

Group-Assignment: EDA or Pre-processing Techniques(s)

PARALLEL COORDINATES PLOT

Presented by

ชญานนท์ มานะกิจจานนท์ 6510503298

ศุภกิตต์ ววศ์โต 6510503816

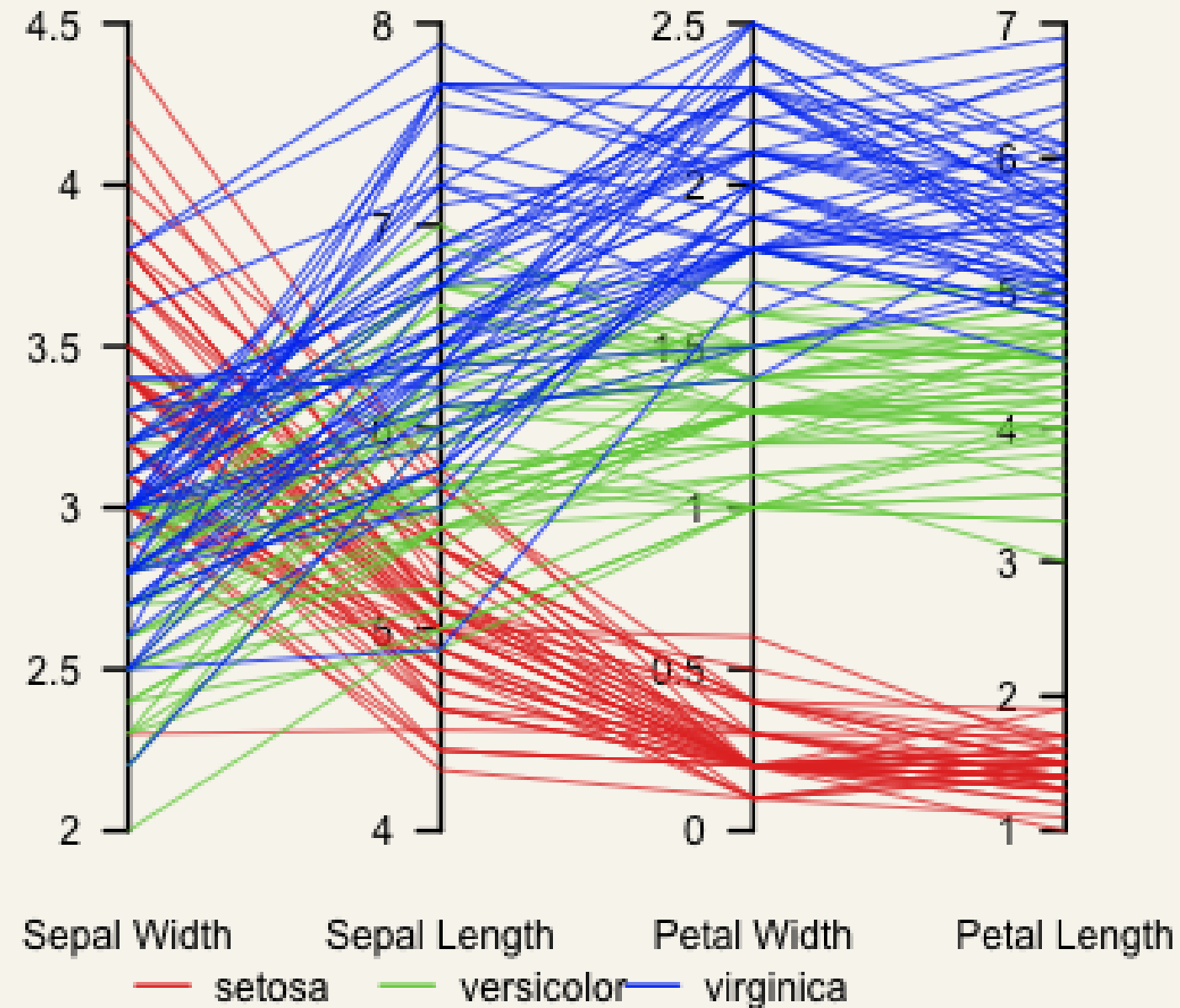
อธิบายเทคนิค

Parallel Coordinate Plot คือวิธีการ plot graph สำหรับ dataset ที่ข้อมูลมีหลายมิติ ที่จะแสดงข้อมูลที่เราสนใจออกมาเป็นเส้นที่ลากผ่านหลายๆ แกน โดยที่หนึ่งเส้นจะแทนข้อมูลหนึ่งตัว และแต่ละแกนจะแทนตัวแปรหนึ่งตัวที่เราสนใจ ทำให้สามารถมองเห็นภาพความสัมพันธ์ของข้อมูลกับหลายๆตัวแปรที่เราสนใจใน dataset ได้ง่ายยิ่งขึ้น

โดยวิธีการนี้จะมีความคล้ายคลึงกับการทำ line chart แต่จะเป็นการลากเส้นผ่านหลายๆแกนและแสดงผลออกมาพร้อมกัน

