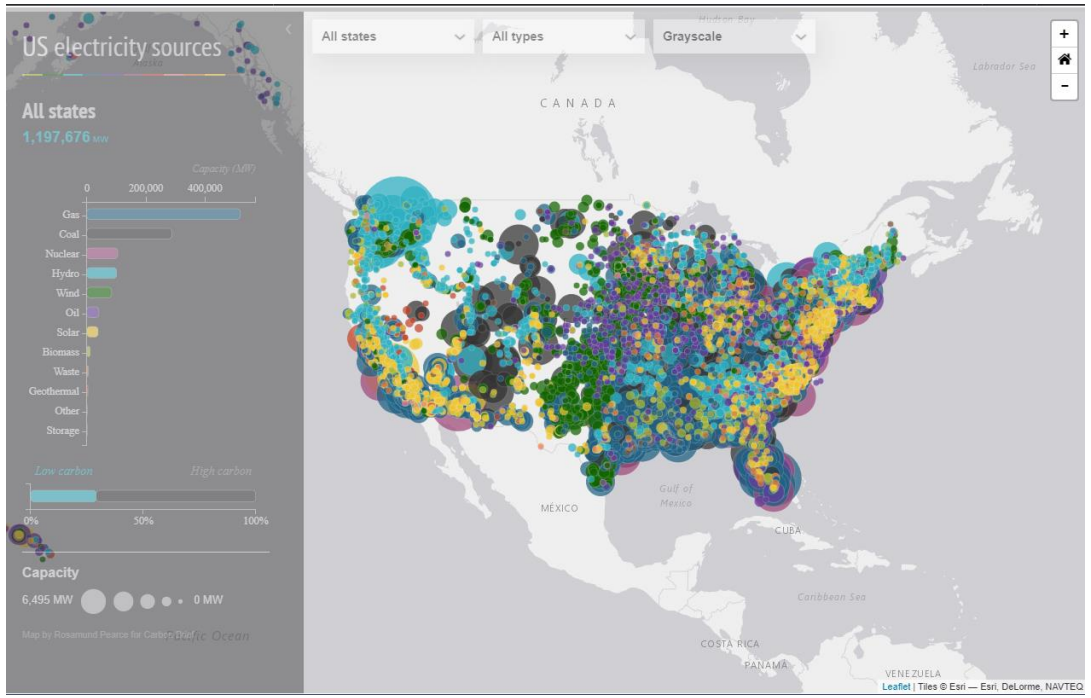
탐색적 데이터 분석 2 : 기초 시각화

Basic Data Visualization using Matplotlib, Pandas, Seaborn

부산대학교 정보·의생명공학대학
정보컴퓨터공학부

What is Visualization?

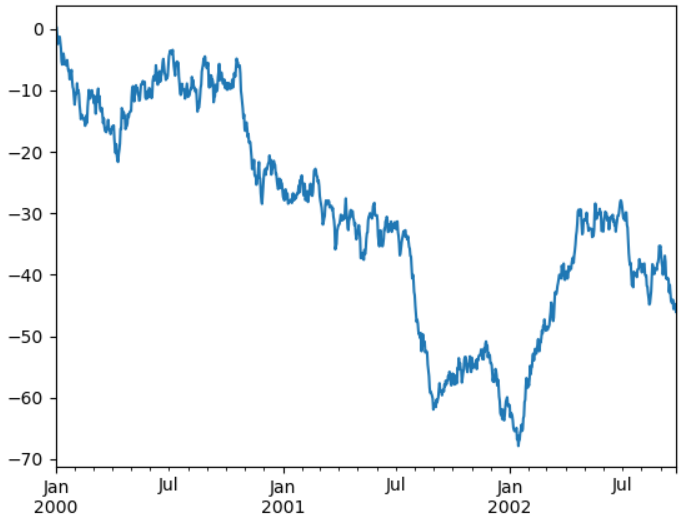
- “Transformation of the symbolic into the geometric,” McCormick et al. 1987
- “... finding the artificial memory that best supports our natural means of perception,” Bertin 1967
- “The use of computer-generated, interactive, visual representations of data to amplify cognition,” Card, M ackinlay, & Shneiderman 1999

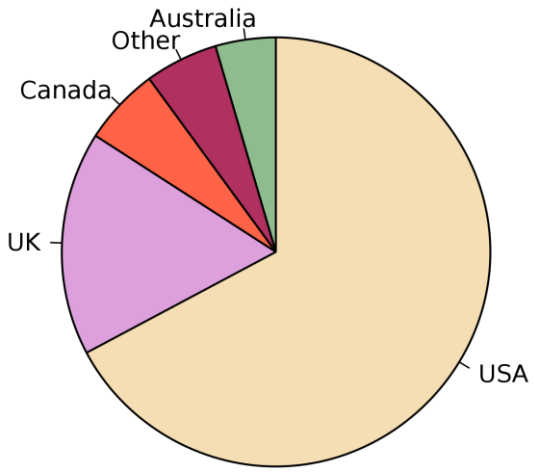


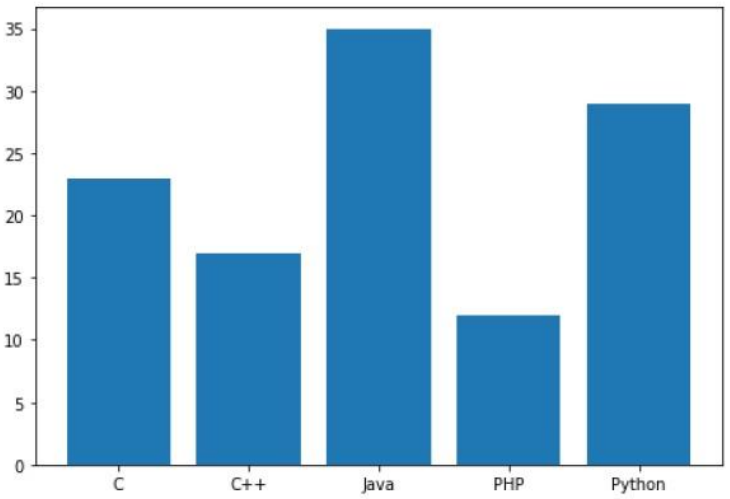
https://www.carbonbrief.org/mapped-how-the-us-generates-electricity

