# Introduction to Data Visualization and Web Scraping

https://github.com/kaopanboonyuen/GISTDA2023

# Outline

- Part 0: Basic Python and Pandas
- Part 1: Data Visualization using Looker Studio
- Part 2: Web Scraping in Python With BeautifulSoup and Twitter Scraping

All python codes/notebooks/slides will be posted here:

https://github.com/kaopanboonyuen/GISTDA2023

# Part 0: Basic Python and Pandas

## Module 0: Basic Python and Pandas

- Python Recap:  Open in Colab
- Pandas:  Open in Colab

GISTDA

Python

pandas

Looker Studio

Selenium

# Part 1: Data Visualization using Looker Studio

**Module 1: Google Data Studio (Looker Studio):** **https://lookerstudio.google.com/**

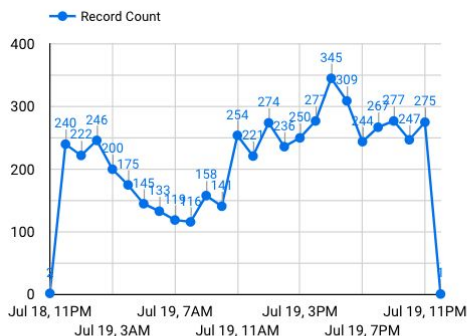- Disaster Tweets (Data Set):

  https://github.com/kaopanboonyuen/GISTDA2023/raw/main/dataset/visualize/disaster_text.csv
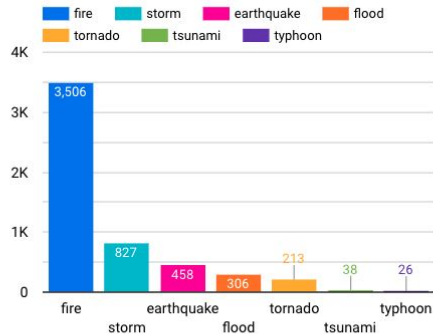
- Med Resource (Data Set):

  https://github.com/kaopanboonyuen/GISTDA2023/raw/main/dataset/visualize/med_resources_text.csv
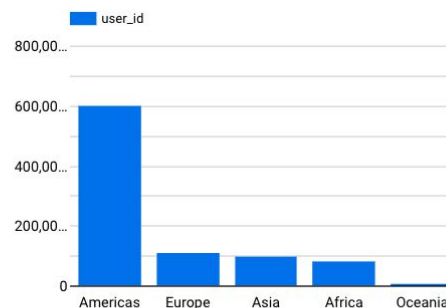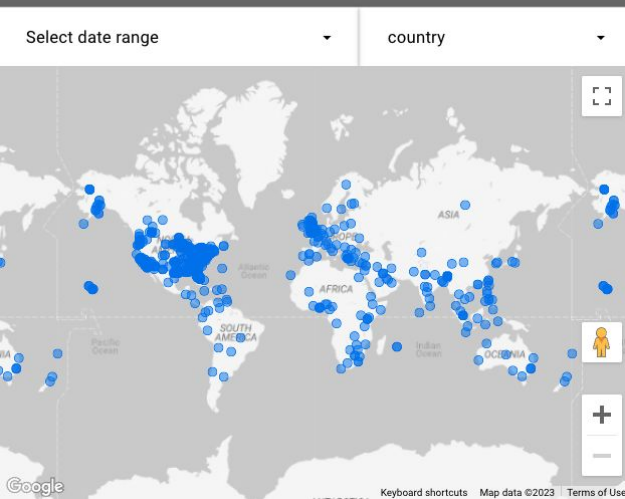
# Disastor Monitoring Dashboard



## # Tweet by Datetime

Record Count

400 — 300 — 200 — 100 — 0

345
309
274 277 275
254 250 267 264 247
240 246 236
222 200
175
145 133 111 116 141 158

Jul 18, 11PM    Jul 19, 7AM    Jul 19, 3PM    Jul 19, 11PM
        Jul 19, 3AM    Jul 19, 11AM    Jul 19, 7PM

