Data Exploration

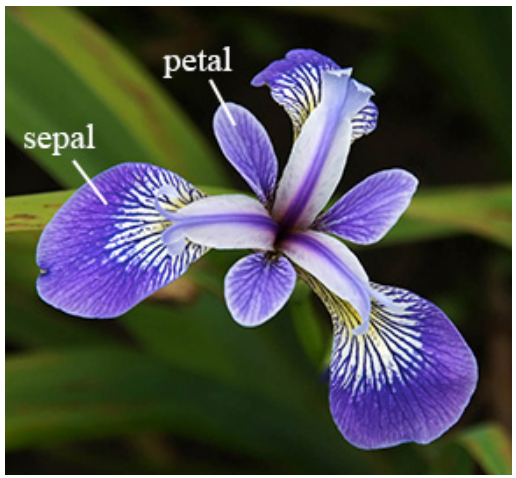
Huiping Cao

Outline

- Types of data
- Data quality
- Measurement of proximity
- Data pre-process
- Data exploration
  - Summary statistics
  - Data visualization

The Iris dataset

- University of California at Irvine (UCI) Machine Learning Repository
  `https://archive.ics.uci.edu/ml/datasets/Iris`
- 150 Iris flowers, 50 each from one of species: Setosa, Versicolour, and Virginica
- Sepals are the outer structure that protect the more fragile parts of the flower, such as the petals.
- 5 attributes
  1. sepal length in centimeters
  2. sepal width in centimeters
  3. petal length in centimeters
  4. petal width in centimeters
  5. class (Setosa, Versicolour, Virginica)



Summary Statistics

- Summary statistics are quantities such as the mean and standard deviation.
- They capture various characteristics of a large set of values with a single number or a small set of numbers.
- Example:
  - average household income
  - the fraction of college students who complete an undergraduate degree in four years

Descriptive statistics (1) - Frequency and mode

- Given a categorical attribute x which can take values $\{v_1, \cdots, v_i, \cdots, v_k\}$ and a set of m objects
- The frequency of a value v_i

$$frequency(v_i) = \frac{\text{number of objects with attribute value } v_i}{m}$$

- The mode of a categorical attribute is the value that has the highest frequency.

Descriptive statistics (1) - Example

- A set of students who have an attribute *class*. The values of class can be *freshman*, *sophomore*, *junior*, *and senior*.
- The below table shows the number of students for each value of the *class* attributes.
- The mode of the *class* attribute is *freshman* with the highest frequency 0.33.

class	size	frequency
freshman	200	0.33
sophomore	160	0.27
junior	130	0.22
senior	110	0.18

Descriptive statistics (2) - Percentile

- Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile x_p is a value of x such that $p\%$ of the observed values of x are less than x_p.
- Example: the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- Quartiles: $Q1$ (25th percentile), $Q3$ (75th percentile)
- Interquartile range: IQR = Q3 - Q1

Descriptive statistics (2) - Example

- Percentiles for sepal length, sepal width, petal length, and petal width. (All values are in centimeters.)

Percentile	Sepal length	Sepal Width	Petal length	Petal width
0	4.3	2.0	1.0	0.1
10	4.8	2.5	1.4	0.2
20	5.0	2.7	1.5	0.2
30	5.2	2.8	1.7	0.4
40	5.6	3.0	3.9	1.2
50	5.8	3.0	4.4	1.3
60	6.1	3.1	4.6	1.5
70	6.3	3.2	5.0	1.8
80	6.6	3.4	5.4	1.9
90	6.9	3.6	5.8	2.2
100	7.9	4.4	6.9	2.5

Percentile example - Python code

```
# get a list from 0.0 to 1.0 with step 0.1
index = np.arange(0.0,1.1,0.1)

# [0.  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1. ]
print(index)

# print the given percentile of all the columns
df.quantile(index)
```

