Data Preprocessing
CS1D6: Introduction to data and statistics

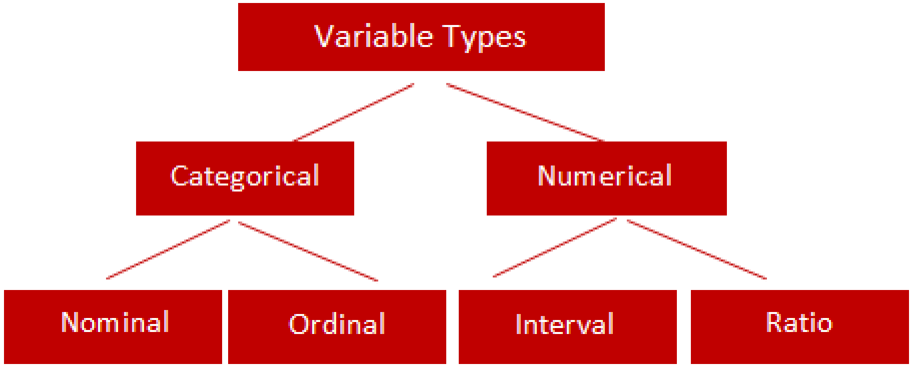
Dr. Fayyaz Minhas

Department of Computer Science

University of Warwick

Types of variables

- Based on the nature of a variable



https://towardsdatascience.com/data-types-in-statistics-347e152e8bee

Categorical Variables

- Qualitative Data
  - Nominal
    - Qualitative variables without any ordering defined on them
      - Male, Female
      - Football, Cricket, Tennis
  - Ordinal
    - Qualitative variables without some ordering defined on them
      - High, Medium, Low
  - More examples?

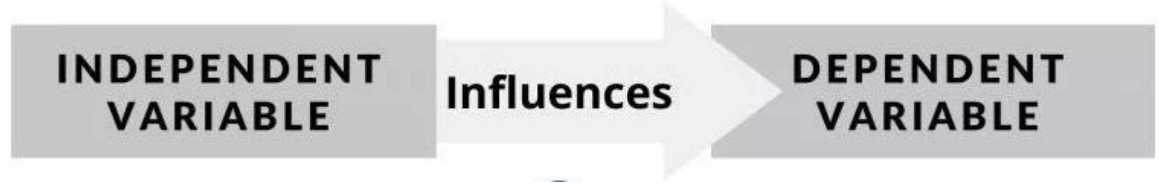
Numerical Variables

- Numbers
- Interval
  - When the value of the variable is assigned based on the interval (out of an ordering of equal intervals) in which the phenomenon that the variable is measuring falls in
    - With no concept of a natural zero
    - For example: Temperature in Celsius
- Ratio
  - With a concept of a natural zero or nothing
    - For example: (Temp in K) 0K, (weight) 0Kg, height (0m)

Other concepts

- Independent Variables

- Dependent Variables

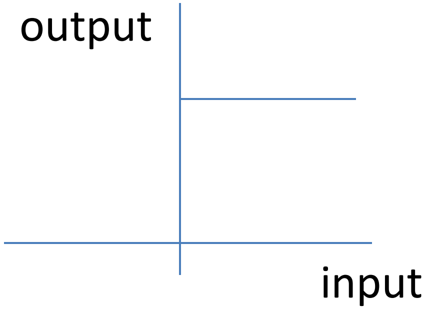


Preprocessing: Encoding

- Encoding: Converting to numbers
  - Indicator Variables (One-Hot Encoding)
    - Nominal
      - A single nominal variable will need to be converted to multiple indicator variables
      - Your Favourite Game
        - » Football: (1,0,0)
        - » Cricket: (0,1,0)
        - » Tennis: (0,0,1)
  - Ordinal
    - Since there is an ordering, we can use numbers directly):
      - (Low, Med, High): (0,1,2)

Data Transformations

- Transforming a variable
  - Binarization
  - Binning/Discretization
  - Min-Max Scaling
  - Standardization (Mean-Stdev Scaling)
  - Log-transformation
  - Power-law transformation
  - Rank-transformation
  - Smoothing
  - Normalization
  - Logit/Sigmoid mapping
  - Detrending
  - Whitening
  - …



General Form: $x' = \phi(x)$

Can be for a single variable
Or a single sample
Or even multiple samples

Typically: Invertible

https://en.wikipedia.org/wiki/Data_transformation_(statistics)

Why Data Transformations?

