Data Preparation for Data Mining

Natasha Balac, Ph.D.

MAS DSE

April 2016

Outline

- Motivation and Goals
- What is data?
- Data Preparation:
    - Organizing data (structural issues)
    - Preprocessing (data value issues)
    - Exploring Variables and Descriptive Statistics
    - Exploring the Data Matrix
    - Outliers, Anomalies, and Visualizations

The Importance of Data Prep

- “Garbage in, garbage out”
- A crucial step of the DM process
- Could take 60-80% of the whole data mining effort

Working Definition

- Data Preparation:
  - Cleaning, filtering, transforming, and organizing the data
  - Preparing data for modeling
  - Data Munging
  - Feature Engineering

Prerequisites

- Data Understanding:
  - Descriptors, values, ranges, labels
- Data History
- Domain Knowledge
  - Meaning and data relations
- Questions to be addressed

Input - Output

- Inputs:
  - raw data
- Outputs:
  - two data sets: training and test (if available)
  - Training further broken into training and validation

End Product: Quality Data

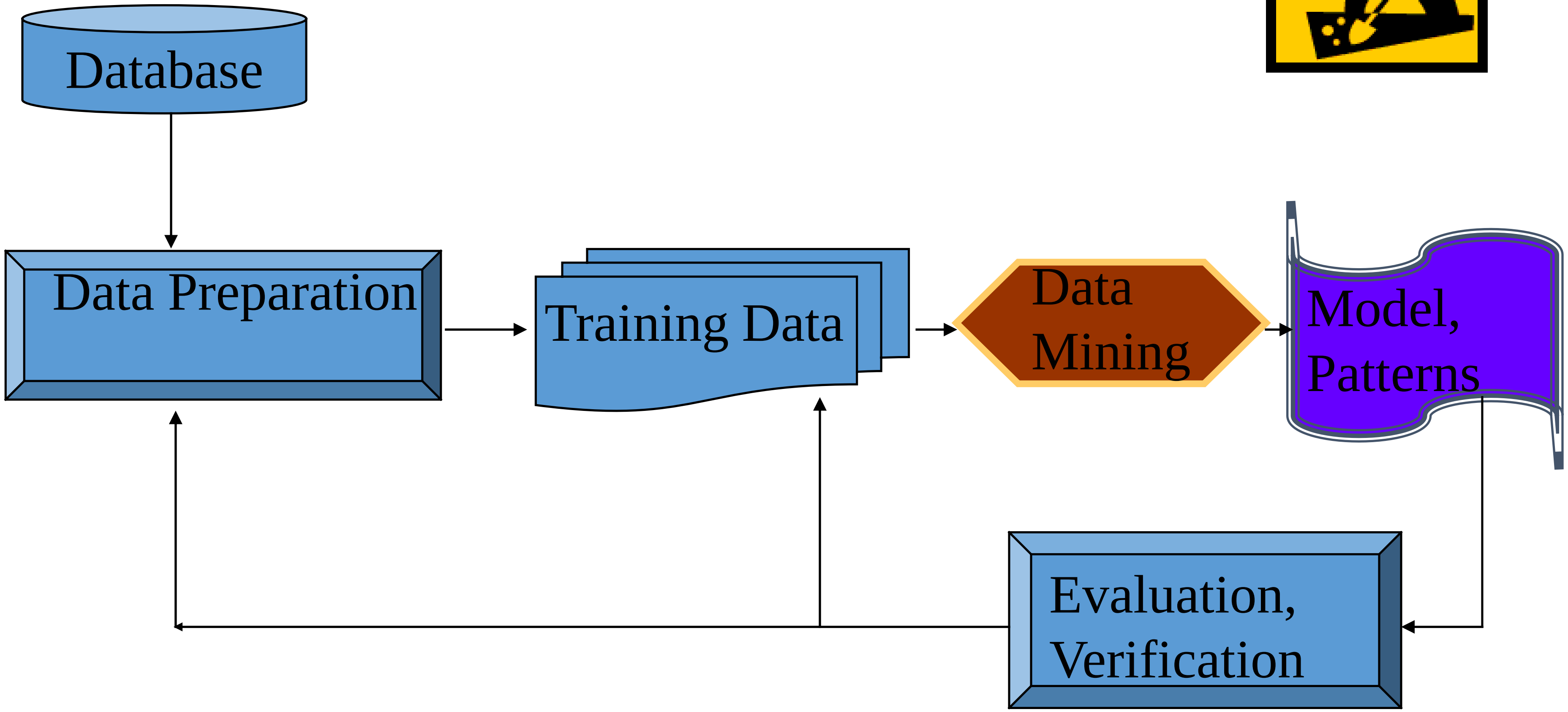
- Accurate
- Complete
- Consistent
- Interpretable

In other words: Good data →Better results!

Outline

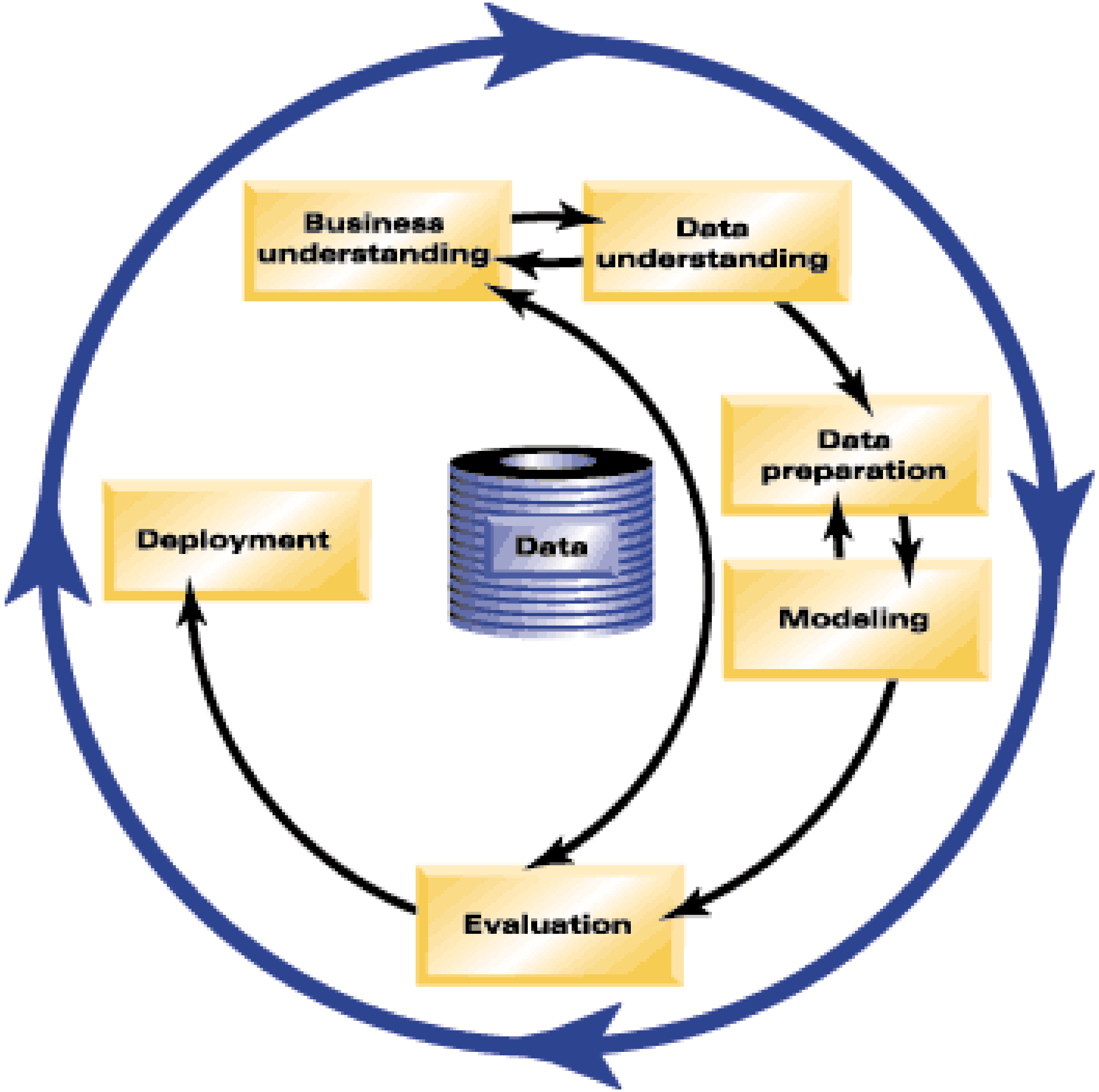
- Motivation and Goals
- What is data?
- Data Preparation:
    - Organizing data (structural issues)
    - Preprocessing (data value issues)
    - Exploring Variables and Descriptive Statistics
    - Exploring Data Matrix
    - Outliers, Anomalies, and Visualizations