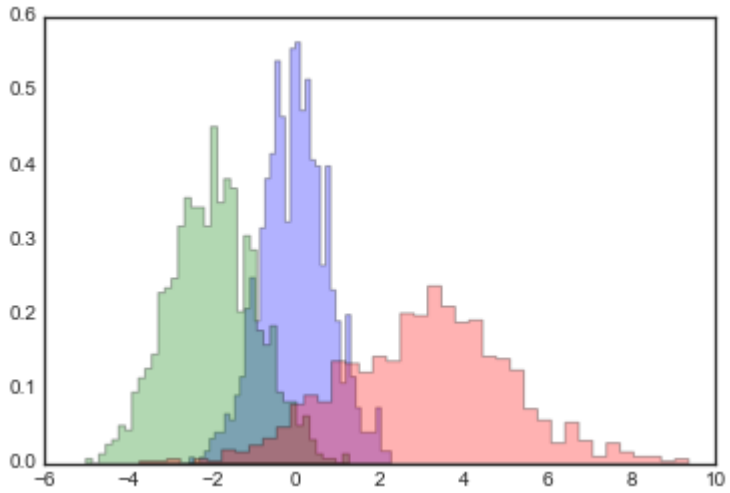
https://youtu.be/0ksaAnu9kog?t=75

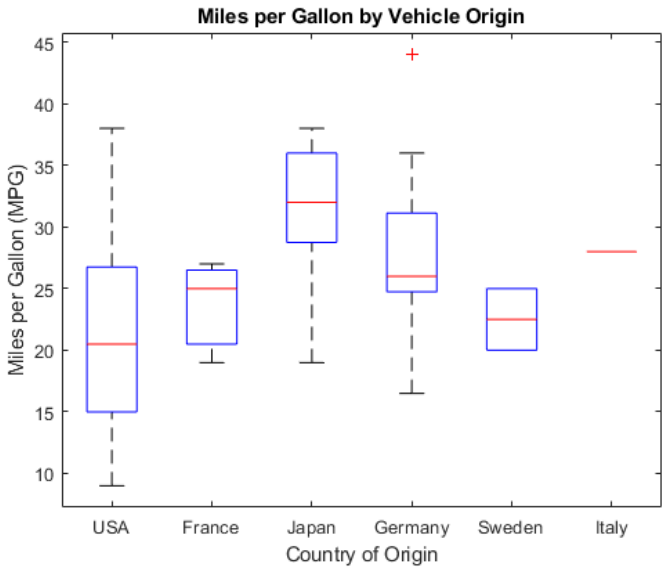
Frequently Used Graphs

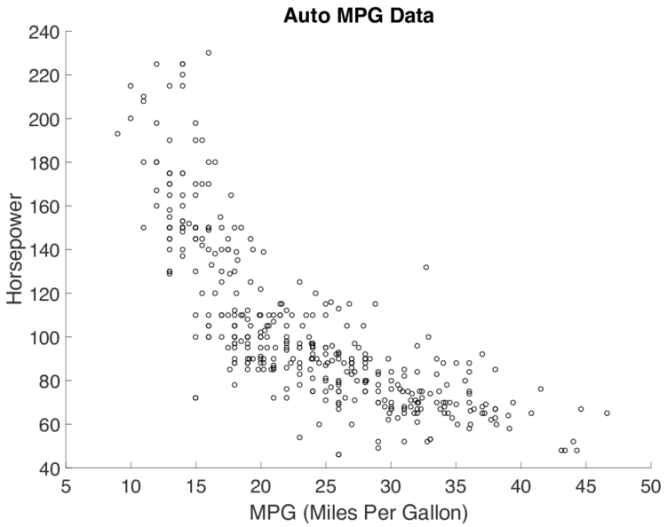




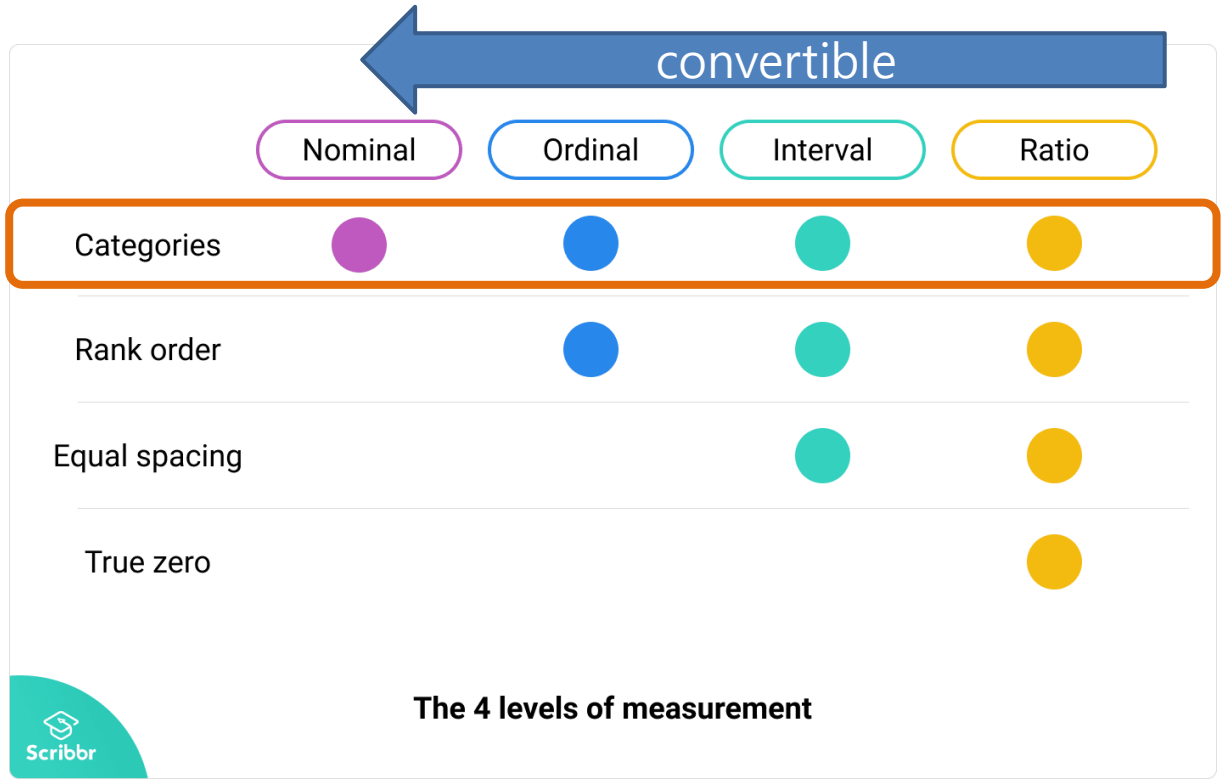


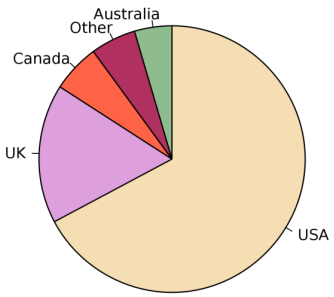




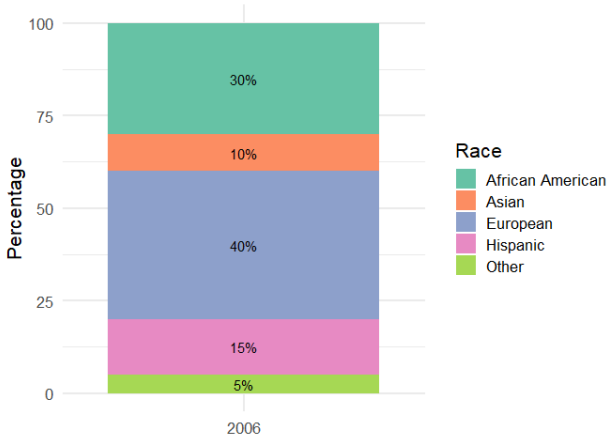


Visualizing Categorical Data

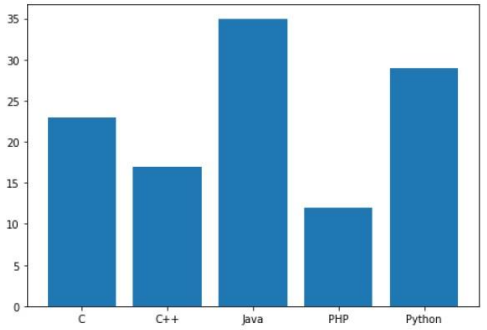




Pie Chart



Bar Chart



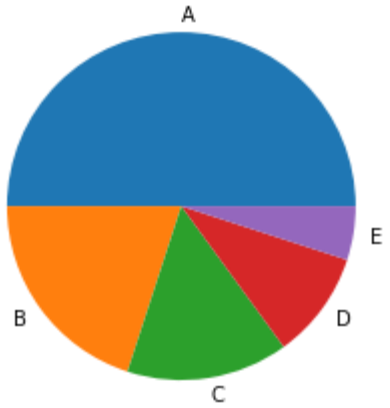
Code Examples – Pie Chart

❖ **1. list → pie chart**

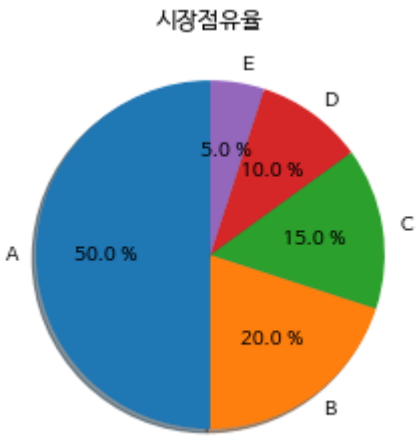
❖ 2. csv → pandas DataFrame → 도수 분포표 → pie chart

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

labels = ['A', 'B', 'C', 'D', 'E']
x = np.array([50, 20, 15, 10, 5])
plt.pie(x, labels=labels)
plt.show()
```




```
plt.rc('font', family='NanumBarunGothic')
plt.pie(x, labels=labels, shadow=True, \
        startangle=90, autopct='%0.1f %%')
plt.title('시장점유율')
plt.show()
```



Code Examples – 한글 사용 준비

그래프에서 한글을 사용하려면 우선 한글 폰트를 설치해야함.
아래 코드를 코랙 시작시 실행.