- Scale matching
- Improve Interpretation
- Visualization
- Pre-processing

Discretization

- Convert a continuous variable into discrete values

- Example: Divide people based on their heights into short, medium, tall

Standardization

- Make the mean of the variable zero and scale it by its standard deviation
  - Common requirement for a number of pattern recognition models esp. when using multiple variables

$$x' = \frac{x - \mu_x}{\sigma_x}$$

  - Example: Heights of all people in a class can be scaled or standardized using the above approach

Min-Max Scaling

- When you want the transformed variable to be in the range [0,1]

$$x' = \frac{x - x_{min}}{x_{max}}$$

Rank Transformation

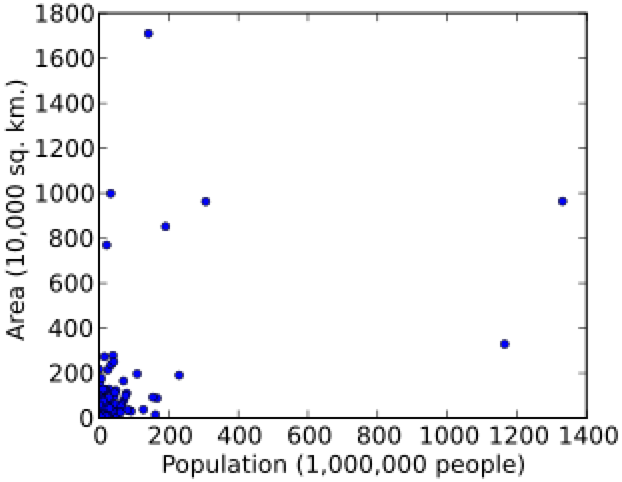
- When only the rank, rather than the true value, of a variable matters we may want to do a rank transform

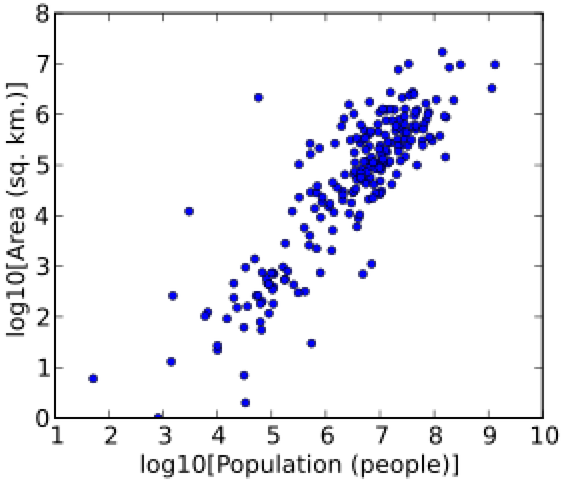
- Example: Assign an ordering to individuals in a class based on their heights

