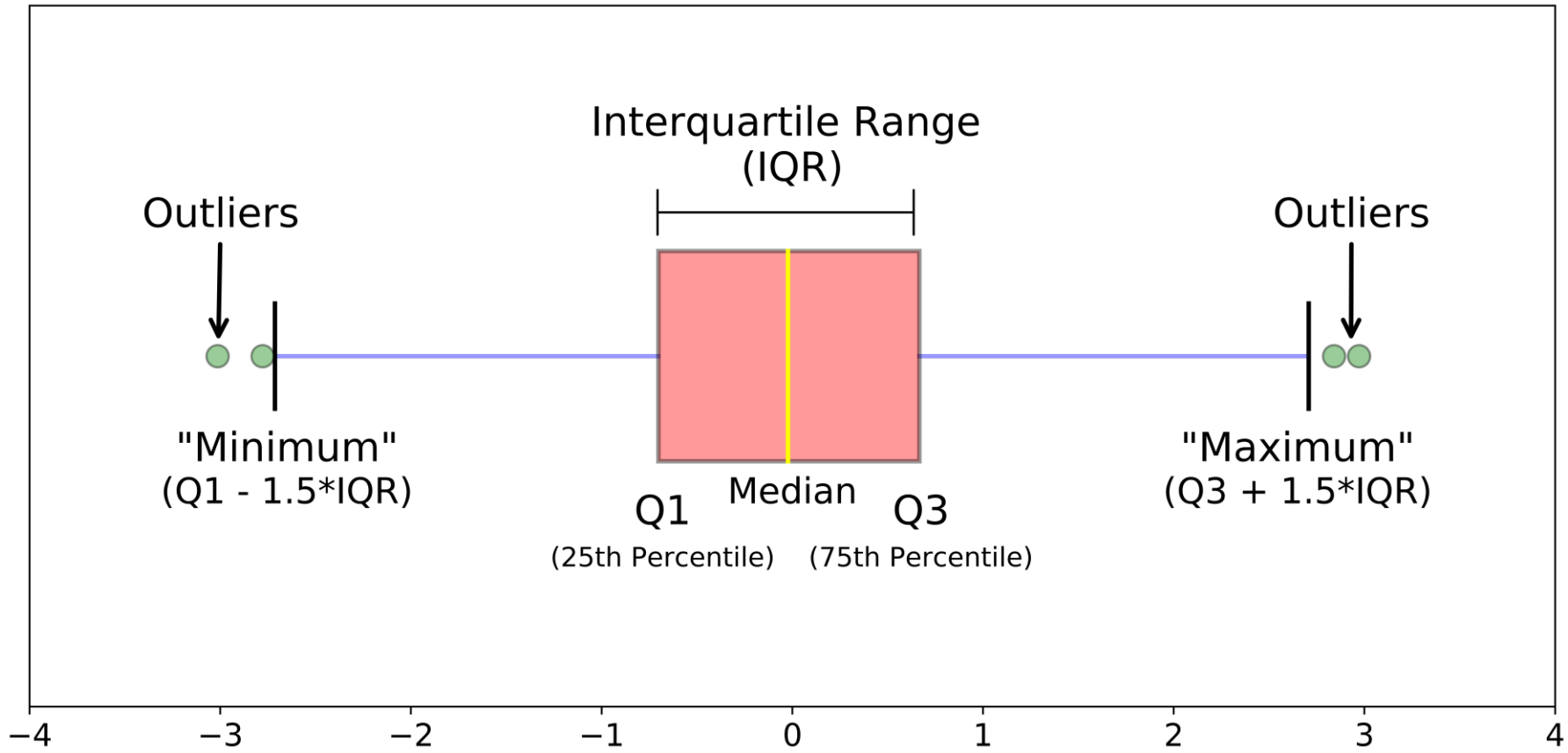

# Class period 15

บทที่ 6 การแสดงผลการกระจายของข้อมูล (ต่อ)