```
!sudo apt-get install -y fonts-nanum
!sudo fc-cache -fv
!rm ~/.cache/matplotlib -rf
```

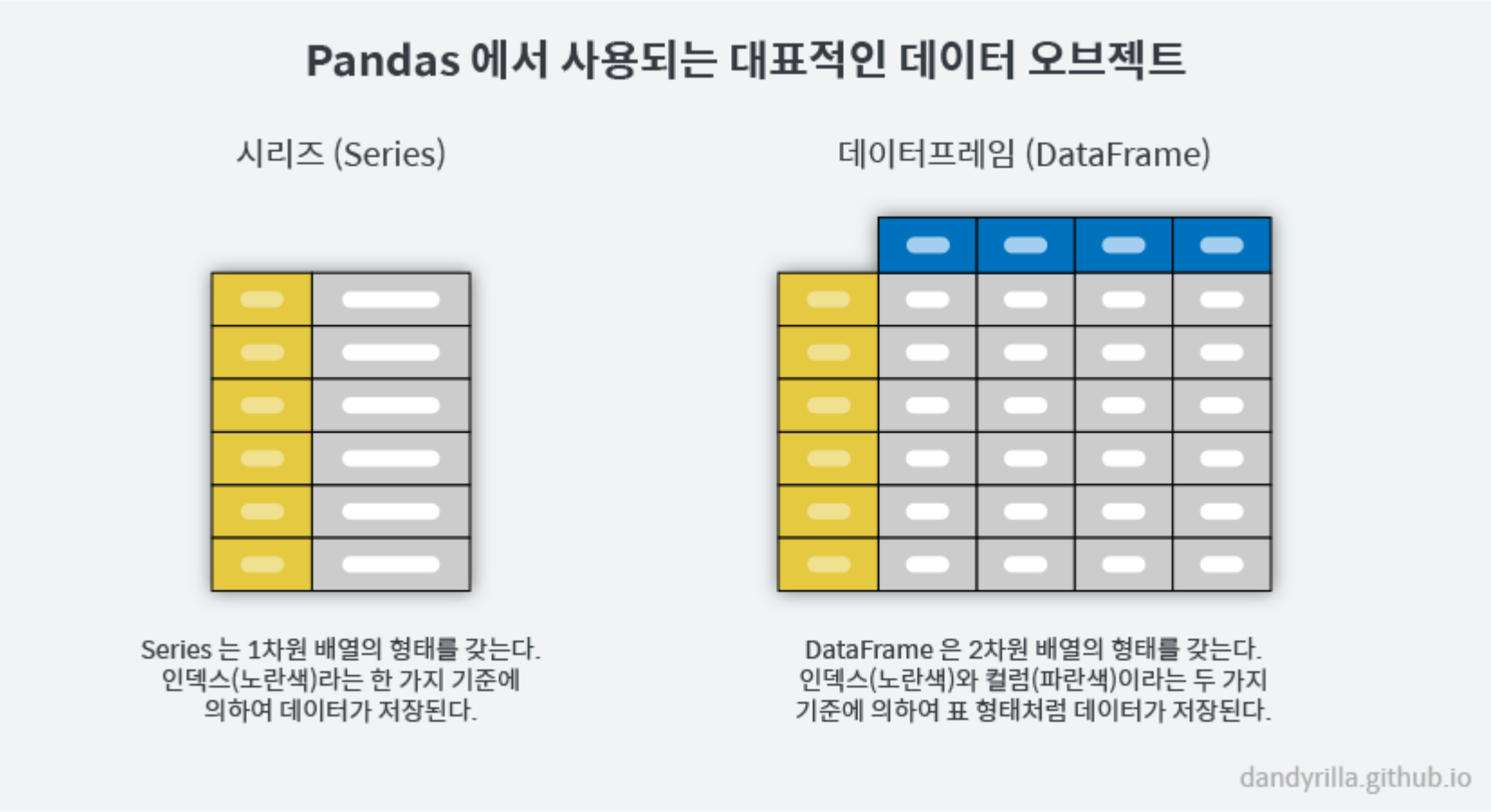
그래프를 그리기 전에 우선 폰트를 지정해주어야 함.

```
import matplotlib.pyplot as plt
plt.rc('font', family='NanumBarunGothic')
```

Code Examples – Pie Chart

❖ 1. list → pie chart

❖ **2. csv → pandas DataFrame → 도수 분포표 → pie chart**



https://dandyrilla.github.io/2017-08-12/pandas-10min/

https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html

Code Examples – Pie Chart

❖ 1. list → pie chart

❖ **2. csv → pandas DataFrame → 도수 분포표 → pie chart**

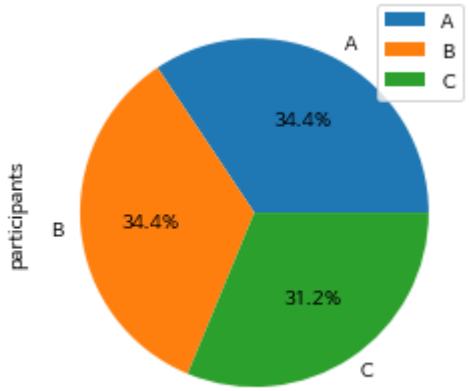
```
import pandas as pd
myDF = pd.read_csv('https://raw.githubusercontent.com/mlee-pnu\
/IDS/main/data1.csv')
myDF.head()
```

	pid	group	gender
0	1	3	1
1	2	2	1
2	3	1	1
3	4	2	2
4	6	1	1

```
table = pd.crosstab(index=myDF["group"].values, \
                    colnames=["group"],columns='participants')
table.index = ["A","B","C"]
print(table)
```

```
group  participants
A                 11
B                 11
C                 10
```

```
table.plot.pie(y='participants', autopct='%0.1f%%')
```



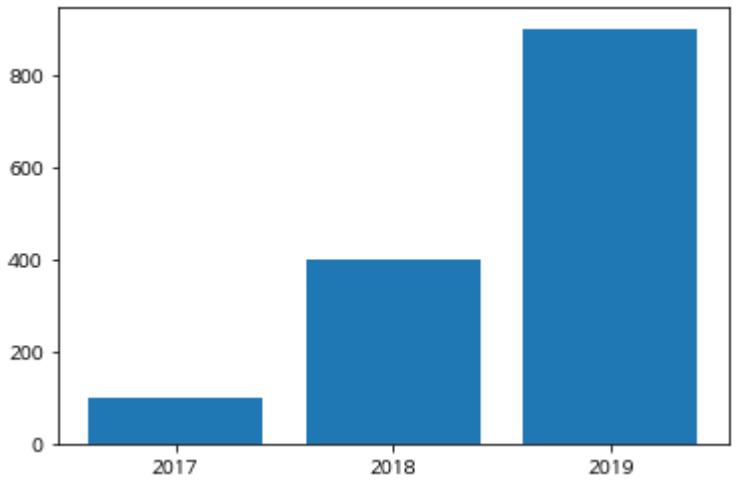
Code Examples – Bar Chart

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

x = np.arange(3) # 0, 1, 2
years = ['2017', '2018', '2019']
values = [100, 400, 900]
```

```
plt.bar(x, values)
# displaying years on x axis
plt.xticks(x, years)
plt.show()
```

```
plt.bar(years,values)
plt.show()
```



Code Examples – Bar Chart

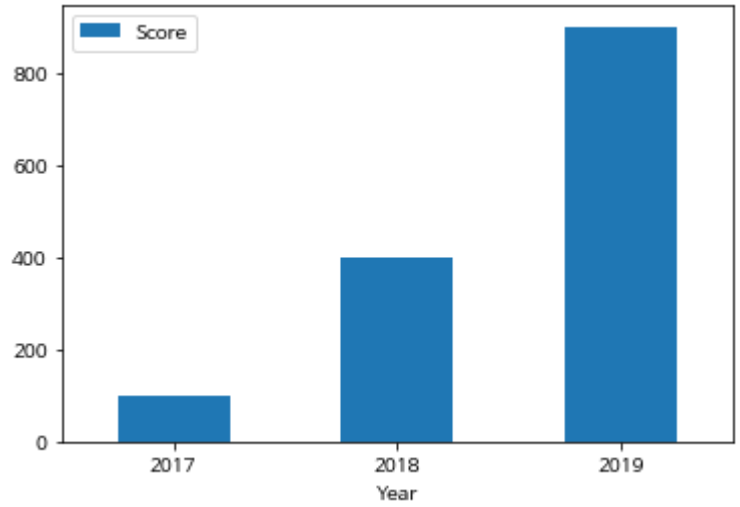
```
import matplotlib.pyplot as plt
import numpy as np

x = np.arange(3) # 0, 1, 2
years = ['2017', '2018', '2019']
values = [100, 400, 900]
```

```
import pandas as pd
myDF = pd.DataFrame({'Year':years,
                     'Score':values})
myDF.head()
```

	Year	Score
0	2017	100
1	2018	400
2	2019	900

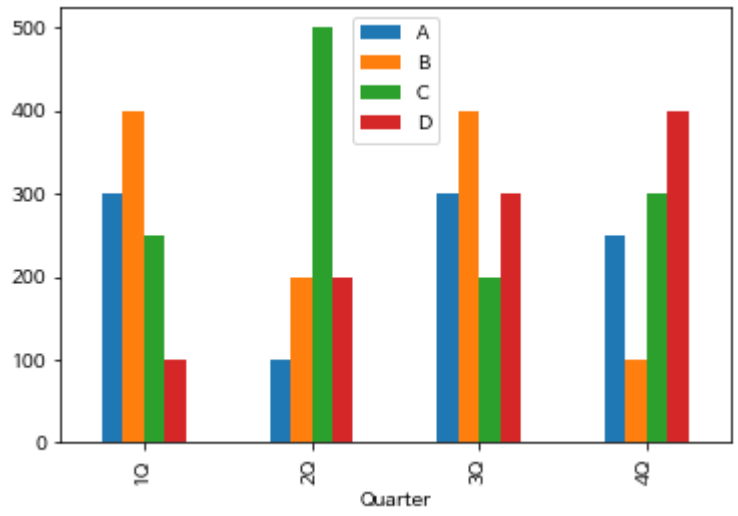
```
myDF.plot.bar(x='Year', y='Score', rot=0)
```