$$x' = Rank(x; X)$$

Log transformation

$$x' = log(x)$$



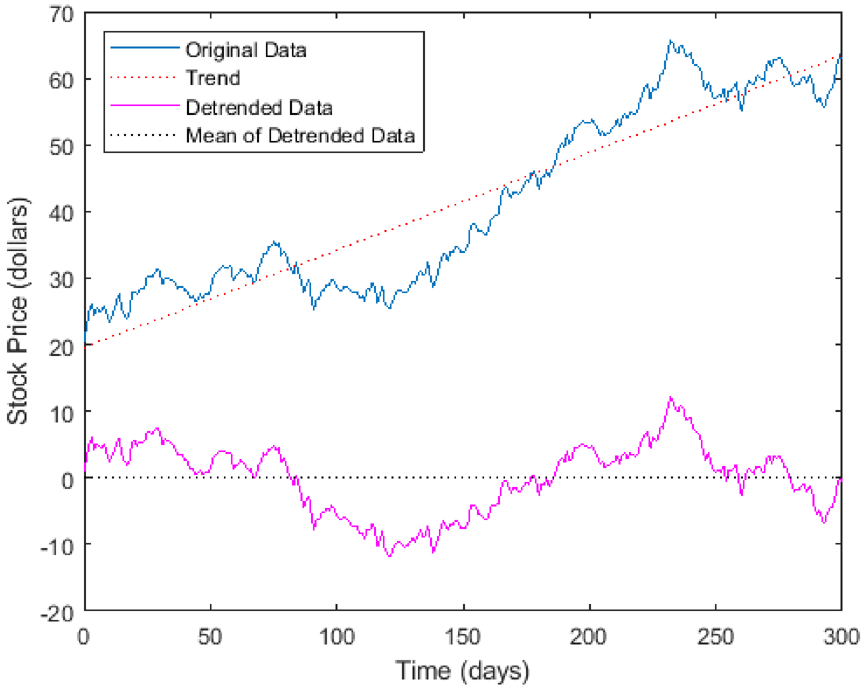


Example: Suppose we have a scatterplot in which the points are the countries of the world, and the data values being plotted are the land area and population of each country. If the plot is made using untransformed data (e.g. square kilometers for area and the number of people for population), most of the countries would be plotted in tight cluster of points in the lower left corner of the graph. The few countries with very large areas and/or populations would be spread thinly around most of the graph's area. Simply rescaling units (e.g., to thousand square kilometers, or to millions of people) will not change this. However, following logarithmic transformations of both area and population, the points will be spread more uniformly in the graph.

https://en.wikipedia.org/wiki/Data_transformation_(statistics)