Recall the KDD Process

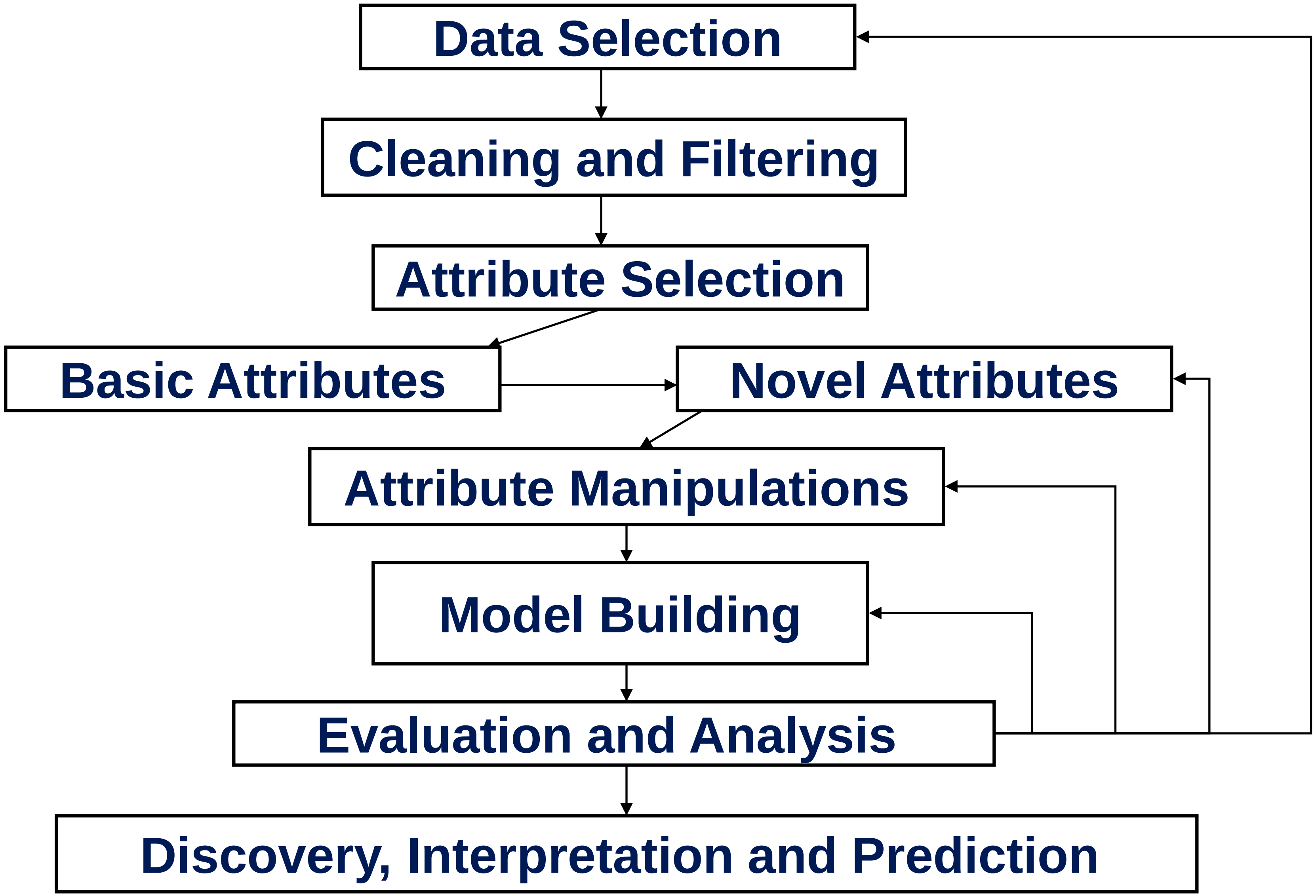


CRISP-DM Methodology

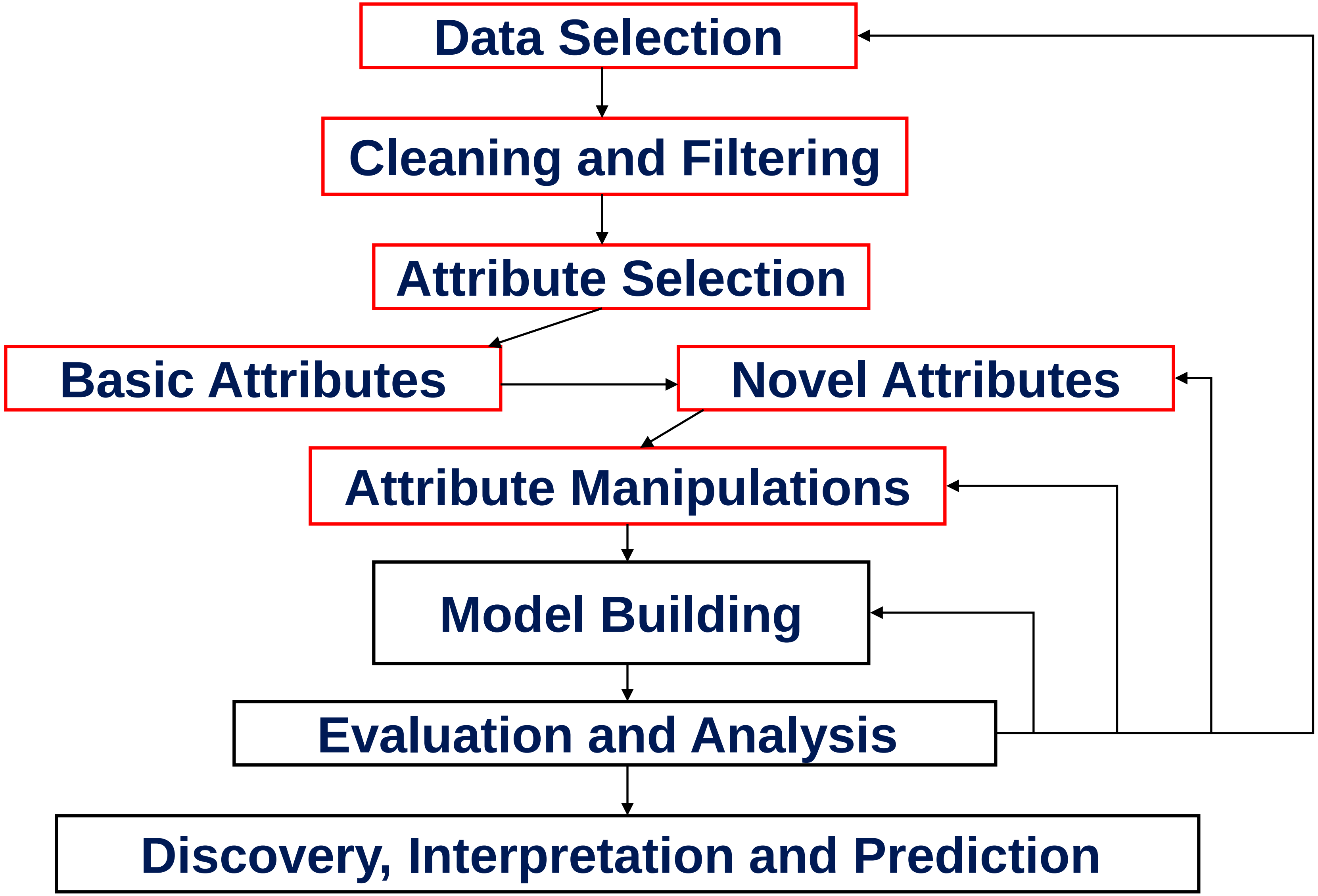
- Cross Industry Standard Process for Data Mining
  - http://www.crisp-dm.org/
- Six Phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

The Details of the DM Process



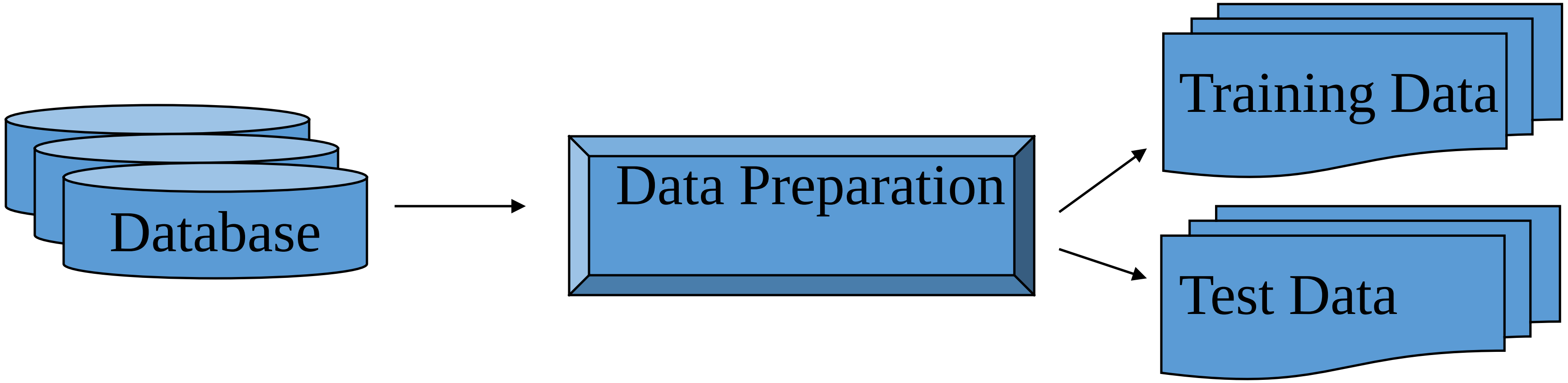
Data Preparation in the DM Process



The Data Mining Process

- Iterative Nature
- Exploratory Process
- Highly tailored to the dataset
- Need for Fine-Tuning
- Need for Model Revision from time to time

From Data Source To Algorithm Input



User Decides:

- Selection Criteria –
- Joins => denormalize
- How much data?

Depends on needs and domain knowledge about what’s relevant

User Performs:

- Cleaning data and Transformations

Depends on domain knowledge, data itself and possibly on algorithms

Data Terminology

- Data consists of:
  Examples, observations, measurements, events, transactions, records..

- Data can be:
  Structured (e.g. database rows) or unstructured (e.g. text)

What **Algorithms Consume?**

- Instance = specific example
  - thing to be classified, associated, or clustered
  - instances may be labeled as a class, or as an outcome
  - If no labels available you can either do unsupervised learning or try to get labels

- Set of instances comprise the input dataset
  - Often represented as a single flat file or *data matrix*

Algorithm Input Detail

- Each instance described by a predefined set of “attributes” or “variables”

- Attributes’ values, or it’s existence, may or may not be dependent on each other
  - e.g. height and weight may be correlated
  - e.g. spouse name depends on marital status

What’s a concept?

- Styles of learning:
  1. Classification learning: predicting a discrete class
  2. Association learning: detecting associations between features
  3. Clustering: grouping similar instances into clusters
  4. Numeric prediction: predicting a numeric quantity
- Concept: thing to be learned
- Concept description: output of learning scheme

What’s in an example?

- Instance: specific type of example
  - Thing to be classified, associated, or clustered
  - Individual, independent example of target concept
  - Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances/dataset
  - Represented as a single relation/flat file
- Rather restricted form of input
  - No relationships between objects
- Most common form in practical data mining

Generating a flat file

- Process of flattening called "denormalization"
  - Several relations are joined together to make one
- Possible with any finite set of finite relations
- Problematic: relationships without pre-specified number of objects
  - Example: concept of nuclear-family
- Denormalization may produce spurious regularities that reflect structure of database
  - Example: "supplier" predicts "supplier address"

What’s in an attribute?

- Each instance is described by a fixed predefined set of features, its “attributes”
- Number of attributes may vary in practice
  - Possible solution: “irrelevant value” flag
- Related problem: existence of an attribute may depend of value of another one
- Possible attribute types (“levels of measurement”):.
  - Nominal, ordinal, interval and ratio
  - Nominal (categorical) vs. numeric (continuous)

Terms from database to math

TABLE

Attributes (i.e. columns) ….

Rows

…

…

Cells …		

DATA MATRIX

Variables ….

Instances

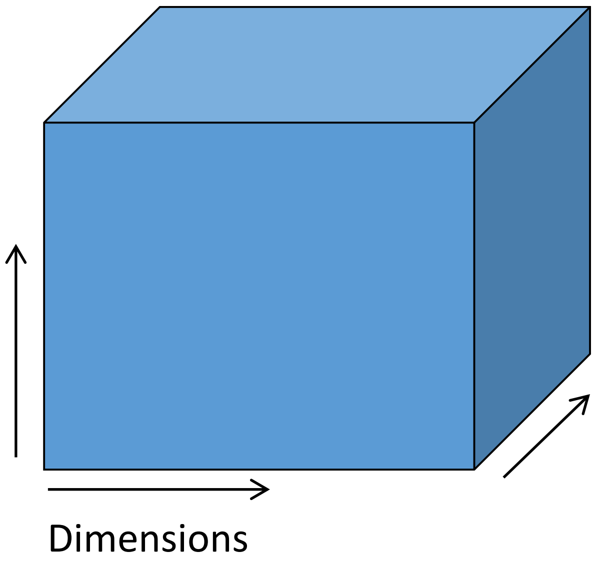
…

…

attributes in the database relate to variables in the data matrix

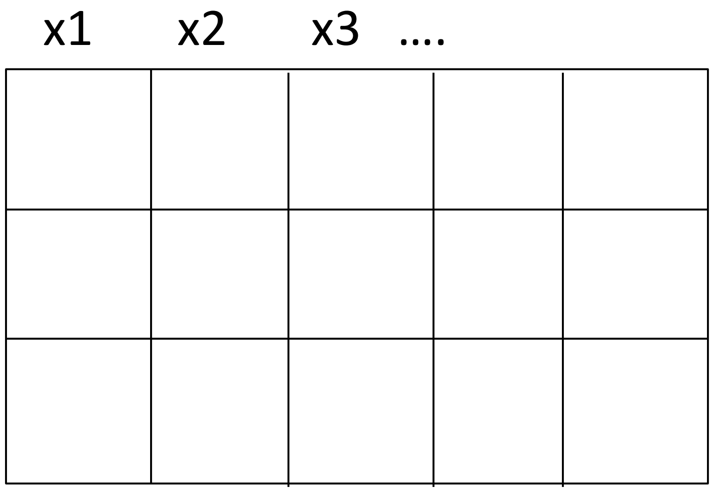
Terms database to math

TABLES can 2 or more dimensions (multi-way) given by discrete attributes called Factors



In DATA MATRIX each variable is a dimension in some coordinate space

Row Vector is a Coordinate Pt. ….



- Matrix Variables can also be Factors
- Factor Tables can also be treated mathematically

Variables and Features terms

- Variables and their transformations are features
- Instance labels are outcomes or dependent variables (as in supervised learning)
- No instance labels available then use unsupervised learning

Outline

- Motivation and Goals
- What is data?
- Data Preparation:
    - Organizing data (structural issues)
    - Preprocessing (data value issues)
    - Exploring Variables and Descriptive Statistics
    - Exploring Data Matrix
    - Outliers, Anomalies, and Visualizations

Database to Data Matrix

- Goal: gather all relevant information into each instance in one data matrix
  - Typical models are: *instance outcomes = F(row values)*

- Key: the functions you model and questions you pose determine what variables are bought together and how they are presented

Organizing data example

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

Customer	Zip
John	99000
Jane	11000
Fred	99000

2 tables, keyed on customer id

Simple descriptive queries

Customer	Total Spent
John	110
Jane	140
Fred	15

A data matrix using Aggregation Levels

Relevant Questions involve customers and totals

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

- What would the data matrix be for a relationship question:

How similar are zip codes?

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

- Coding Issues among variables
  - implicit domain knowledge: customers buy items
  - large number of categorical values: number of items bought
  - spurious regularities, e.g. “item” predicts “supplier”
  - usual data issues, e.g. date/time, composite fields, entity resolution, etc..

Database to Data Matrix

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

How similar are zip codes?

'similar' wrt to what entities?

'similar' implies a comparison?

An approach: instances are transpose of items, cell values are counts

Customer Zip	Acme Mower	Ace Mower	Acme Wrench	Ace Wrench	…	(last item)
99000	1	0	1	0		
11000	0	1	0	0		
…						

Get related measurements down row into separate columns of the same instance

How do zip codes compare?
What items go together?
How do they impact purchases?

Instance are counts, but aggregated across item types

Customer Zip	Mower	Wrench	Rake	Hammer	…	(last item)
99000	1	1	1	1		
11000	1	0	0	0		
…						

What questions can we ask now?
Should we include customer name and zip code?

Customer	Zip
John	99000
Jane	11000
Fred	99000

Customer	Item	Price	Date
John	Acme Mower	100	Jan 2000
John	Acme Wrench	10	Sept 2000
Jane	Ace Mower	120	Mar 2003
Jane	Ace Rake	20	Mar 2003
Fred	Ace Hammer	15	July 2002

Can also compare customer-item pairs

	Mower	Wrench	Rake	Hammer	…	(last item)
John	1	1	0	0		
Jane	1	0	1	0		
Fred	0	0	0	1		

Would John buy a Rake too?

