



データセット

- 都道府県を宿泊施設の種類が似ているかでクラスタリングしてみましょう。
  - 宿泊施設数の多い20都道府県について
  - 種類
    - 旅館
    - リゾートホテル
    - ビジネスホテル
    - シティホテル
    - 簡易宿所
    - 会社、団体の宿泊所
- 準備
  - Spyderを起動し、[新規]-[新規ファイル]を作成します。
  - 今回のデータは、python.xlsxの「宿泊施設数」シートになります。

	A	B	C	D	E	F	G
1	都道府県	旅館	リゾート	ビジネス	シティホ	簡易宿所	会社・団体
2	東京都	1668200	367810	24030850	11909250	1713830	416380
3	北海道	3638330	2213410	9971510	4252600	420410	88770
4	大阪府	269970	1189350	11740500	6217540	683610	119020
5	愛知県	583160	279110	8524180	2467780	339580	116920
6	神奈川県	971920	1136460	6519380	2378860	721880	504520

データの取り込みと確認

階層的クラスタリングに必要なメソッドをインポート

```
import pandas as pd
import numpy as np
from scipy.spatial.distance import pdist
from scipy.cluster.hierarchy import linkage, dendrogram
import matplotlib.pyplot as plt

plt.rcParams["font.family"] = "IPAexGothic"
plt.rcParams["font.size"] = 15

file = pd.ExcelFile("C:/Users/user/Downloads/python.xlsx")
data = file.parse("宿泊施設数")
print(data)
```

userというユーザのDownloadsフォルダに保存されていると仮定（userは環境に合わせて書き換えてください）

	都道府県	旅館	リゾートホテル	ビジネスホテル	シティホテル	簡易宿所
	会社・団体の宿泊所					
0	東京都	1668200	367810	24030850	11909250	1713830	416380
1	北海道	3638330	2213410	9971510	4252600	420410	88770
2	大阪府	269970	1189350	11740500	6217540	683610	119020
3	愛知県	583160	279110	8524180	2467780	339580	116920

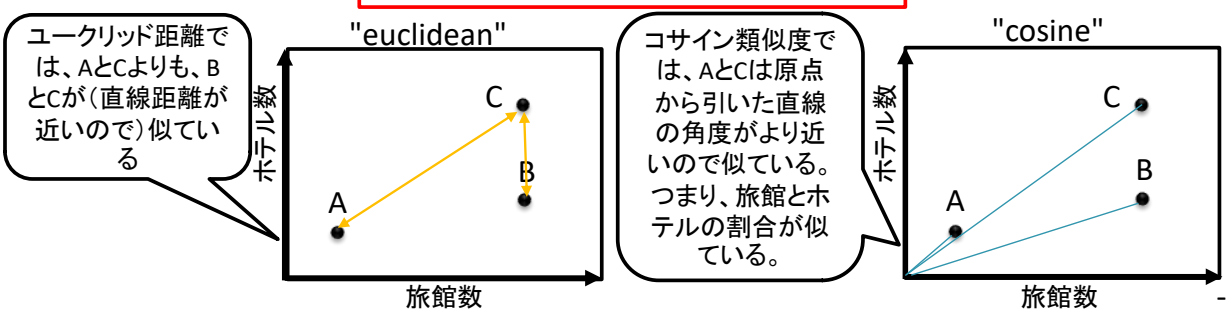
データ間の距離計算

- まず、データ間の距離をpdistメソッドで計算します。
  - 都道府県名の列を除いた数値だけをpdistメソッドに渡します。
  - また、距離の計算方法を指定します。"euclidean"は、ユークリッド距離で計算するように指示しています。
    - "euclidean"：ユークリッド距離を指定。データ同士を結んだ直線の距離のこと。
    - "cosine"：コサイン類似度を指定。データに含まれる項目の割合がどれだけ似ているか。

```
dist = pdist(data.iloc[:, 1:], "euclidean")
```

	都道府県	旅館	リゾートホテル	ビジネスホテル	シティホテル	簡易宿所
	会社・団体の宿泊所					
0	東京都	1668200	367810	24030850	11909250	1713830	416380
1	北海道	3638330	2213410	9971510	4252600	420410	88770
2	大阪府	269970	1189350	11740500	6217540	683610	119020
3	愛知県	583160	279110	8524180	2467780	339580	116920

