Data pre-processing operations

M. Cristina

SCC5836/0252 Visualização Computacional

Data sources

- Sensors, IoT, measurements, collection, simulations, computations, digital transactions, ...
- **Raw data**: as obtained/collected
- **Curated data**: went through processes for organization, rescaling, smoothing, ...

Data transformations

- Select from the table the attribute(s) you want/need to use in a visualization

- Representing the values as required/suitable by the visualization technique

Data transformations

- Select from the table the attribute(s) you want/need to use in a visualization

- Representing the values as required/suitable by the visualization technique

  – an overview of which are the most common data transformations/data processing strategies

Terminology: data variables

- Phenomena are described by multiple variables/attributes
  - Independent variable: its values are not determined/affected by another variable
  - Dependent variables: its values are determined/affected by other variable(s)

Terminology: data variables

- Example: you have data from a study designs to test whether changes in room temperature have an effect on math test scores.
  - The **independent variable** is the temperature of the room. The experimenter varied the room temperature by making it cooler for half the participants, and warmer for the other half.
  - The **dependent variable** is math test scores. The experimenter measured the math skills of all participants using a standardized test and checked whether their grades differ based on room temperature.

Terminology: data variables

- Example: you have data from a study designed to test whether changes in room temperature have an effect on math test scores.
  - The **independent variable** is the **temperature** of the room. The experimenter varied the room temperature by making it cooler for half the participants, and warmer for the other half.
  - The **dependent variable** is **math test scores**. The experimenter measured the math skills of all participants using a standardized test and checked whether their grades differ based on room temperature.

Terminology: metadata

- Data about the data
  - Units, resolution
  - Description,
  - ..
- Semantics and interpretation
  - Ex.
  - Scientific paper: which is data and which is metadata?
  - Table: which is data and which is metadata?
  - Image: which is data and which is metadata?

Data transformation/preprocessing

- Typically, we do not visualize the raw data: some preprocessing/transformation is likely necessary

- ??

Data preprocessing

- Typically, we do not visualize the raw data: some preprocessing is likely necessary
  - Sampling, normalization, handling missing values, handling outliers, handling errors, interpolation, attribute selection, identifying correlation between attributes, reducing dimensionality...

Data preprocessing

- Typically, we do not visualize the raw data: some preprocessing is likely necessary

- Data profiling: `know´ about the data, e.g., type, organization, statistical distribution of variables/attributes, correlations,

Steps in data preprocessing

- Ex. to build a classification model from the data...
  - Ex. creditcard.csv
  - Kaggle: credit card transactions
  - Task is binary classification: fraud/not fraud
  - https://towardsdatascience.com/data-pre-processing-techniques-you-should-know-8954662716d6

- “Here are the steps I have followed:”
  1. Import libraries
  2. Read data
  3. Checking for missing values
  4. Checking for categorical data
  5. Standardize the data
  6. PCA transformation
  7. Data splitting

Data pre-processing tasks

- Checking for & handling missing values
- Checking for & handling categorical data (data mining)
- Verify distribution of variables - check for anomalies and outliers
- Feature scaling (normalization, standardization)
- Assessing data (dis)similarity
- Assess correlation between variables/attributes
- Dimension reduction (e.g. PCA transformation)
- Data sub-sampling
- Data aggregation
- Data interpolation
- Data splitting (e.g. for data mining model learning)

Data pre-processing tasks

- Checking for & handling missing values

- What if a value is missing?
  - Discard item?
  - Signal with a `sentinel´ (do not include in computations!)
  - Data imputation: assign a value

