Foundations of Data Science & Analytics: Data Exploration and Visualization

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

What is data exploration?

- A preliminary exploration of a data set to better understand its characteristics
  - "Getting your hands on the data"
- Key motivations for data exploration include
  - Selecting the right tool for preprocessing or analysis
  - Making use of human abilities to recognize patterns
    - People can recognize patterns not found immediately by data analysis tools
- In our discussion of data exploration, we focus on
  - Summary Statistics
  - Visualization

Summary Statistics

- For **discrete** features, we first want to look at summary statistics, including:
  - Frequency/count of each value
  - Mode (most frequently occurring value)
- When examining **continuous** features, we look at:
  - Location
    - Mean/median
    - Range of values: [min, max]
  - Spread
    - Variance/standard deviation

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute **Soda** and a representative population of soda drinkers, the value **Coke** occurs about 17% of the time.
- The mode of an attribute is the most frequently occurring attribute value
- The notions of frequency and mode are typically used with discrete data

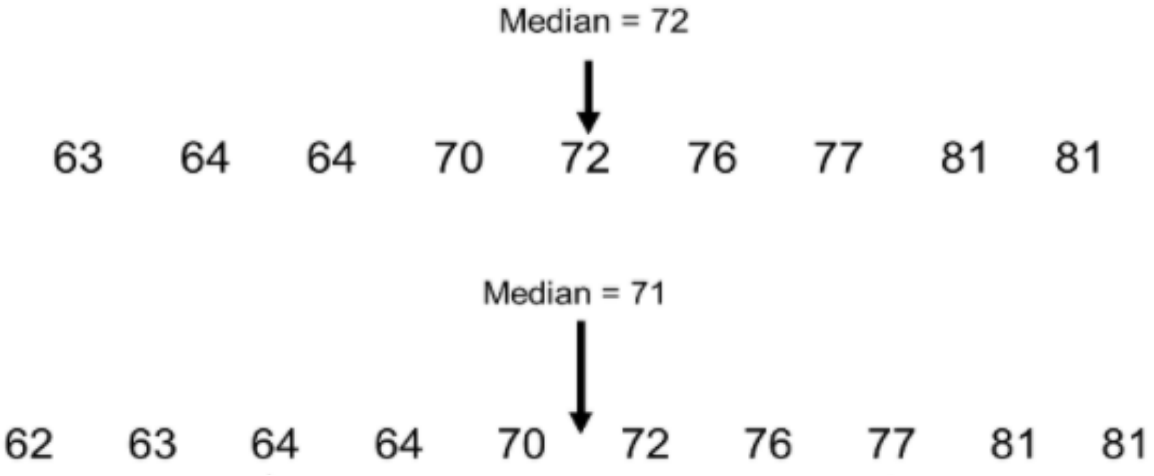
Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median is commonly used.
  - Note: the data must be sorted to compute the median!

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

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

Median Examples



Measures of Spread: Range and Variance

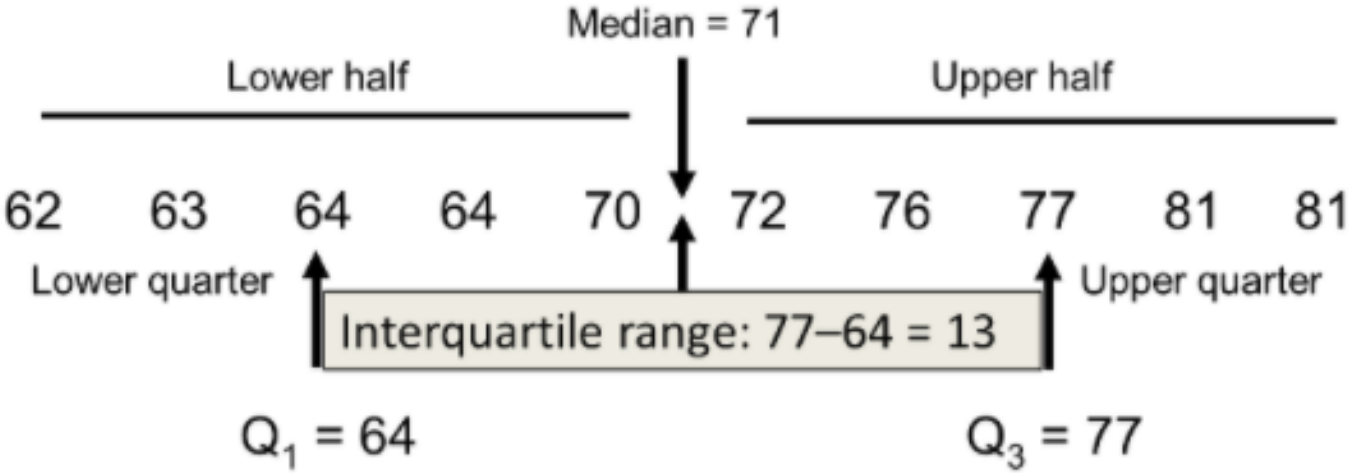
- Range is defined as the difference between the maximum and minimum values in the data set
- Variance and standard deviation are another common measure of the spread of a set of points.