## # Tweet by DisasterType

■ fire   ■ storm   ■ earthquake   ■ flood
■ tornado   ■ tsunami   ■ typhoon

4K
3K — 3,506
2K
1K — 827
        458
0 —          306   213   38   26

fire    earthquake    tornado    typhoon
   storm    flood    tsunami

## # Tweet by Region

■ user_id

800,00
600,00
400,00
200,00
0

Americas   Europe   Asia   Africa   Oceania

## # Tweet by Location

| Select date range ▼ | country ▼ |



## # Most Retweeted Tweets

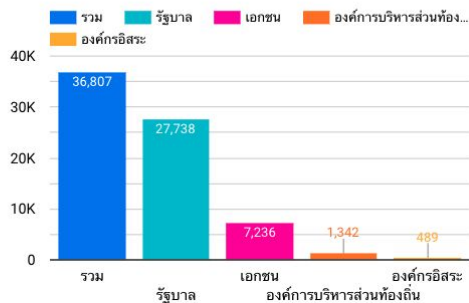| | clean_text | retweet_count ▼ |
|---|---|---|
| 1. | ghana 🇬🇭🇬🇭🇬🇭 you should be proud of the shatt... | 2,202 |
| 2. | join hands for helping the flood victims of assam a... | 1,354 |
| 3. | heavy police and fire presence at 4400 holden in we... | 1,027 |
| 4. | amazing video shot of the tornado east of carmang... | 982 |
| 5. | m5.3 earthquake σεισμός strikes 23 km nw of athe... | 905 |
| 6. | excited to finally announce that i am joining the fire... | 647 |
| 7. | finally after all the stop and starts delays unforesee... | 590 |
| 8. | 11 37pm fireworks then tear gas deployed puertoric... | 434 |
| 9. | m5.1 earthquake σεισμός strikes 23 km nw of athe... | 372 |
| 10. | if unsealepstein is "the happening" wouldn't that be ... | 243 |
| 11. | m5.3 earthquake σεισμός strikes 23 km nw of athe... | 180 |
| 12. | found our biggest offset yet for the fault that cause... | 169 |

1 - 100 / 5055   ‹  ›

# Thailand Physician Resource

## #Total Physicians

จำนวนแพทย์
**36,807**

จำนวนสถานพยาบาล
**1,271**

จำนวนเตียง
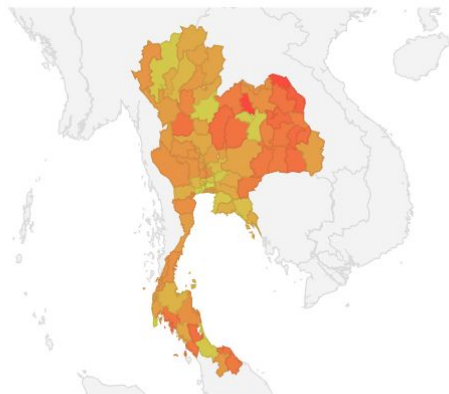**149,641**

## #Physician by Department

Legend: รวม | รัฐบาล | เอกชน | องค์การบริหารส่วนท้องถิ่น... | องค์กรอิสระ

- รวม: 36,807
- รัฐบาล: 27,738
- เอกชน: 7,236
- องค์การบริหารส่วนท้องถิ่น: 1,342
- องค์กรอิสระ: 489

## #Physician by Region

Legend: รวม | รัฐบาล | เอกชน | องค์การบริหารส่วนท้องถิ่น | องค์กรอิสระ

- กรุงเทพมหานคร: 9,753 | 3,693 | 3,824 | 1,252 | 295
- ภาคกลาง: 9,295 | 7,863 | 1,812 | 194
- ภาคตะวันออกเฉียงเหนือ: 8,075 | 7,863 | 880 | 32 | 0
- ภาคเหนือ: 6,008 | 5,308 | 684 | 14 | 0
- ภาคใต้: 4,429 | 3,870 | 536 | 23 | 0
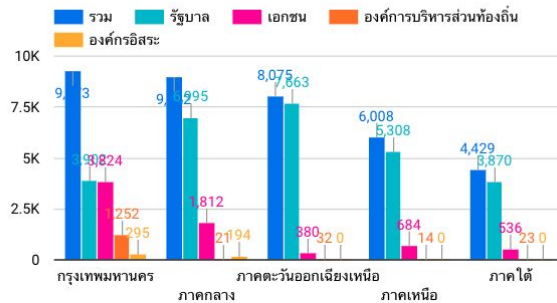
## #Population per Physician

ภาค  ▼          จังหวัด  ▼



601 ▬▬▬ 4,965

## #Highest Population Per Physician

| จังหวัด | สัดส่วนประชากรต่อแพทย์ ❶ ❷ ▼ |
|---|---|
| 5. กาฬสินธุ์ | 3,731 |
| 6. เพชรบูรณ์ | 3,719 |
| 7. ศรีสะเกษ | 3,697 |
| 8. พัทลุง | 3,669 |
| 9. อำนาจเจริญ | 3,666 |

1 - 77 / 77  ‹ ›

## #Lowest Population Per Physician

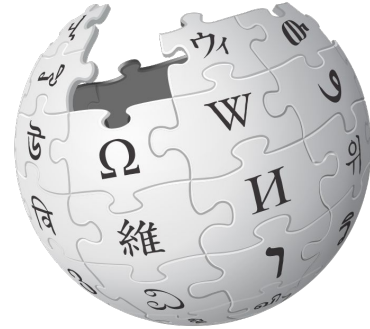| จังหวัด | สัดส่วนประชากรต่อแพทย์ ▲ |
|---|---|
| 1. กรุงเทพมหานคร | 601 |
| 2. ภูเก็ต | 901 |
| 3. สมุทรสาคร | 1,069 |
| 4. พิษณุโลก | 1,086 |