Visualize_Data_Distribution part4

โดย ผศ. ดร.ธนพงศ์ อินทระ สาขาวิชาสถิติ มหาวิทยาลัยขอนแก่น

# Box-plot




โดย ผศ. ดร.ธนพงศ์ อินทระ สาขาวิชาสถิติ มหาวิทยาลัยขอนแก่น

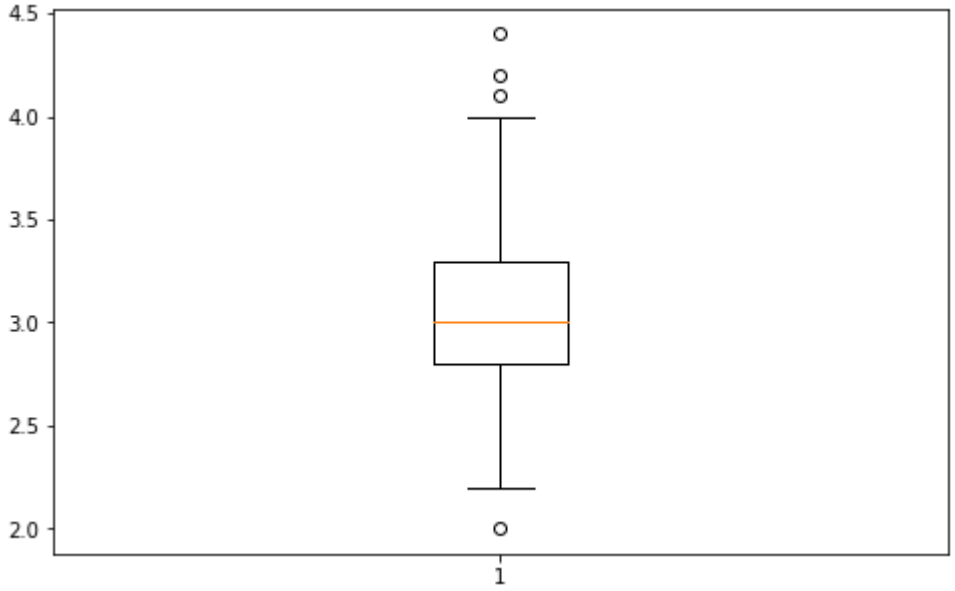
# Box-plot

ใช้ดูการกระจายของข้อมูลและ outliers โดย box-plot สามารถ return ดูค่าที่ต้องการ เช่น whiskers, caps, boxes, medians, fliers, means จากการวาดกราฟได้

วาด Box-plot ใช้คำสั่ง plt.boxplot('ข้อมูลคอลัมน์x') เช่น ใช้ข้อมูลดอกไม้ iris

```
plt.boxplot(df['SepalWidth'])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7a8794dea440>,
  <matplotlib.lines.Line2D at 0x7a8794dea170>],
 'caps': [<matplotlib.lines.Line2D at 0x7a8794deb490>,
  <matplotlib.lines.Line2D at 0x7a8794deb370>],
 'boxes': [<matplotlib.lines.Line2D at 0x7a8794deb520>],
 'medians': [<matplotlib.lines.Line2D at 0x7a8794de9a50>],
 'fliers': [<matplotlib.lines.Line2D at 0x7a8794de9870>],
 'means': []}
```