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

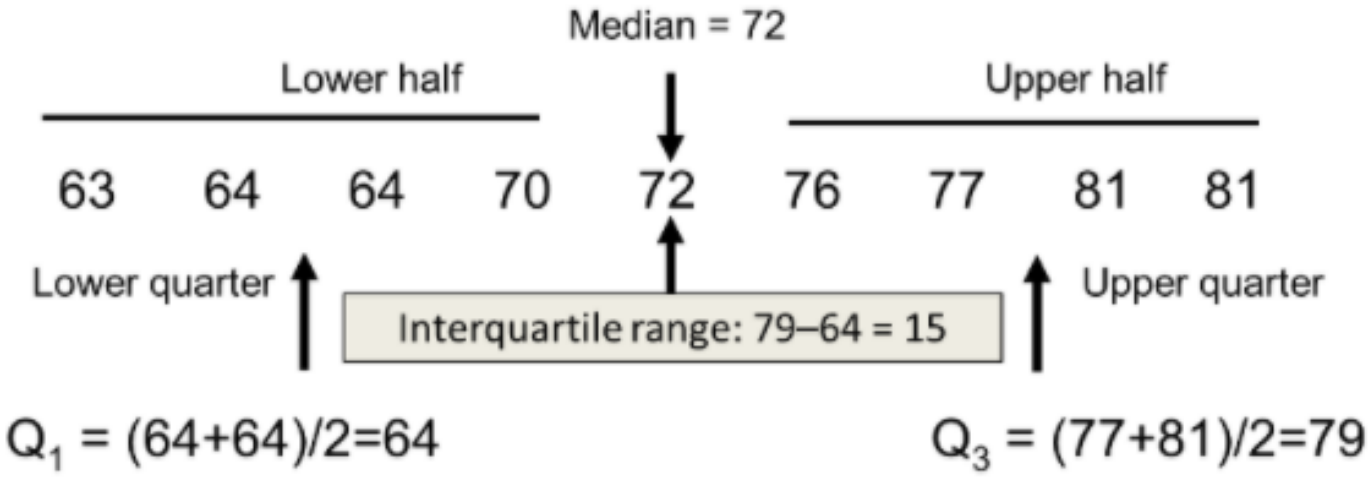
- However, both are also sensitive to outliers, so IQR (interquartile range) is often used.

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

IQR w/Even Number of Samples

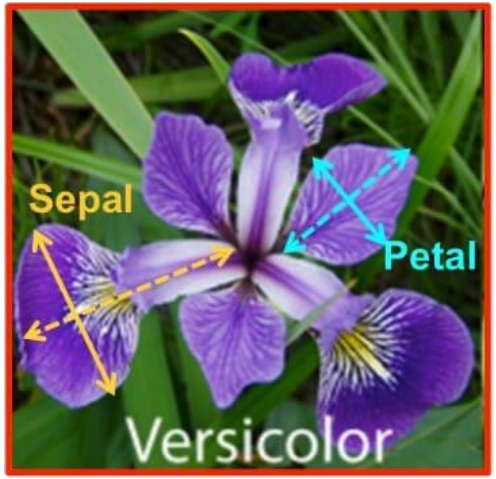


IQR w/Odd Number of Samples



Iris Data Set

- Contains 150 data instances
- Divided into three flower types (classes):
  - Setosa
  - Virginica
  - Versicolor
- Has four (non-class) features
  - Sepal width and sepal length
  - Petal width and petal length



Iris Data Set Summary Statistics

Name	Min	Max	Mean	Std
petal_length	1	6.9000	3.7587	1.7644
petal_width	0.1000	2.5000	1.1987	0.7632
sepal_length	4.3000	7.9000	5.8433	0.8281
sepal_width	2	4.4000	3.0540	0.4336

count_virginica	50
count_versicolor	50
count_setosa	50

Summary statistics don't tell the whole story. We need to visualize the data to understand it.