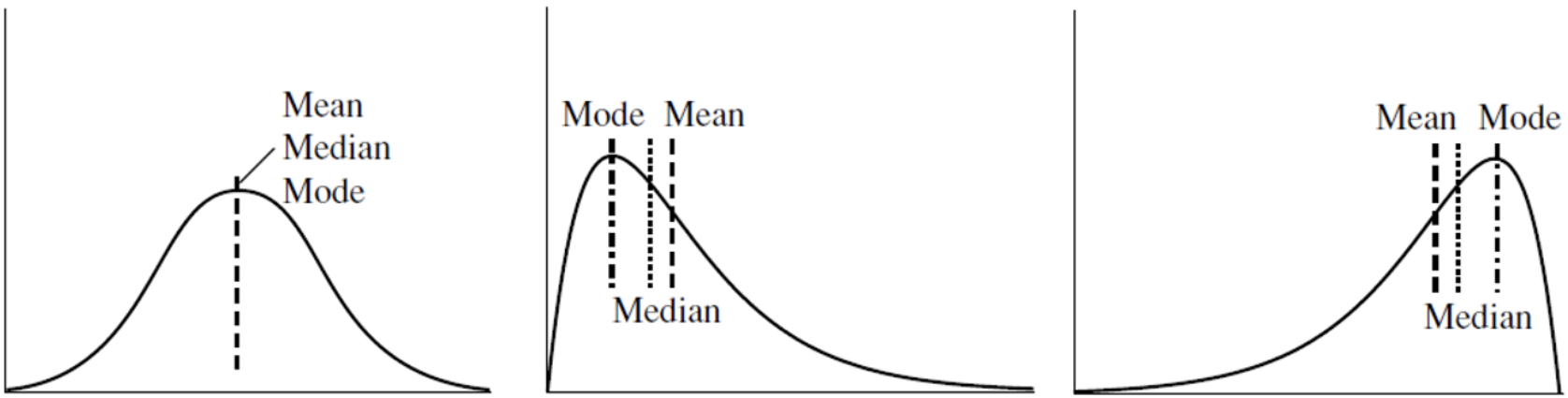
Descriptive statistics (3) - Mean and Median

- Consider a set of m objects and an attribute x. Let $\{x_1, \cdots, x_m\}$ be the attribute values of x for these m objects.
- Definitions

$$mean(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$median(x) = \begin{cases} x_{r+1} & \text{if } m \text{ is odd, i.e., m=2r+1} \\ x_r & \text{if } m \text{ is even, i.e., m=2r} \end{cases}$$

Example - Symmetric vs. Skewed Data



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Descriptive statistics (4) - Range and Variance

- Measures of spread
- Given an attribute x with a set of m values $\{x_1, \cdots, x_m\}$, is defined as
- Definitions