Code Examples – Stacked Bar Chart

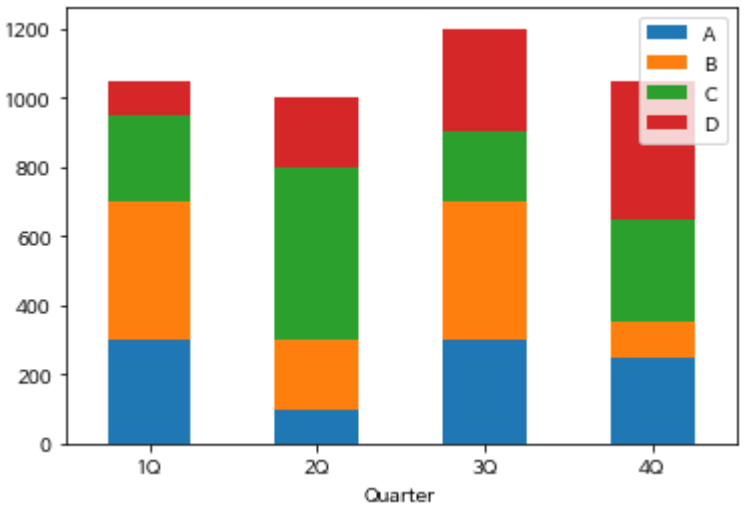
```
import pandas as pd
df = pd.DataFrame()
df['Quarter'] = ['1Q','2Q','3Q','4Q']
df['A'] = [300,100,300,250]
df['B'] = [400,200,400,100]
df['C'] = [250,500,200,300]
df['D'] = [100,200,300,400]
df.head()
```

	Quarter	A	B	C	D
0	1Q	300	400	250	100
1	2Q	100	200	500	200
2	3Q	300	400	200	300
3	4Q	250	100	300	400

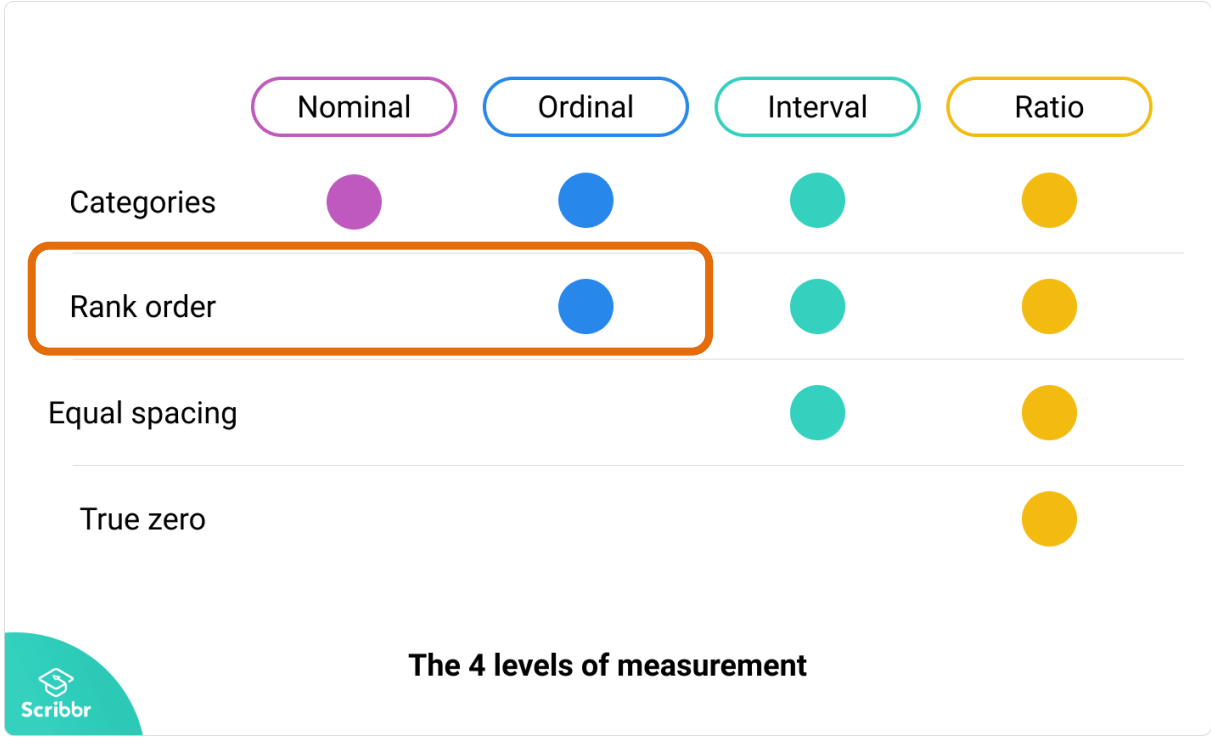



```
ax = df.plot.bar(x = 'Quarter')
```

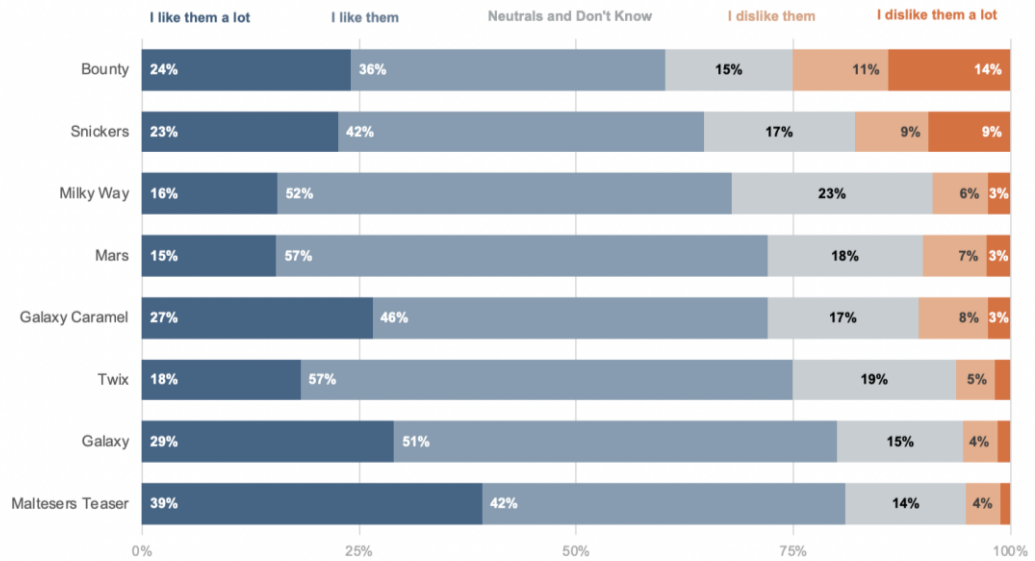
```
ax = df.plot.bar(stacked = True,\
                 x = 'Quarter', rot=0)
```

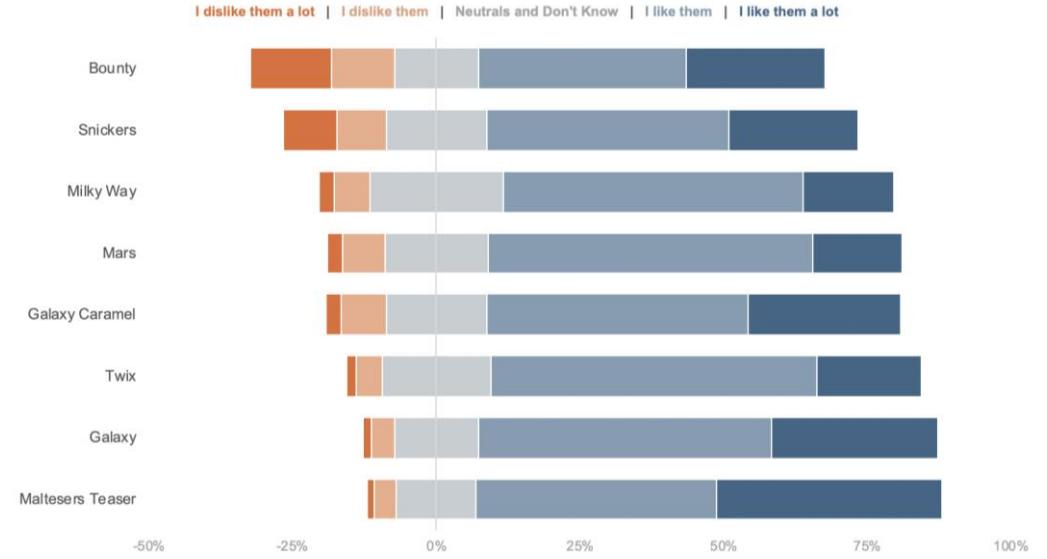


Stacked Bar for Ordinal data (e.g., Likert Scale)