Should 0 indicate ‘not yet bought’?

We can compare customers, or products.
Can we use customer-item pairs collaboratively?

Data Wrangling Cautions

- Beware of data integration:

  different names for same data

  different data for same names

Outline

- Motivation and Goals
- What is data?
- Data Preparation:
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

4 Preprocessing data values and QA

- Preprocessing involves:
  - Cleansing data
  - Missing data
  - Exploring variable characteristics
  - Re-representing variables (normalizing, discretizing, transforming)

Because real data is incomplete, inconsistent, noisy, etc…

Data Preparation is Variable Prep

- Know the meanings (domain knowledge!)
- Know types of variables
- Know statistical properties
- Do QA (clean, fill-in, fix errors)
- Do enhance or re-represent
  - add more data as needed
  - apply domain knowledge to ease the work of the tool

Types of Measurements

- Nominal (names)
- Categorical (zip codes)

$\left.\right\}$ Qualitative (unordered, non-scalar)

- Ordinal (H,M,L)
- Real Numbers
  - May or may not have a Natural Zero Point?
    - If not comparisons are OK but not multiplication (e.g. dates)

$\left.\right\}$ Quantitative (ordered, scalar)

Know variable properties

- Explore characteristics of each variable:
  - typical values, min, max, range etc.
  - entirely empty or constant variables can be discarded
  - explore variable dependencies
- Sparsity
  - missing, N/A, or 0?
- Monotonicity
  - increasing without bound, e.g. dates, invoice numbers
  - new values not in the training set
- Visualize the distribution
    - Check skews, outliers

Noise in Data

- Noise is unknown error source
  - sometimes assumed to be independent and random

- Approaches to Address Noise
  - Detect suspicious values and remove outliers
  - Smooth by averaging with neighbors
    - but then how many neighbors?
  - Smooth by fitting the data with other variables

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equal-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

Data Errors and Noise

- Incorrect attribute values
  - data collection errors
  - data entry errors
  - duplicate records
  - Etc..

- Approaches to Address Problems
  - apply domain knowledge to replace values
  - model error process to reverse engineer correct value
    - e.g. common misspellings and typos

Missing Data

- Data values not present

  e.g. customer income in sales data not easy to get

  e.g. sensor malfunction

- Or data available but missing due to
  - deletions
  - not entered

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value

Missing Data

- Important: review statistics of a missing variable
  - Are missing cases random?
  - Are missing cases random but dependent on other variable(s)?
  - Are other variables missing data in same instances?
  - Is there a relation between missing cases and outcome variable?
  - What is frequency of missing cases?

Quick Approaches to Handle Missing Data

- If there's enough data and missing seems random
  - Delete instances with missing attribute values
  - Delete attributes with high "missingness"
- Use the attribute mean to fill in (impute) the missing value
- Use the attribute mean for all samples belonging to the same class

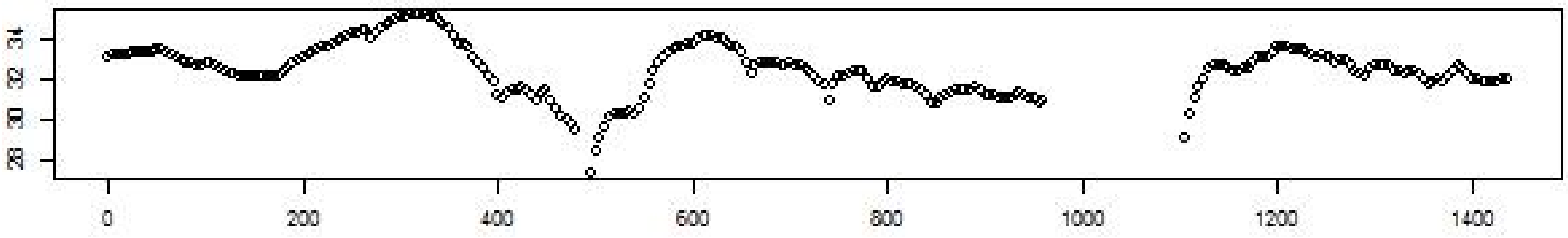
Additional Approaches to Handle Missing Data

- Use a model (based on other attributes) to infer missing value
- Use a global constant to fill in the missing value, e.g. “unknown”, and let algorithms figure it out (e.g. Decision Trees)
- Add a new indicator variable (1 or 0) to indicate missing and let algorithms figure it out (e.g Linear Models)

Missing Data Example

Time series of glucose measurements over 24hours

raw data of glucose level



Time (minutes)

Can we ignore missing values?
Should we fill it in with a constant (eg last value)? Or with a mean? Or a model?

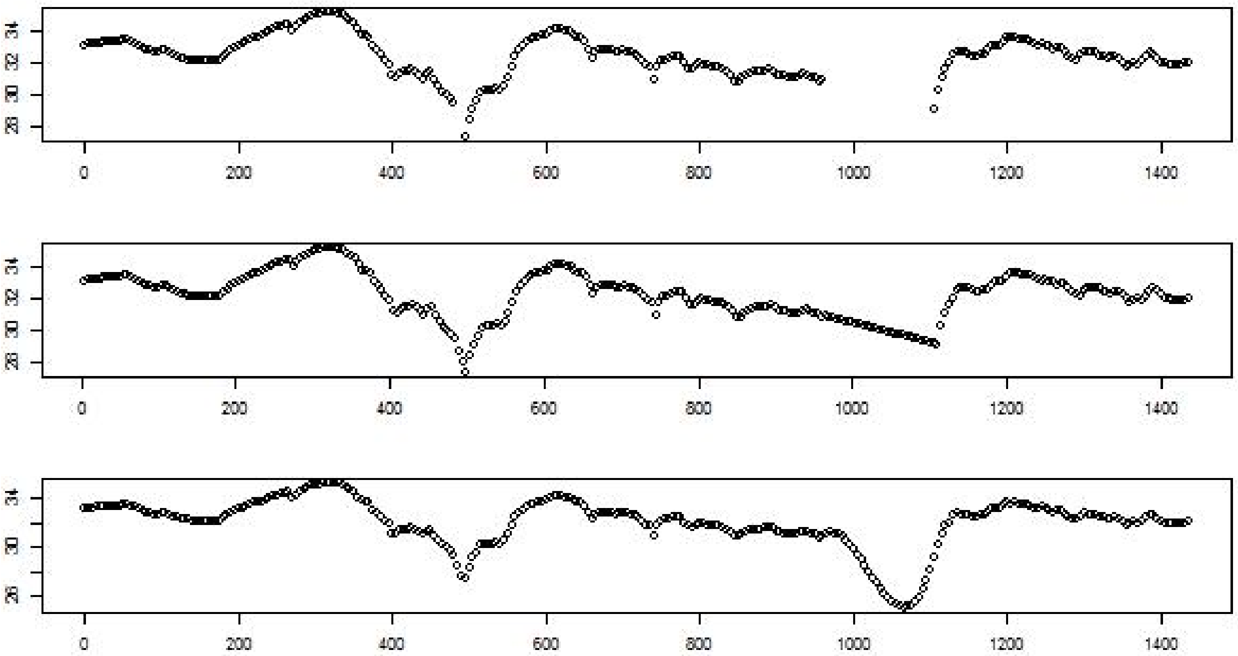
Missing Data Example

Time series of glucose measurements

raw data

linear interpolation
(too linear)

polynomial interpolation
(too nonlinear)

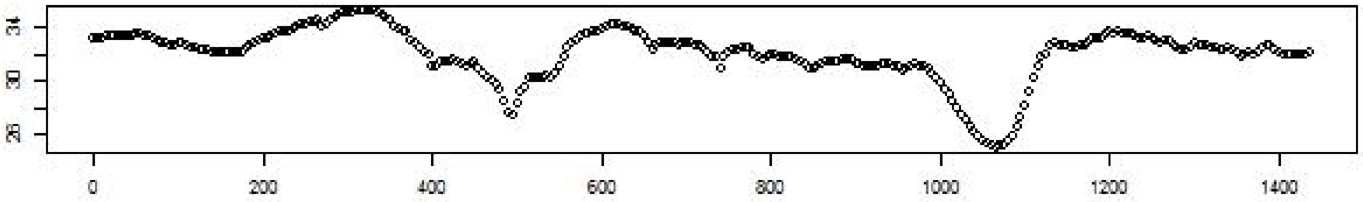


Time (minutes)

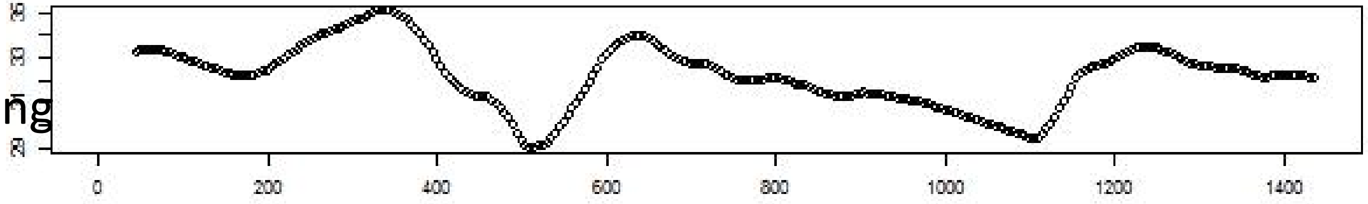
Missing Data Example

Time series of glucose measurements

polynomial
interpolation
(too nonlinear)



polynomial
interpolation then
smoothed by averaging
over windows
(better, but trade offs?)



Time (minutes)

Variable Transformations

- Why transform data?
  - **Combine attributes**

    ratios can be more useful
  - **Normalizing data**

    to same scale
  - **Simplifying data**

    discrete data is often more intuitive for user and algorithm and helps the algorithms

Feature Engineering is Variable Enhancement

- Use Domain and world knowledge to help model
- Example: variables exist that represent date and location of doctor visits
  - deduce a new variable for Number-of-1st-time-visits
  - deduce a new variable for Number-of-visits-over-25-miles
  - deduce a new variable for Amount-of-time-between-visits

Adding Information As Variable Enhancement

- Example: zip codes
  - Change ZIP to latitude and longitude
  - Change ZIP to miles to a reference point
  - Change ZIP to known category (H,M,L income)
  - Change ZIP to set of indicator variables (1 per ZIP)

Discretization/Binning May Enhance Data

- Discretization
  - A continuous attribute divided into intervals and replaced by Interval labels
  - E.g. replace age by functional concepts (such as young, middle-aged, or senior) which may have better predictive value

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning:
  - It divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Discretization/Binning Options

- E.g. Equal-width (distance) partitioning:
  - *N* intervals of equal size, but outliers skew range

64 65 68 69	70 71 72	75	80 81 83	85
Yes No Yes Yes	Yes No No Yes	Yes Yes	No Yes Yes	No

- **E.g. Equal-depth (frequency) partitioning:**
  - *N* intervals, of equal sample frequency, can help scale data

64 65	68 69 70	71 72	75 80	81 83 85
Yes No	Yes Yes Yes	No No Yes	Yes Yes No	Yes Yes No

Is 85 special?