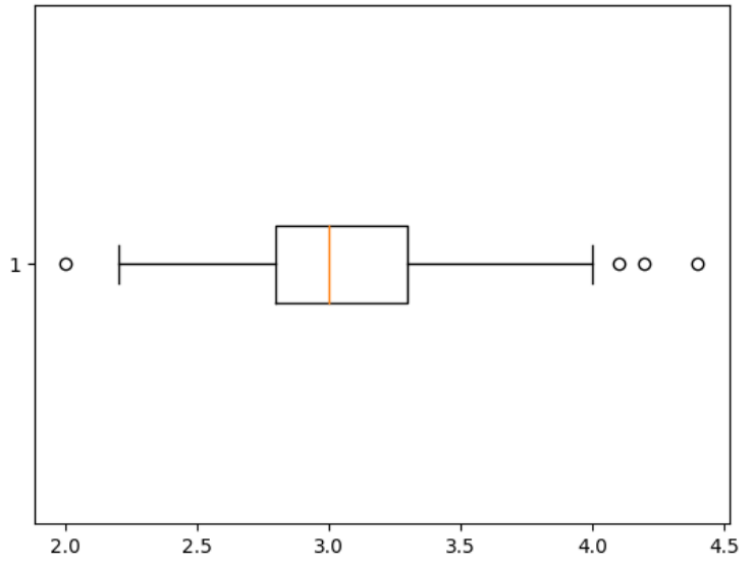


# การ return ค่าจาก Box-plot


- กำหนดตัวแปรที่จะใช้เก็บค่ากราฟ Box-plot
- ตัวแปร Box-plot จากนั้นเลือกค่าที่ต้องการ return ตามด้วย .get_ydata() หรือ .get_xdata() เลือกแนวแกนที่ต้องการดูค่า เช่น ต้องการดูค่า fliers(outliers) ในแนวแกน y
- `o = plt.boxplot(df['SepalWidth'])`
- `o['fliers'][0].get_ydata()`
- ผลลัพธ์จะได้ค่า fliers ของกราฟ Box-plot ในตัวแปร o
- `array([2. , 4.4, 4.1, 4.2])`

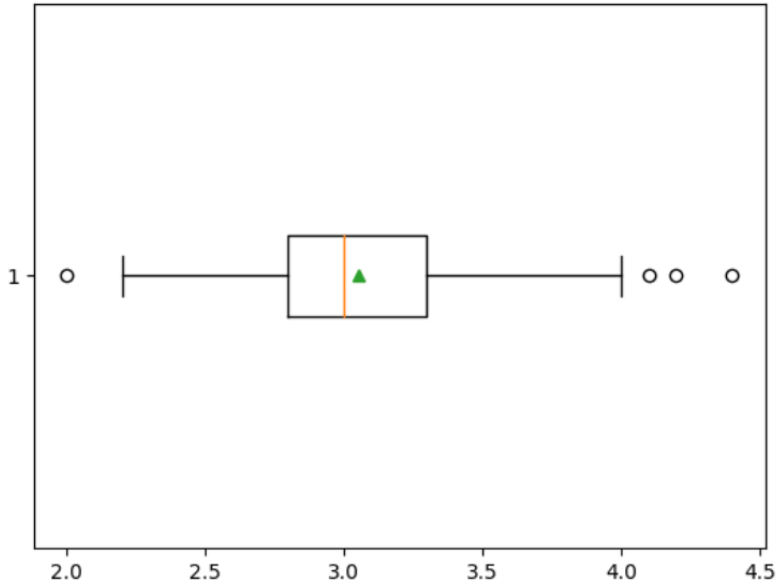
# Parameter: vert=False ปรับกราฟเป็นวาดแนวแกน x

- สามารถวาดกราฟในแนวแกน x ได้โดยการใส่และกำหนด parameter: vert=False (default=True) เช่น
- `ybp = plt.boxplot(df['SepalWidth'],vert=False)`