1 - 77 / 77  ‹ ›

# Part 2: Web Scraping in Python With BeautifulSoup and Twitter Scraping

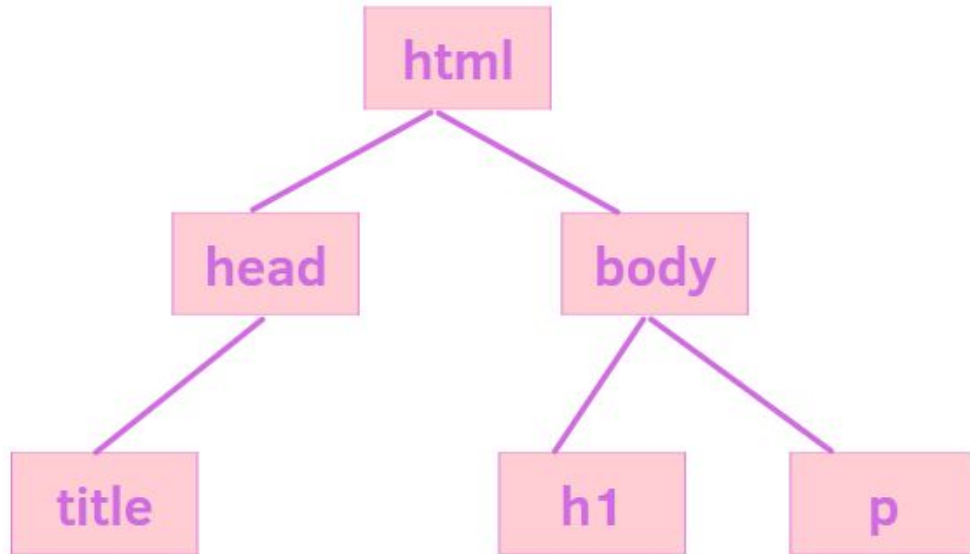## Module 2: Web Scraping and Twitter Scraping

- Web Scraping: [CO Open in Colab]
- Twitter Scraping: [CO Open in Colab]

# The Russian journalist refusing to be silenced

Nobel Peace Prize laureate Dmitry Muratov is looking for hope in his country's younger generation.

🕐 1h | Europe

- **Journalist's Nobel medal sells for $103.5m**
- ▶ **Nobel Peace Prize journalists share win joy**
- **Nobel winneer doused with paint**

## Pope Francis in hospital with respiratory infection

The pontiff, 86, will stay in hospital for a few days but does not have Covid, the Vatican says.

🕐 1m | Europe

## The 70s nuclear relic that may be about to open at last

Finished in 1986, the Bataan plant in the Philippines has never produced a kilowatt of electricity.
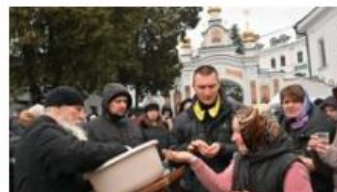
🕐 3h | Asia

## Video of deadly Mexico fire causes outrage

Footage appears to show officers failing to open a cell door as the fire erupted.

🕐 5h | Latin America & Carib…

## Orthodox monks refuse to leave Kyiv monastery

Top clergy in Ukraine's Orthodox Church's are suspected of continuing ties with Russia's Orthodox Church.

🕐 8h | Europe

## Melissa Joan Hart: I helped kids flee shooting

The Sabrina The Teenage Witch actress says she was near Monday's deadly Nashville school shooting.

🕐 6h | US & Canada

headline: The Russian journalist refusing to be silenced
description: Nobel Peace Prize laureate Dmitry Muratov is looking for hope in his country's younger generation.
url: https://www.bbc.com//news/world-europe-65119595

headline: Hotel Rwanda hero Paul Rusesabagina arrives in US
description: The outspoken critic of Rwanda's government arrives in America days after being released from prison.
url: https://www.bbc.com//news/world-africa-65120307

headline: Orthodox monks refuse to leave Kyiv monastery
description: Top clergy in Ukraine's Orthodox Church's are suspected of continuing ties with Russia's Orthodox Church.
url: https://www.bbc.com//news/world-europe-65117269

headline: Video 1 minute 57 secondsWatch: King Charles speaking German in Berlin
description: The King joked with his hosts as he impressed with his language skills at a lavish banquet.
url: https://www.bbc.com//news/uk-65118773

headline: Swimmers in Hawaii accused of harassing dolphins
description: Swimming with dolphins is popular with tourists in Hawaii but federal law bans getting too close.
url: https://www.bbc.com//news/world-us-canada-65114336

headline: Burrito linked to firebombing of anti-abortion group
description: The suspect was arrested at Boston's airport before boarding a one-way fight to Guatemala.
url: https://www.bbc.com//news/world-us-canada-65119859

headline: Swimmers in Hawaii accused of harassing dolphins
description: Swimming with dolphins is popular with tourists in Hawaii but federal law bans getting too close.
url: https://www.bbc.com//news/world-us-canada-65114336

headline: Burrito linked to firebombing of anti-abortion group
description: The suspect was arrested at Boston's airport before boarding a one-way fight to Guatemala.
url: https://www.bbc.com//news/world-us-canada-65119859

```python
# Set up the search query
search_term = "ลุงตู่"
since_date = "2022-01-01"
until_date = "2023-01-31"
#geocode = "13.736717,100.523186, 50km" # search within 50 km of bangkok

# Setting variables to be used below
maxTweets = 500

# Creating list to append tweet data to
tweets_list = []

# create the search query
query = f"{search_term} since:{since_date} until:{until_date}"

# Using TwitterSearchScraper to scrape data and append tweets to list
for i,tweet in enumerate(sntwitter.TwitterSearchScraper(query).get_items()):
    if i > maxTweets:
        break
    tweets_list.append([tweet.date, tweet.id, tweet.content, tweet.username])
```