Data pre-processing tasks

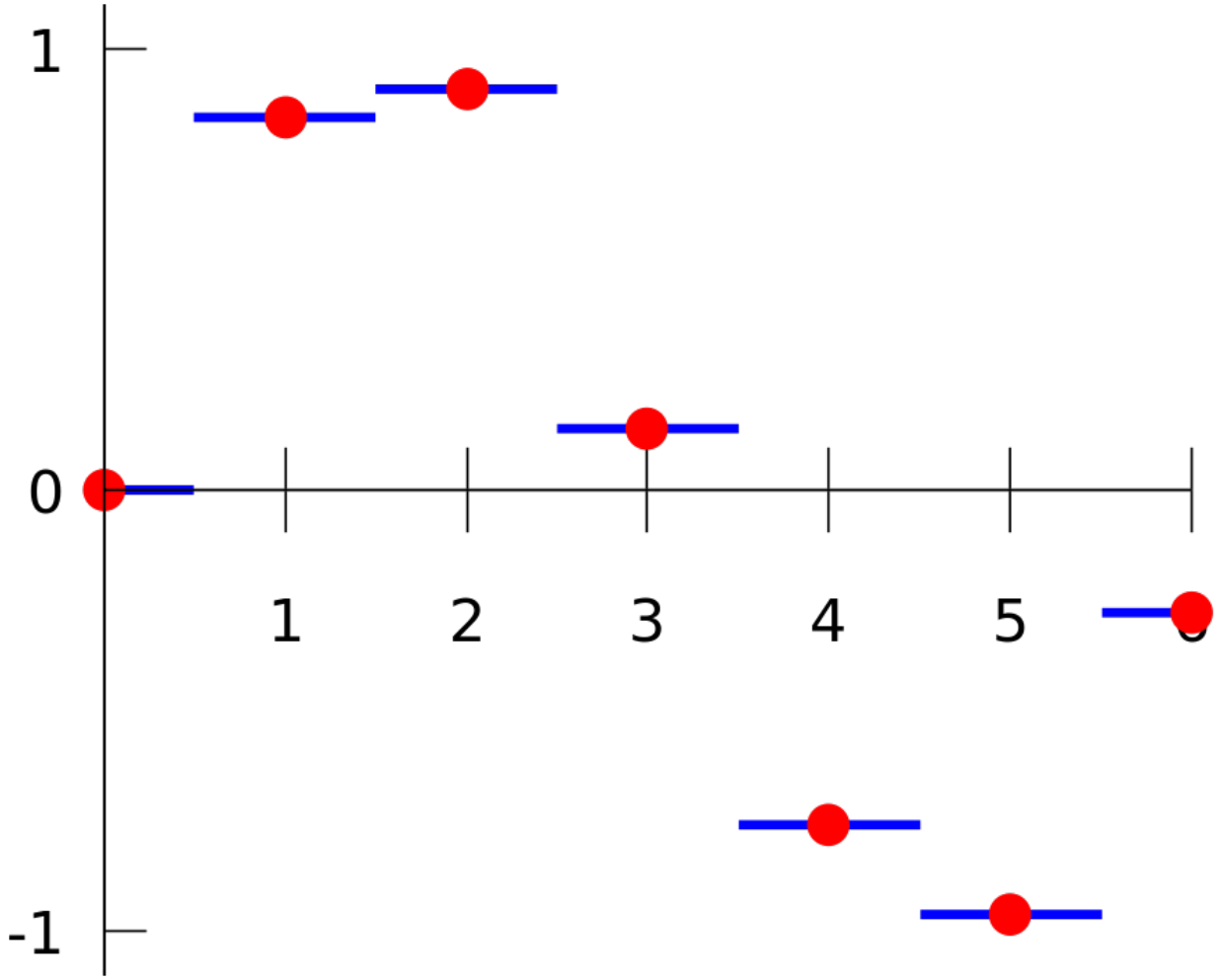
- Checking for & handling missing values

- What if a value is missing?
  - Discard item?
  - Signal with a `sentinel´ (do not include in computations!)
  - Data imputation: replace missing value with...
    - average?
    - median?
    - most frequent value?
    - interpolated value? (k-nearest neighbor interpolation? linear interpolation? higher-order interpolation?)

Missing values

- Interpolation
  - Assuming the data is generated by a smooth (spatial) function, it is possible to interpolate between known data points in order to infer the value of an unknown data point
  - Nearest-neighbor interpolation
  - Linear interpolation

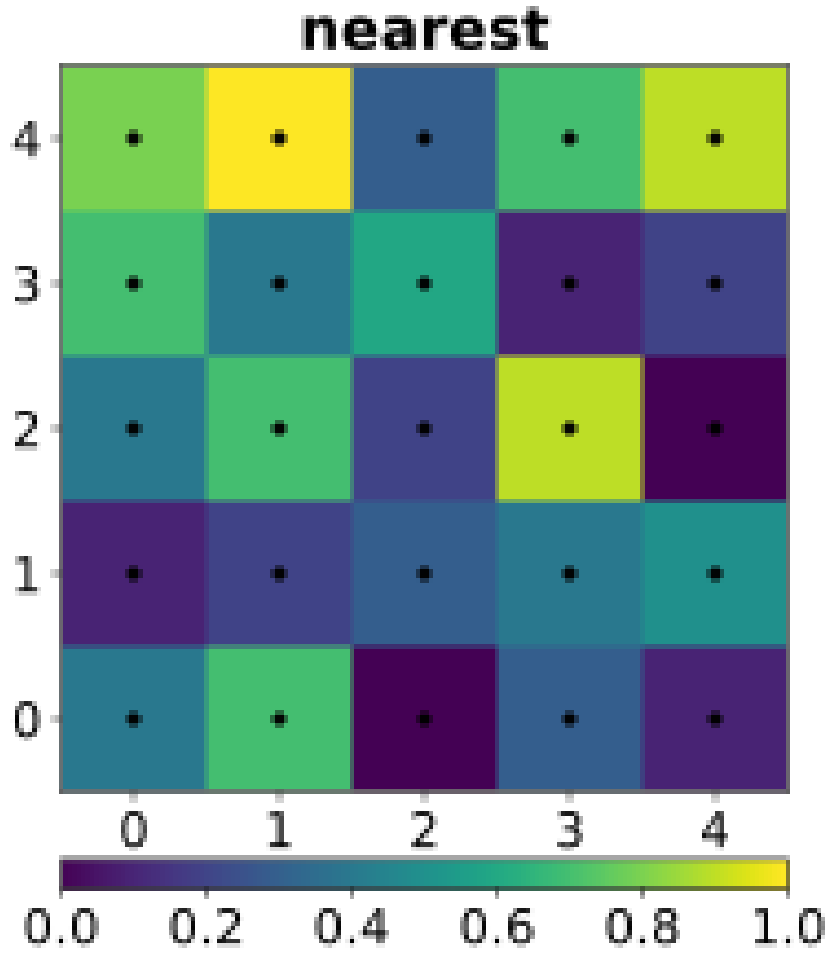
Nearest-neighbor interpolation



Nearest neighbor interpolation (blue lines) in one dimension on a (uniform) dataset (red points).