Variable Transformation Summary

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Introduce/re-label/categorize variable values
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction

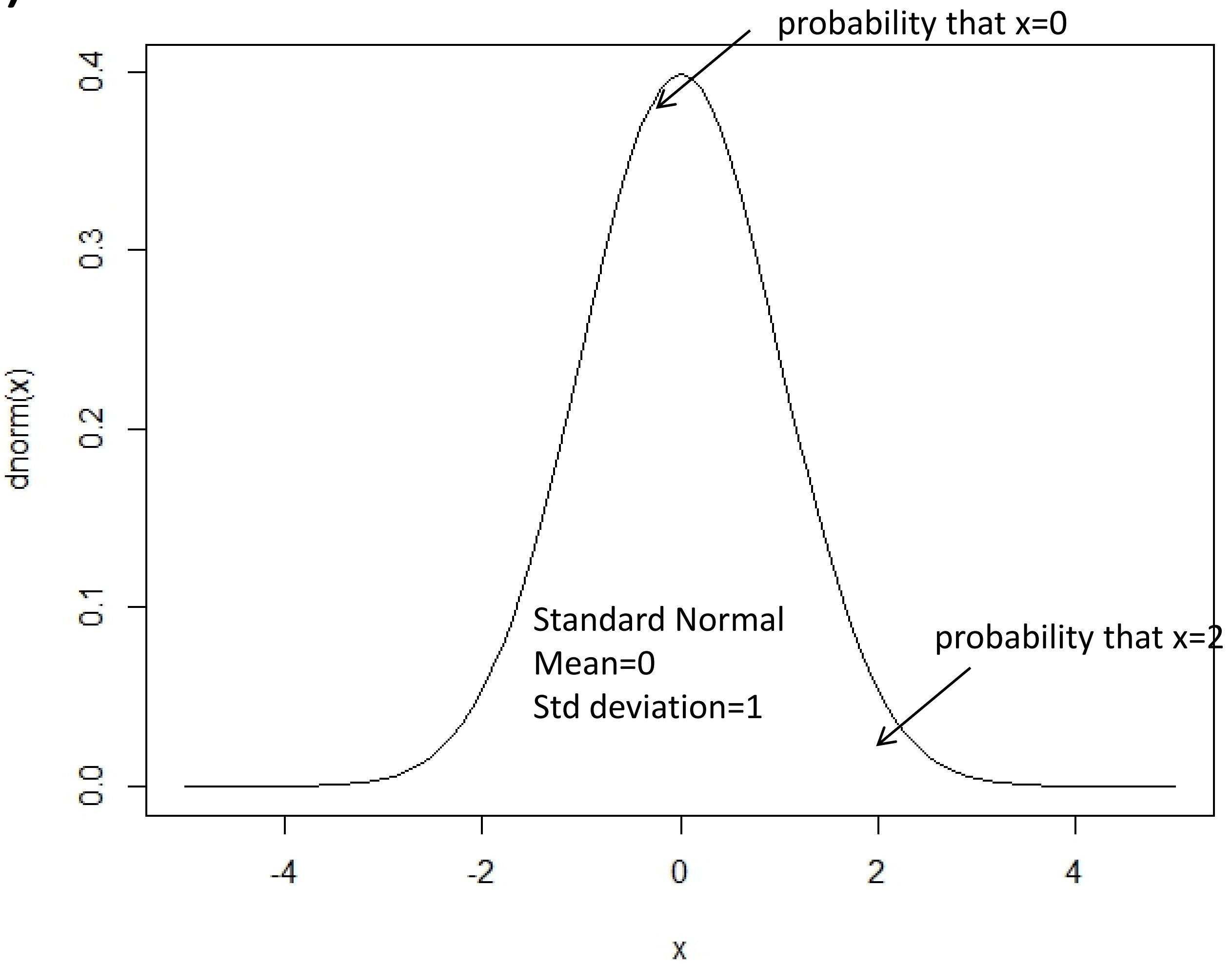
Outline

- Motivation and Goals
- What is data?
- Data Preparation:
  - Organizing data (structural issues)
  - Preprocessing   (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

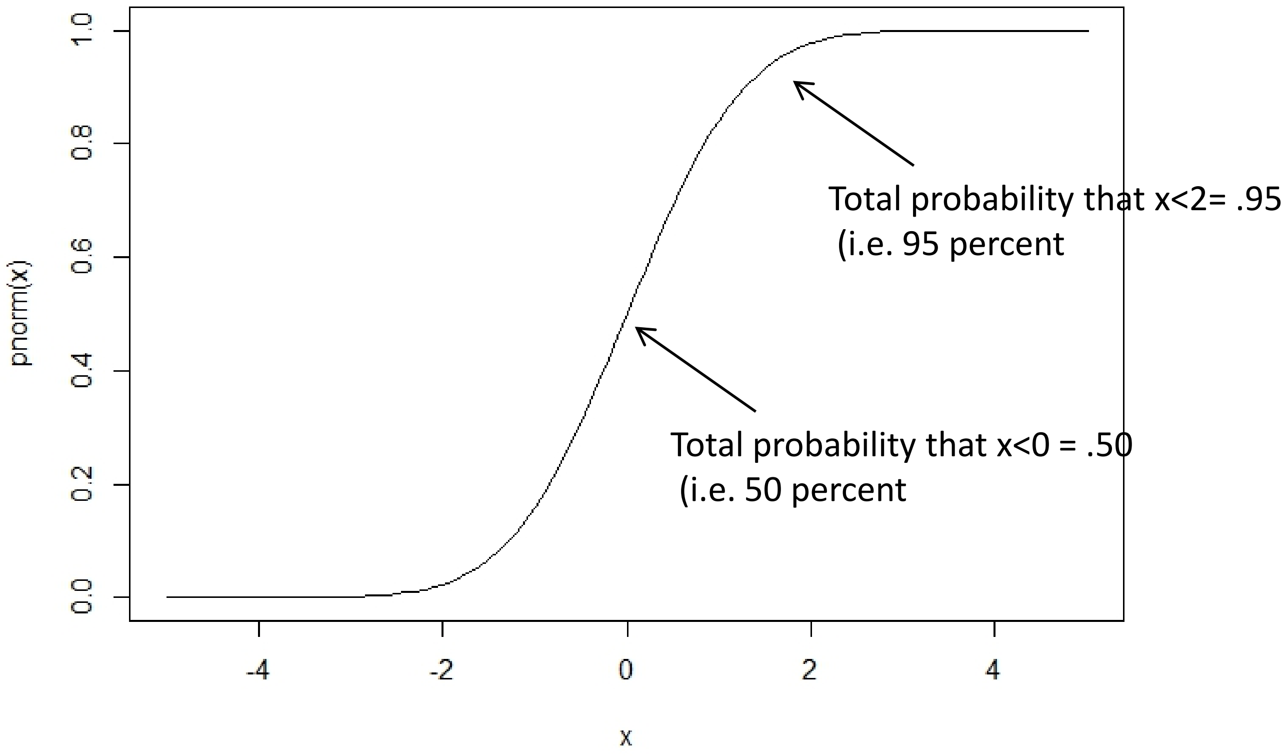
Stats for Data Preprocessing

- Distributions and histograms
  - Continuous variables (functions and graphs)
  - Discrete variables (sets and counting)
- Normalizations
- Correlations

Normal Probability Density Function (PDF)

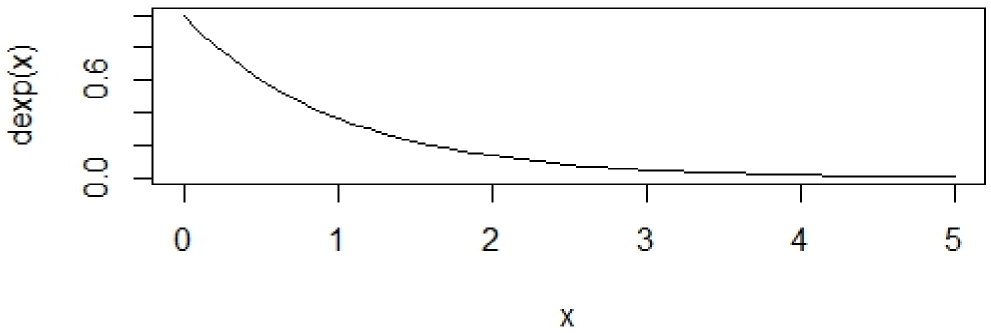


Normal Cumulative Distribution

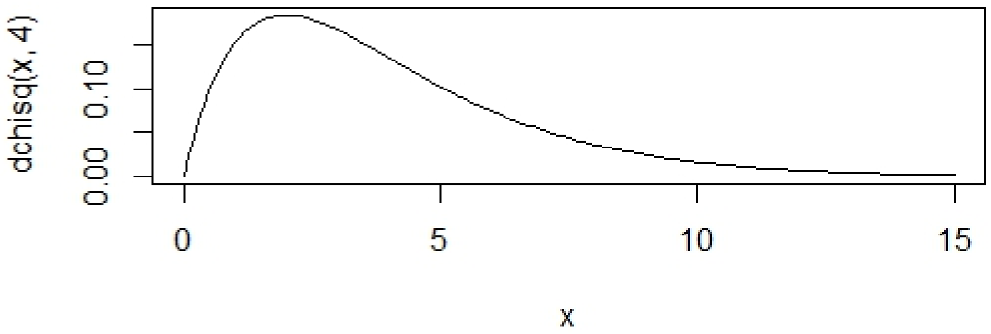


Exponential and Chi-squared density functions

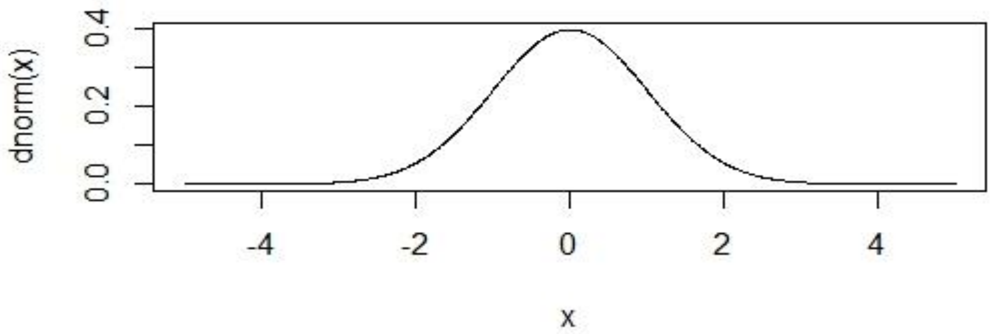
Exponential is good for ‘counts’, ‘events’, etc…,
ie, items that are >0, usually near 0, and higher values more rare



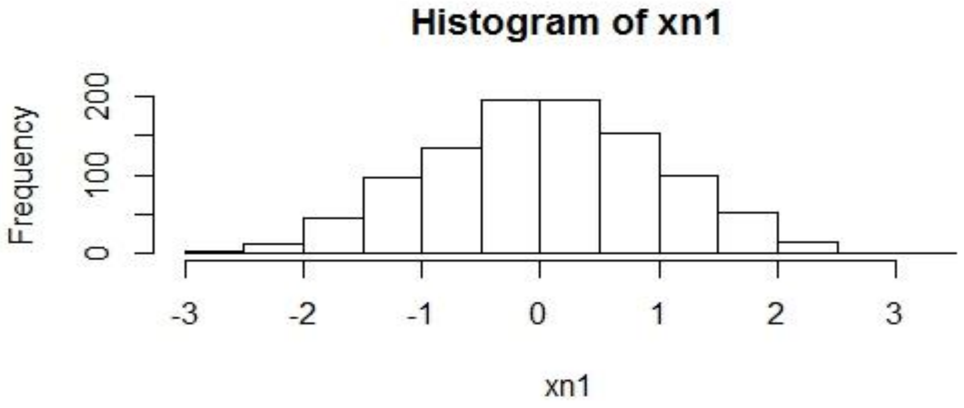
Chi Square is good for ‘costs’, ‘rates’, ‘salaries’, etc…,
ie, items that are > 0, usually not near 0, and higher values more rare



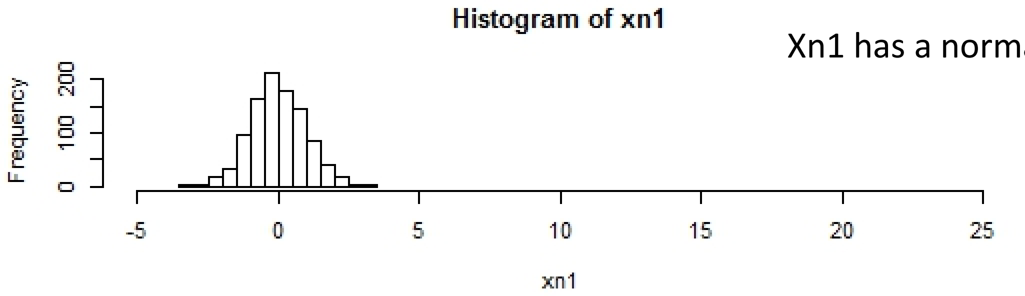
Histogram is a sample PDF



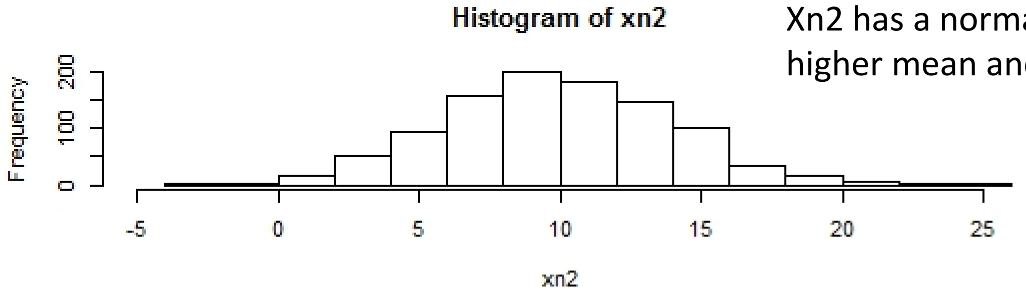
Frequency count ~
probability times sample size