| | datetime | tweet id | text | username |
|---|---|---|---|---|
| 0 | 2023-01-30 15:20:29+00:00 | 1620079509126135810 | @jieng2019 ซุปเปอร์ไฮเวย์เชียงใหม่-ลำพูน 6 เลน... | jitkasame |
| 1 | 2023-01-30 15:18:42+00:00 | 1620079063200325632 | ถ้าใครไปเชียงราย-เชียงของ-เชียงแสน ถนนเลียบแม่... | jitkasame |
| 2 | 2023-01-30 15:16:10+00:00 | 1620078424399417349 | @kirin_alicha @Kea_New เลือกทั้งบ้านแน่นอนครับ... | GiJOE2538 |
| 3 | 2023-01-30 15:11:23+00:00 | 1620077218209878017 | เหนื่อยหน่อยนะ อยู่กับฉัน\nเมื่อสิ่งที่ฝันนั้น... | NuuNuu_Official |
| 4 | 2023-01-30 15:01:33+00:00 | 1620074745516658691 | #ลุงตู่ บอกที่ทนอยู่ต่อ เพราะรักประชาชน \nคนไท... | 31102551 |
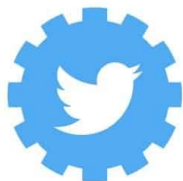


Twitter API

# Image Scraping

```python
1  c = 0
2  for i, url in enumerate(image_urls):
3      try:
4          response = requests.get(url)
5          with open(f'image_scraping_results/image_{i}.jpg', 'wb') as f:
6              f.write(response.content)
7              print(f"-- {c} we found the.jpg format and scrape it")
8          c+=1
9      except:
10         print("!! it is not .jpg format")
11
```

```
!! it is not .jpg format
!! it is not .jpg format
-- 0 we found the.jpg format and scrape it
-- 1 we found the.jpg format and scrape it
-- 2 we found the.jpg format and scrape it
-- 3 we found the.jpg format and scrape it
-- 4 we found the.jpg format and scrape it
-- 5 we found the.jpg format and scrape it
-- 6 we found the.jpg format and scrape it
-- 7 we found the.jpg format and scrape it
-- 8 we found the.jpg format and scrape it
-- 9 we found the.jpg format and scrape it
-- 10 we found the.jpg format and scrape it
-- 11 we found the.jpg format and scrape it
-- 12 we found the.jpg format and scrape it
-- 13 we found the.jpg format and scrape it
-- 14 we found the.jpg format and scrape it
-- 15 we found the.jpg format and scrape it
-- 16 we found the.jpg format and scrape it
-- 17 we found the.jpg format and scrape it
-- 18 we found the.jpg format and scrape it
-- 19 we found the.jpg format and scrape it
-- 20 we found the.jpg format and scrape it
```



webdesignerdepot.com/2009/01/the-evolution-of-apple-design-between-1977-2008/

## Apple II — 1977

1977

saw the invention of both the Apple II and the famous rainbow Apple logo. Steve Jobs added the colours to the logo to reflect the Apple II's superior colour output. Colour graphics set the Apple II apart from its rivals on the market. Image: Wikipedia

# Seaborn: Statistical Data Visualization

Seaborn helps to visualize the statistical relationships, To understand how variables in a dataset are related to one another and how that relationship is dependent on other variables, we perform statistical analysis.

This Statistical analysis helps to visualize the trends and identify various patterns in the dataset.

- Line Plot
- Scatter Plot
- Box plot
- Point plot
- Count plot
- Violin plot
- Swarm plot
- Bar plot
- KDE Plot

https://www.geeksforgeeks.org/data-visualization-with-python-seaborn/

```
1 !pip install -q seaborn
```

```
1 # load the csv
2 data = pd.read_csv("https://github.com/kaopanboonyuen/GISTDA2023/raw/main/dataset/visualize/nba.csv")
3
4 # show first 5 column
5 data.head()
```

|   | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|------|------|--------|----------|-----|--------|--------|---------|--------|
| 0 | Avery Bradley | Boston Celtics | 0.0 | PG | 25.0 | 6-2 | 180.0 | Texas | 7730337.0 |
| 1 | Jae Crowder | Boston Celtics | 99.0 | SF | 25.0 | 6-6 | 235.0 | Marquette | 6796117.0 |
| 2 | John Holland | Boston Celtics | 30.0 | SG | 27.0 | 6-5 | 205.0 | Boston University | NaN |
| 3 | R.J. Hunter | Boston Celtics | 28.0 | SG | 22.0 | 6-5 | 185.0 | Georgia State | 1148640.0 |
| 4 | Jonas Jerebko | Boston Celtics | 8.0 | PF | 29.0 | 6-10 | 231.0 | NaN | 5000000.0 |