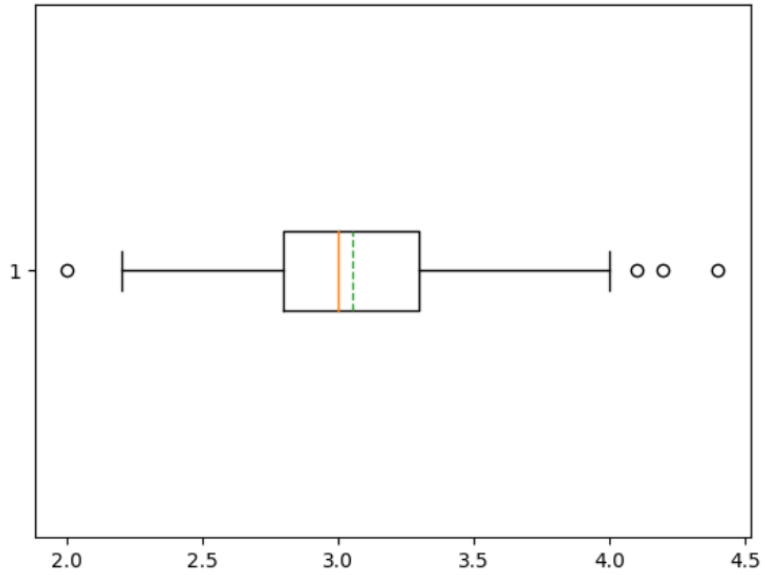
# Parameter: showmeans

- การแสดง means บนกราฟ box-plot สามารถแสดงได้โดยใช้ parameter: showmeans=True (default=False ) เช่น
- `ybp = plt.boxplot(df['SepalWidth'],vert=False,showmeans=True)`



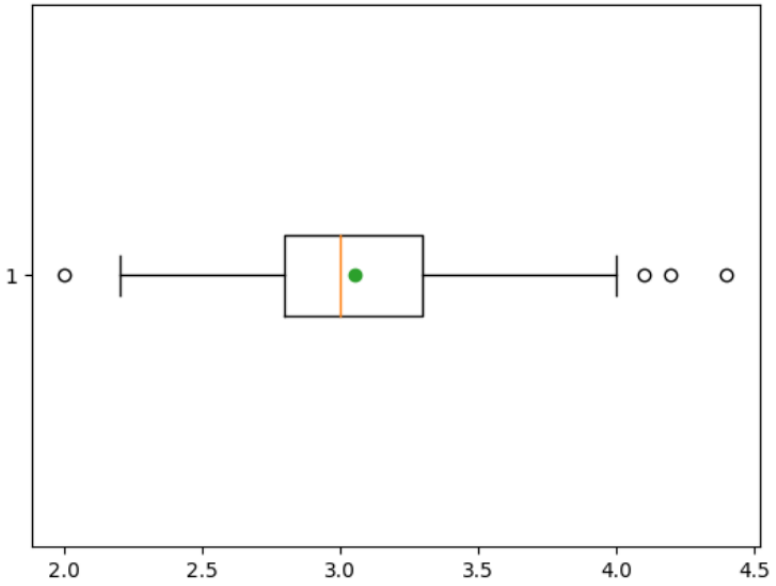
# Parameter: meanline

- แสดง means เป็นเส้นเพื่อง่ายต่อการเปรียบเทียบ ใช้ parameter: meanline=True (default=False) เช่น
- `ybp = plt.boxplot(df['SepalWidth'],vert=False,showmeans=True)`



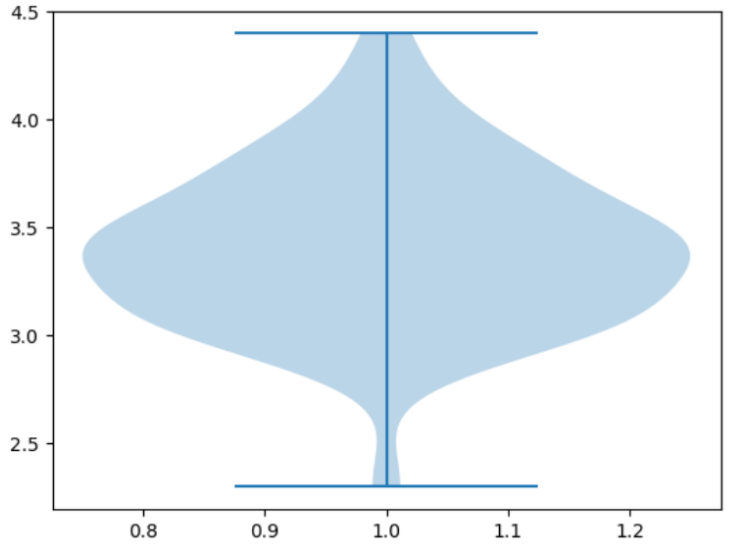
# Parameter: meanprops เปลี่ยนหน้าตา merker ของ mean