ตัวอย่าง PARALLEL COORDINATES PLOT

Parallel coordinate plot, Fisher's Iris data



DATASET

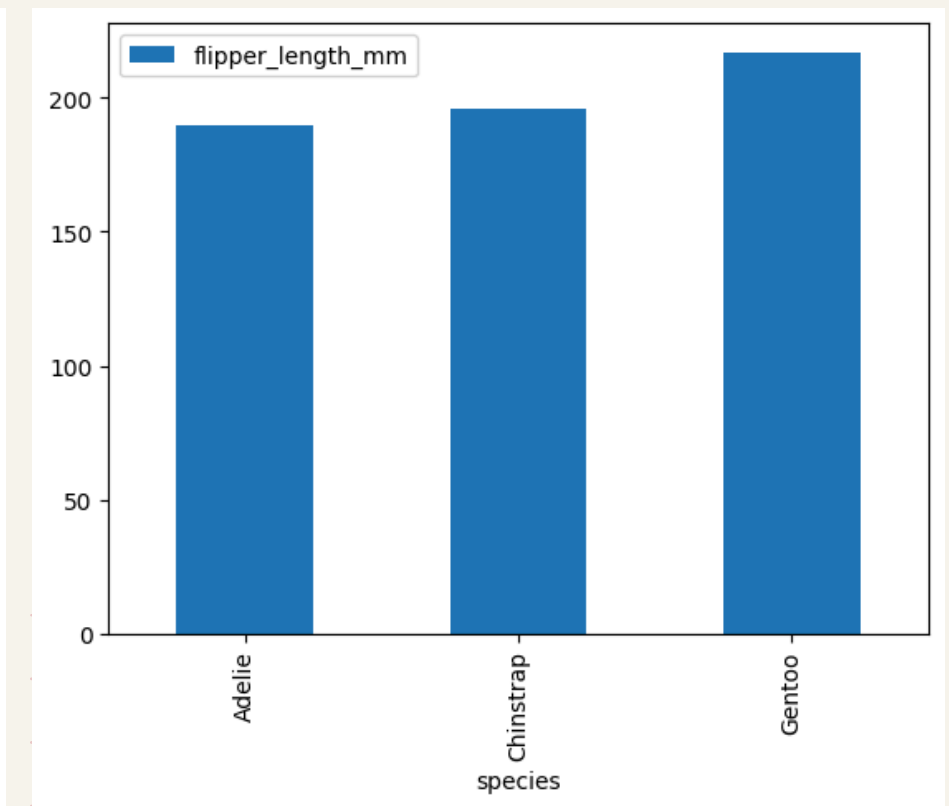