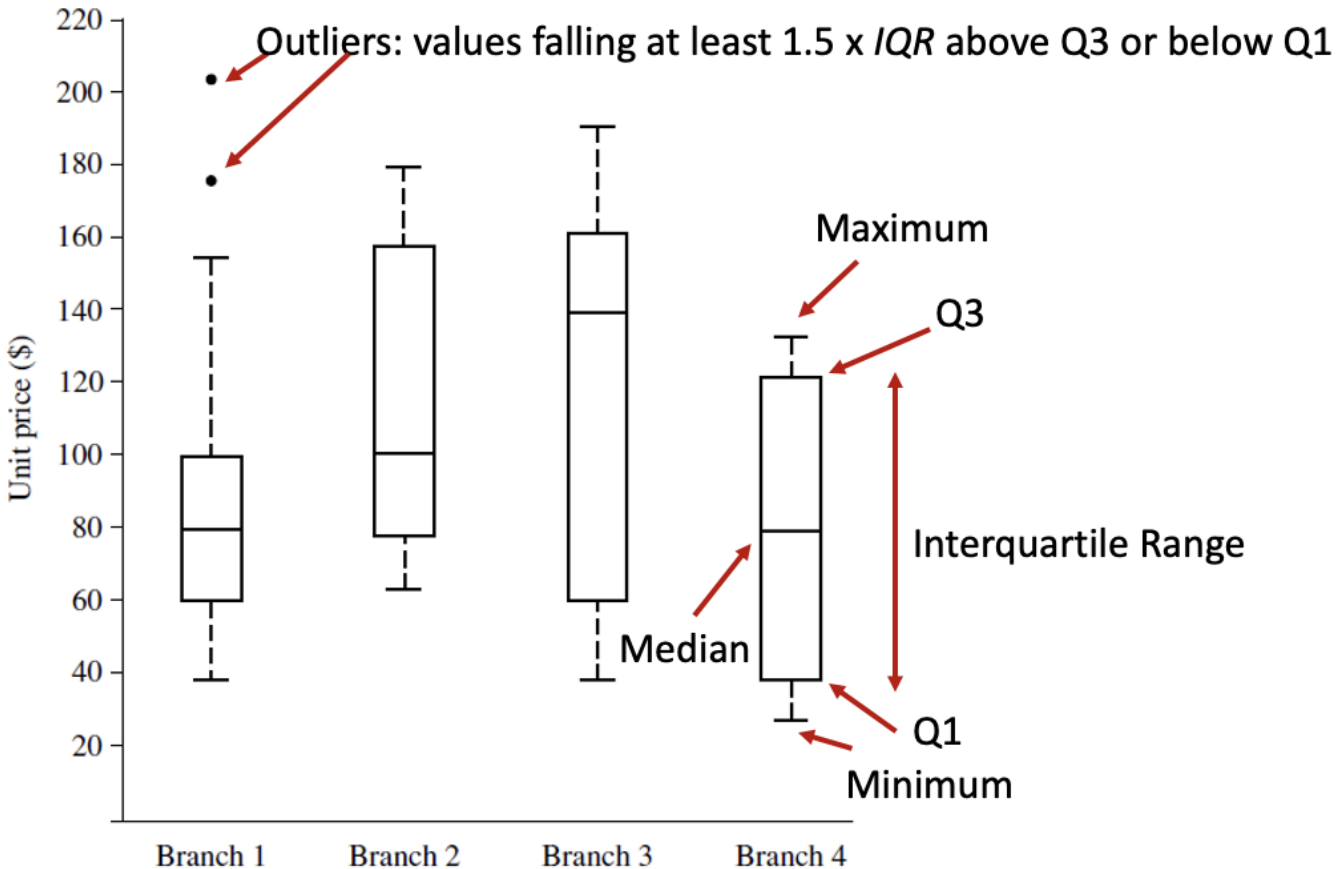
$$range(x) = max(x) - min(x)$$

$$variance(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$$

Visualizing Descriptive Statistics

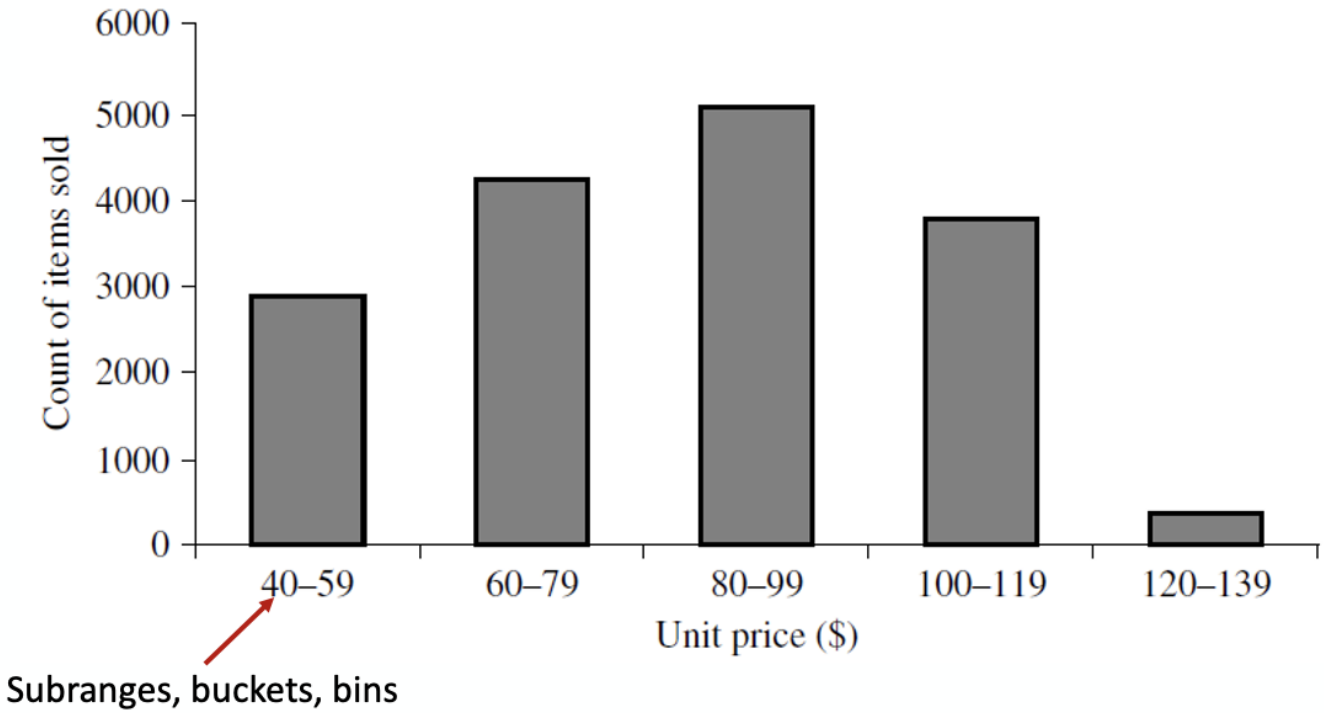
- Boxplot: graphic display of five-number summary.
- Histogram: x-axis are values, y-axis represent frequencies.
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane.