# Line plot:

Lineplot Is the most popular plot to draw a relationship between x and y with the possibility of several semantic groupings.
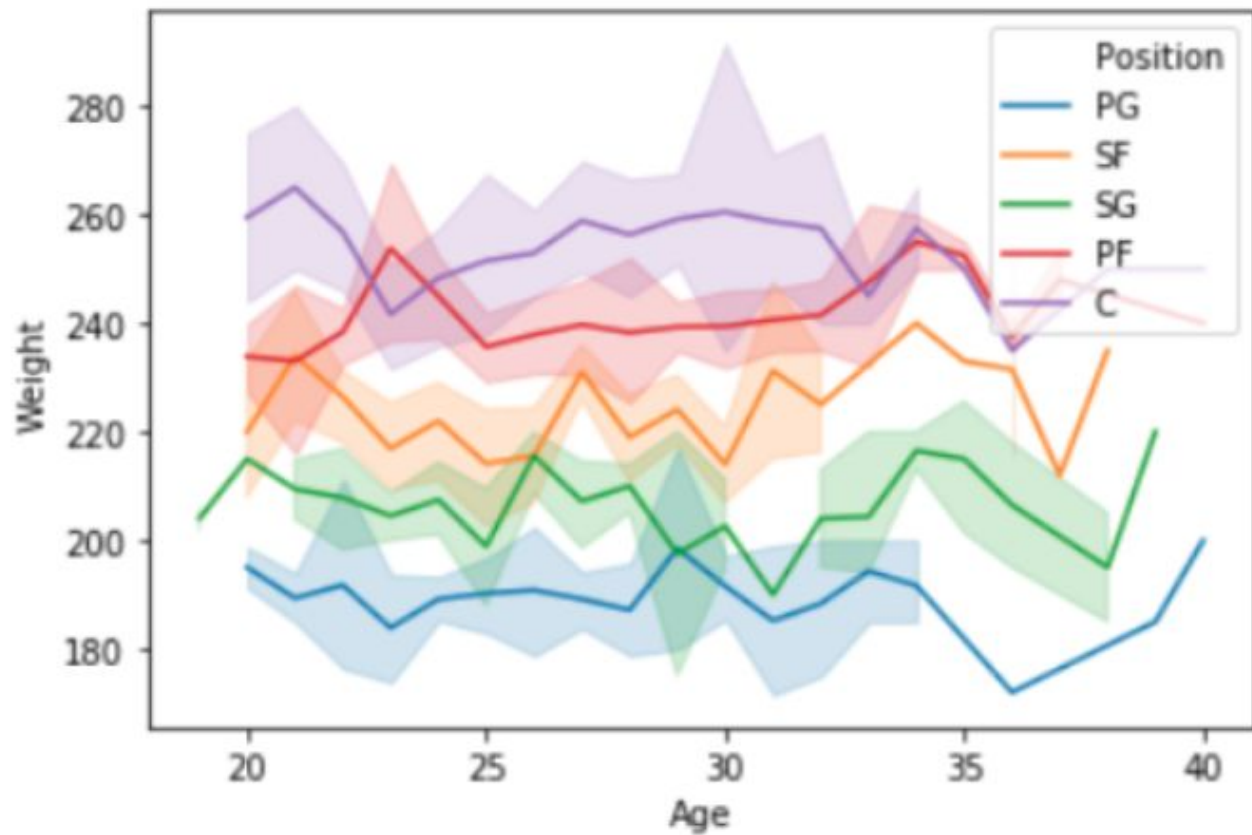
> **Syntax :** *sns.lineplot(x=None, y=None)*
>
> **Parameters:**
>
> **x, y:** *Input data variables; must be numeric. Can pass data directly or reference columns in data.*

**Output:**
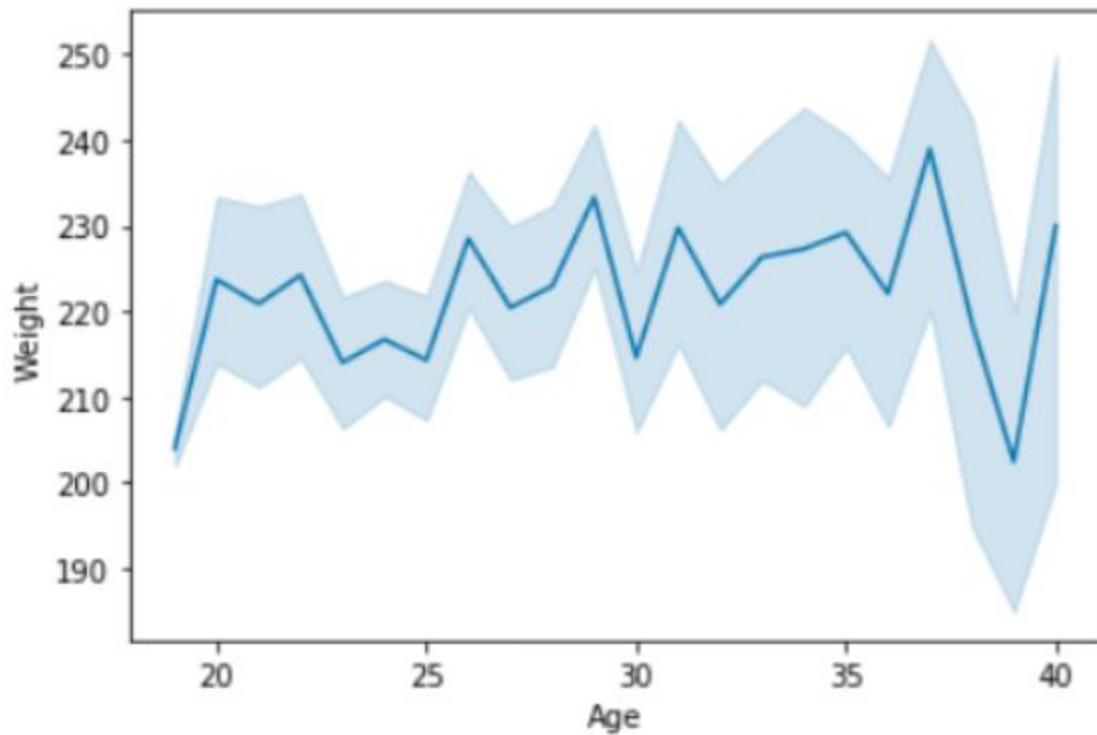
**Output:**

## Scatter Plot:

[Scatterplot](#) Can be used with several semantic groupings which can help to understand well in a graph against continuous/categorical data. It can draw a two-dimensional graph.