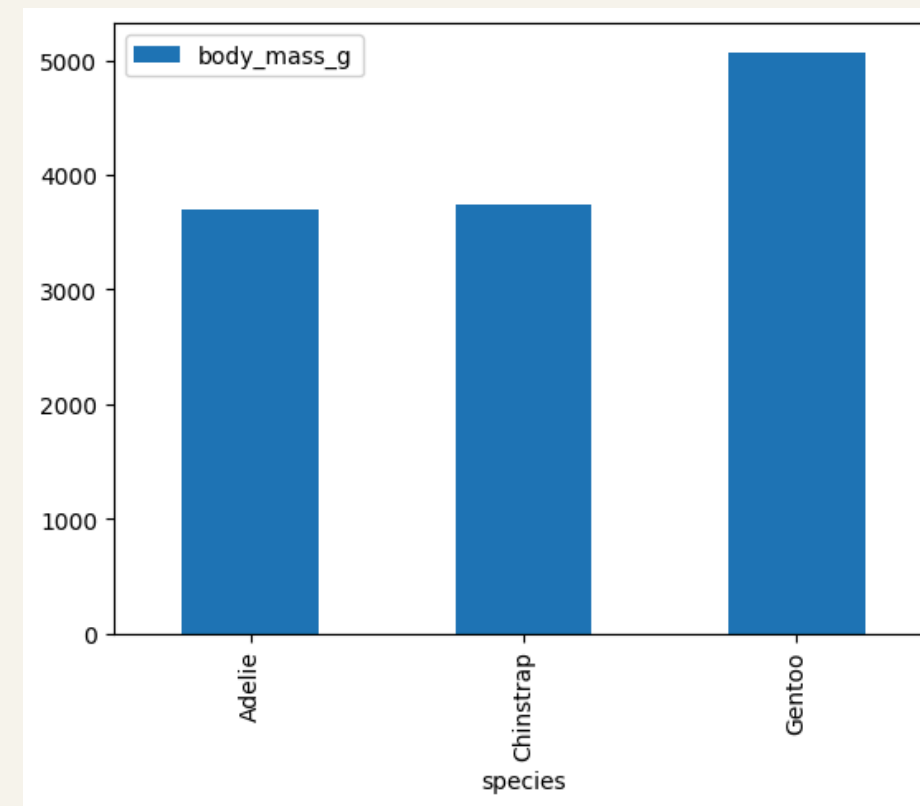
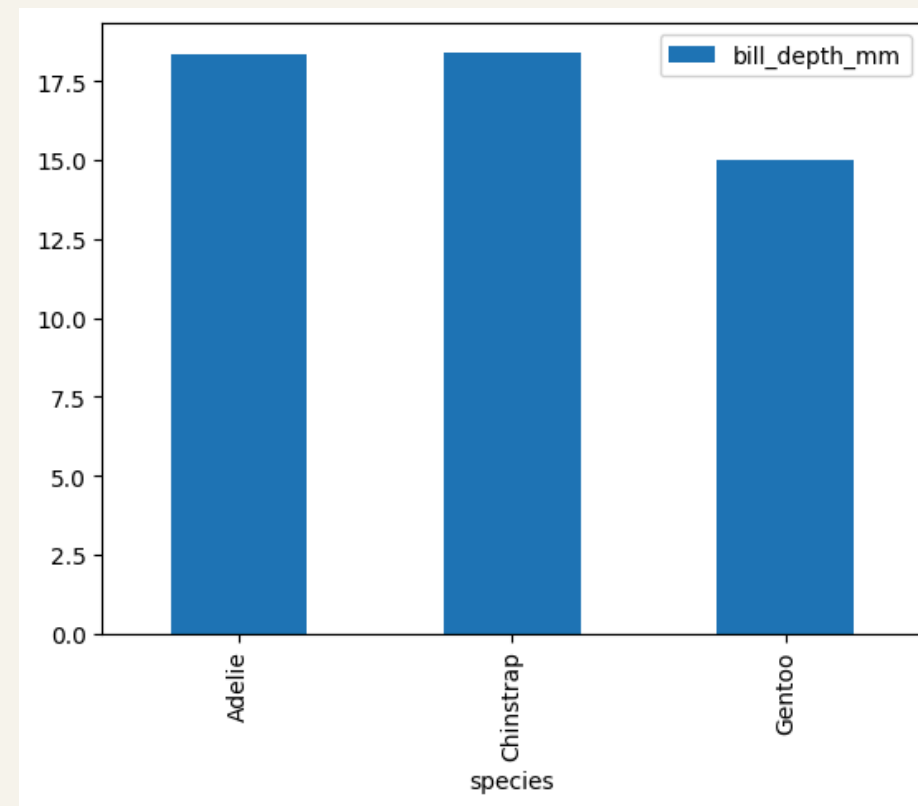
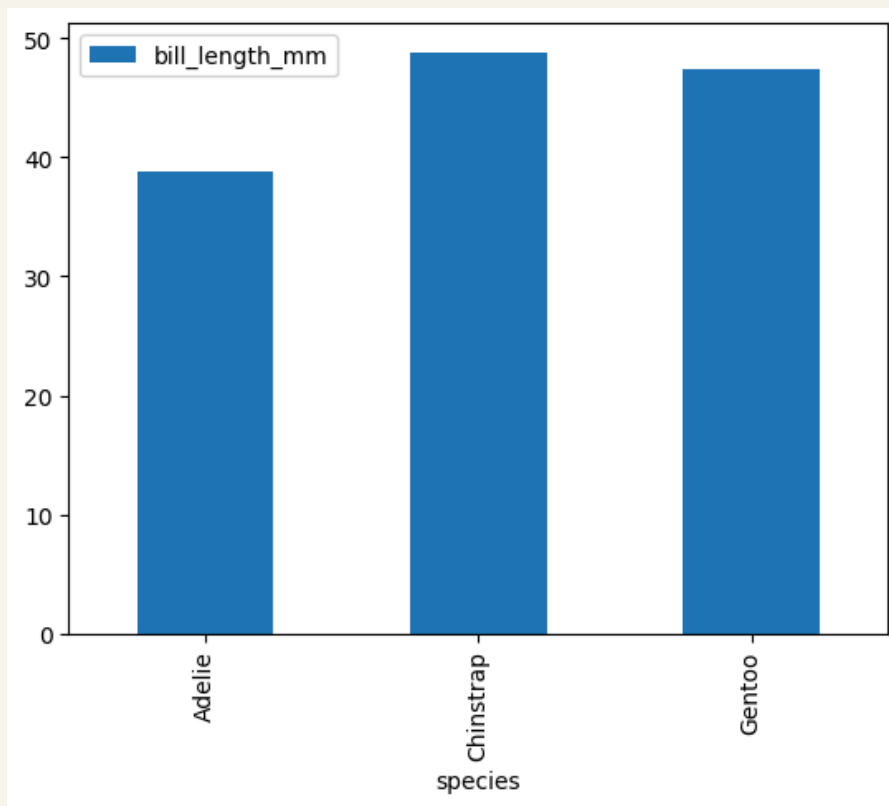