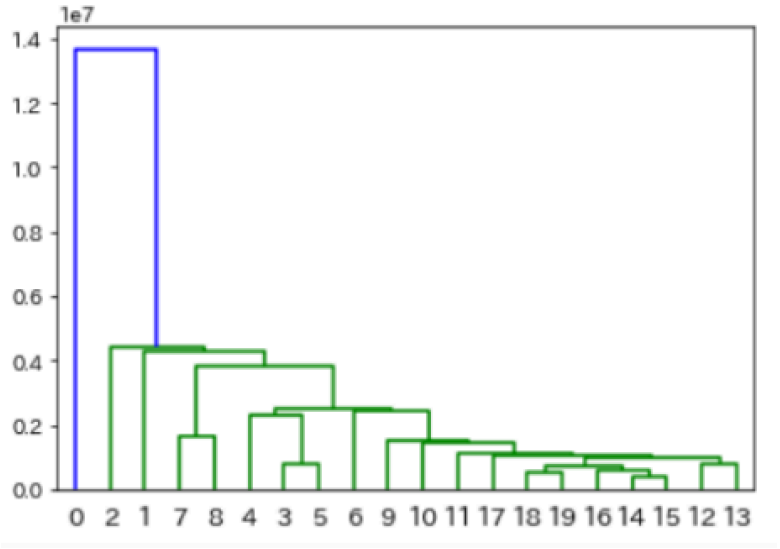
data.iloc[:,1:]



階層的クラスタリングを可視化

- linkageメソッドで階層的クラスタリングを実行し、dendrogramメソッドで樹形図として可視化します。

```
result = linkage(dist)
dendrogram(result)
```



樹形図にラベルを表示

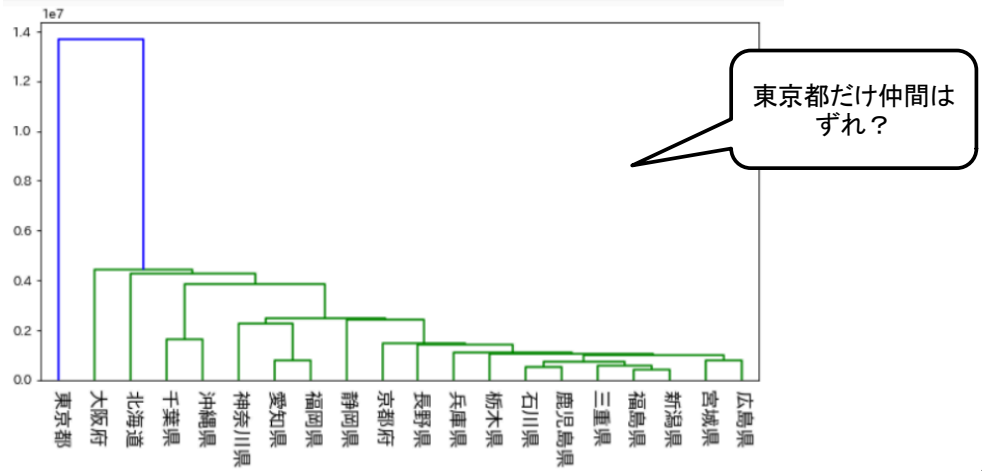
- 図のサイズをより大きくし、ラベルを表示しましょう。

dataの１列目（都道府県名）をtolistメソッドでリストに変換してラベルに設定

```
result = linkage(dist)
plt.subplots(figsize=(10,5))
dendrogram(result, labels=data.iloc[:, 0].tolist(), leaf_rotation=-90,
leaf_font_size=15)
plt.show()
```