Boxplot



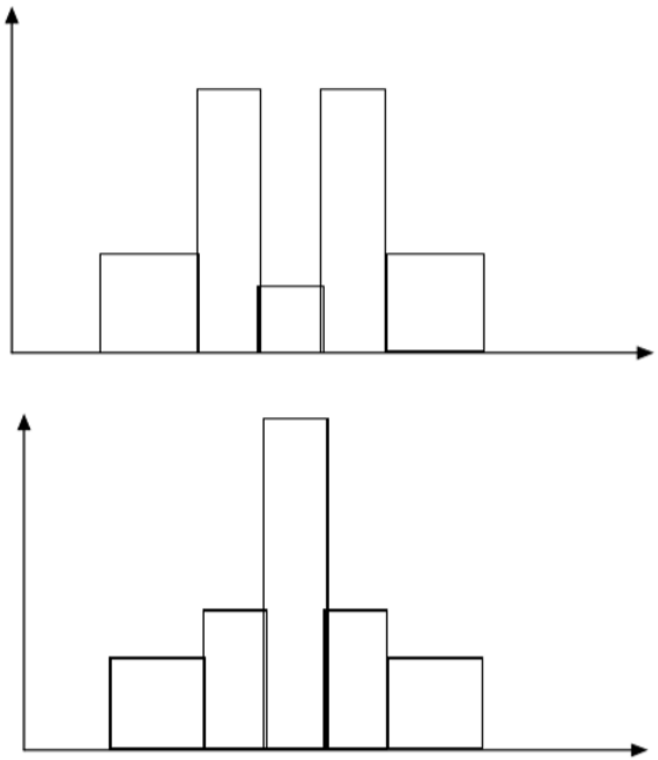
Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

Histogram


6000
5000
4000
3000
2000
1000
0
Count of items sold
40–59
60–79
80–99
100–119
120–139
Unit price ($)
Subranges, buckets, bins