penguins.csv มีตัวแปรต่างๆ ซึ่งเกี่ยวข้องกับข้อมูลนกเพนกวิน ดังนี้

- species ขอนกเพนกวิน ซึ่งมีทั้งหมด 3 สปีชีส์ คือ Adelie, Gentoo และ Chinstrap
- island หรือเกาะที่อาศัย ซึ่งมีทั้งหมด 3 ตัวแปร คือ Biscoe, Dream และ Torgersen
- bill_length_mm คือความยาวของปากนก ในหน่วยมิลลิเมตร
- bill_depth_mm คือความกว้างของปากนก ในหน่วยมิลลิเมตร
- flipper_length_mm คือความยาวของปีกนก ในหน่วยมิลลิเมตร
- body_mass_g คือน้ำหนักของนก ในหน่วยกรัม
- sex คือเพศของนกเพนกวิน

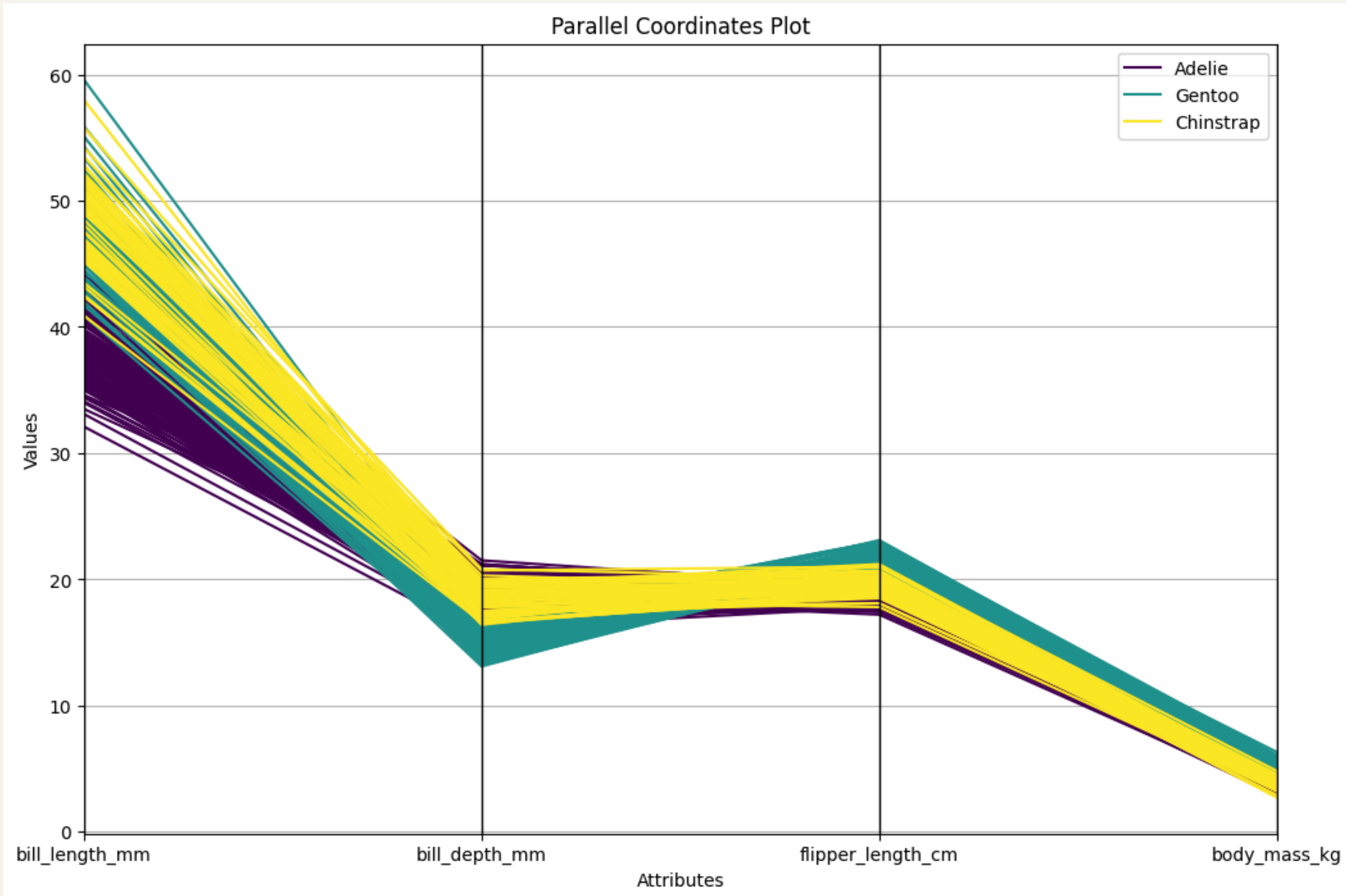
Parallel coordinates plot

DATASET(S) BEFORE PRE-PROCESSED



DATASET(S) AFTER PRE-PROCESSED

10



DEMO WITH TOOL OR LIBRARIES OR CODING

- สำหรับการทำ Parallel coordinates plot นั้นมี library อยู่ภายใน pandas สามารถนำมาใช้ได้เลย

วิธีการใช้ฟังก์ชัน `plotting.parallel_coordinates`

```
import pandas as pd
```

```
pd.plotting.parallel_coordinates(frame, class_column, cols=None, ax=None,  
color=None, use_columns=False, xticks=None, colormap=None,  
axvlines=True, axvlines_kwds=None, sort_labels=False)
```

DEMO WITH TOOL OR LIBRARIES OR CODING

● Import Library and dataset

▼ Importing the libraries

```
[ ] 1 import numpy as np
    2 import matplotlib.pyplot as plt
    3 import pandas as pd
    4 import seaborn as sns
```

▼ Importing the dataset

```
1 df = pd.read_csv('penguins.csv')
2 df.head(10)
```

| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---------|-----------|----------------|---------------|-------------------|-------------|--------|
| 0 | Adelie | Torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | male |
| 1 | Adelie | Torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | female |
| 2 | Adelie | Torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | female |
| 3 | Adelie | Torgersen | NaN | NaN | NaN | NaN | NaN |
| 4 | Adelie | Torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | female |
| 5 | Adelie | Torgersen | 39.3 | 20.6 | 190.0 | 3650.0 | male |