- สามารถเปลี่ยนลักษณะหน้าตาของ merker mean บนกราฟได้ตามที่ต้องการ เช่น
- `plt.boxplot(df['SepalLength'],vert=False,showmeans=True,meanprops={'marker':'o'})`



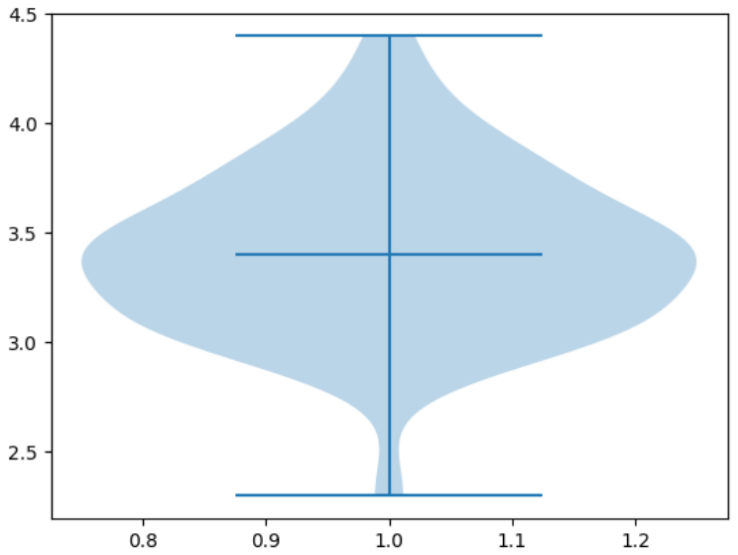
# Violin plot

- เป็นกราฟแสดงการกระจายตัวของข้อมูล
- สามารถใช้งานได้โดยใช้คำสั่ง plt.violinplot(‘ข้อมูลคอลัมน์x’) เช่น
- `plt.violinplot(df['PetalLength'][:50])`



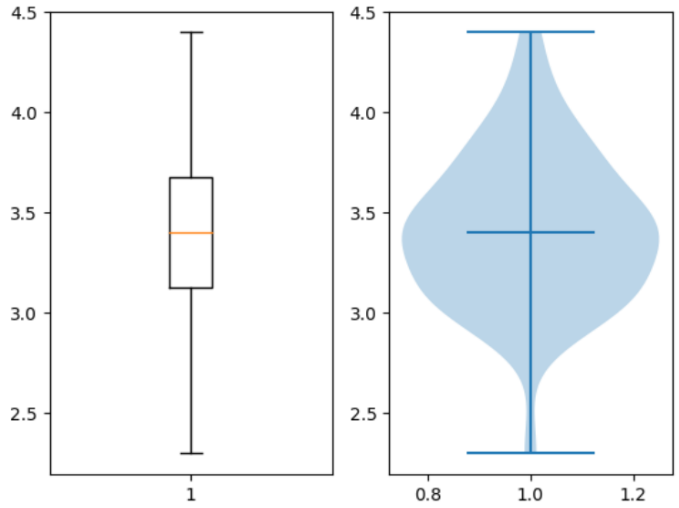
# Parameter: showmedians

- การแสดง medians บนกราฟ violin-plot สามารถแสดงได้โดยใช้ parameter: showmedians =True (default=False ) เช่น
- `vi = plt.violinplot(df['SepalWidth'][:50],showmedians=True)`