フォントサイズ15を指定

ラベルを-90度回転



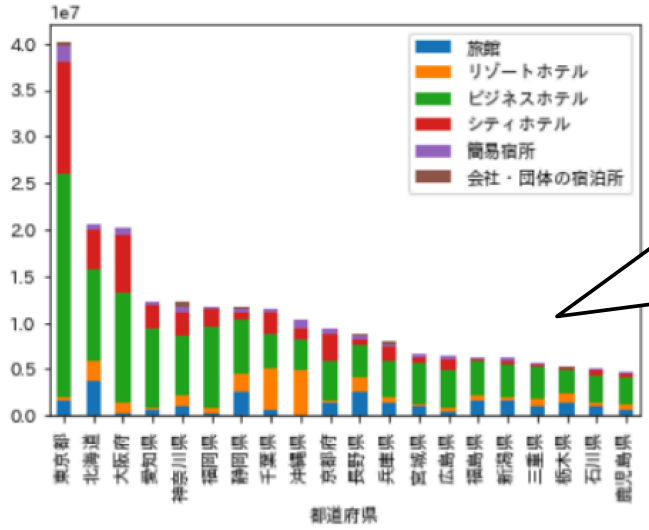
東京都だけ仲間はずれ？

宿泊施設の割合を棒グラフで確認

```
data.loc[:, "旅館":].plot.bar(stacked=True)
plt.show()
```

旅館から右のデータで積み上げ棒グラフを描く

	都道府県	旅館	リゾートホテル	ビジネスホテル	シティホテル	簡易宿所
0	東京都	1668200	367810	24030850	11909250	1713830
1	北海道	3638330	2213410	9971510	4252600	420410
2	大阪府	269970	1189350	11740500	6217540	683610



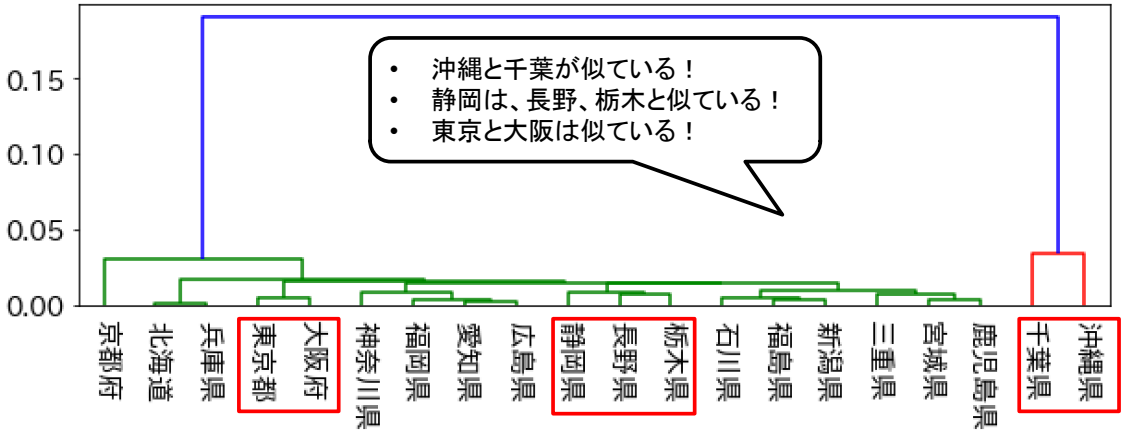
東京都だけ宿泊施設数が圧倒的に多い！
ユークリッド距離では、宿泊施設の絶対数がクラスタ分類にかなり影響している。

距離計算をコサイン類似度に変更

- 宿泊施設の絶対数ではなく、構成比が似ているかでクラスタリングしたい！

```
dist = pdist(data.iloc[:, 1:], "cosine")

result = linkage(dist)
plt.subplots(figsize=(10,5))
dendrogram(result, labels=data.iloc[:, 0].tolist(), leaf_rotation=-90,
leaf_font_size=15)
plt.show()
```