[See Nearest-neighbor interpolation - Wikipedia]

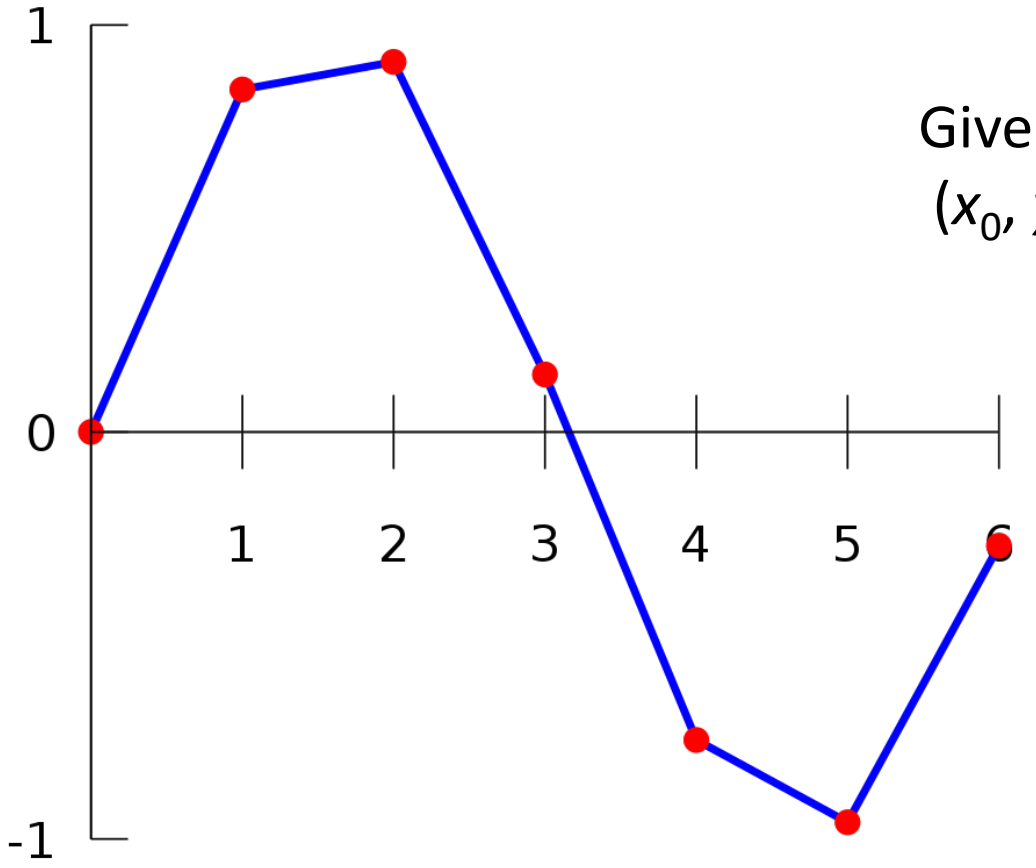
Nearest-neighbor interpolation



Nearest neighbor interpolation on a uniform 2D grid (black points). Each coloured cell indicates the area in which all the points have the black point in the cell as their nearest black point.

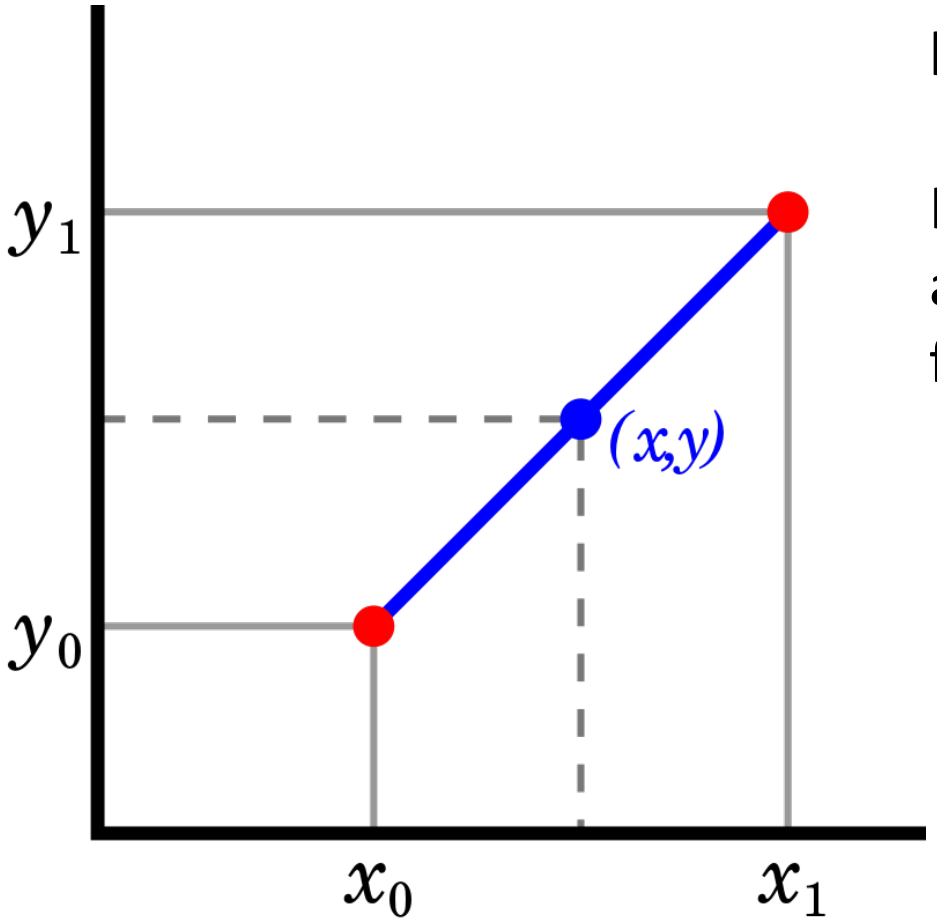
[See Nearest-neighbor interpolation - Wikipedia]