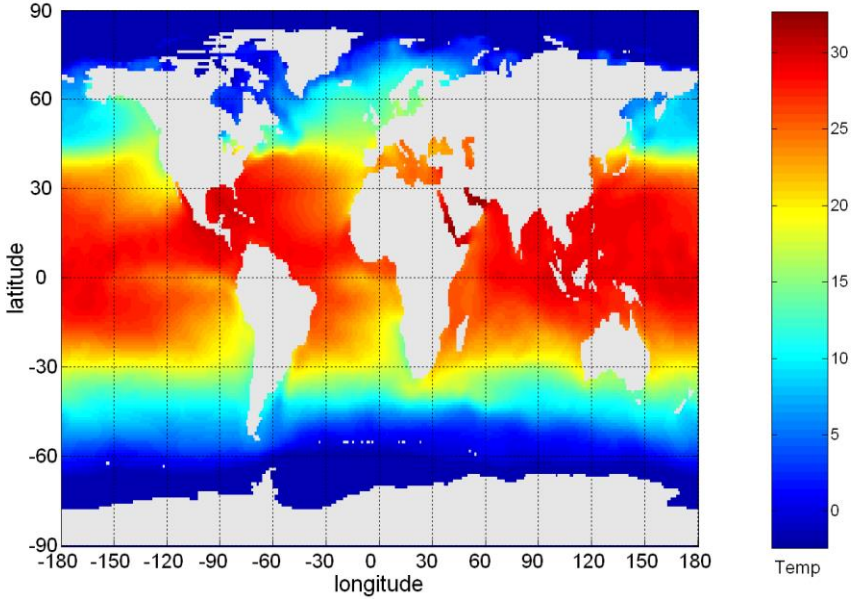
Visualization

- Conversion of data into a visual form so that the characteristics of the data and the relationships among data instances or attributes can be analyzed or reported.

- Visualizing data is one of the most powerful and techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Classic Visualization Example

- Sea Surface Temperature for July 1982
  - Tens of thousands of data points are summarized in a single figure