Parallel coordinates plot

DEMO WITH TOOL OR LIBRARIES OR CODING

● Checking for missing data and handling missing data

```
1 df.isnull().sum()
```

| | |
|-------------------|----|
| | 0 |
| species | 0 |
| island | 0 |
| bill_length_mm | 2 |
| bill_depth_mm | 2 |
| flipper_length_mm | 2 |
| body_mass_g | 2 |
| sex | 11 |

dtype: int64

```
[ ] 1 numerical_column = ['bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', 'body_mass_g']
2
3 for col in numerical_column:
4     df[col].fillna(df[col].mean(), inplace=True)
5
6 categorical_column = ['sex']
7
8 for col in categorical_column:
9     df[col].fillna(df[col].mode()[0], inplace=True)
```

Parallel coordinates plot

DEMO WITH TOOL OR LIBRARIES OR CODING

- Rescale some columns to make graph visualization better

```
1 df['body_mass_kg'] = df['body_mass_g']/1000
2 df['flipper_length_cm'] = df['flipper_length_mm']/10
3 df.head(10)
```

| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex | body_mass_kg | flipper_length_cm |
|---|---------|-----------|----------------|---------------|-------------------|-------------|--------|--------------|-------------------|
| 0 | Adelie | Torgersen | 39.10000 | 18.70000 | 181.000000 | 3750.000000 | male | 3.750000 | 18.10000 |
| 1 | Adelie | Torgersen | 39.50000 | 17.40000 | 186.000000 | 3800.000000 | female | 3.800000 | 18.60000 |
| 2 | Adelie | Torgersen | 40.30000 | 18.00000 | 195.000000 | 3250.000000 | female | 3.250000 | 19.50000 |
| 3 | Adelie | Torgersen | 43.92193 | 17.15117 | 200.915205 | 4201.754386 | male | 4.201754 | 20.09152 |
| 4 | Adelie | Torgersen | 36.70000 | 19.30000 | 193.000000 | 3450.000000 | female | 3.450000 | 19.30000 |
| 5 | Adelie | Torgersen | 39.30000 | 20.60000 | 190.000000 | 3650.000000 | male | 3.650000 | 19.00000 |
| 6 | Adelie | Torgersen | 38.90000 | 17.80000 | 181.000000 | 3625.000000 | female | 3.625000 | 18.10000 |
| 7 | Adelie | Torgersen | 39.20000 | 19.60000 | 195.000000 | 4675.000000 | male | 4.675000 | 19.50000 |
| 8 | Adelie | Torgersen | 34.10000 | 18.10000 | 193.000000 | 3475.000000 | male | 3.475000 | 19.30000 |
| 9 | Adelie | Torgersen | 42.00000 | 20.20000 | 190.000000 | 4250.000000 | male | 4.250000 | 19.00000 |