Linear interpolation of a data set



Given data points:
$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$

Linear interpolation



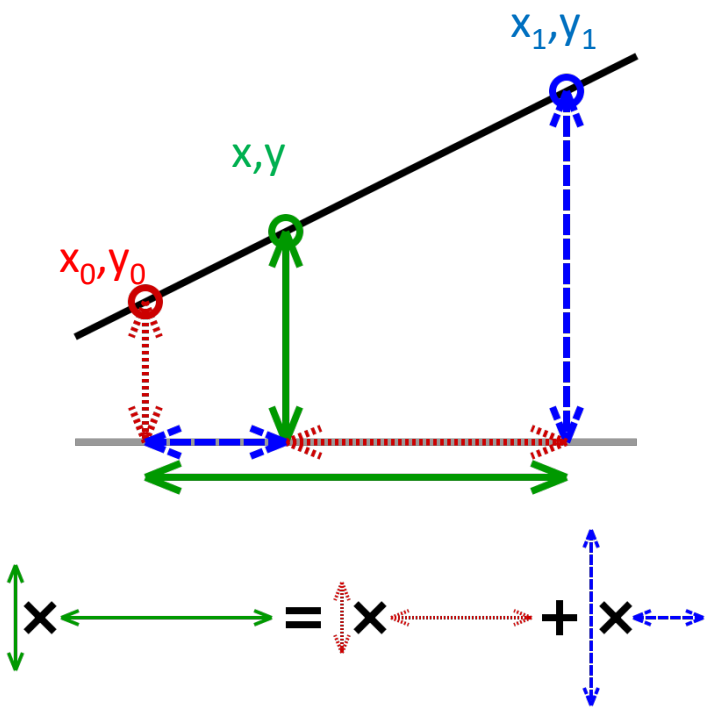
known data points: $(x_0, y_0), (x_1, y_1)$

Infer value of *y* corresponding to *x*, assuming a line as the interpolant function

$$\frac{y - y_0}{x - x_0} = \frac{x_1 - x_0}{y_1 - y_0}$$

https://en.wikipedia.org/wiki/Linear_interpolation

Linear interpolation



$$y = y_0 \frac{x_1 - x}{x_1 - x_0} + y_1 \frac{x - x_0}{x_1 - x_0}$$

Geometric interpretation: the value at the green point as a weighted average of the values at the red and blue points: the weights are inversely related to the distance from the end points to the unknown point, i.e., the closer point has more influence than the farther point.

Data pre-processing tasks

- Checking for & handling missing values

- What if a value is missing?
  - Discard item?
  - Replace missing value with...

- Decision depends on many factors and involves several concepts
  - ... the more you know about the data, the more likely you will be able to make a sensible choice.

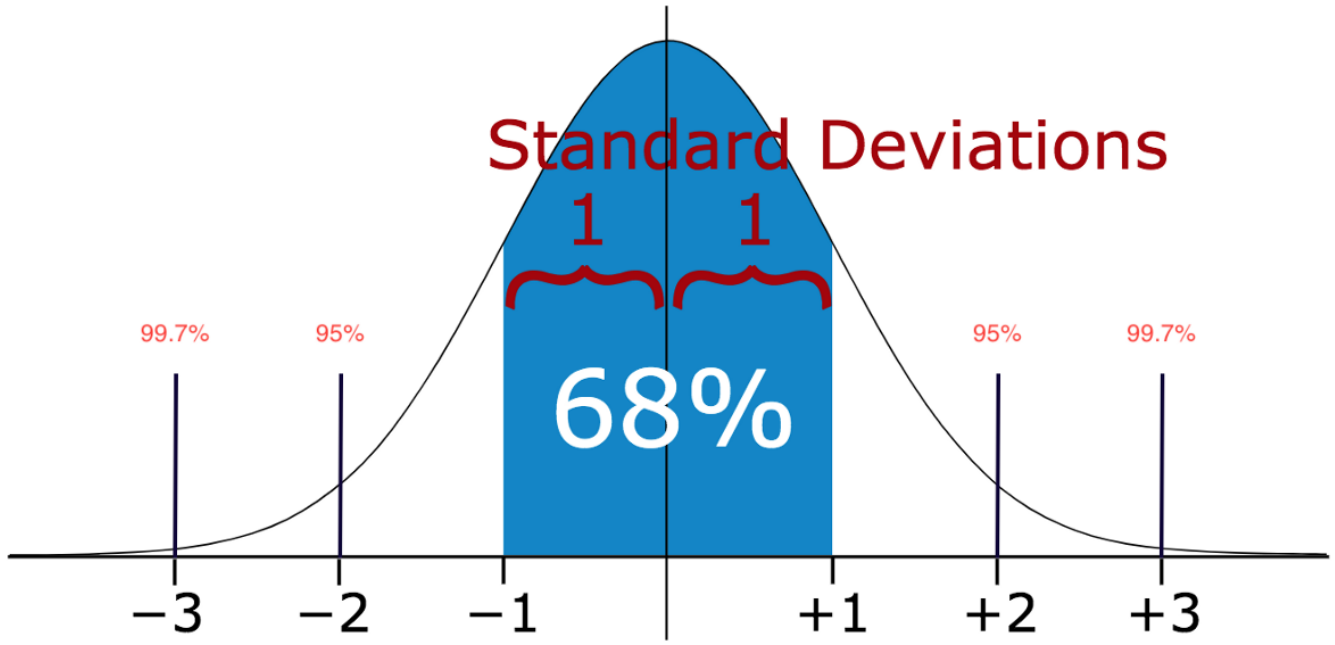
Data pre-processing tasks

- Compute some descriptive statistics, observe data dispersion and distribution

- Metrics and graphs!
  - Average, deviation
  - Range, quartiles (Q1, Q2, Q3)