> *Syntax:* *seaborn.scatterplot(x=None, y=None)*
>
> *Parameters:*
>
> *x, y:* *Input data variables that should be numeric.*
>
> *Returns:* *This method returns the Axes object with the plot drawn onto it.*

**Output:**
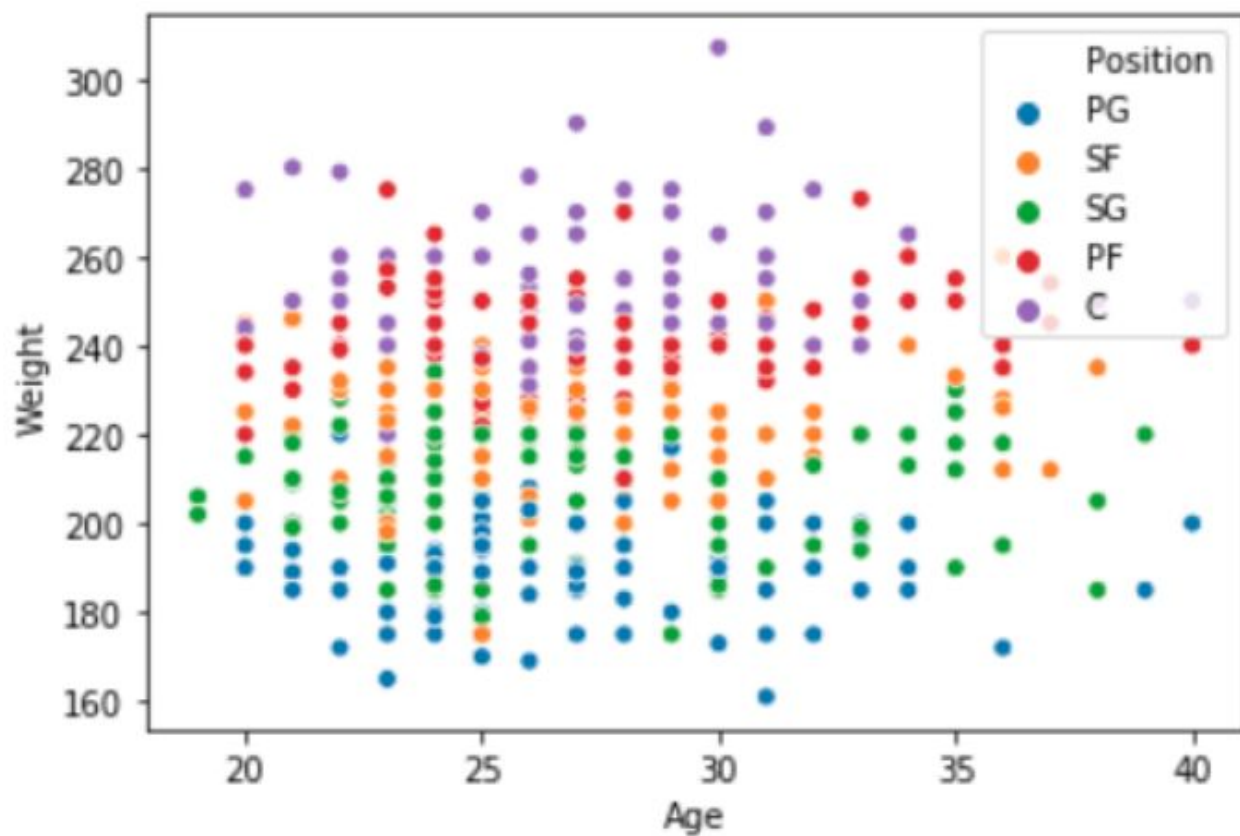
A box plot (or box-and-whisker plot) s is the visual representation of the depicting groups of numerical data through their quartiles against continuous/categorical data.

A box plot consists of 5 things.