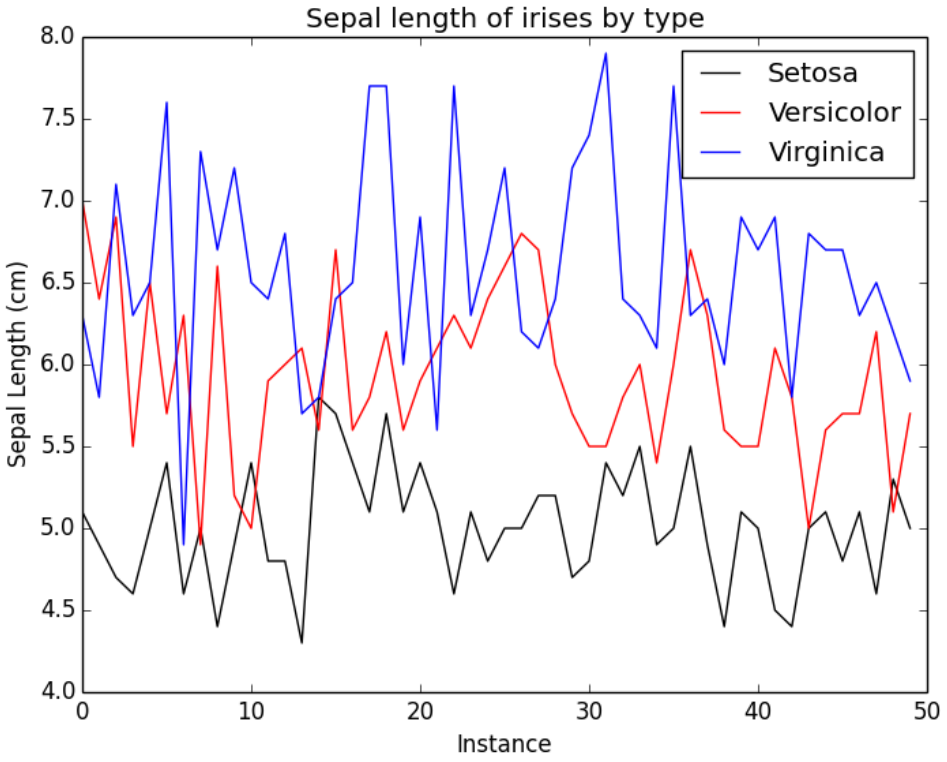


Basic Data Visualizations

- Line plot
- Scatter plot
- Histogram
- Box plot

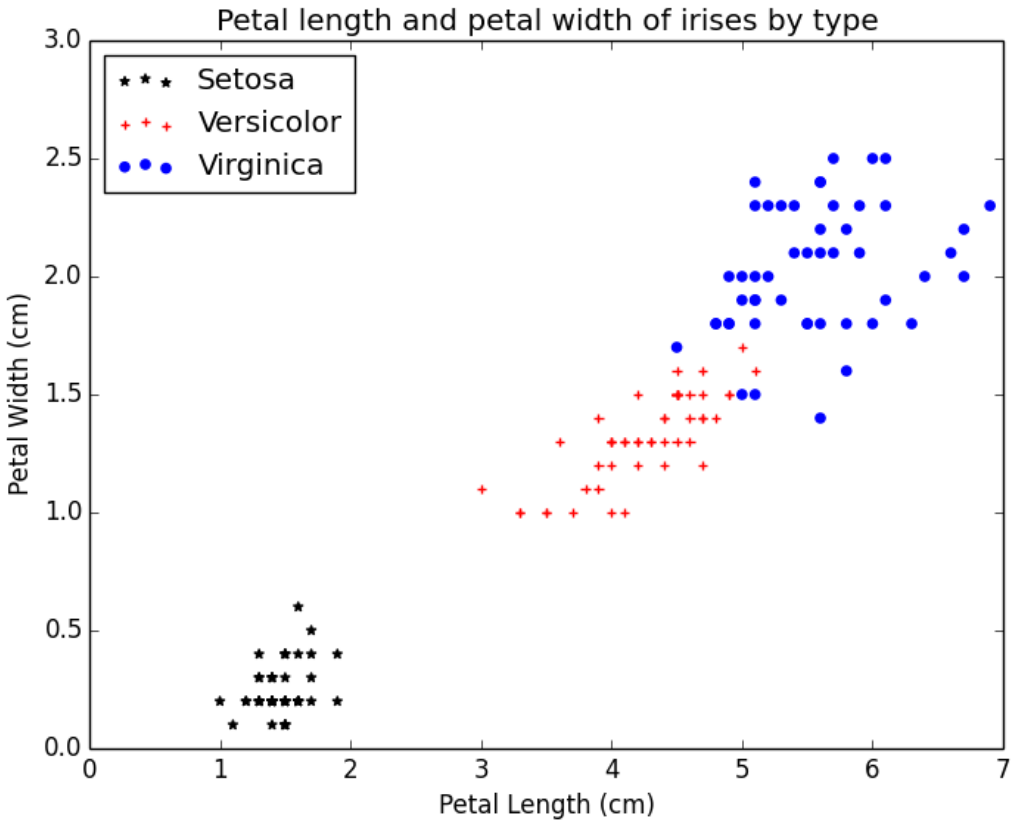
Line Plot

Useful for plotting one feature at a time!