  - Box plots, histograms

3-sigma rule distribution



Symmetric normal distribution

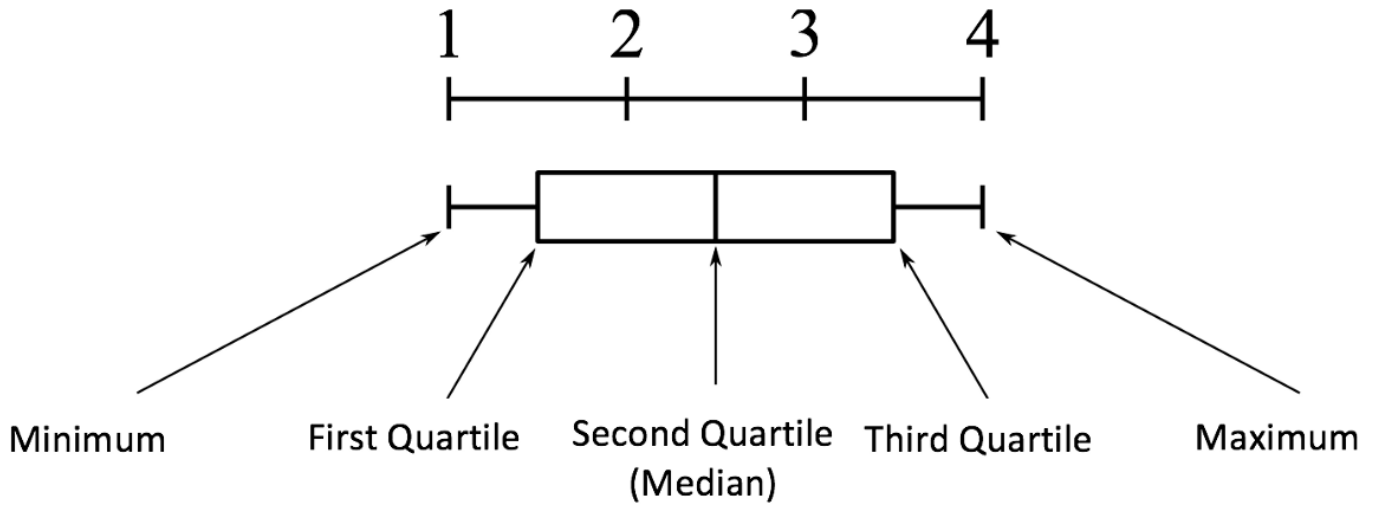
Data pre-processing tasks

- Compute some descriptive statistics, observe data dispersion and distribution

- Metrics and graphs!
  - Range, quartiles
  - 4.5.1 Calculating the range and interquartile range (statcan.gc.ca)
  - Min, Max, Q1, Q2, Q3

Five number summary: box plot

min = 1, max = 4, Q1 = 1,5, Q2 = 2,5, Q3 = 3,5



Data pre-processing tasks

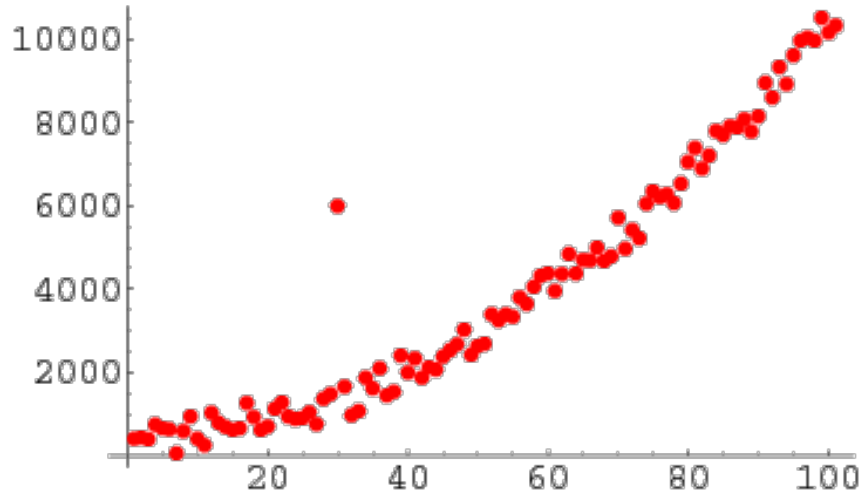
- Compute some descriptive statistics, observe data dispersion and distribution

- Metrics and graphs!

  – Box plots, histograms
  – What are Histograms? Analysis & Frequency Distribution | ASQ
  – Box Plot (Box and Whiskers): How to Read One & How to Make One in Excel, TI-83, SPSS - Statistics How To

Outlier identification

- Outliers are data points that don’t belong to a certain population, an observation that diverges from otherwise well-structured data.