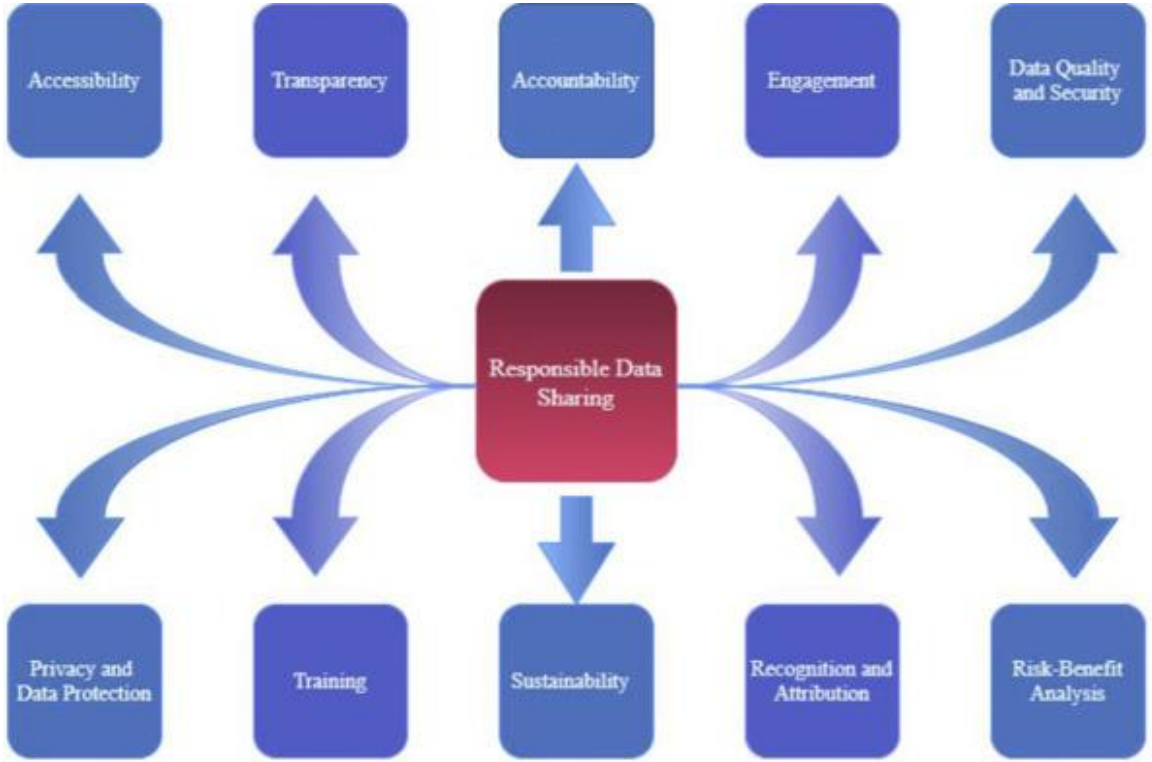
# เปรียบเทียบระหว่าง box-plot กับ violin-plot

- `plt.subplot(1,2,1)`
- `bb = plt.boxplot(df['SepalWidth'][:50])`
- `plt.subplot(1,2,2)`
- `vi = plt.violinplot(df['SepalWidth'][:50],showmedians=True)`



# Health Data Sharing และ Data Privacy

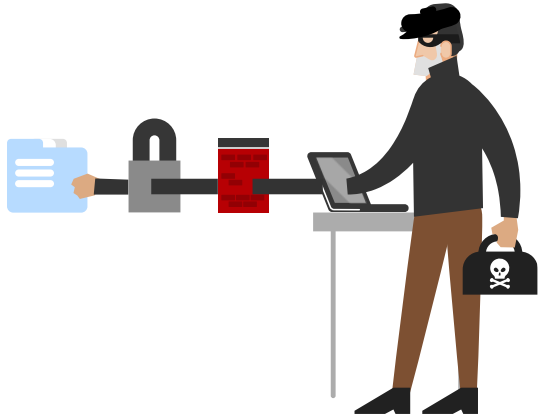
- ปัญหาที่เกิดจากการแบ่งปันข้อมูลสุขภาพและความเป็นส่วนตัวของข้อมูลในช่วงการระบาดของ COVID-19 เช่น
- การรั่วไหลของข้อมูลส่วนบุคคล
- การใช้ข้อมูลสุขภาพเพื่อวัตถุประสงค์อื่น
- ความกังวลเกี่ยวกับการเก็บข้อมูลระยะยาว
- ความไม่เท่าเทียมกันในการเข้าถึงข้อมูลสุขภาพ