- 沖縄と千葉が似ている！
- 静岡は、長野、栃木と似ている！
- 東京と大阪は似ている！

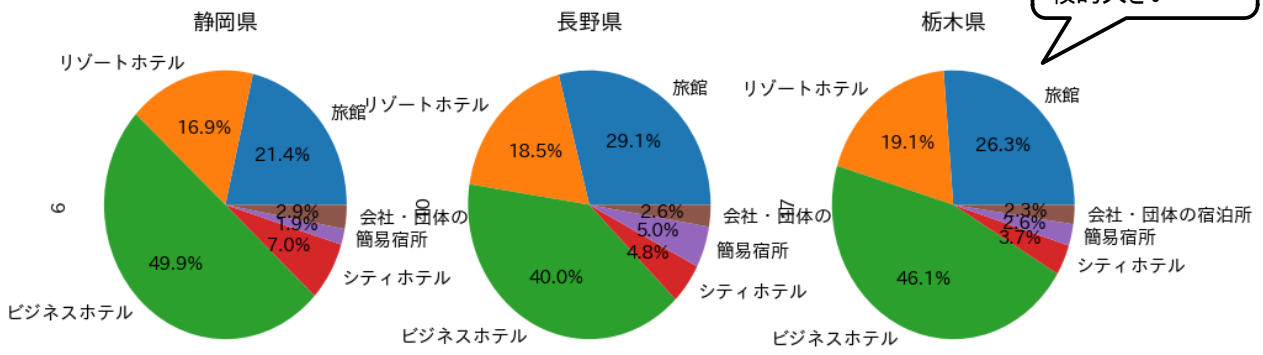
円グラフで構成比を確認

- 静岡県、長野県、栃木県の宿泊施設の構成比を描いて確認してみましょう。

都道府県が静岡県の行を取り出し、その旅館より右側を取り出し、その0行目を取り出し、plotしている

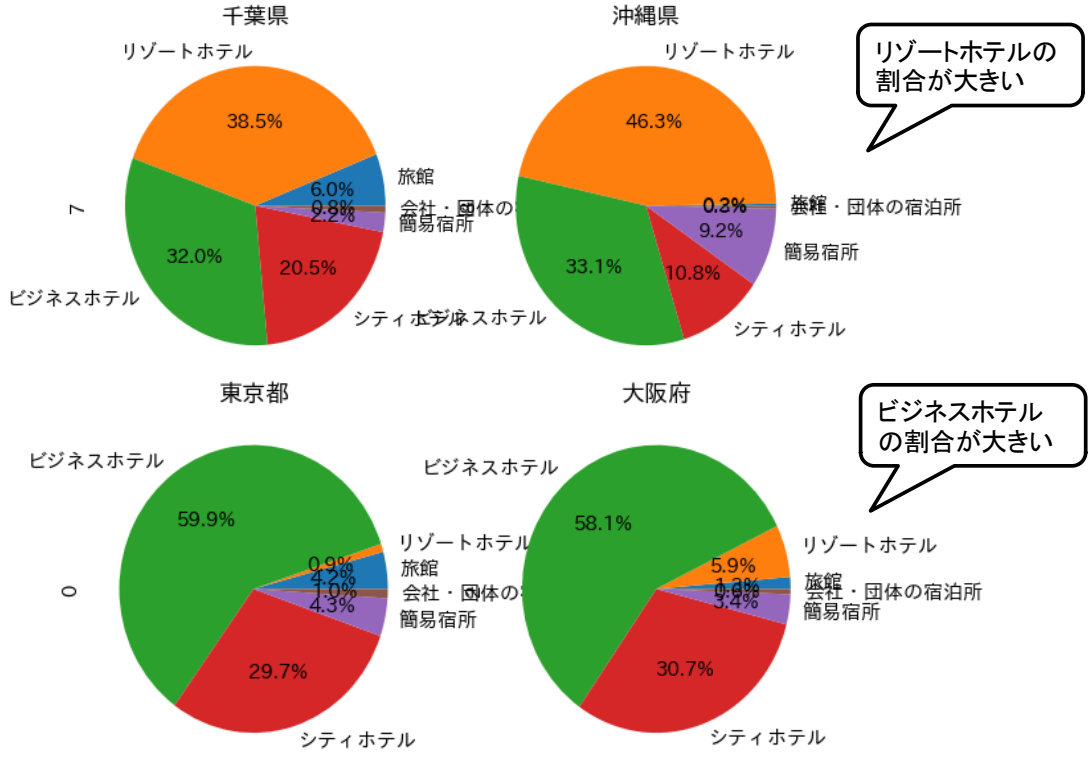
```
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15, 5))
data[data.都道府県=="静岡県"].loc[:, "旅館":].iloc[0].plot.pie(ax=axes[0],
autopct="%.1f%%", title="静岡県")
data[data.都道府県=="長野県"].loc[:, "旅館":].iloc[0].plot.pie(ax=axes[1],
autopct="%.1f%%", title="長野県")
data[data.都道府県=="栃木県"].loc[:, "旅館":].iloc[0].plot.pie(ax=axes[2],
autopct="%.1f%%", title="栃木県")
plt.show()
```

旅館の割合が比較的大きい



課題10

- 千葉県と沖縄県、東京都と大阪府の円グラフも描いてください。



リゾートホテルの割合が大きい

ビジネスホテルの割合が大きい

まとめ

- 本講座では、Pythonの開発環境、基本文法、簡単なデータの集計と可視化について学びました。
- Pythonには、今回紹介した以外にも膨大なモジュールが用意されています。
- あまりにも膨大なため、その全貌をすべて理解するのは不可能に近いでしょう。
- まずは、身近な・興味のある使い道から、学びを拡大していってください。