[20,24,22,19,29,18,**4300**,30,18]

Outlier identification

- Outliers may be valid and important
  - detecting anomalies in heartbeat data can help in predicting heart diseases.
  - anomalies in traffic patterns can help in predicting accidents.
  - can indicate bottlenecks in network infrastructure and traffic between servers…
- They may also result from errors in measurement or collection

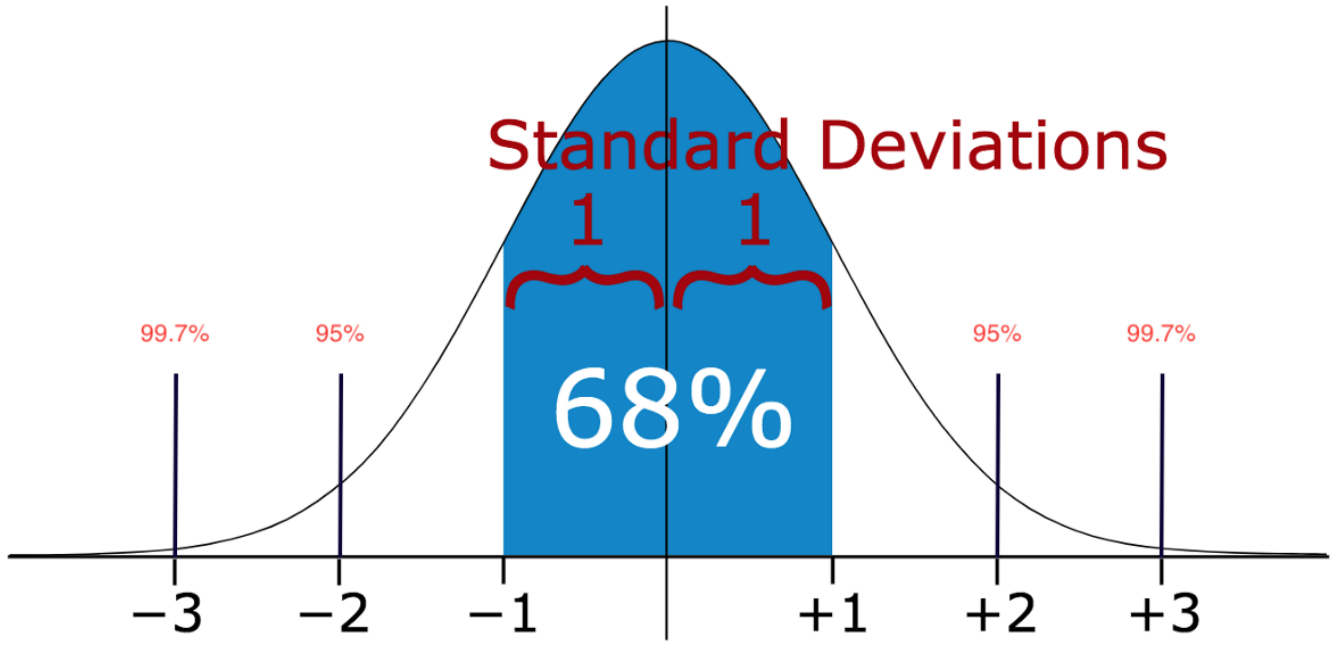
https://towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623