- Minimum
- First Quartile or 25%
- Median (Second Quartile) or 50%
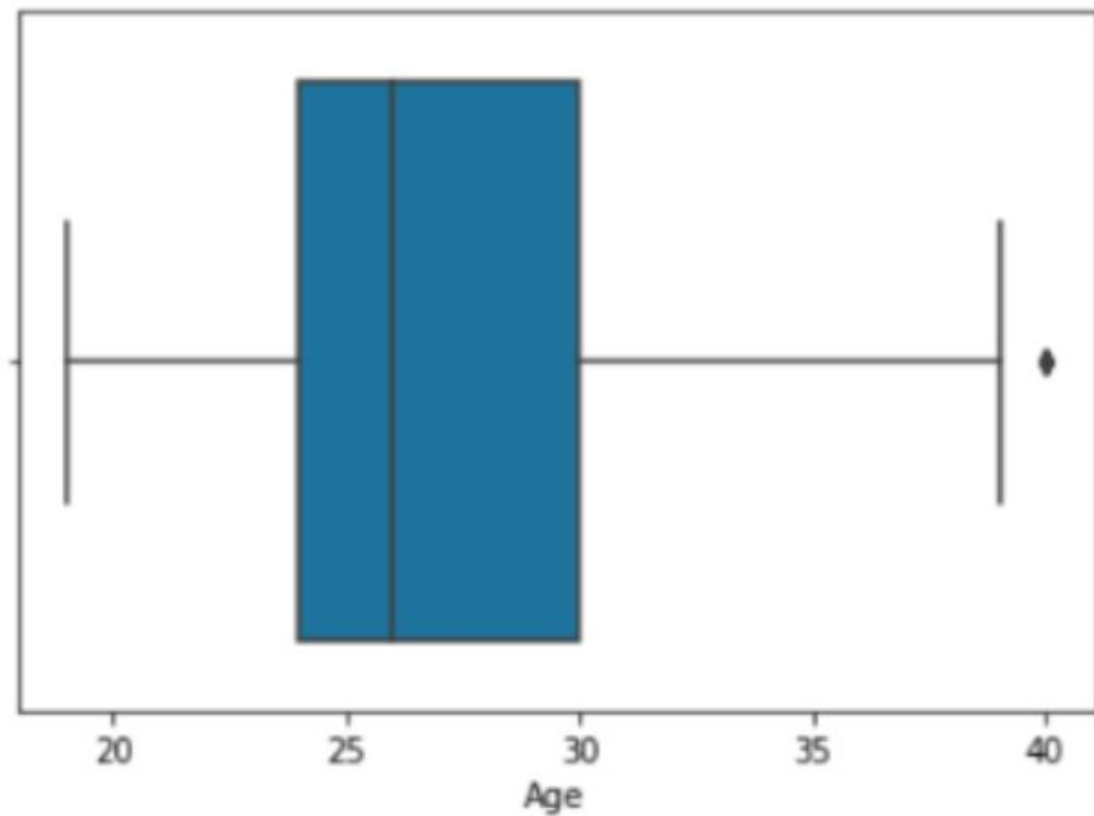- Third Quartile or 75%
- Maximum

*Syntax:*

*seaborn.boxplot(x=None, y=None, hue=None, data=None)*

*Parameters:*

- *x, y, hue: Inputs for plotting long-form data.*
- *data: Dataset for plotting. If x and y are absent, this is interpreted as wide-form.*

*Returns: It returns the Axes object with the plot drawn onto it.*
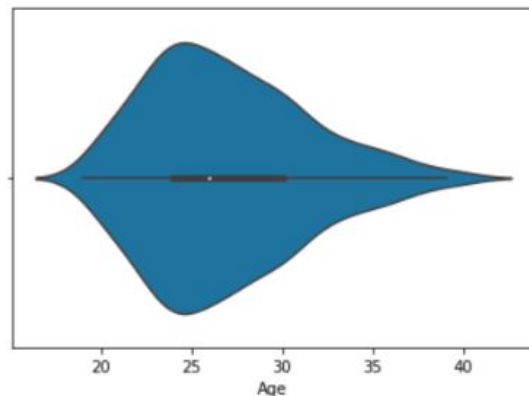
**Output:**

## Violin Plot:

A violin plot is similar to a boxplot. It shows several quantitative data across one or more categorical variables such that those distributions can be compared.

*Syntax:* *seaborn.violinplot(x=None, y=None, hue=None, data=None)*

*Parameters:*

- **x, y, hue:** *Inputs for plotting long-form data.*
- **data:** *Dataset for plotting.*
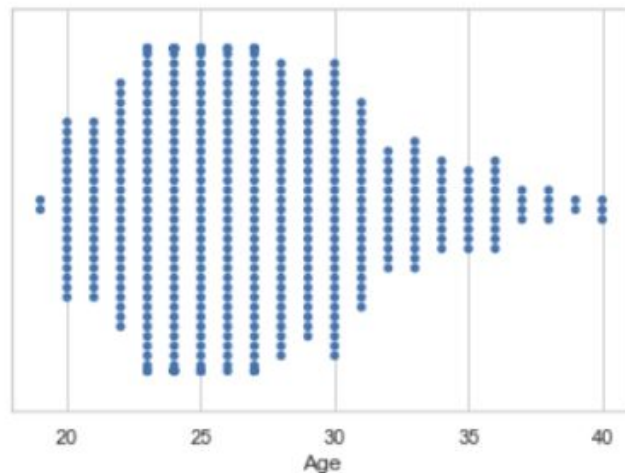
**Output:**

## Swarm plot:

A swarm plot is similar to a strip plot, We can draw a swarm plot with non-overlapping points against categorical data.