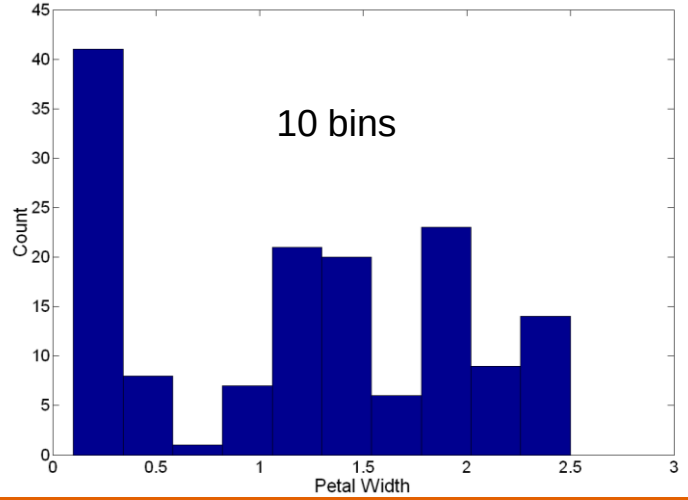
Scatter Plot

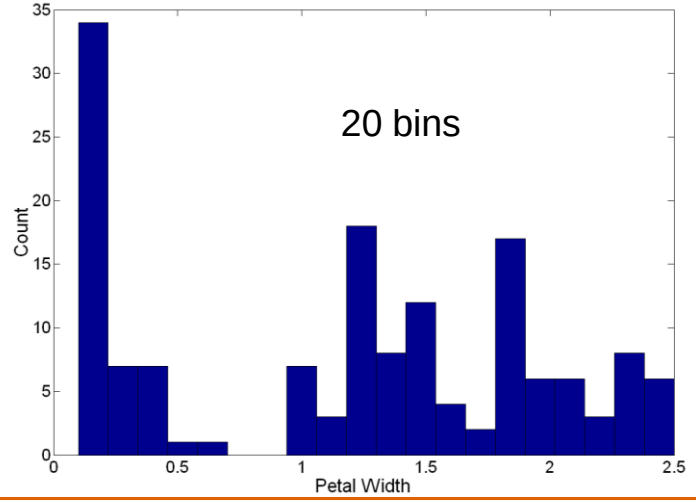
Useful for plotting two features at a time!



Histograms

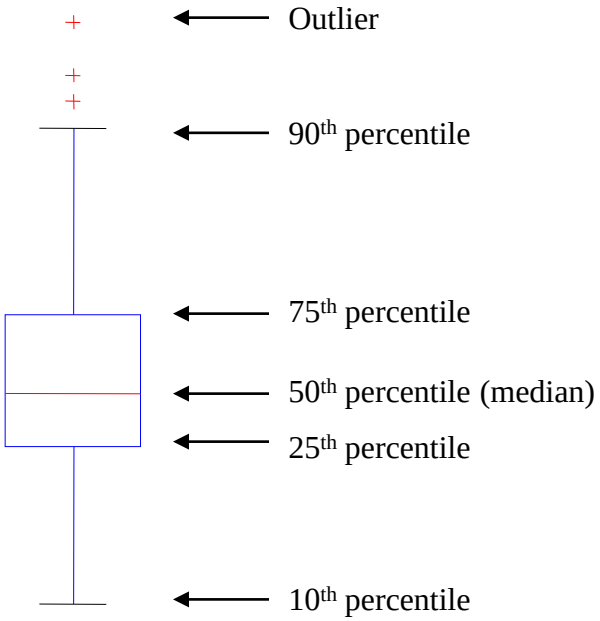
- Shows the distribution of values for a single feature
- Divide the values into bins and show a bar plot of the number of objects in each bin. The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins





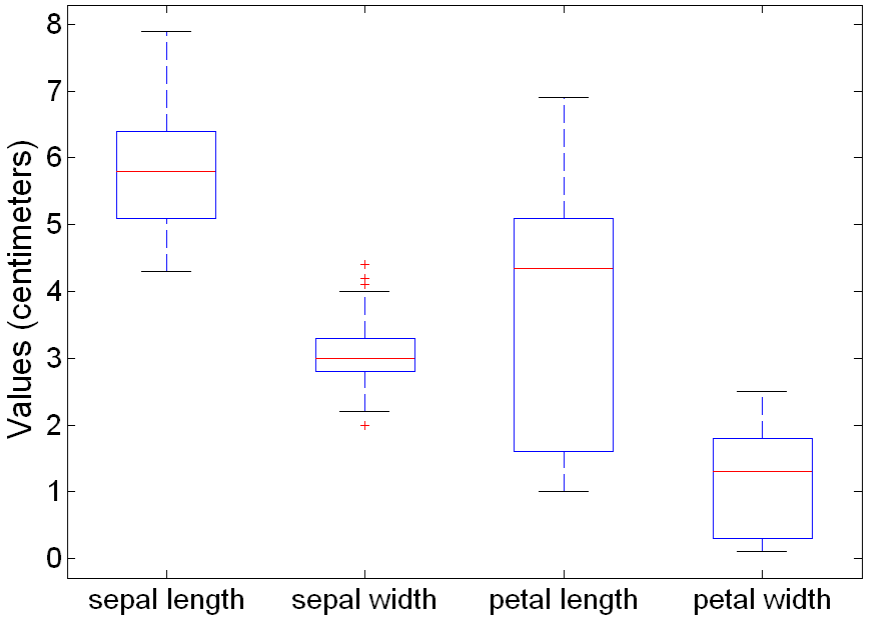
Box Plots

- Provides a simple graphical depiction of interquartile range and outliers



Box Plot Example

- Box plots can be used to compare features across classes or across the entire data set



Data Visualization in Python

- Visualization libraries are available in Python, including matplotlib and scikit-learn

- See the Data Visualization.ipynb for examples