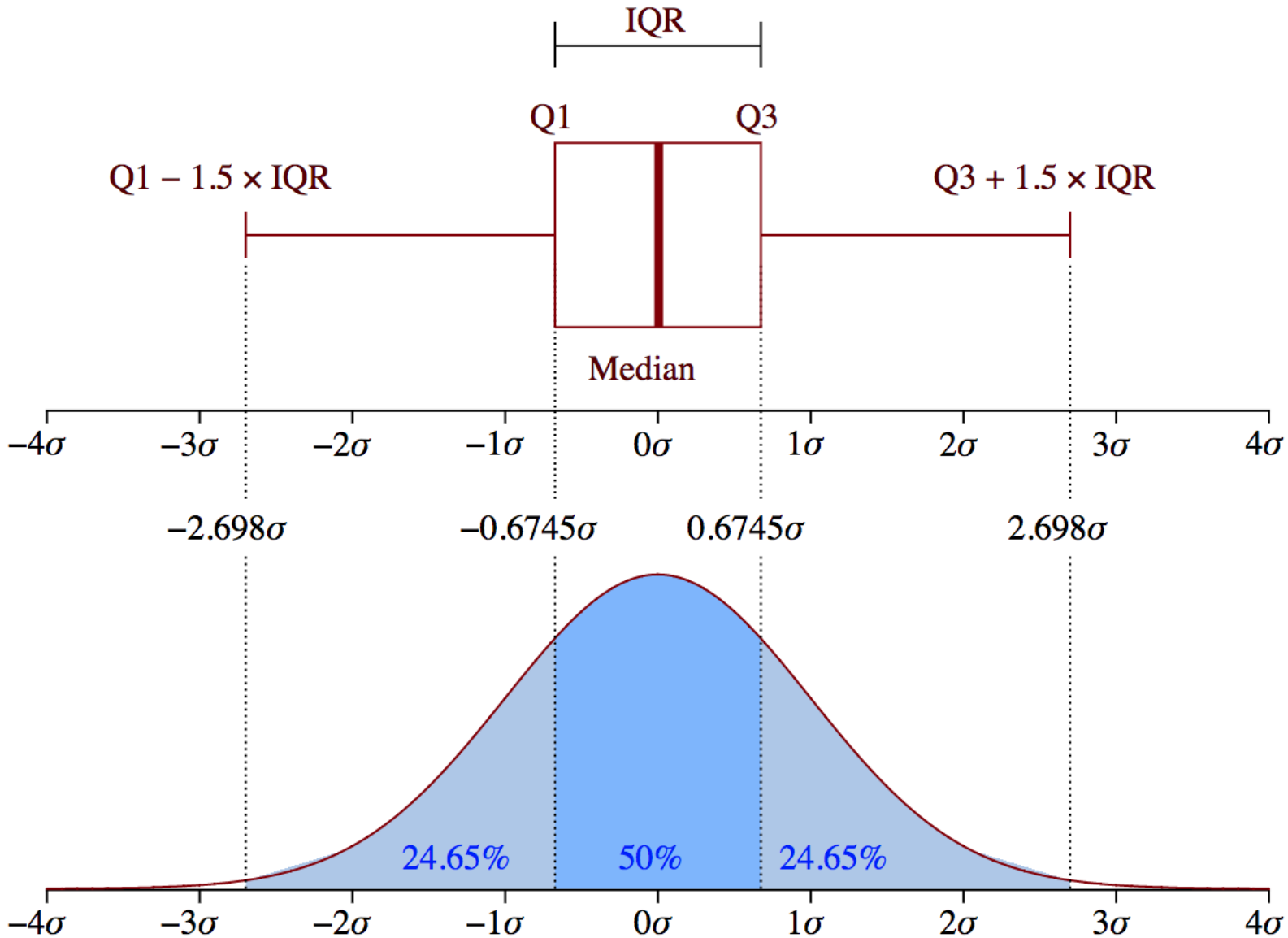
Outlier identification

- In a data distribution approximately normal
  - about 68% of the data values lie within one standard deviation of the mean
  - about 95% are within two standard deviations
  - **about 99.7%** lie within three standard deviations

- Therefore, any data point that is more than 3 times the standard deviation is very likely to be anomalous or an outlier

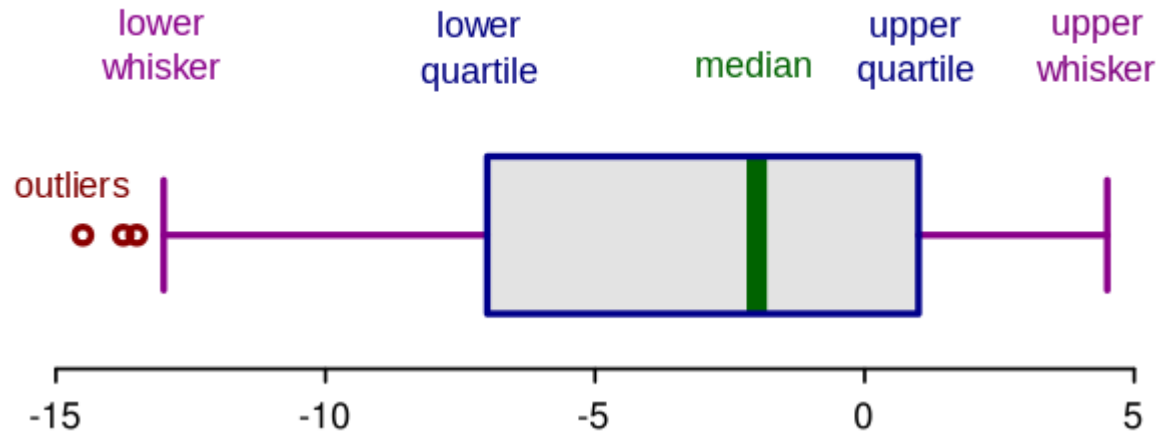
Outlier identification




IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ

Outlier identification

- Boxplots are a simple and effective way to visualize outliers:
  - the lower and upper whiskers are the boundaries of the data distribution, any data points that show outside the whiskers can be considered outliers or anomalous.



Outlier identification

- Example
  - Understanding and using Box and Whisker Plots | Tableau

Feature Scaling

- Some ML algorithms & Vis techniques are sensitive to the features' scale

- Important to avoid bias towards variables of higher magnitudes (in Vis and ML)

Feature Scaling

- Distance-Based algorithms like KNN, K-means, SVM, multidimensional visualizations (e.g., multidimensional projections) are most affected by the range of features
  - This is because behind the scenes **they are using distances between data points to determine their similarity.**

Feature Scaling

For example, let's say we have data with high school scores of students (ranging from 0 to 5) and their (predicted) future incomes (in thousand $):

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Since both the features have different scales, there is a chance that higher weigh is given to features with higher magnitude, causing the algorithm performance to be biased towards one feature

Feature Scaling

For example, let's say we have data with high school scores of students (ranging from 0 to 5) and their future incomes (in thousand $):

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Let´s compute Euclidean distances between students

Distance(S1,S2) = 20
Distance(S2,S3) = 1

Scaling will ensure features have the same importance and distances are more comparable

Feature Scaling

- Absolute maximum scaling
- Find the absolute maximum value of the feature in the dataset
- Divide all the values in the column by that maximum value
  - If we do this for all the numerical columns (features/attributes), then all their values will lie between -1 and 1.