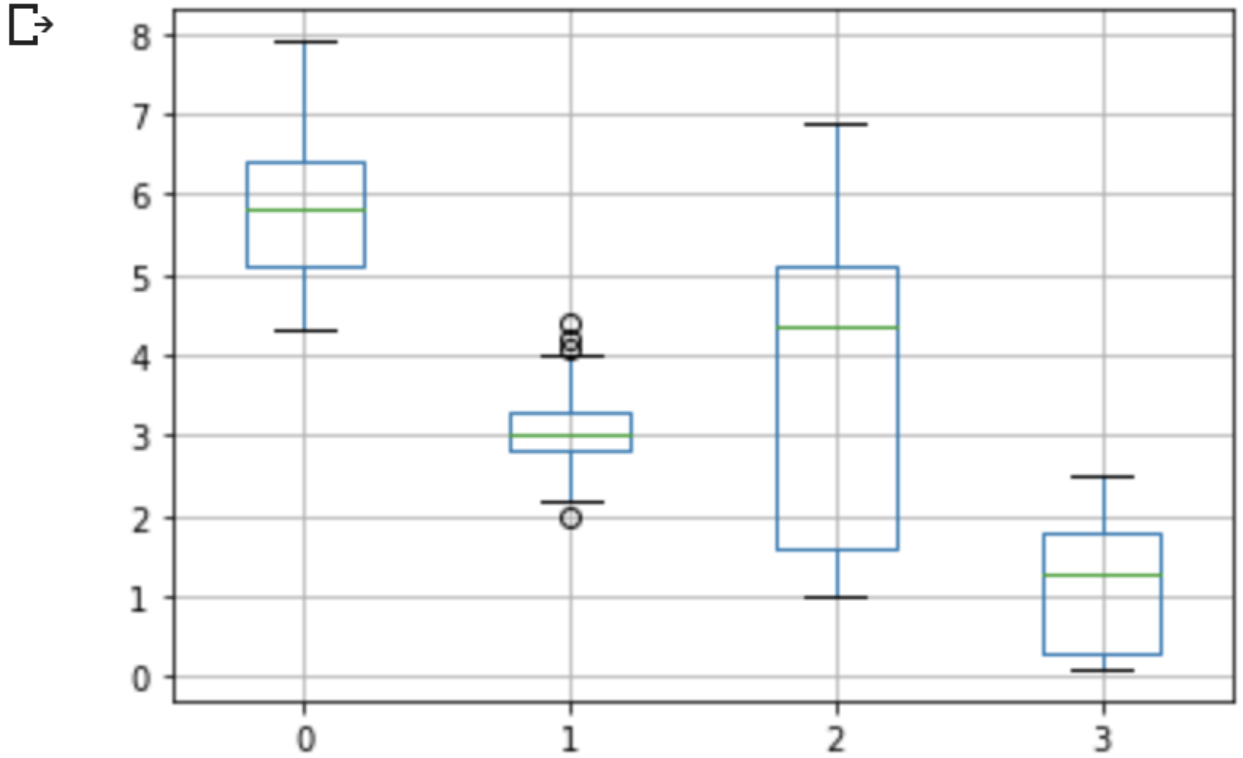
Histogram vs. Boxplot

They have the same: min, Q1, median, Q3, max However, the data distribution looks different.