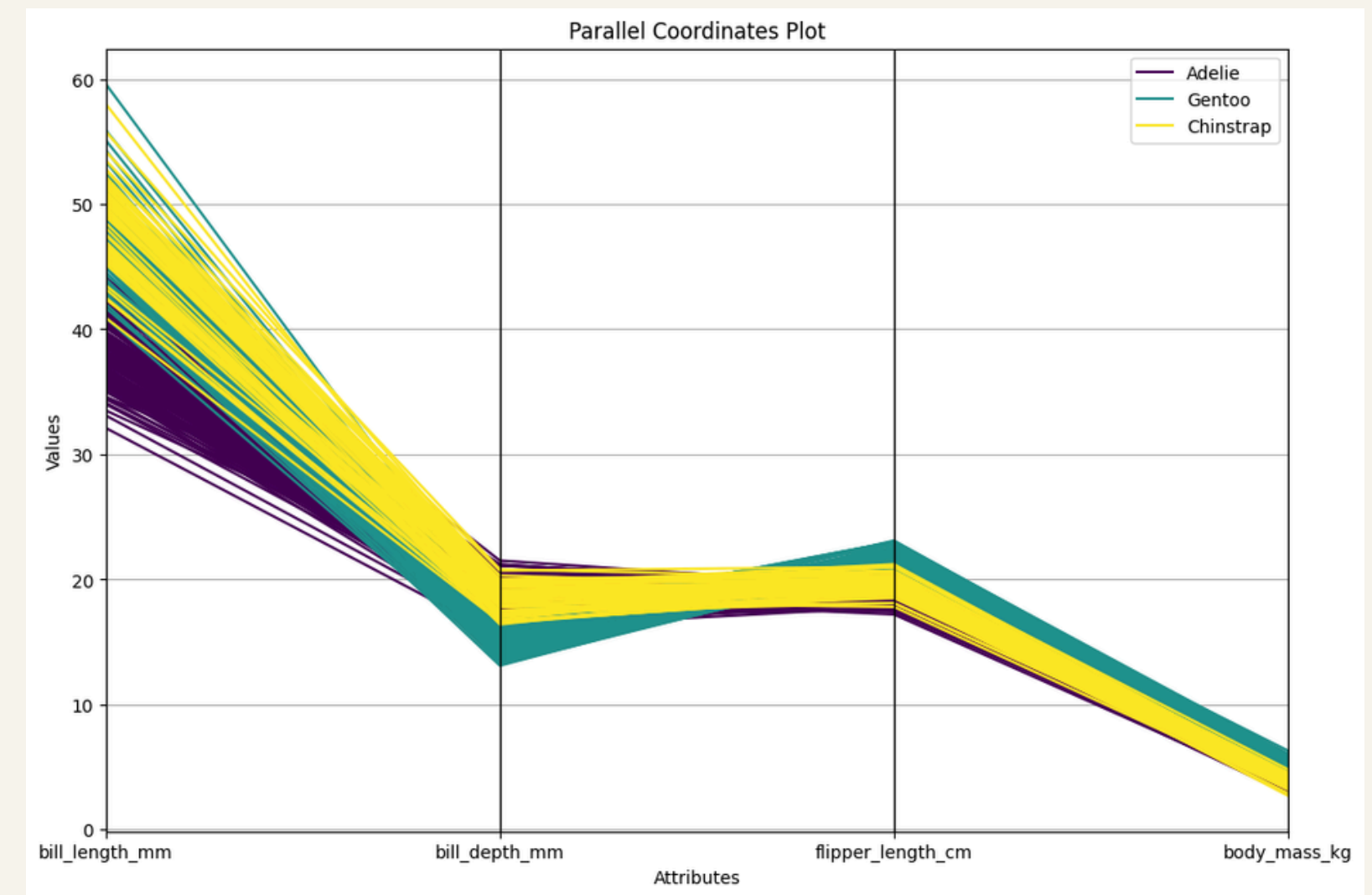
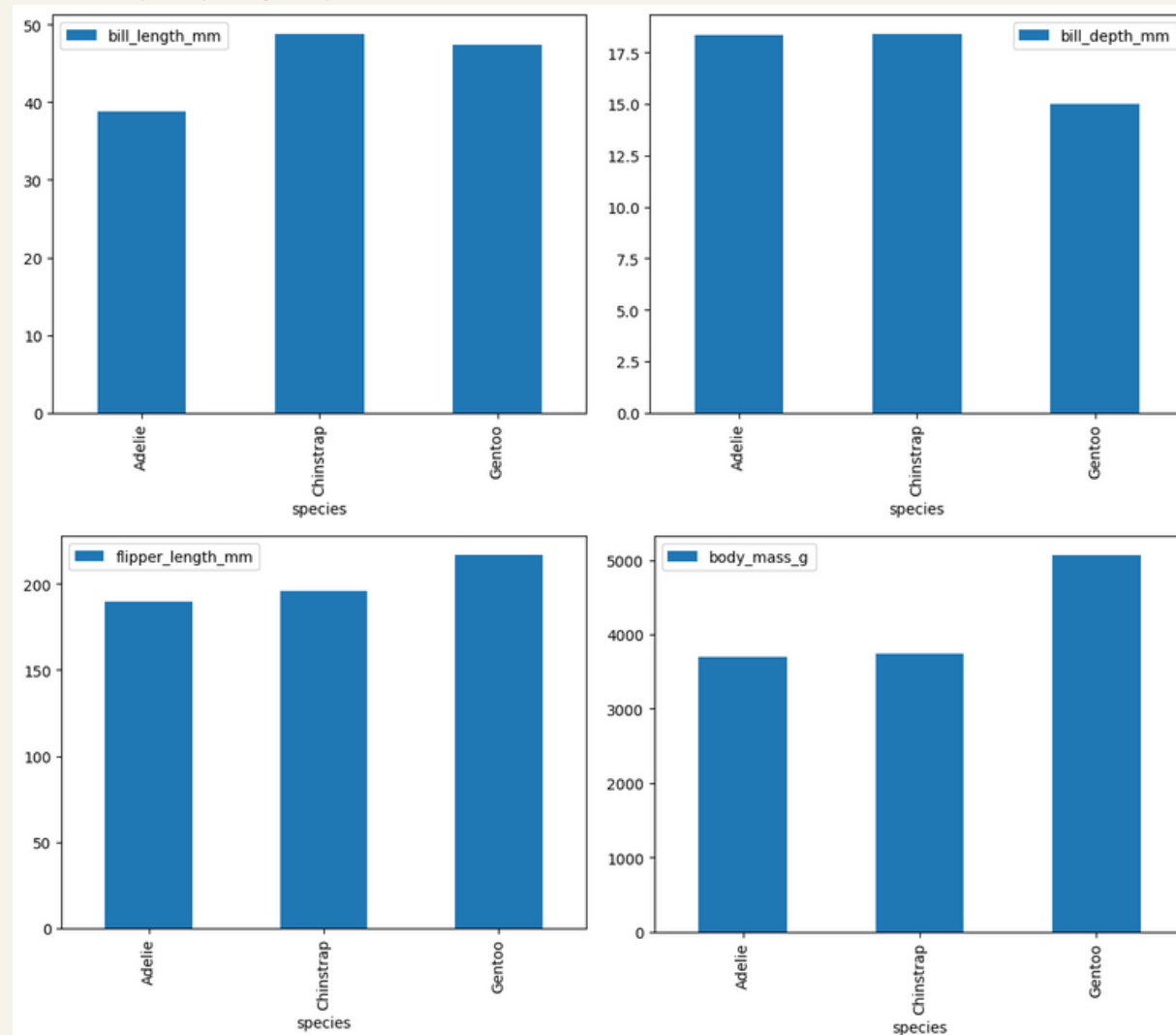
Parallel coordinates plot