One histogram as mixture

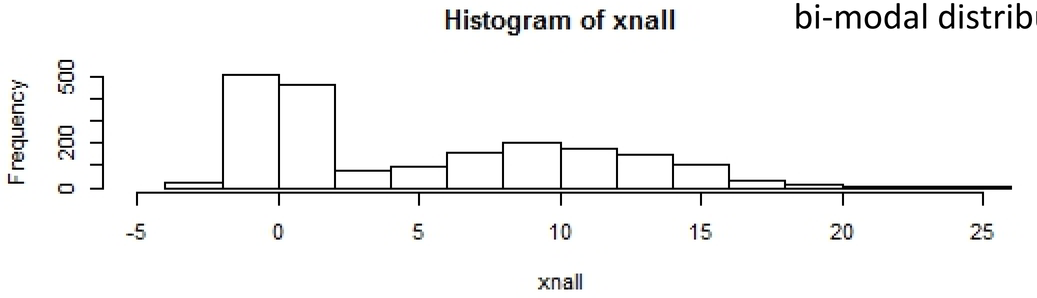


Xn1 has a normal distribution



Xn2 has a normal distribution with higher mean and higher variance

Xn1 + Xn2 has a bi-modal distribution



Descriptive Statistics

- Mean and Std Dev summarize variables

$$\mathrm{std}(x, y) = \sqrt{\mathrm{mean}((x - \mathrm{mean}(x))^2)}$$

- Transformations and Functions also summarize
  - E.g. take the highest amount charged for customers in a zip code, take that for each zip code and get a new distribution
  - E.g. take the difference of 75th to 25th percentile of all customers in a zip code, take that for each zip code and get a new distribution

Data Transformation: Normalizations (to help with scaling)

- Mean center

$$x_{new} = x - \mathrm{mean}(x)$$

- z-score

$$z - score = \frac{x - \mathrm{mean}(x)}{\mathrm{std}(x)}$$

- Scale to [0...1]

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- log scaling

$$x_{new} = \log(x)$$

More Descriptive Statistics

- Covariance between 2 variables

$$\operatorname{cov}(x, y) \sim \operatorname{mean}((x - \operatorname{mean}(x))(y - \operatorname{mean}(y)))$$

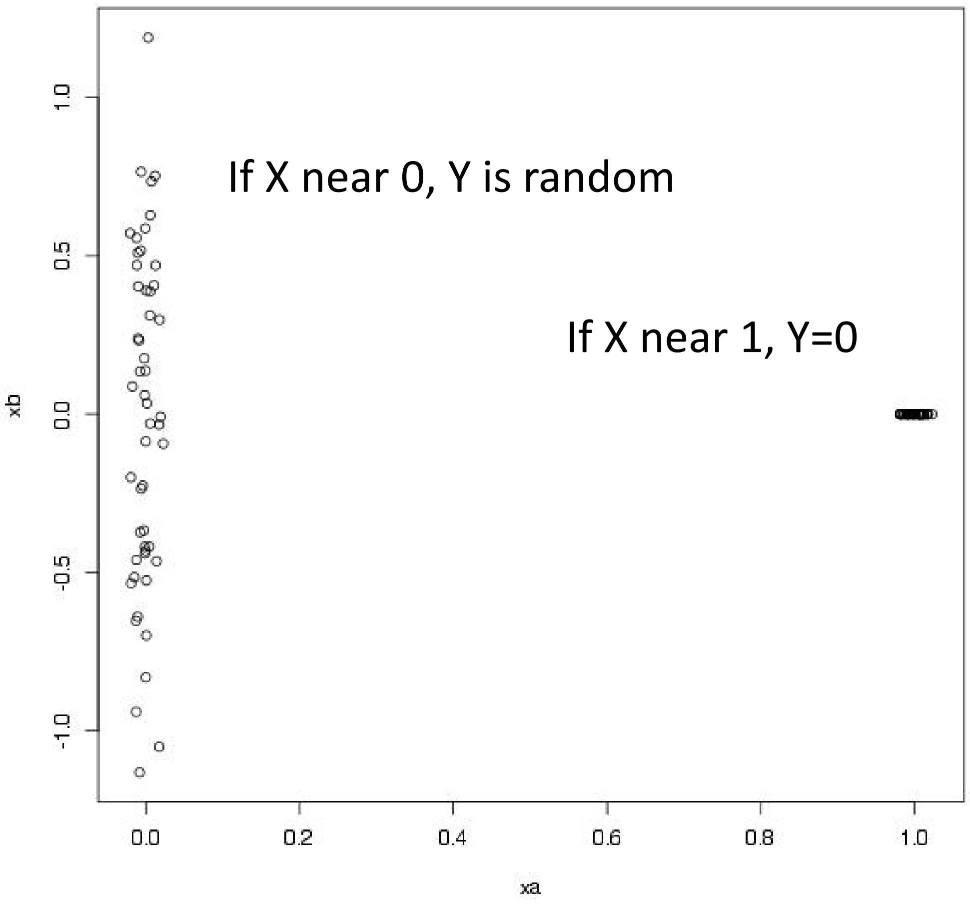
- Correlation between 2 variables

$$\operatorname{corr}(x, y) \sim \frac{\operatorname{cov}(x, y)}{\operatorname{std}(x)\operatorname{std}(y)}$$

  - Ranges -1 to 1
  - Represents linear relationship

Correlation vs. Independence

- No Correlation => Independence



Correlation = .021
But Y depends on X

More Descriptive Statistics

- (Spearman) Rank correlation between 2 variables
  - Rank the instances of each variable
    (now there are 2 ordinal rank variables)
  - Take correlation coefficient of ranks
  - Represents monotonic relationship
- Confidence interval wrt mean or percentiles

$$\text{mean}\,(x) - \text{std}(x), \text{mean}\,(x) + \text{std}(x)$$

$$15th\ \text{percentile}, 85th\ \text{percentile}$$

Outline

- Motivation and Goals
- What is data?
- Data Preparation:
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

Exploratory Stats for More Variables

- Descriptive Statistics Guidelines
    - Get means and variances, do histograms…
    - Feature engineering with summary statistics and functions
- But for many variables need other steps/tools
    - More stats
    - Large P variable selection
    - Large P dimension reduction
    - Sampling

Many Variables

- More variables => more information, but also more noise and more ways of interactions

- 2 ways to handle many variables
  - Variable Selection
  - Dimension reduction methods

Variable Selection vs. Dimensionality Reduction

- Prior to algorithm, depends on data

  - For large P, with noise particular to variables, try variable selection

  - For large P, diffuse noise, try dimension reduction

Variable Selection

- Some algorithms do it already – e.g. random forests will search attribute subsets
  - Select a minimum possible set of features
  - reduce # of features in the patterns, easier to understand

- Heuristic methods (due to large # of choices):
  - remove variables with low correlations to outcome
  - try adding/deleting 1 variable at a time and test algorithm(s)

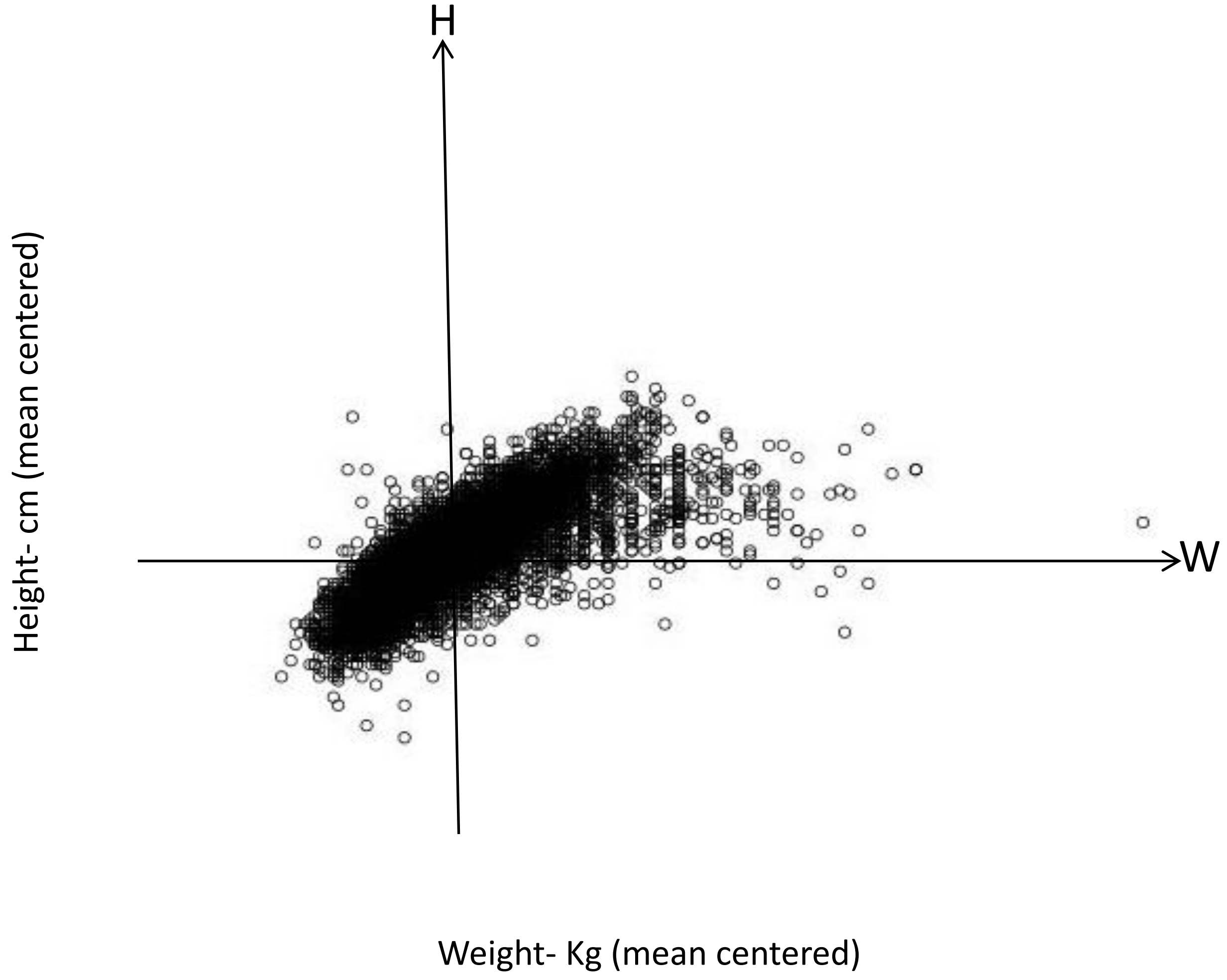
Dimensionality Reduction via Principle Components

- Idea: Given *N* points and *P* features (aka dimensions), can we represent data with fewer features:
  - Yes, if features are constant
  - Yes, if features are redundant
  - Yes, if features only contribute noise (conversely, want features that contribute to variations of the data)

Dimensionality Reduction via Principle Components

- PCA:
  - Find set of *k* vectors (aka factors) that describe data in alternative way
  - First component is the vector that maximizes the variance of data projected onto that vector
  - *K*-th component is orthogonal to all *k*-1 previous components

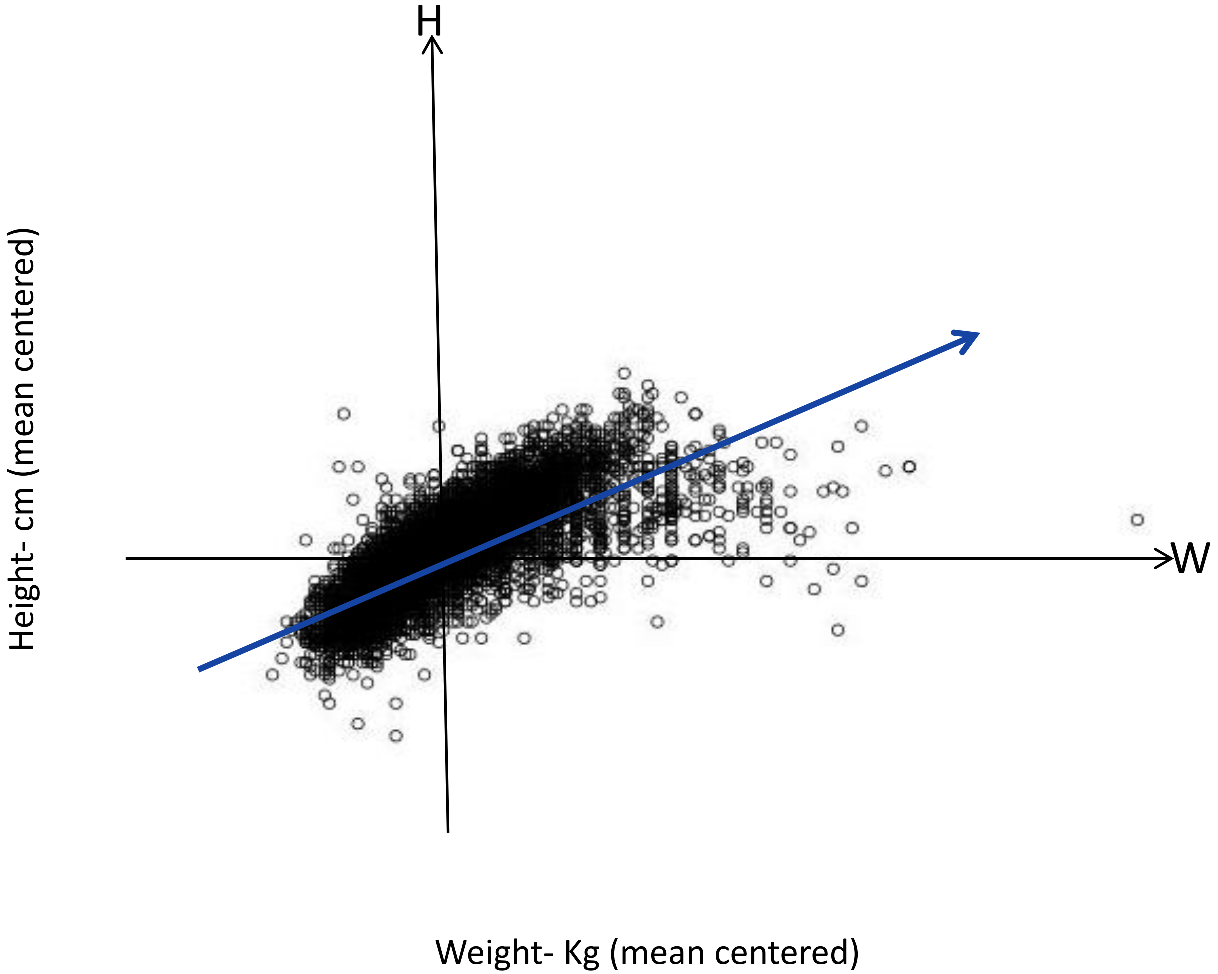
PCA on 2012 Olympic Althetes Height by Weight scatter plot



Idea:
Can you rotate the axis so that the data lines up on one axis as much as possible?