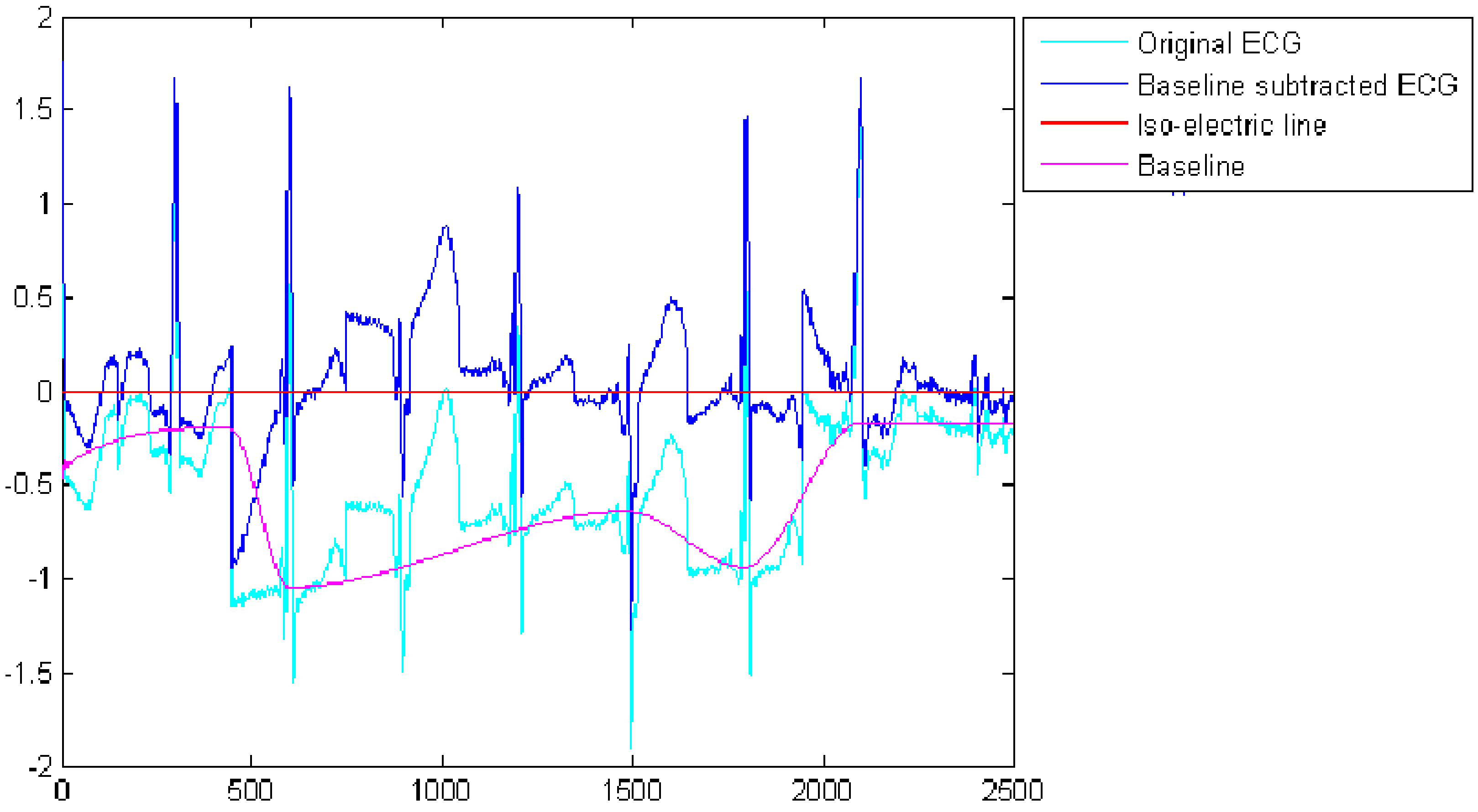
Data Detrending

- Identification and removal of the trend in a time series or other data

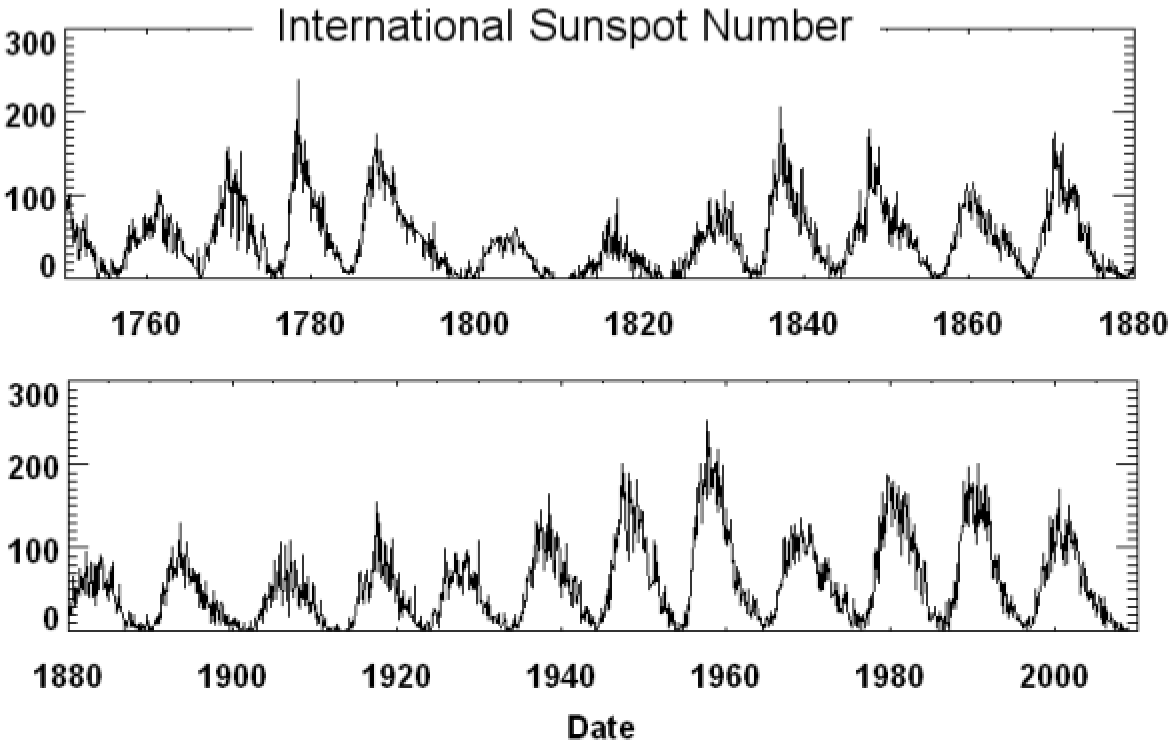


https://uk.mathworks.com/help/matlab/data_analysis/detrending-data.html
https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.detrend.html