Boxplot - Python code

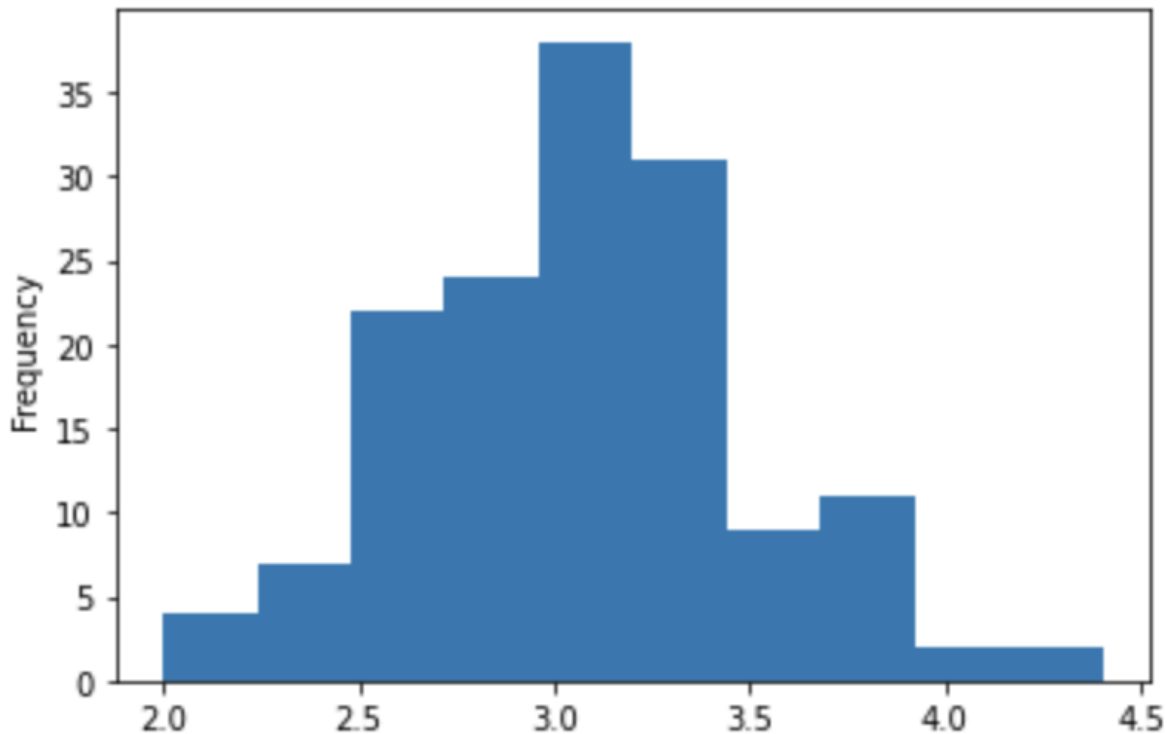
```
boxplot = df.boxplot(column=[0, 1, 2, 3])
```



Histogram - Python code

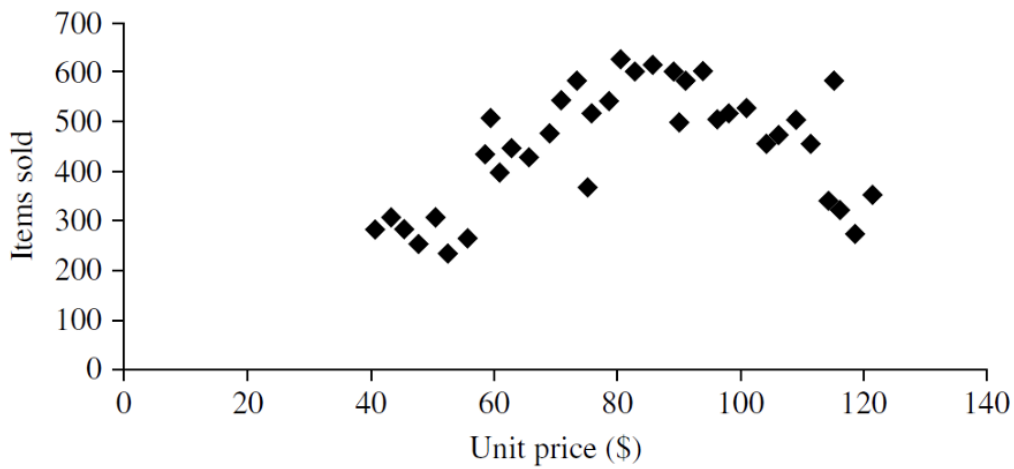
```
print(df[1].min(), df[1].max())
ax = df[1].plot.hist(bins=10)
```

```
2.0 4.4
```



Scatter Plots and Data Correlation

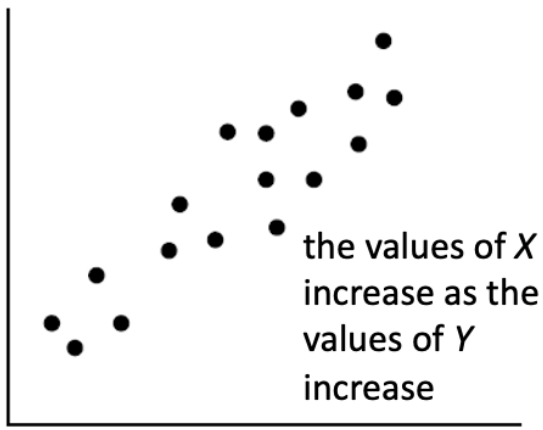
- The scatter plot is used to:
  - provide a first look at bivariate data to see clusters and outliers
  - explore the possibility of correlation relationships
  - each pair of values is treated as a pair of coordinates and plotted as points in the plane