Start with one new axis
(e.g. find the one direction that aligns with data)

PCA on 2012 Olympic Athletes' Height by Weight scatter plot

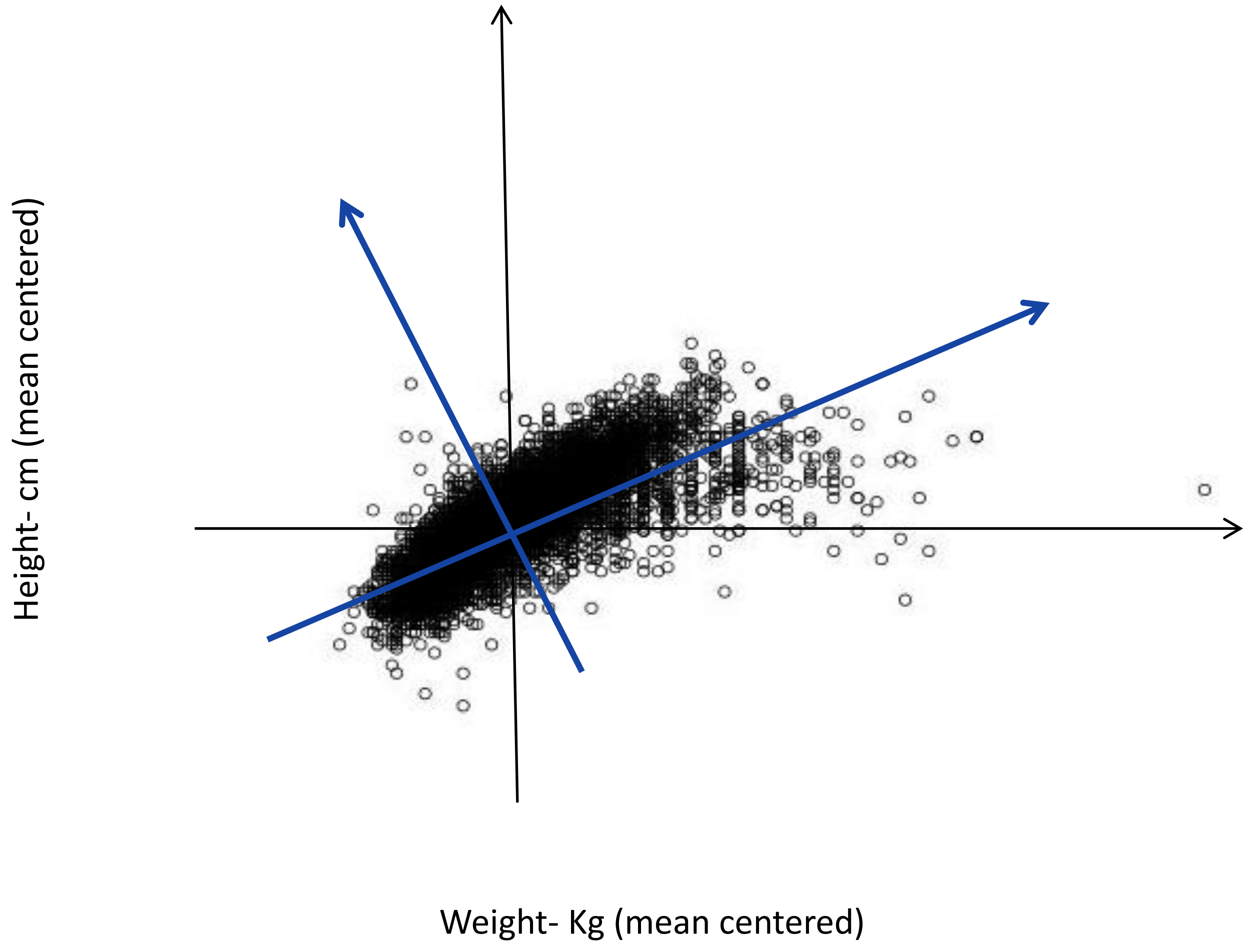


E.g. A new axis:
the line H=.8*W

See how points line up on that line and call that the 1st coordinate – aka project onto line H-.8W=0

Algebraically, the new values will be functions of old H and W axis

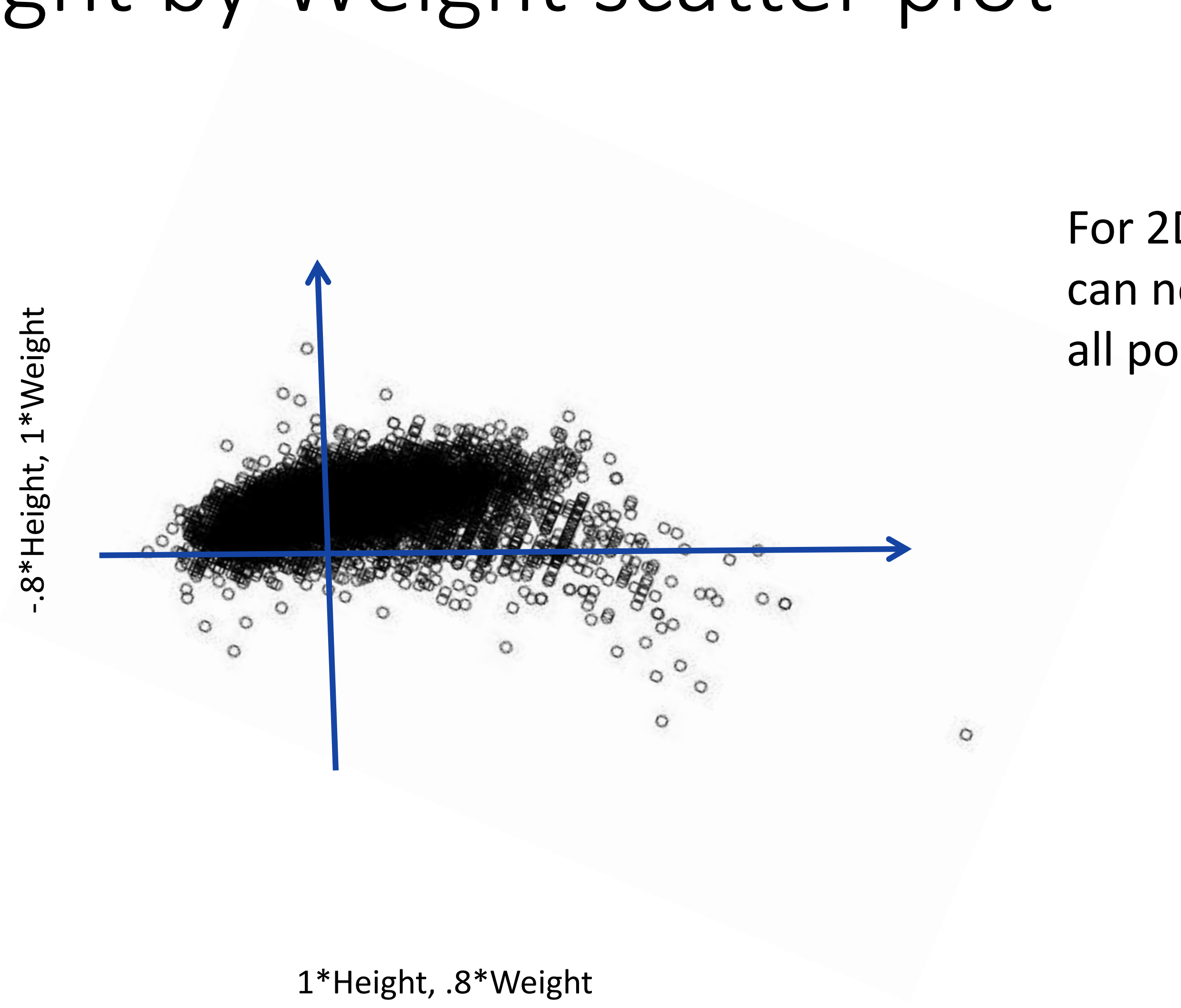
PCA on 2012 Olympic Athletes' Height by Weight scatter plot



If you stop with 1 axis the 2D points are now 1D.

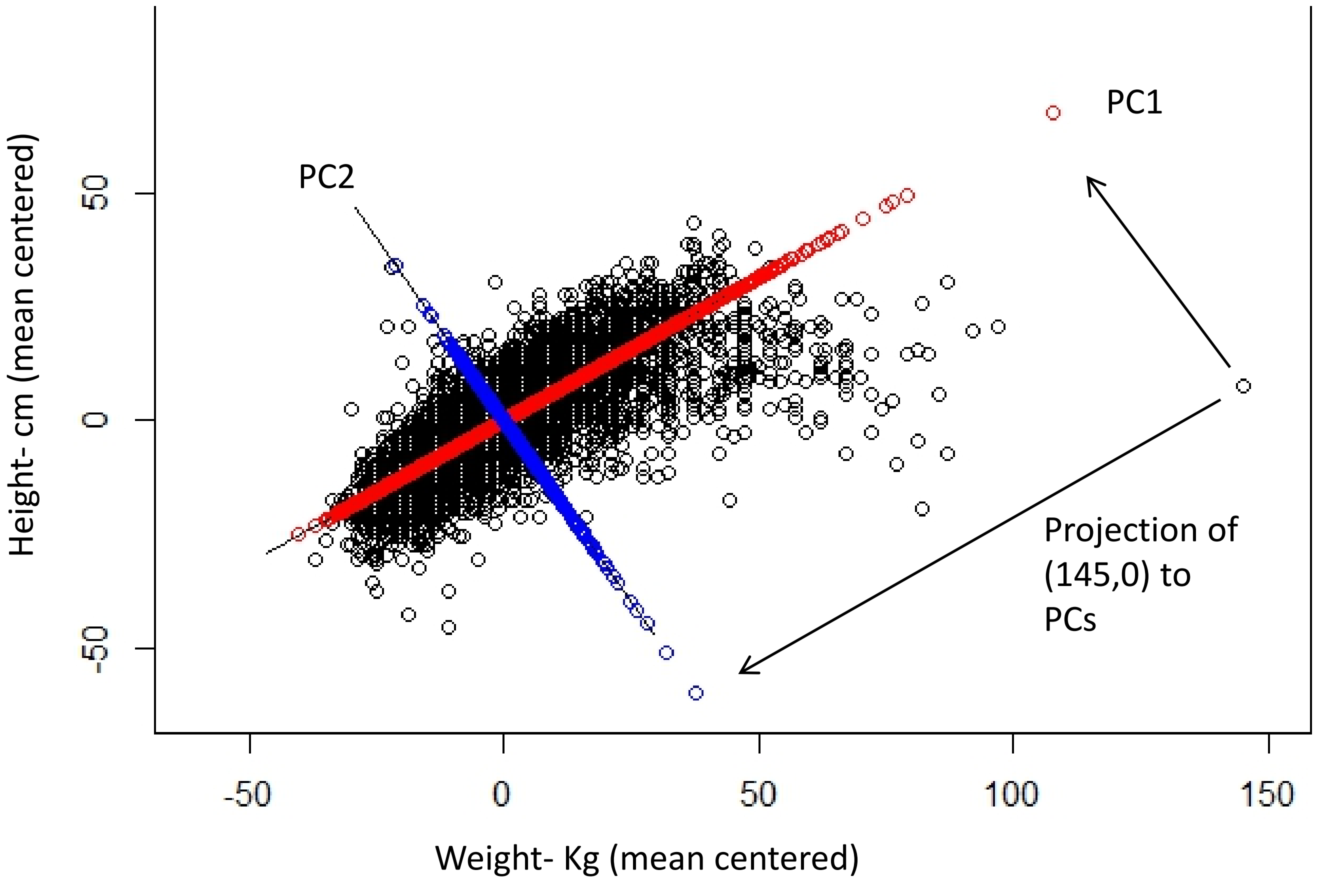
Or, take next axis orthogonal to 1st, continue