DEMO WITH TOOL OR LIBRARIES OR CODING

● Plotting Parallel coordinates graph

```
1 from pandas.plotting import parallel_coordinates
2
3 columns_of_interest = ['bill_length_mm', 'bill_depth_mm', 'flipper_length_cm', 'body_mass_kg', 'species']
4
5 df_plot = df[columns_of_interest]
6
7 # Create a parallel coordinates plot
8 plt.figure(figsize=(12, 8))
9 parallel_coordinates(df_plot, class_column= "species", colormap='viridis',)
10 plt.title('Parallel Coordinates Plot')
11 plt.xlabel('Attributes')
12 plt.ylabel('Values')
13 plt.show()
```

EVALUATION



จาก Histogram ตอนแรก เราจะยังไม่เห็นความสัมพันธ์ของแต่ละตัวแปรอย่างแน่ชัด แต่เมื่อนำตัวแปรทั้ง 4 มาพลอตแบบ Parallel coordinates จะทำให้เห็นรูปแบบที่ชัดเจน ว่าเพนกวินแต่ละสายพันธุ์นั้นมีลักษณะเฉพาะตัวอย่างไร เช่น พันธุ์ Gentoo จะมีปากที่ค่อนข้างยาวและแคบ แต่มีปีกที่ยาวและน้ำหนักที่มาก เป็นต้น

Parallel coordinates plot

ADVANTAGES

12

สามารถใช้กับ Data ที่มีหลาย Dimension หรือหลายตัวแปรได้

เนื่องจากการพลอตที่สามารถแสดงหลายตัวแปรพร้อมๆ กันได้ ทำให้สามารถสังเกตความสัมพันธ์ระหว่างตัวแปร และเปรียบเทียบแต่ละตัวแปรได้ง่ายกว่าวิธีอื่นๆ

ช่วยแสดงถึงรูปแบบและการกระจุกตัวของกลุ่มข้อมูล

ช่วยให้เห็นข้อมูลในภาพรวม ทำให้สามารถสังเกตเห็นถึงแนวโน้ม รูปแบบต่างๆ ของข้อมูล เห็นว่าข้อมูลส่วนใดมีการกระจุกตัว และมีข้อมูลส่วนใดที่เป็น outlier

สามารถใช้ได้ทั้งกับ numerical และ categorical data

มีความยืดหยุ่น สามารถใช้กับทั้งข้อมูลทั้งสองประเภท แต่สำหรับ categorical data ต้อง encoding ก่อน ถึงจะสามารถนำมาพลอตได้

DISADVANTAGES

13

ข้อมูลที่มีปริมาณมหาศาล

เนื่องจากการพลอตที่สามารถแสดงหลายตัวแปรพร้อมกัน ทำให้ข้อมูลที่มีปริมาณมากและมีค่าซ้ำซ้อนเยอะจะทำให้สังเกตความสัมพันธ์ของกราฟได้ยากกว่า

ต้องทำการ Normalization

ข้อมูลที่จะนำมาพลอตอาจจะอยู่ในหลาย scale หลายหน่วย ทำให้การนำมาพลอตโดยไม่ normalize จะทำให้กราฟสังเกตเห็นแนวโน้มได้ยาก และค่าผิดเพี้ยนไป