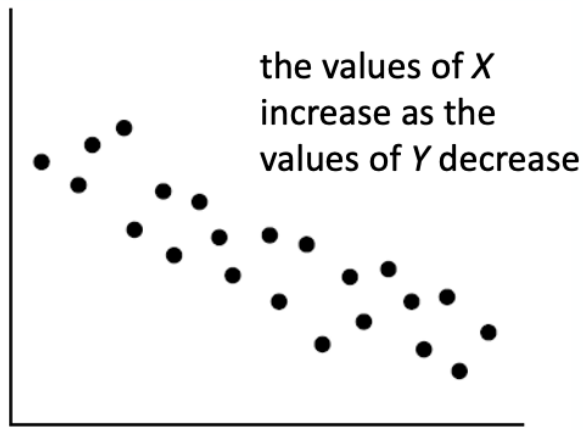
Data Correlation

- Two attributes, X, and Y, are correlated if one attribute implies the other.

- Correlations can be positive, negative, or null (uncorrelated)



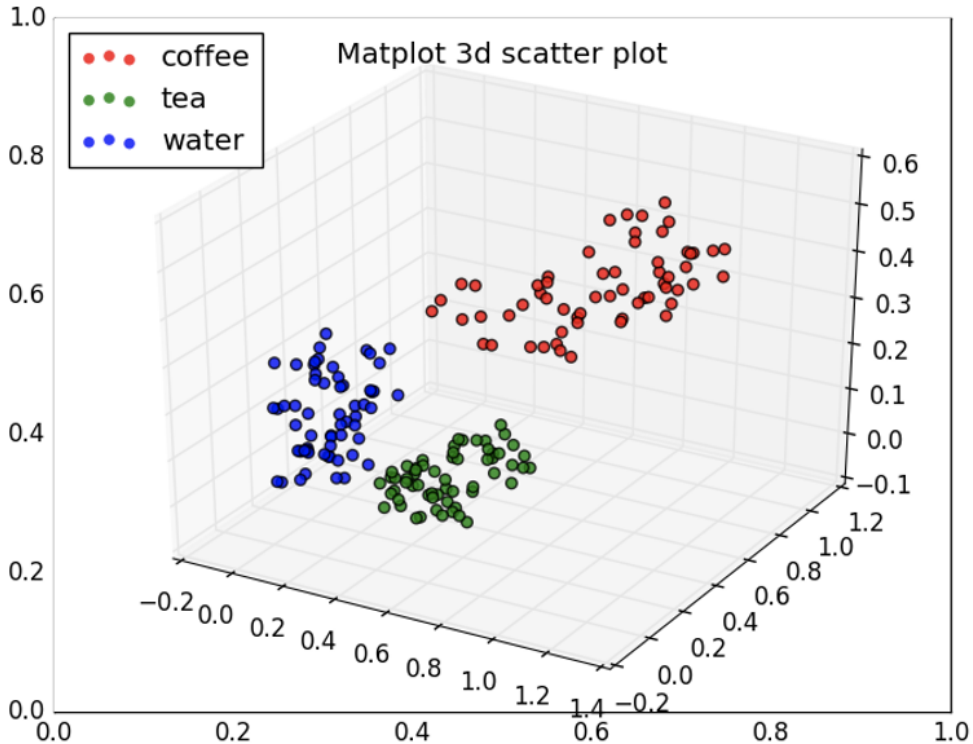
(a)



(b)

How about these scatter plots?