PCA on 2012 Olympic Athletes' Height by Weight scatter plot



For 2D data, two new axis can now fully reproduce all points in new space

PCA on Height by Weight scatter plot



Total Variance Conserved:

Var in Weight +
Var in Height
=
Var in PC1 +
Var in PC2

In general:

Var in PC1>
Var in PC2>
Var in PC3…

Principle Components

- Can choose *k* heuristically as approximation improves, or choose *k* so that 95% of data variance accounted
- aka Singular Value Decomposition

  PCA on square matrices only

  SVD gives same vectors on square matrices
- Works for numeric data only
- For higher dimensional data, use PCA to visualize 2 factors at a time

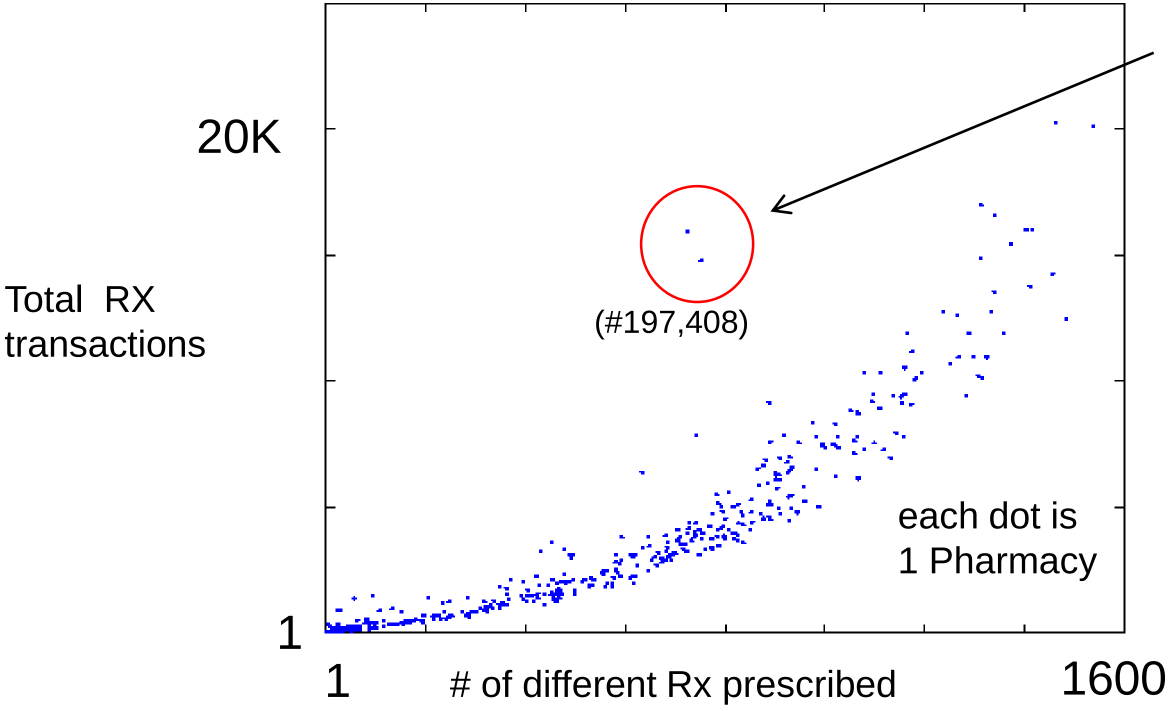
Outline

- Motivation and Goals
- What is data?
- Data Preparation:
  - Organizing data (structural issues)
  - Preprocessing (data value issues)
  - Exploring Variables and Descriptive Statistics
  - Exploring Data Matrix
  - Outliers, Anomalies, and Visualizations

Anomalies

- 3 working definitions of an anomaly
  - statistical outlier (far from mean)
  - distance based (farthest point to its neighbors)
  - deviance based (model quantity, take biggest error to model)
- Making decisions and cutoffs
  - anomalies can be ranked
  - but decisions depend on some cutoff

The importance of normalization and varieties of deviance



Not an outlier in
Rx or in total Rx

Outliers for #Rx/total Rx

Far from others

Deviant wrt main trend

Visualizations

- For communication and exploration
- MultiDimensional Scaling (MDS)
  - Find points in 2D that preserve relative distances in P-dimensions of full data matrix
  - In some cases similar to PC1 and PC2
- Plotting relations between variables
- Heat Maps over vectors
  - Discretize into bins and labeled by a few colors

Ten golden rules

1. Select clear problem with tangible benefit

2. Specify required solution

3. Define how solution is implemented

4. Understand the domain

6. Stipulate assumptions

5. Let the problem drive the modeling

7. Refine the model iteratively

8. Make the model as simple as possible (but no simpler)

9. Find areas of instability

10. Find areas of uncertainty

Summary

- Data preparation is a key issue for mining
- Lots of techniques
- Partly an art that depends on data and algorithm knowledge
- Partly a science that depends on statistical principles

Reading Material

- **Data Preparation for Data Mining** by Dorian Pyle
  - http://www.ebook3000.com/Data-Preparation-for-Data-Mining_88909.html
- **Data mining – Practical Machine learning tools and techniques** by Witten & Frank
  - http://books.google.com
- Paper: “Tidy Data” by Hadley Wickham; Journal of statistical software

Exercise in Weka

Exploring Variable characteristics

Adding Variables

Viewing Correlations

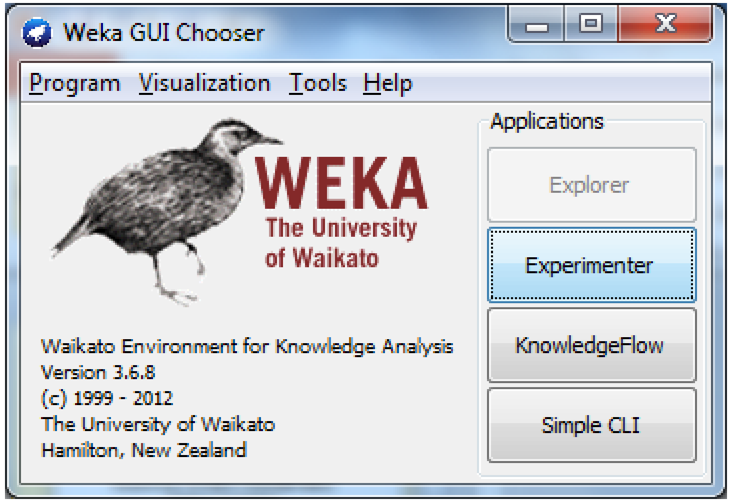
Explore Variables in Weka

- Dataset of London 2012 Olympians

  download AHW_1.CSV
- View histograms
- View correlation (visualization)
- Adding new variables
- Consider Filters and Transformations
- Get missing data stats

Weka Exercise

- Open Weka
- Choose Explorer



Weka Exercise

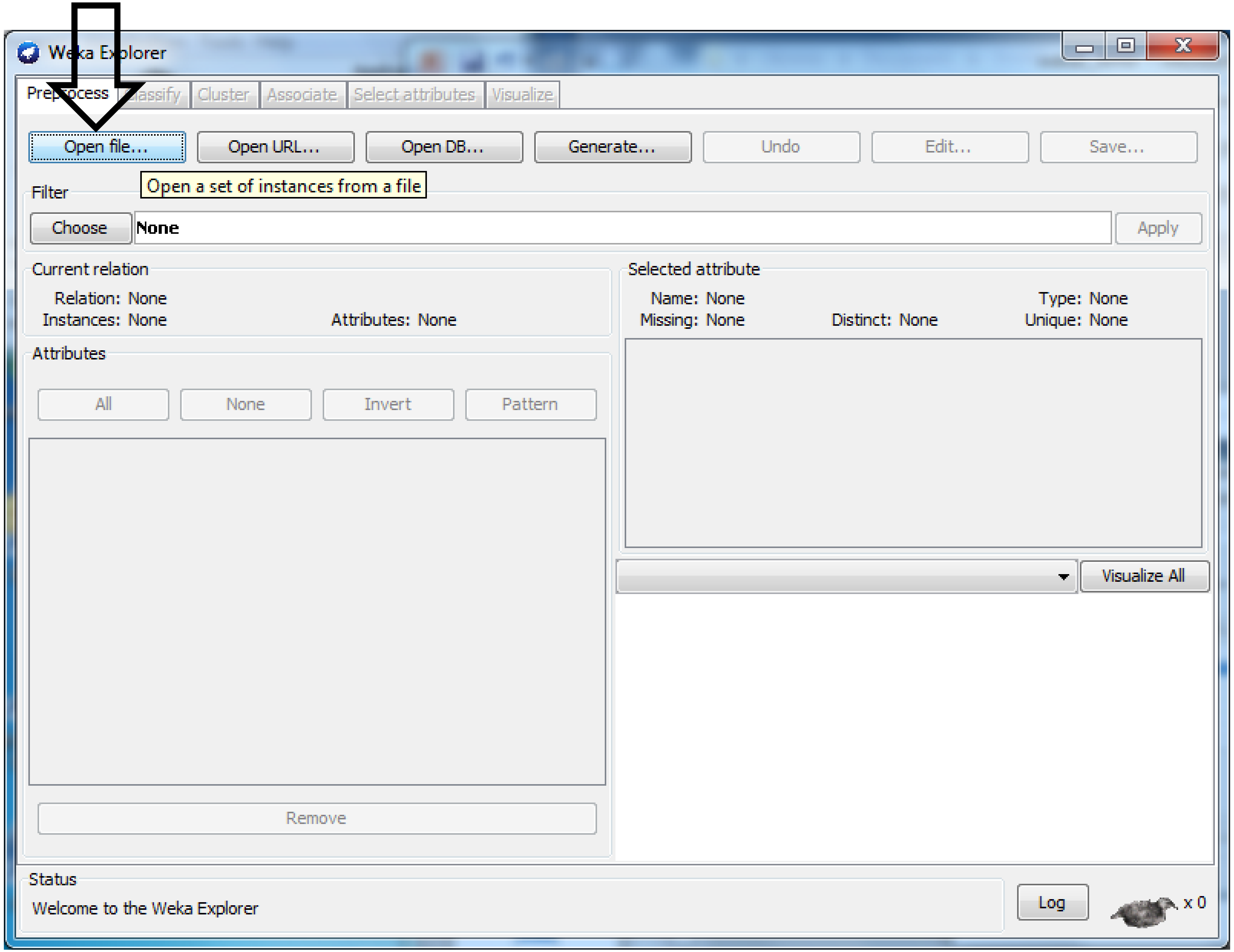
- Open Athletes height and weight file (ahw_1.csv)

(preprocessing tab, open file, select csv as type, select ahw_1.csv)

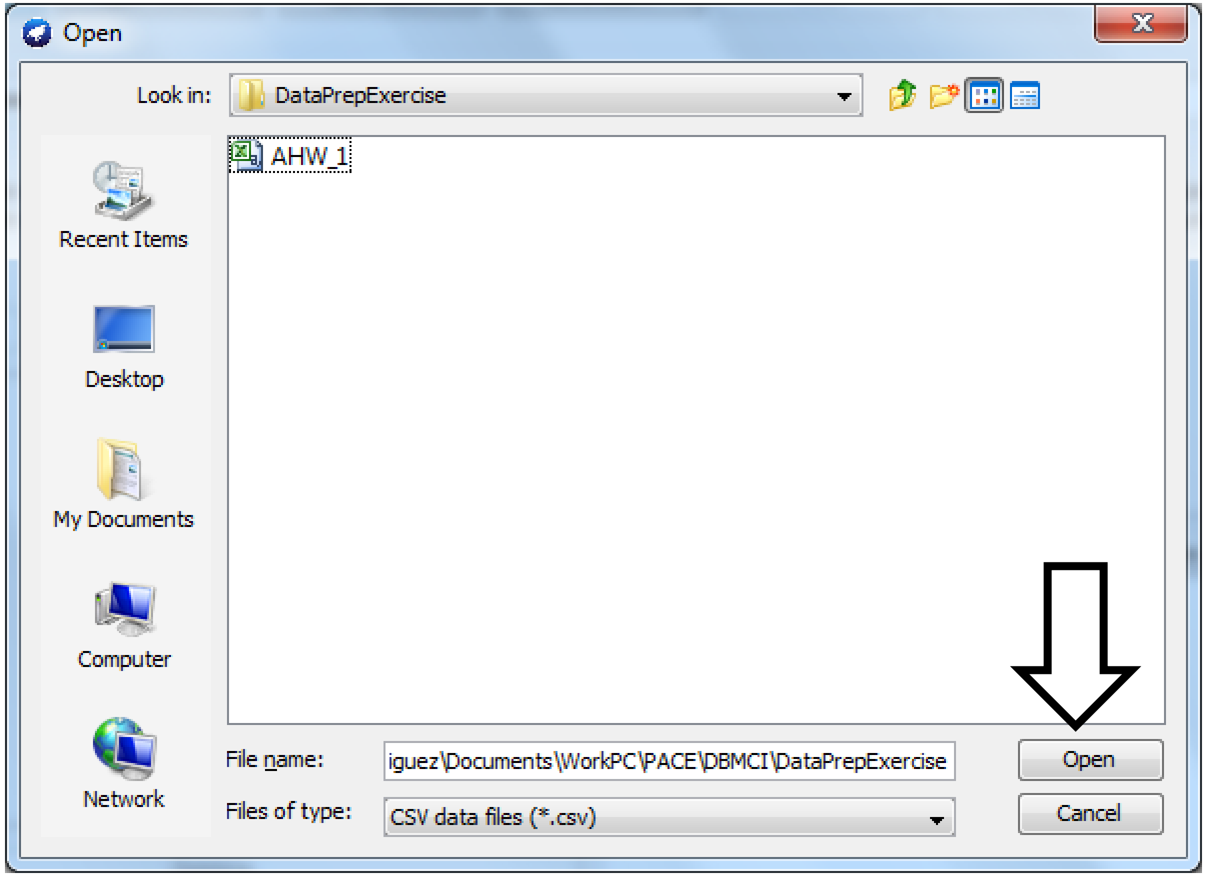
What are the statistical distributions of variables using no class?