> **Syntax:** *seaborn.swarmplot(x=None, y=None, hue=None, data=None)*
>
> **Parameters:**
>
> - **x, y, hue:** *Inputs for plotting long-form data.*
> - **data:** *Dataset for plotting.*

**Output:**

## Bar plot:

[Barplot](#) represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars.
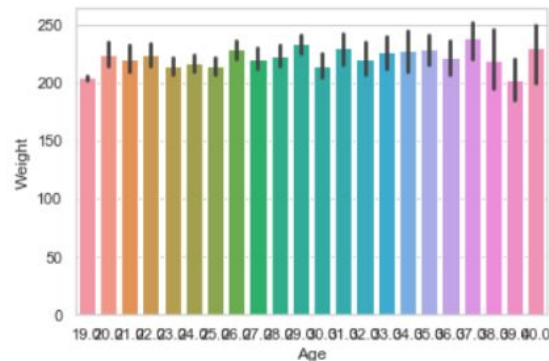
**Syntax :** *seaborn.barplot(x=None, y=None, hue=None, data=None)*

**Parameters :**

- **x, y :** *This parameter take names of variables in data or vector data, Inputs for plotting long-form data.*
- **hue :** *(optional) This parameter take column name for colour encoding.*
- **data :** *(optional) This parameter take DataFrame, array, or list of arrays, Dataset for plotting. If x and y are absent, this is interpreted as wide-form. Otherwise it is expected to be long-form.*

**Returns :** *Returns the Axes object with the plot drawn onto it.*

**Output:**

# Point plot:

**Point plot** used to show point estimates and confidence intervals using scatter plot glyphs. A point plot represents an estimate of central tendency for a numeric variable by the position of scatter plot points and provides some indication of the uncertainty around that estimate using error bars.
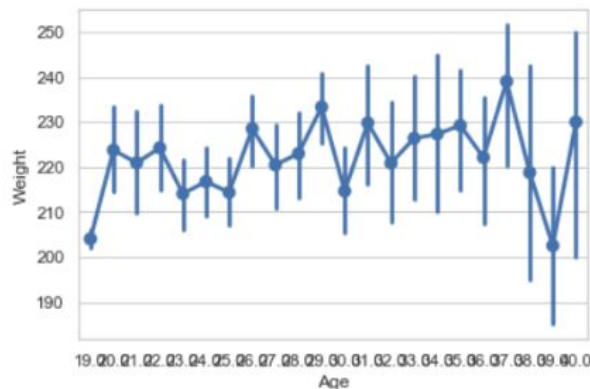
**Output:**



*Syntax:* *seaborn.pointplot(x=None, y=None, hue=None, data=None)*

*Parameters:*

- **x, y:** *Inputs for plotting long-form data.*
- **hue:** *(optional) column name for color encoding.*
- **data:** *dataframe as a Dataset for plotting.*

*Return:* *The Axes object with the plot drawn onto it.*

## Count plot:

Count plot used to Show the counts of observations in each categorical bin using bars.

**Syntax :** *seaborn.countplot(x=None, y=None, hue=None, data=None)*

**Parameters :**

- **x, y:** *This parameter take names of variables in data or vector data, optional, Inputs for plotting long-form data.*
- **hue :** *(optional) This parameter take column name for color encoding.*
- **data :** *(optional) This parameter take DataFrame, array, or list of arrays, Dataset for plotting. If x and y are absent, this is interpreted as wide-form. Otherwise, it is expected to be long-form.*

**Returns:** *Returns the Axes object with the plot drawn onto it.*
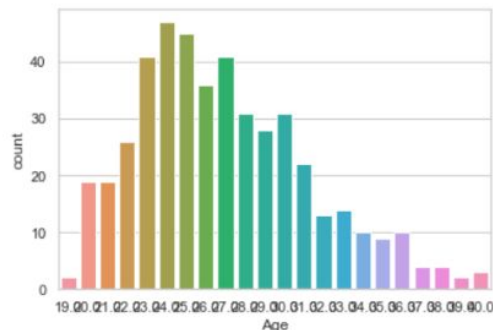
## KDE Plot:

KDE Plot described as **Kernel Density Estimate** is used for visualizing the Probability Density of a continuous variable. It depicts the probability density at different values in a continuous variable. We can also plot a single graph for multiple samples which helps in more efficient data visualization.

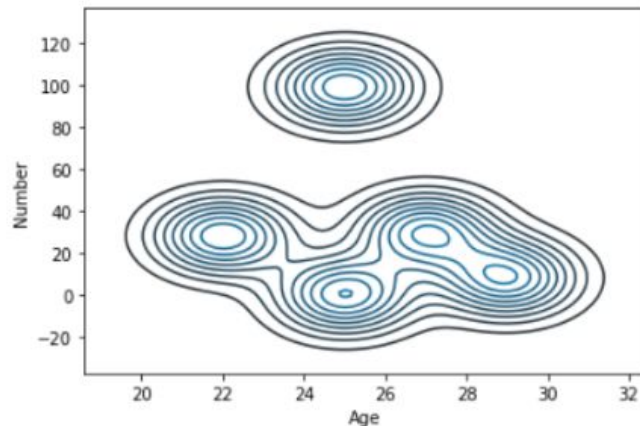**Syntax:** *seaborn.kdeplot(x=None, *, y=None, vertical=False, palette=None, **kwargs)*

**Parameters:**

**x, y :** *vectors or keys in data*

**vertical :** *boolean (True or False)*

**data :** *pandas.DataFrame, numpy.ndarray, mapping, or sequence*

Output:

# Pair Plot