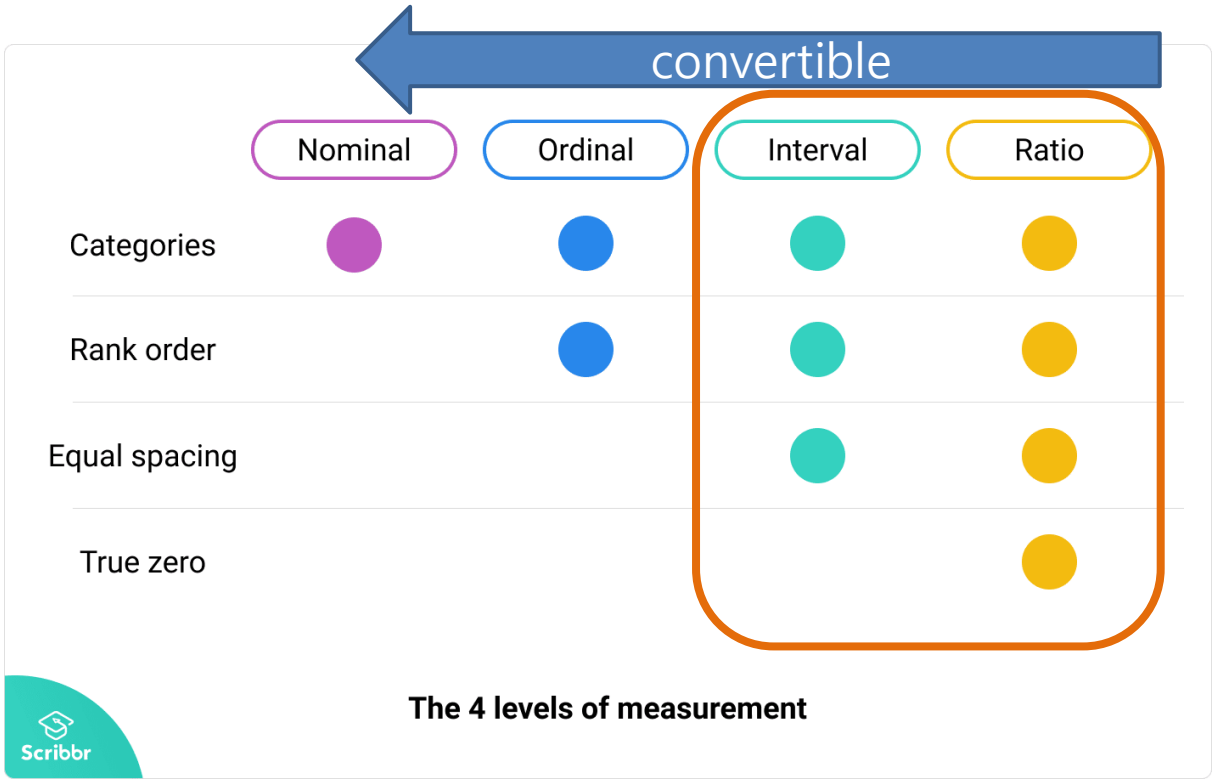


The 4 levels of measurement

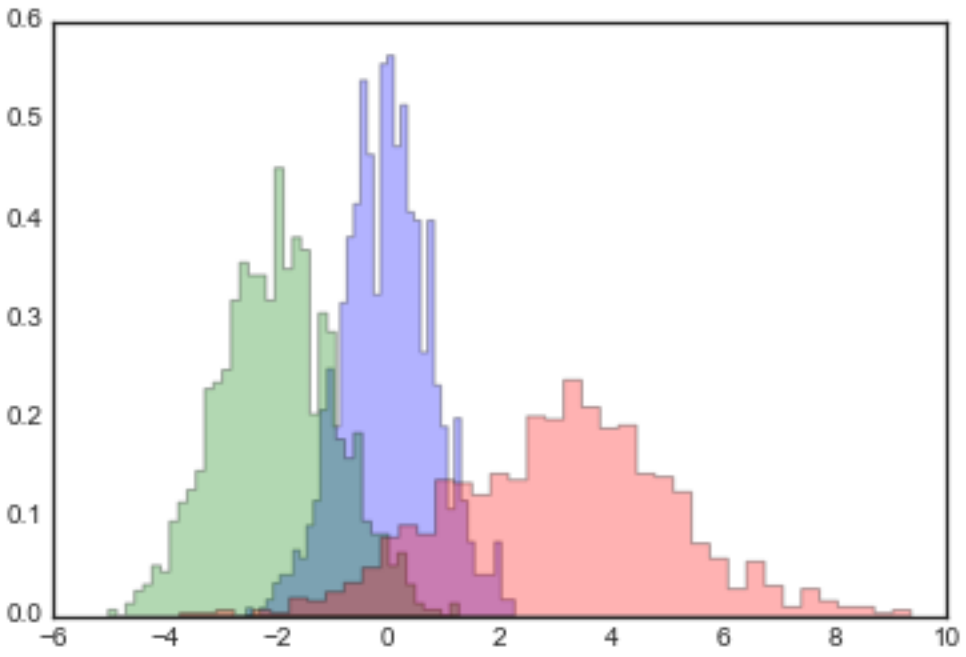




Visualizing frequency distribution table– Histogram



The 4 levels of measurement



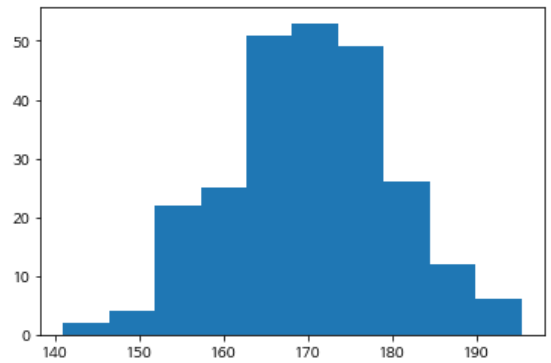
Code Examples – Histogram

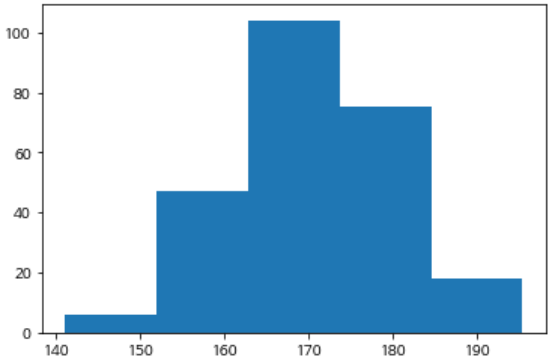
```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(170, 10, 250)
x.mean(), x.std(), x.min(), x.max()
```

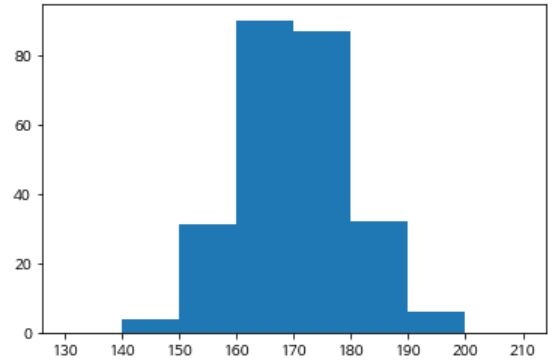
```
plt.hist(x)
plt.show()
```

```
n, edges, patch  = plt.hist(x, bins = 5)
plt.show()
```

```
plt.hist(x, bins=[130,140,150,160,
                  170,180,190,200,210])
plt.show()
```