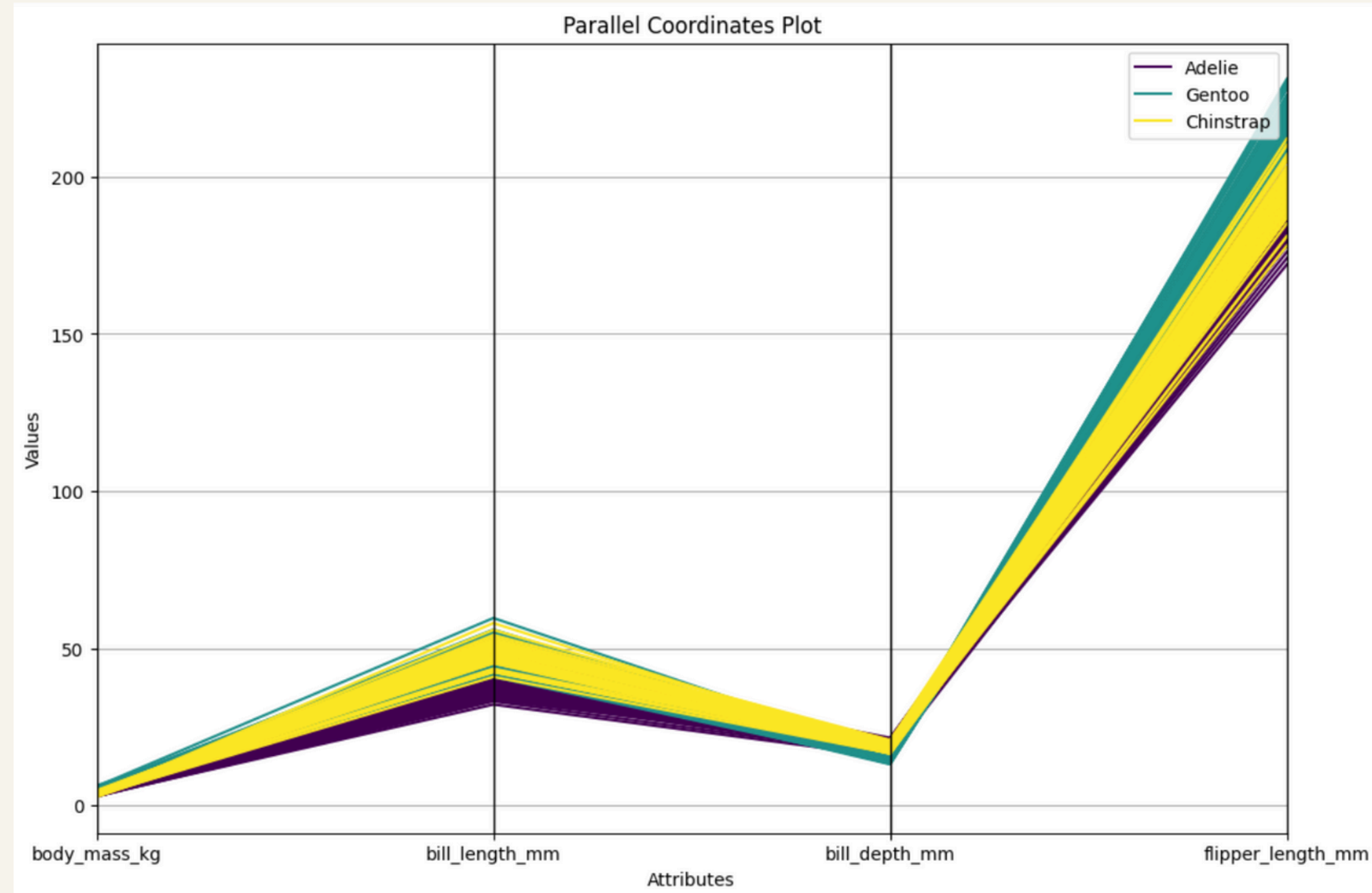
ต้องเรียงลำดับตัวแปร หรือ Dimension ให้เป็นระบบ

ควรต้องจัดระเบียบความสัมพันธ์ของแต่ละแกนให้ถูกต้อง หากไม่เช่นนั้นจะทำให้สังเกตความสัมพันธ์ได้ยาก

DISADVANTAGES

14

Parallel coordinates plot



ตัวอย่างของกราฟที่ไม่ได้ scale และจัดเรียงให้เหมาะสม ทำให้ยากต่อการดูความสัมพันธ์

TECHNIQUE COMPARISON

● Parallel coordinates plot

- สามารถดูข้อมูลหลายตัวแปรพร้อมกันได้
- สามารถมองเห็นความสัมพันธ์ระหว่างแต่ละตัวแปรได้ชัดเจน
- เนื่องจากเป็นกราฟที่แสดงตัวแปรหลายมิติ ทำให้ต้องทำการ Normalize หรือจัดการกับ scale ของตัวแปรแต่ละตัวก่อน ไม่เช่นนั้นจะทำให้ดูยาก
- ไม่เหมาะกับการจัดการข้อมูลที่มีค่าซ้ำซ้อนเยอะ หรือมีจำนวนมาก

● Scatterplot (Pair plot)

- เน้นดูความสัมพันธ์ระหว่างตัวแปร 2 ตัว
- สามารถสังเกต correlation ของคู่ตัวแปรได้ว่าเป็นไปในทางลบ ทางบวก หรือไม่สัมพันธ์กัน
- ไม่ต้องทำการปรับค่า หรือ Normalize ข้อมูล เนื่องจากมีแค่ 2 ตัวแปร ซึ่งมีแกน X และแกน Y แทน scale ของข้อมูลแล้ว
- ไม่มีปัญหาเกี่ยวกับการจัดการกับข้อมูลจำนวนมาก ข้อมูลจำนวนมากยิ่งอาจทำให้เห็นความสัมพันธ์ระหว่างตัวแปรทั้งสองได้ดียิ่งขึ้น

REFERENCES

● Dataset


<https://www.kaggle.com/datasets/ashkhagan/palmer-penguins-datasetalternative-iris-dataset/data>

● Knowledge

https://pandas.pydata.org/docs/reference/api/pandas.plotting.parallel_coordinates.html
<https://python-graph-gallery.com/150-parallel-plot-with-pandas/>

● Photo

<https://www.antarctica.gov.au/about-antarctica/animals/penguins/adelie-penguin/>
https://en.wikipedia.org/wiki/Parallel_coordinates

The background features three vertical stripes on the left: a wide pink stripe, a medium blue stripe, and a narrow beige stripe. The right side of the image is a light beige background with two rectangular areas of a pink dot pattern, one in the top right and one in the bottom right.

THANK YOU