# การรั่วไหลของข้อมูลส่วนบุคคล

- มีกรณีที่ข้อมูลส่วนบุคคลของผู้ติดเชื้อ COVID-19 รั่วไหลสู่สาธารณะ เช่น ชื่อ ที่อยู่ และประวัติการเดินทาง
- การรั่วไหลนี้นำไปสู่การตีตราและการเลือกปฏิบัติต่อบุคคลเหล่านั้น รวมถึงครอบครัวและเพื่อนของพวกเขา
- สิ่งนี้แสดงให้เห็นถึงความสำคัญของการปกป้องความเป็นส่วนตัวของข้อมูลและการมีมาตรการรักษาความปลอดภัยที่เหมาะสม

# การใช้ข้อมูลสุขภาพเพื่อวัตถุประสงค์อื่น



- มีความกังวลเกี่ยวกับการใช้ข้อมูลสุขภาพที่รวบรวมระหว่างการระบาดใหญ่เพื่อวัตถุประสงค์อื่น เช่น การเฝ้าระวังหรือเป้าหมายทางการตลาด
- การใช้ข้อมูลนอกเหนือจากวัตถุประสงค์เดิมโดยไม่ได้รับความยินยอมจากบุคคลนั้นถือเป็นการละเมิดความเป็นส่วนตัวและอาจทำลายความไว้วางใจของสาธารณชน
- จำเป็นต้องมีการป้องกันและข้อจำกัดที่ชัดเจนเกี่ยวกับวิธีการใช้ข้อมูลสุขภาพที่ละเอียดอ่อน

# ความกังวลเกี่ยวกับการเก็บข้อมูลระยะยาว

- การเก็บรวบรวมข้อมูลสุขภาพจำนวนมากระหว่างการระบาดใหญ่ทำให้เกิดคำถามเกี่ยวกับระยะเวลาที่ข้อมูลจะถูกเก็บไว้และใครจะสามารถเข้าถึงได้
- ความกังวลเกี่ยวกับผลกระทบระยะยาวที่อาจเกิดขึ้นจากการเก็บรวบรวมข้อมูลจำนวนมาก เช่น การใช้ในอนาคตเพื่อการเลือกปฏิบัติหรือปฏิเสธโอกาส
- จำเป็นต้องมีนโยบายการเก็บรักษาข้อมูลที่ชัดเจนและกลไกสำหรับบุคคลในการเข้าถึงและควบคุมข้อมูลของตนเอง

# ความไม่เท่าเทียมกันในการเข้าถึงข้อมูลสุขภาพ

- การระบาดใหญ่ส่งผลกระทบต่อชุมชนบางแห่งอย่างไม่เป็นสัดส่วน เช่น ชนกลุ่มน้อยทางเชื้อชาติและชาติพันธุ์ และประชากรที่มีรายได้น้อย
- การเข้าถึงข้อมูลสุขภาพและทรัพยากรอย่างเท่าเทียมกันกลายเป็นข้อกังวลด้านจริยธรรมที่สำคัญ เนื่องจากความไม่เท่าเทียมกันอาจนำไปสู่ผลลัพธ์ด้านสุขภาพที่แย่ลงสำหรับกลุ่มที่มีความเสี่ยง
- จำเป็นต้องมีความพยายามเชิงรุกเพื่อเอาชนะอุปสรรคในการเข้าถึงและรับประกันการกระจายข้อมูลและทรัพยากรด้านสุขภาพอย่างเป็นธรรมในหมู่ประชากรที่หลากหลาย

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8988992/

# Quiz


- 1. เขียน function ที่รับ input เป็น output ของ boxplot แล้ว แสดงค่า min, max, q1, q2, q3
- 2. วาด boxplot เปรียบเทียบ การกระจายตัวของข้อมูล PetalLength ของดอก iris ทั้ง 3 ชนิด