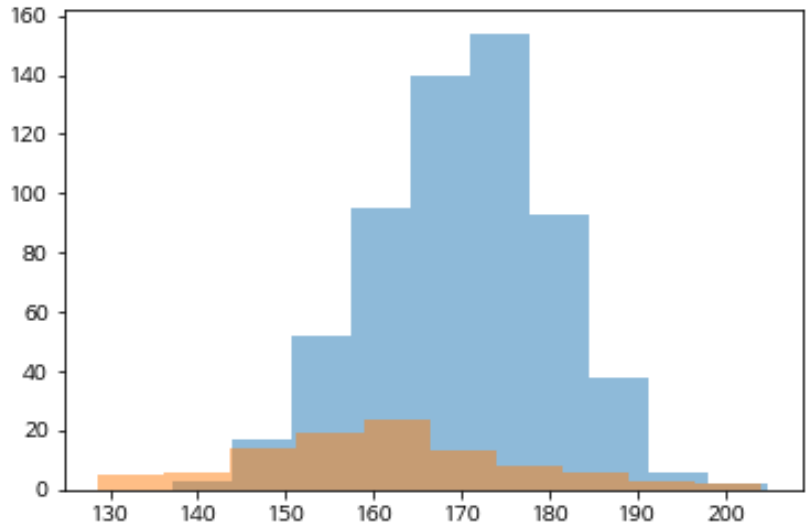




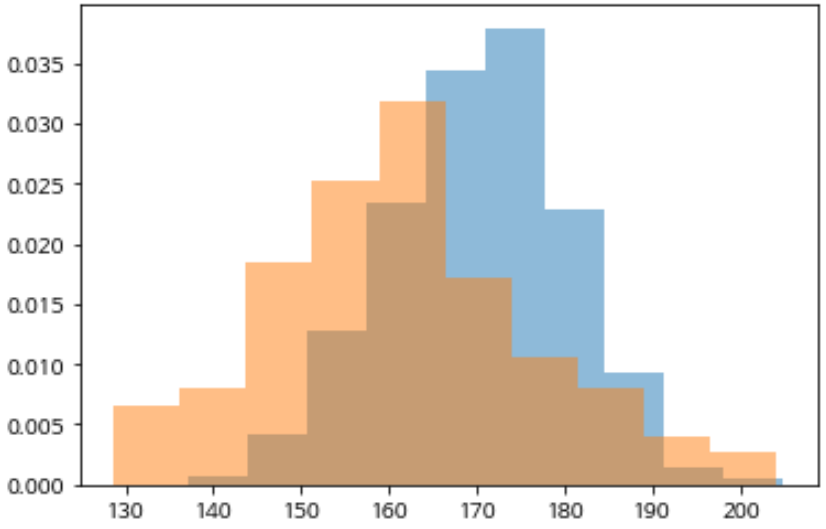


Code Examples – Histogram

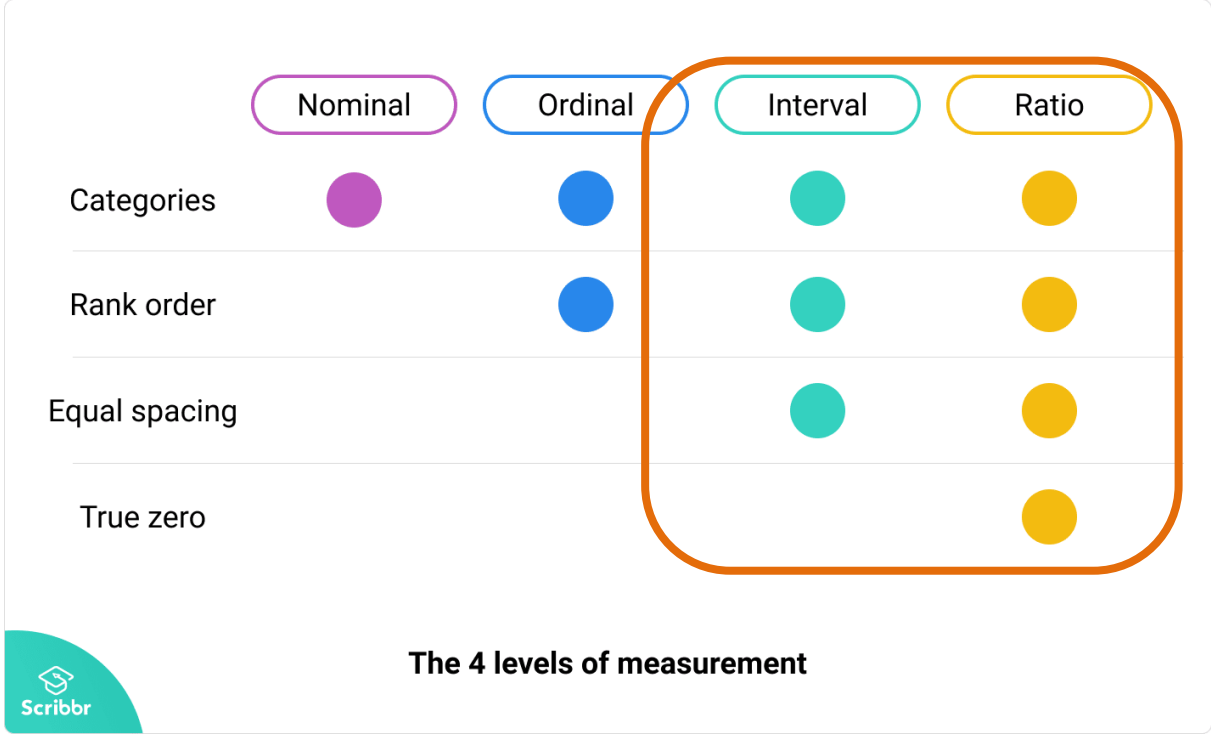
```
a = np.random.normal(170, 10, 600)
b = np.random.normal(160, 15, 100)
plt.hist(a, alpha=0.5)
plt.hist(b, alpha=0.5, density = False)
plt.show()
```

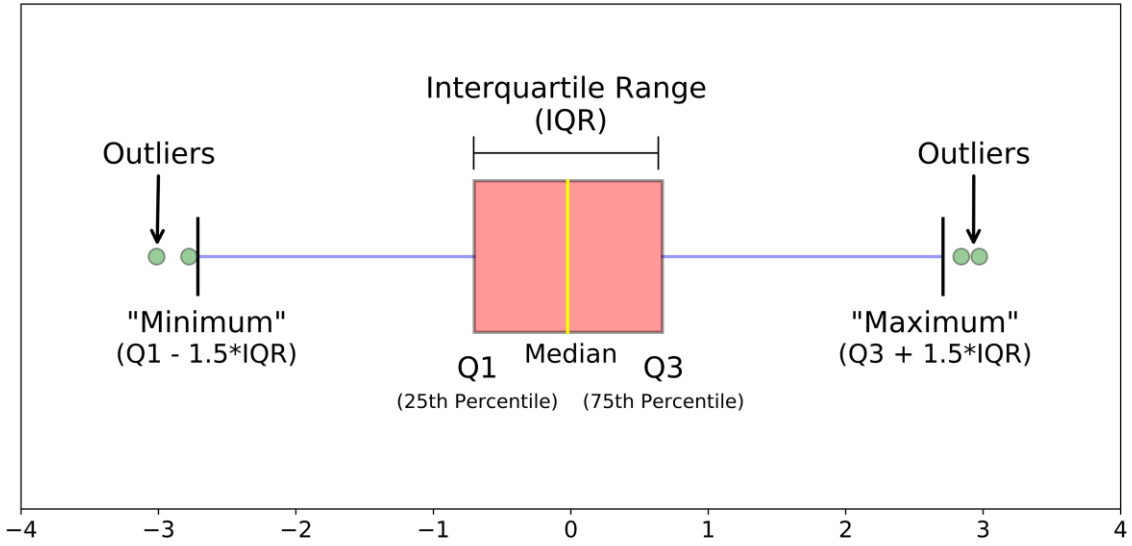



```
plt.hist(a, alpha=0.5, density = True)
plt.hist(b, alpha=0.5, density = True)
plt.show()
```



Box Plot for Numerical data





Code Examples – Boxplot

❖ Load dataset from seaborn

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
# loading dataset
iris = sns.load_dataset('iris')
iris.shape
```

```
(150, 5)
```

```
iris.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
iris.groupby('species').count()
```

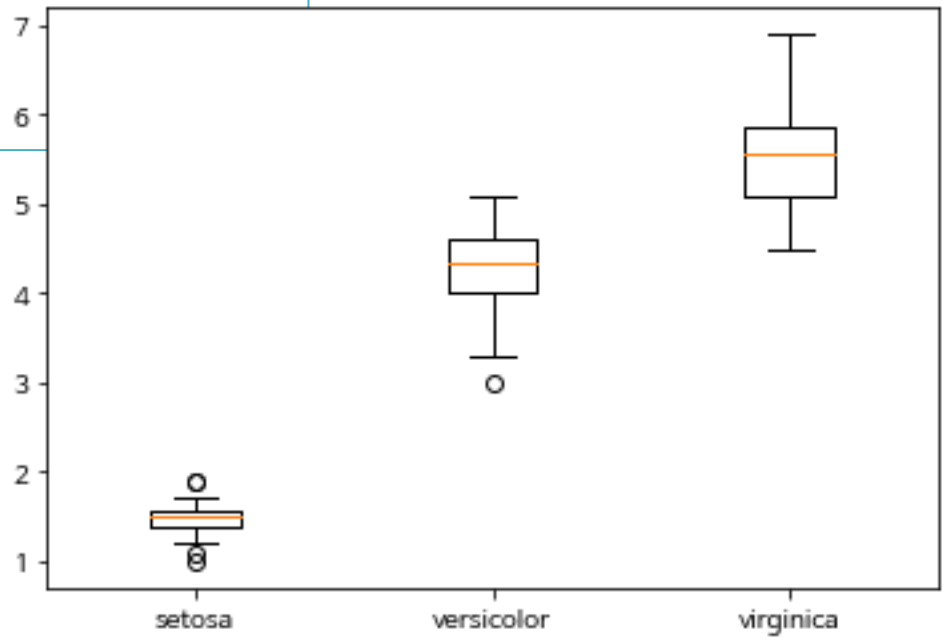
species	sepal_length	sepal_width	petal_length	petal_width
setosa	50	50	50	50
versicolor	50	50	50	50
virginica	50	50	50	50

Code Examples – Boxplot using Matplotlib

❖ Make subsets by each species