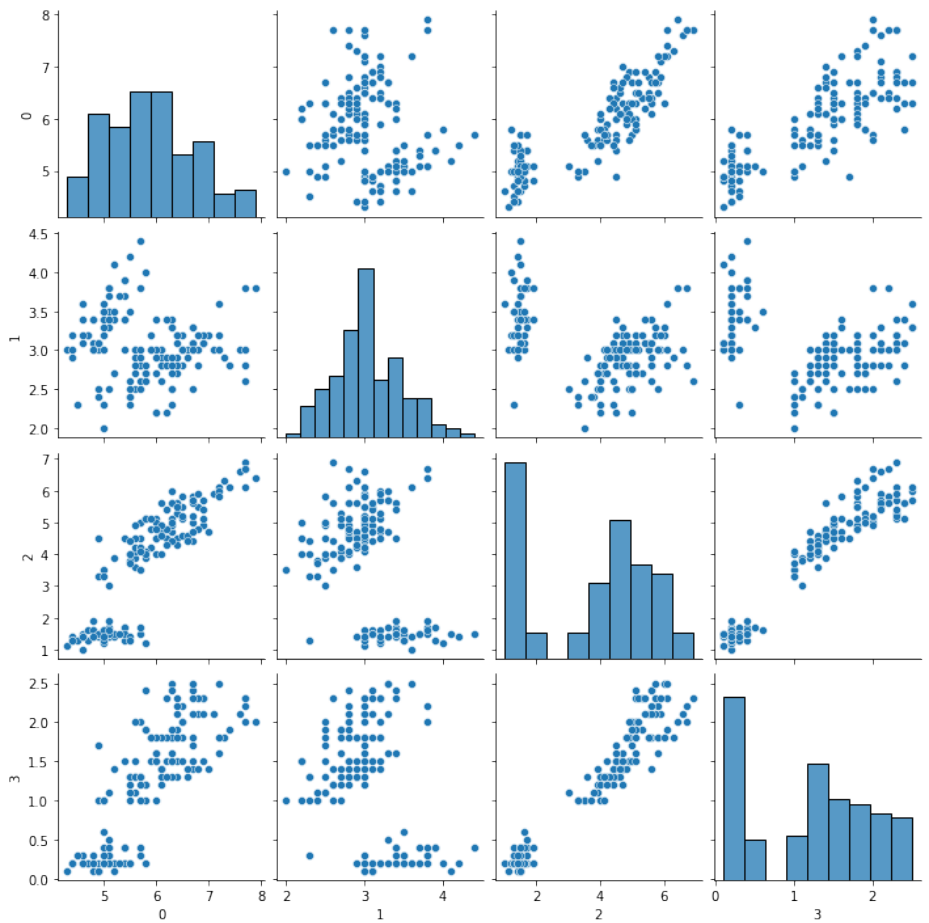
Data Visualization 3D Scatter plot



https://pythonspot.com/3d-scatterplot/

Data Visualization Scatter plot of Iris dataset

```
import seaborn as sns
sns.pairplot(df)
```

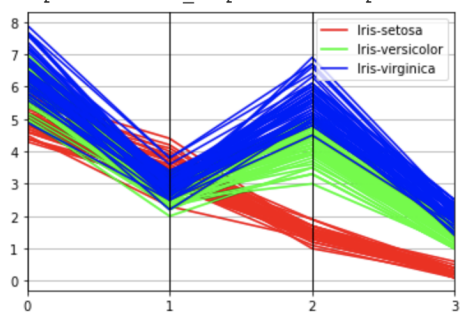


Parallel coordinates

- Parallel coordinates have one coordinate axis for each attribute, but the different axes are parallel to each other.
- One object is represented as a line instead of one point.
- The value of each attribute of an object is mapped to a point on the coordinate axis associated with that attribute, and these points are then connected to form the line that represents the object.

```
pd.plotting.parallel_coordinates(
    df, 4, color=('#FF0000', '#00FF00', '#0000FF')
)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f531c516790>
```



References

- Chapter 2: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar (`https://www-users.cs.umn.edu/~kumar001/dmbook/index.php`)
- Data Exploration: `https://www-users.cs.umn.edu/~kumar001/dmbook/data_exploration_1st_edition.pdf`
- Chapter 2: Data Mining: Concepts and Techniques (3rd Edition) by Jiawei Han, Micheline Kamber, and Jian Pei (`https://hanj.cs.illinois.edu/bk3/`)