Feature Scaling

- Standardization/Min-Max scaling

- Both strategies have the same goal
  - Place all variables values in a comparable scale/magnitude

Standardization

- Z-score

$$z = \frac{x - \mu}{\sigma}$$

- Gaussian distribution with 0 mean and 1 standard deviation

Normalization

- Min-Max transformation

$$\tilde{X} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Transforms the data values to be in the range [0,1]
- Must be careful with outlier values: min-max normalization may introduce distortions

The Big Question – Normalize or Standardize?

- Normalization is good to use when you know that the data distribution does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

The Big Question – Normalize or Standardize?

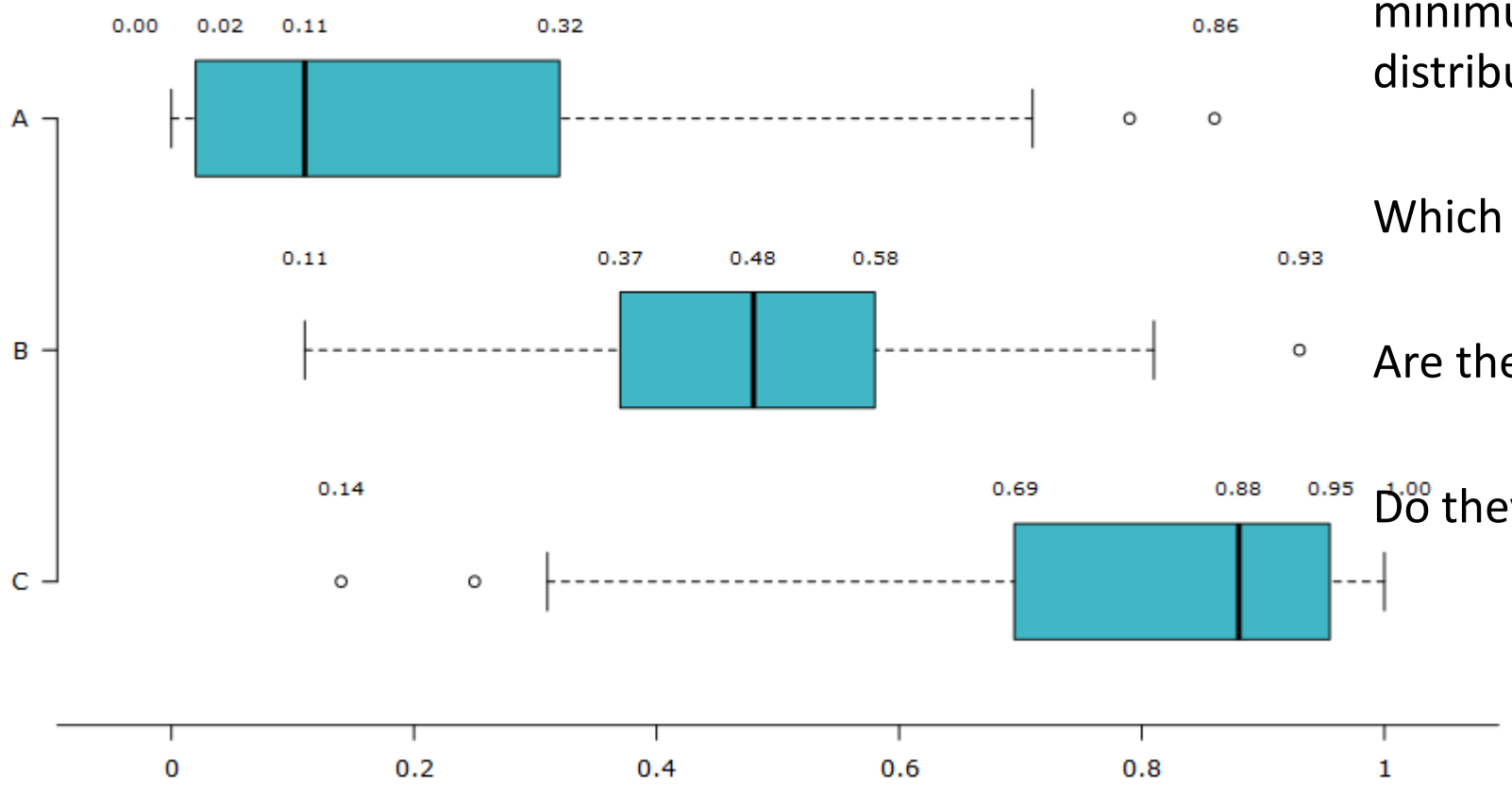
- However, at the end of the day, the choice of using normalization or standardization will depend on your problem and task/algorithm.
- No hard and fast rule to tell you when to normalize or standardize your data.
  - You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

Exercises

- Compute the median, lower and upper quartiles for the data: {1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57}
- Draw a box plot for this data
- Which statistics would you use to summarize the following data, which describes the prices, in US$, of 11 products in a store? {1, 1, 1.5, 0.5, 1, 1, 1, 1, 1, 1, 20}
- Draw a box plot for this data

Exercises

Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C



Inform the values of Q1, Q2, Q3, minimum and maximum for distributions A, B and C

Which is the IQR of each distribution?

Are they symmetric distributions?

Do they include potential outliers?

Exercises

- For the data: {1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57}
  - Normalize with maximum scaling
  - Normalize with min-max scaling
  - Normalize with standardization
- Draw the box plots for the data after each transformation. Compare with that of the original data
- Compute the mean, variance and standard deviation for this data: {2, 7, 3, 12, 9}

Sources/material/biblio

T. Munzner, Visualization Analysis & Design

Information Visualization Fundamentals, Enrico Bertini, online course in Coursera