```
c1 = iris[iris['species'] == 'setosa']
c2 = iris[iris['species'] == 'versicolor']
c3 = iris[iris['species'] == 'virginica']
```

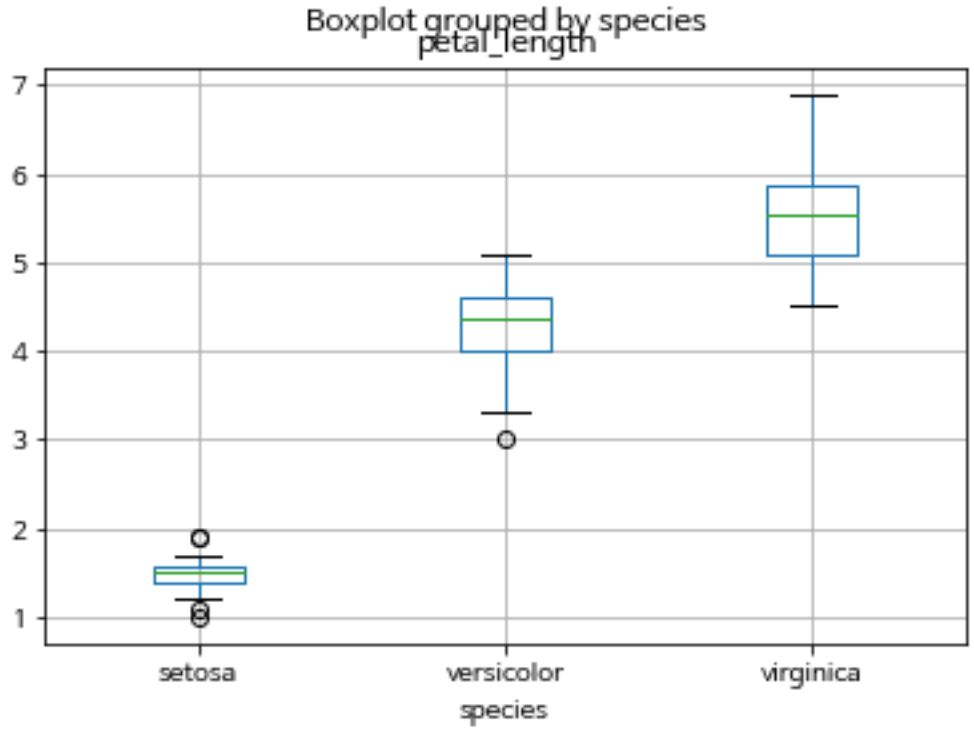
```
plt.boxplot((c1['petal_length'], c2['petal_length'],
             c3['petal_length']))
plt.xticks([1,2,3],['setosa', 'versicolor','virginica']
)
#plt.grid()
plt.show()
```

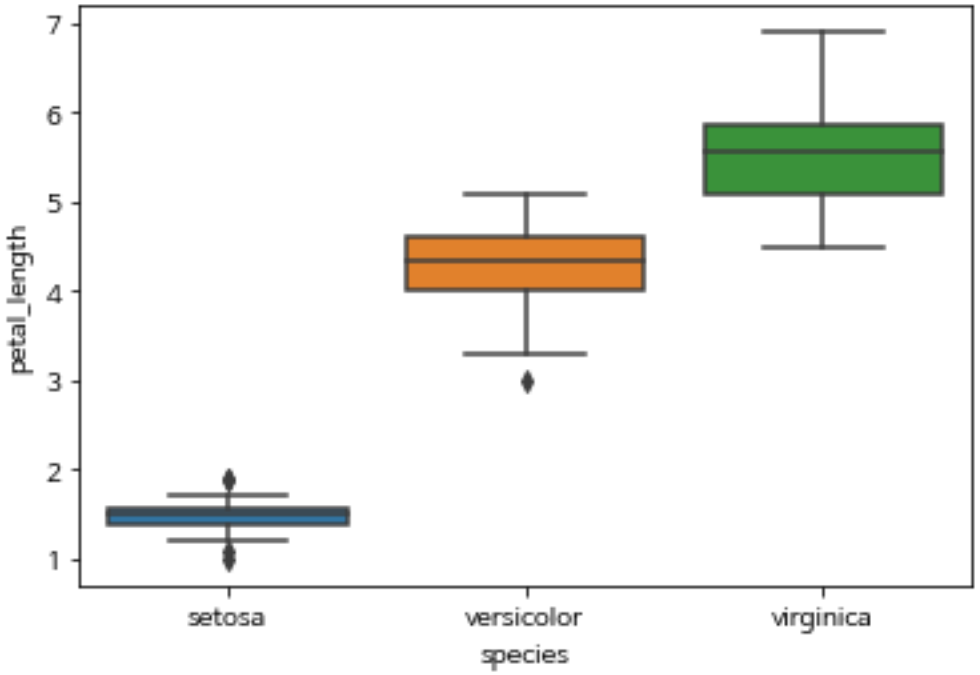


Code Examples – Boxplot using Pandas or Seaborn

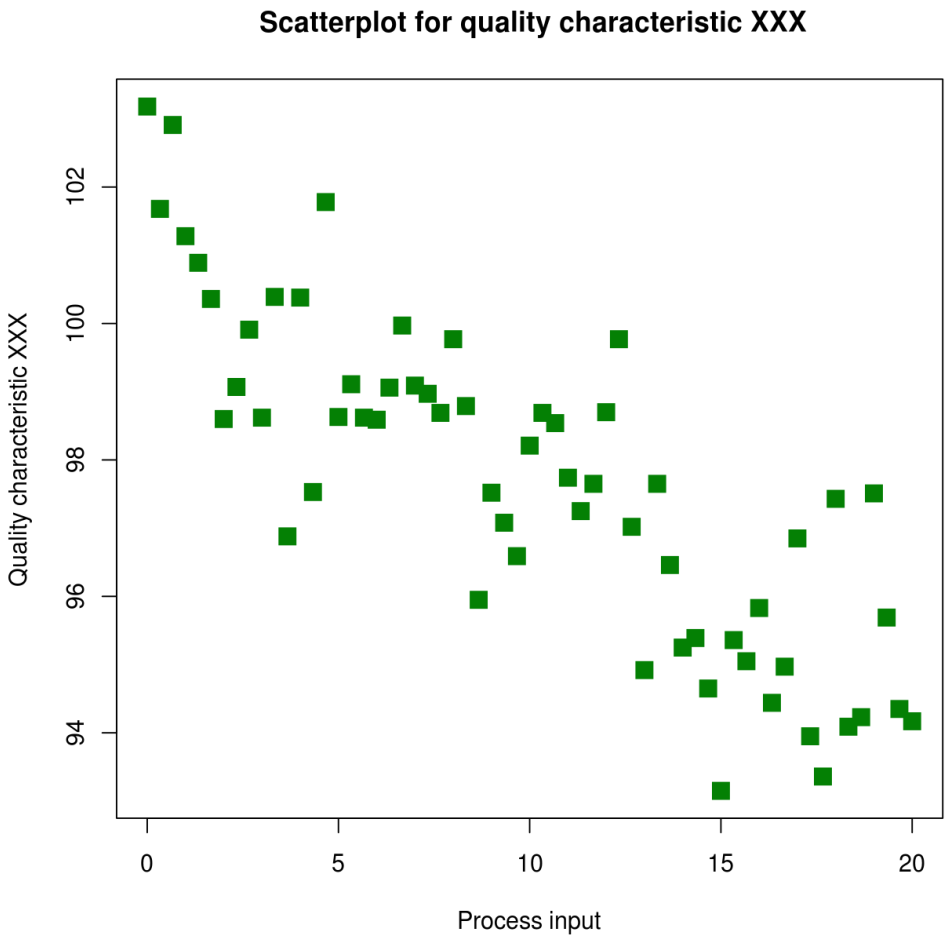
```
boxplot = iris.boxplot(column = 'petal_length', by = 'species')
```

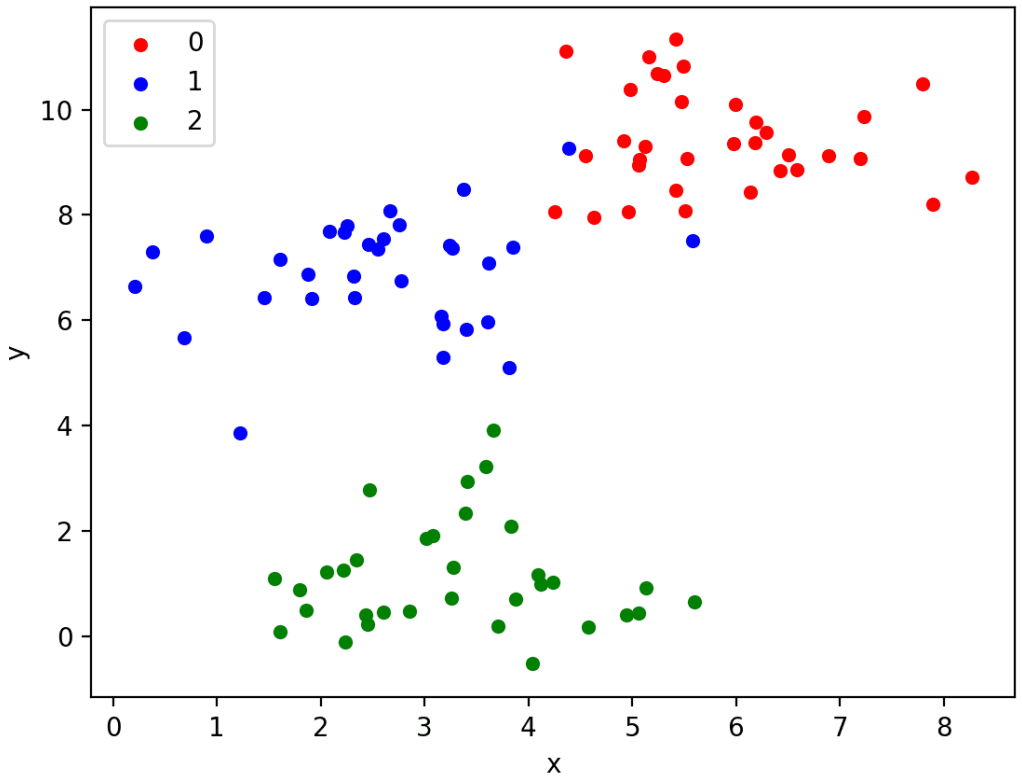
```
boxplot = sns.boxplot(data=iris, x='species', y='petal_length')
```