“Baseline” Detrending in ECG

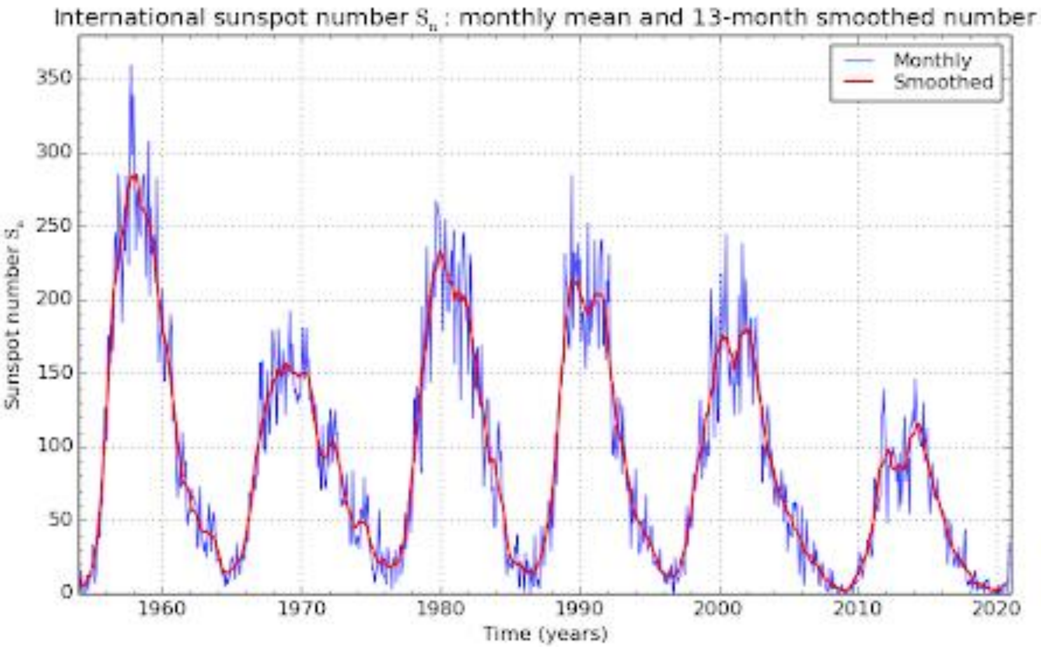


Data Denoising



Data Denoising

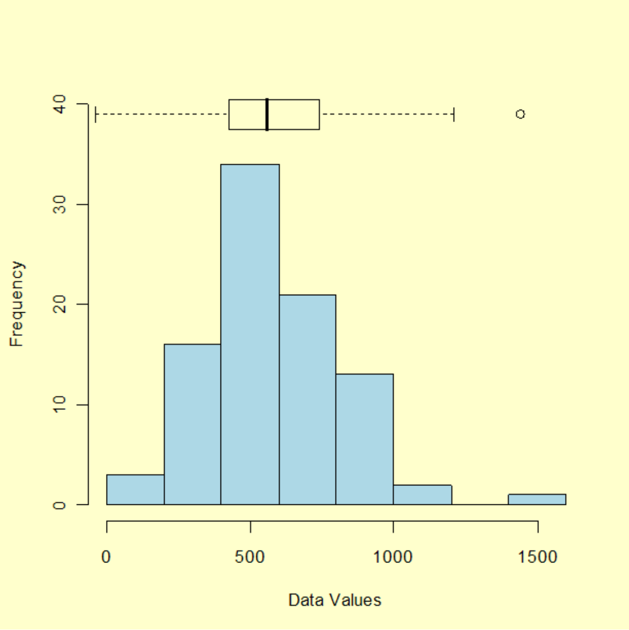
- Typically some form of smoothing or averaging is used
  - For example: replace the current value with the average of 3 samples



SILSO graphics (http://sidc.be/silso) Royal Observatory of Belgium 2020 December 1