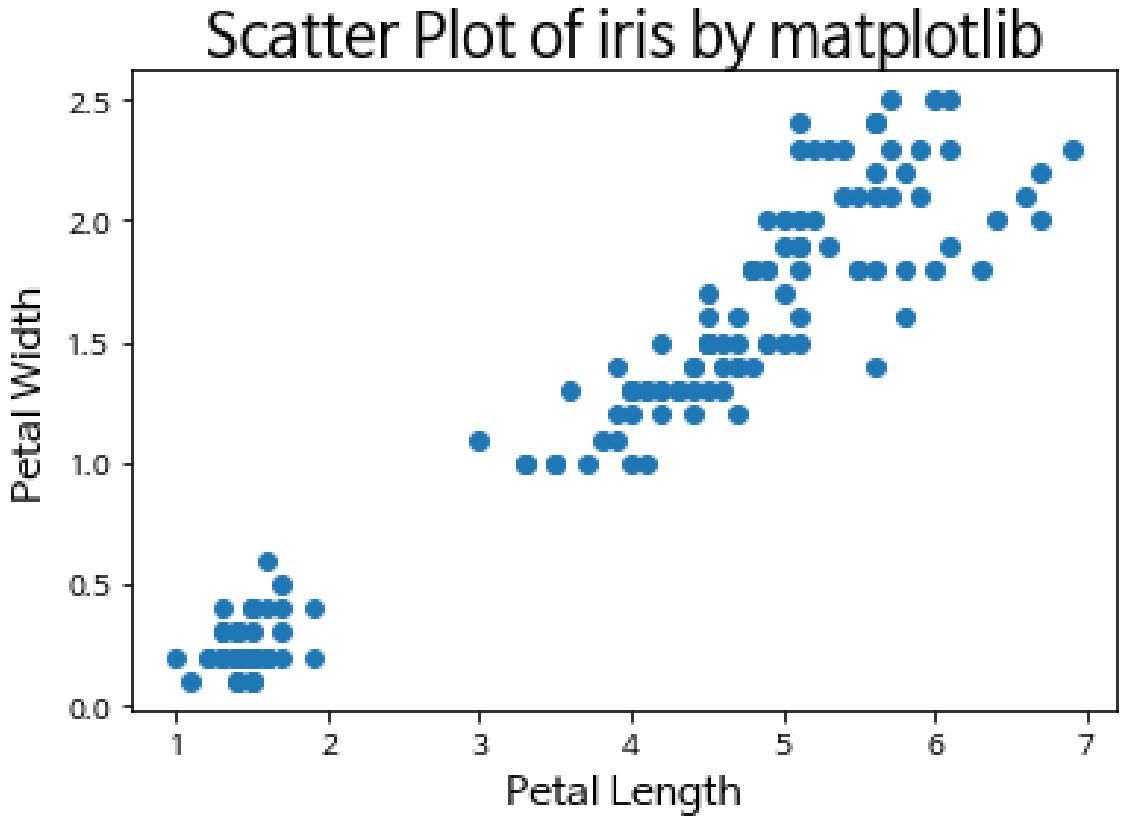
Scatter plot for two variables





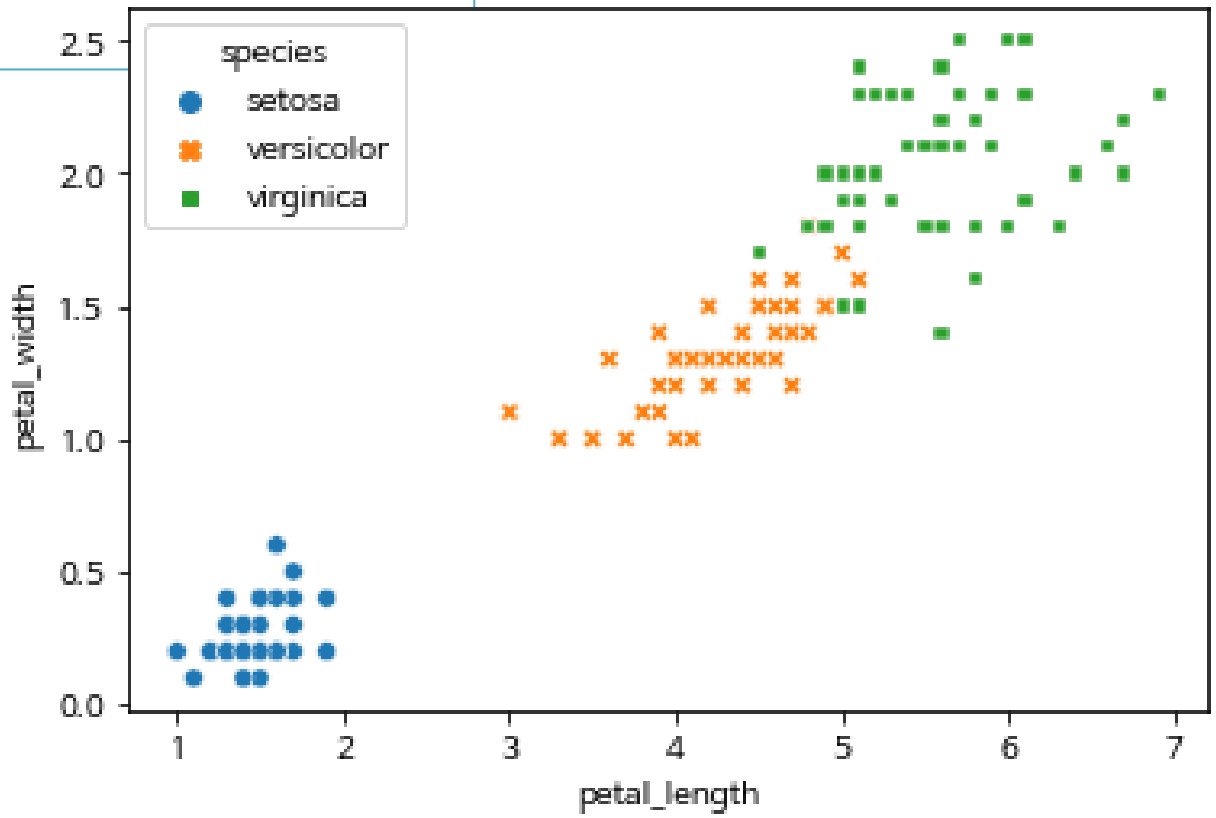
Code Examples – Scatter plot

```
plt.scatter(iris['petal_length'],iris['petal_width'])
plt.title('Scatter Plot of iris by matplotlib', fontsize=20)
plt.xlabel('Petal Length', fontsize=14)
plt.ylabel('Petal Width', fontsize=14)
plt.show()
```



Code Examples – Scatter plot using Seaborn

```
sns.scatterplot(x='petal_length',
                y='petal_width',
                hue='species',
                style='species',
                data=iris)
plt.show()
```



THANK YOU!

The Colab notebook used in this class is provided in Google Classroom.