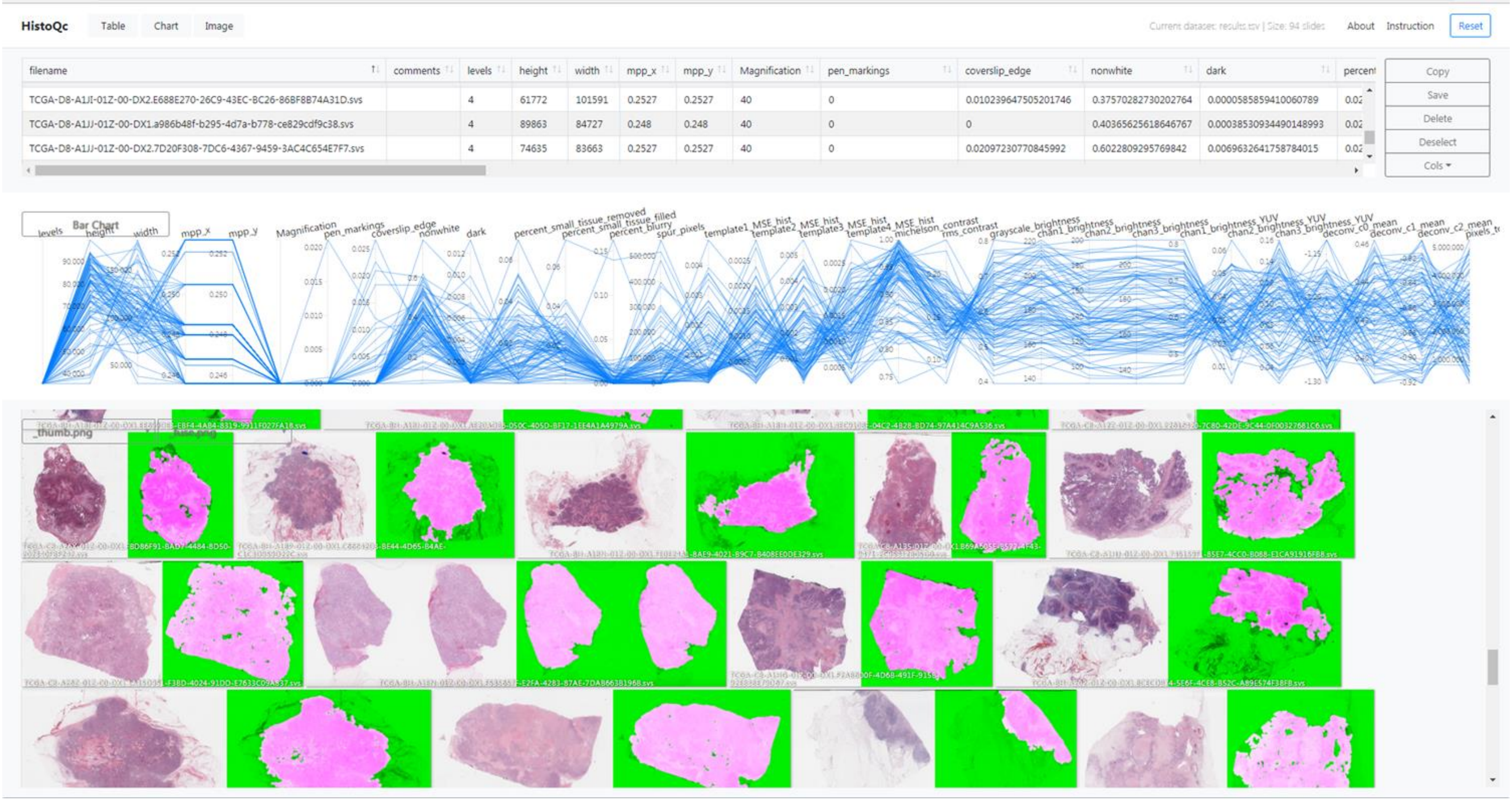
Understanding and Identifying Outliers

- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



Example Application

- Quality Analysis of Pathology Images



filename	comments	levels	height	width	mpp_x	mpp_y	Magnification	pen_markings	coverslip_edge	nonwhite	dark	percent
TCGA-D8-A1JI-01Z-00-DX2.E688E270-26C9-43EC-BC26-86BF8B74A31D.svs		4	61772	101591	0.2527	0.2527	40	0	0.010239647505201746	0.37570282730202764	0.000058585941006789	0.02
TCGA-D8-A1JJ-01Z-00-DX1.a986b48f-b295-4d7a-b778-ce829cdf9c38.svs		4	89863	84727	0.248	0.248	40	0	0	0.40365625618646767	0.00038530934490148993	0.02
TCGA-D8-A1JJ-01Z-00-DX2.7D20F308-7DC6-4367-9459-3AC4C654E7F7.svs		4	74635	83663	0.2527	0.2527	40	0	0.02097230770845992	0.6022809295769842	0.0069632641758784015	0.02

https://github.com/choosehappy/HistoQC

Handling Missing Data: Imputation

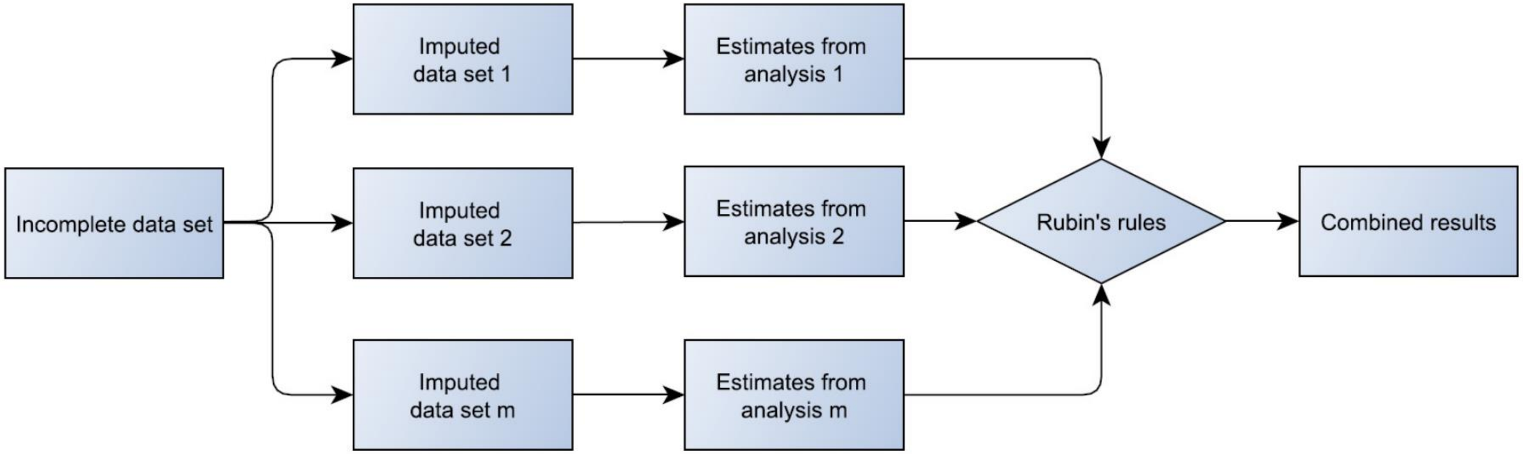
- Imputation is the process of replacing missing data with substituted values.
  - Let’s say we collect the height and weight of a group of individuals. However, for some samples, we observe that one of the two variables have been “missed”
  - What do we do?

Data Imputation

- Mean Imputation: Replace with the mean of other samples
  - Within Class Mean Imputation
    - For example: impute the weight of a male based on the average weight of the male samples in the group
- Pick the unknown value for a given variable for a given sample based on its most similar other sample
  - Can also use weighted average of its neighbours
  - For example: impute the weight of a male based on the weighted of the male samples that are closest in height

Data Imputation

- Multiple imputation by chained equations (MICE)



Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. “Multiple Imputation by Chained Equations: What Is It and How Does It Work?” *International Journal of Methods in Psychiatric Research* 20, no. 1 (February 24, 2011): 40–49. https://doi.org/10.1002/mpr.329.

https://github.com/venkateshavula/Data-Imputation-Strategies

Data Imputation

- Non-negative matrix factorization

Exercise

- Observe the impact of different types of data transformations on the distribution

End of Lecture

We want to make a machine that will be proud of us.

- Danny Hillis