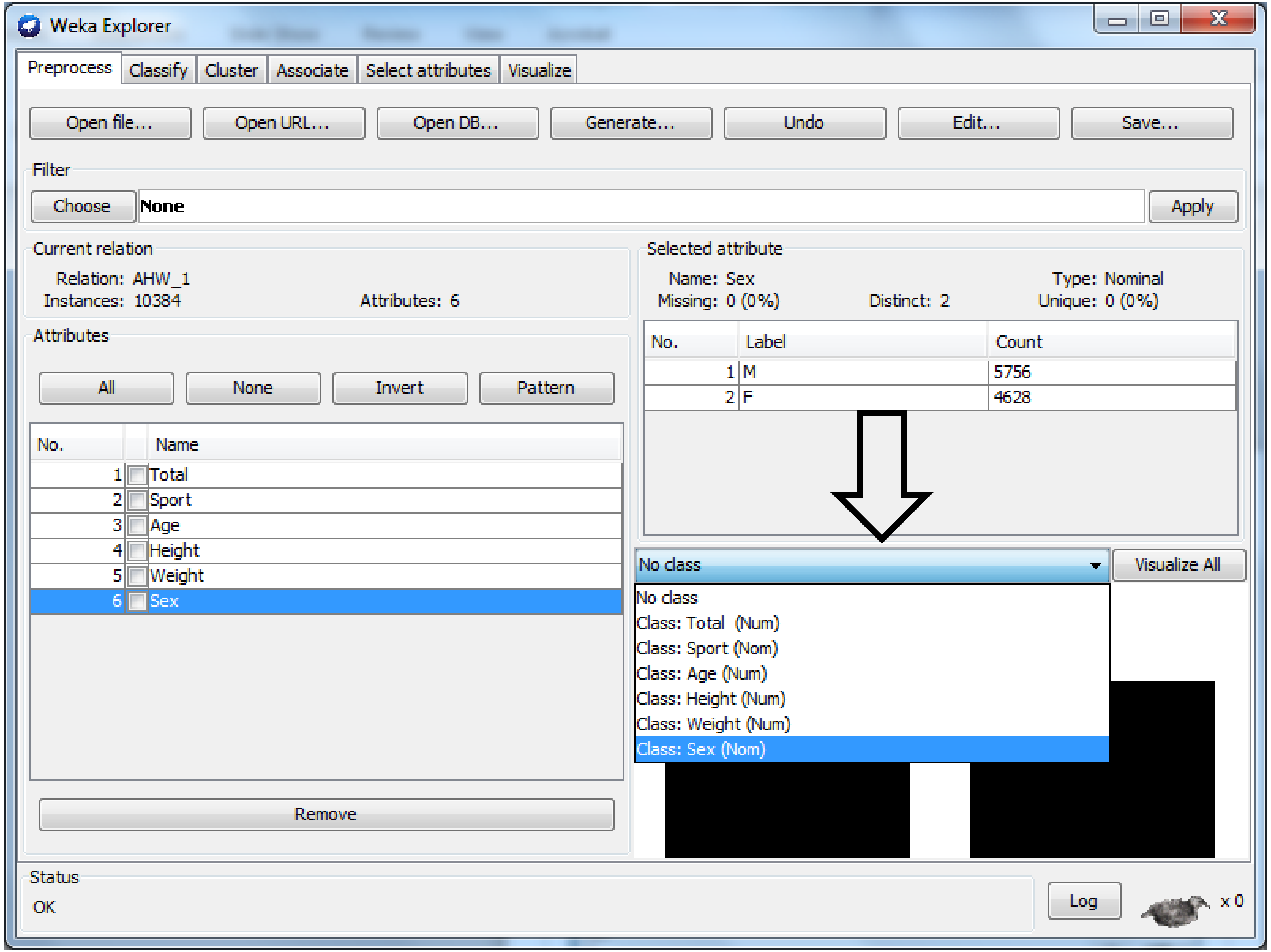
How do distributions differ by sex?

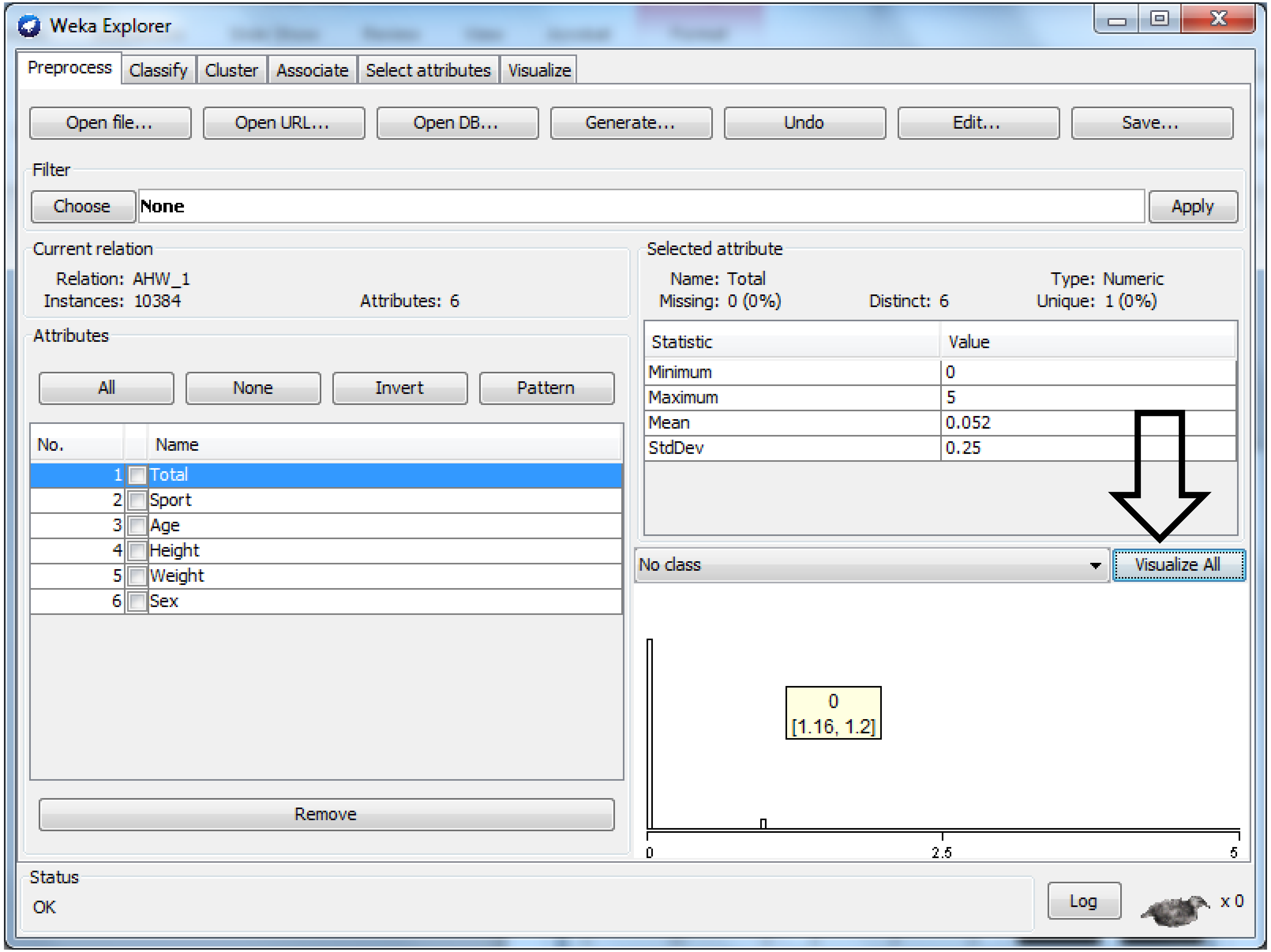
(hint use sex as the class nominal)

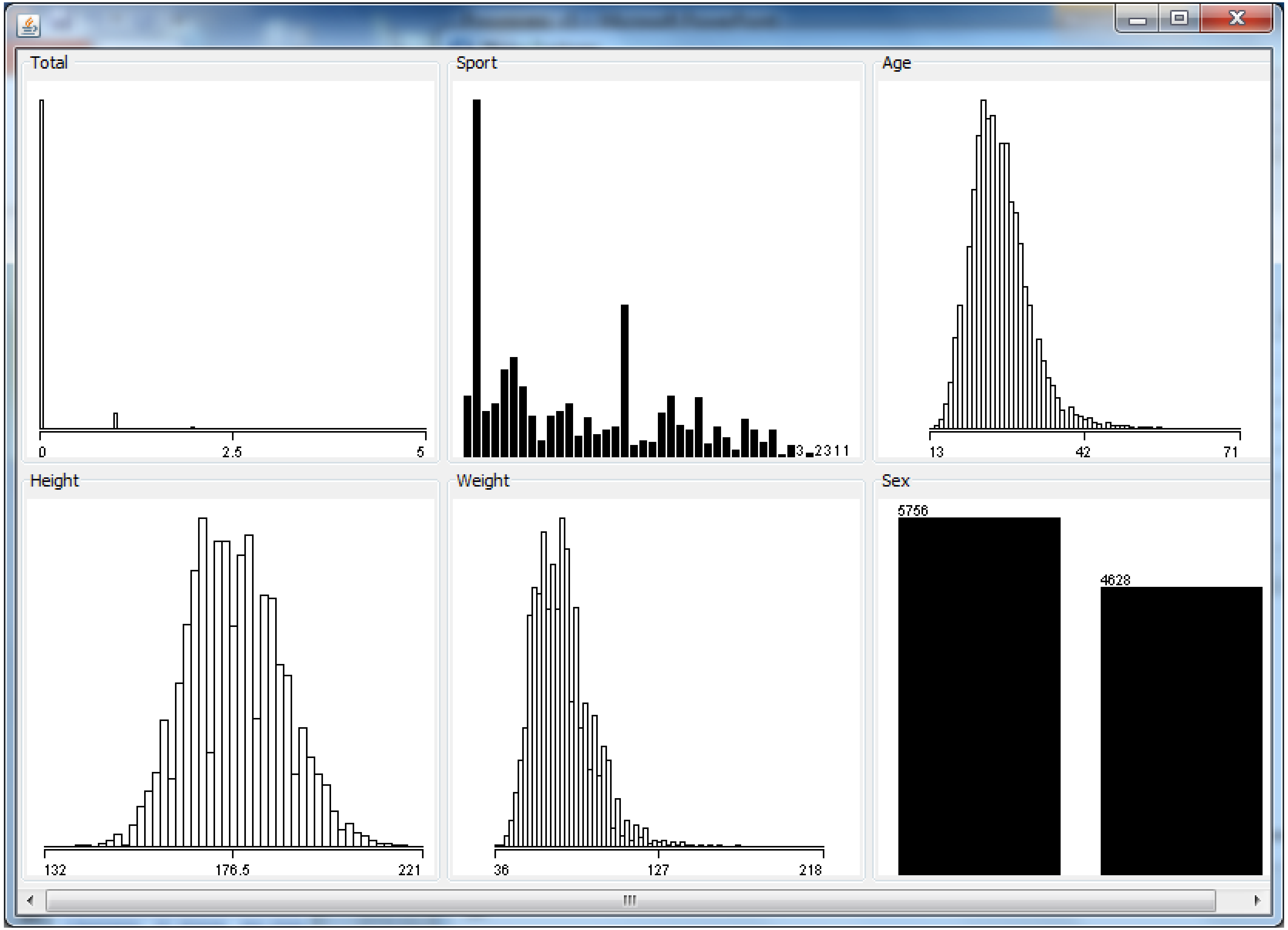


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Open a set of instances from a file
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All
Remove
Status
Welcome to the Weka Explorer
Log
x 0

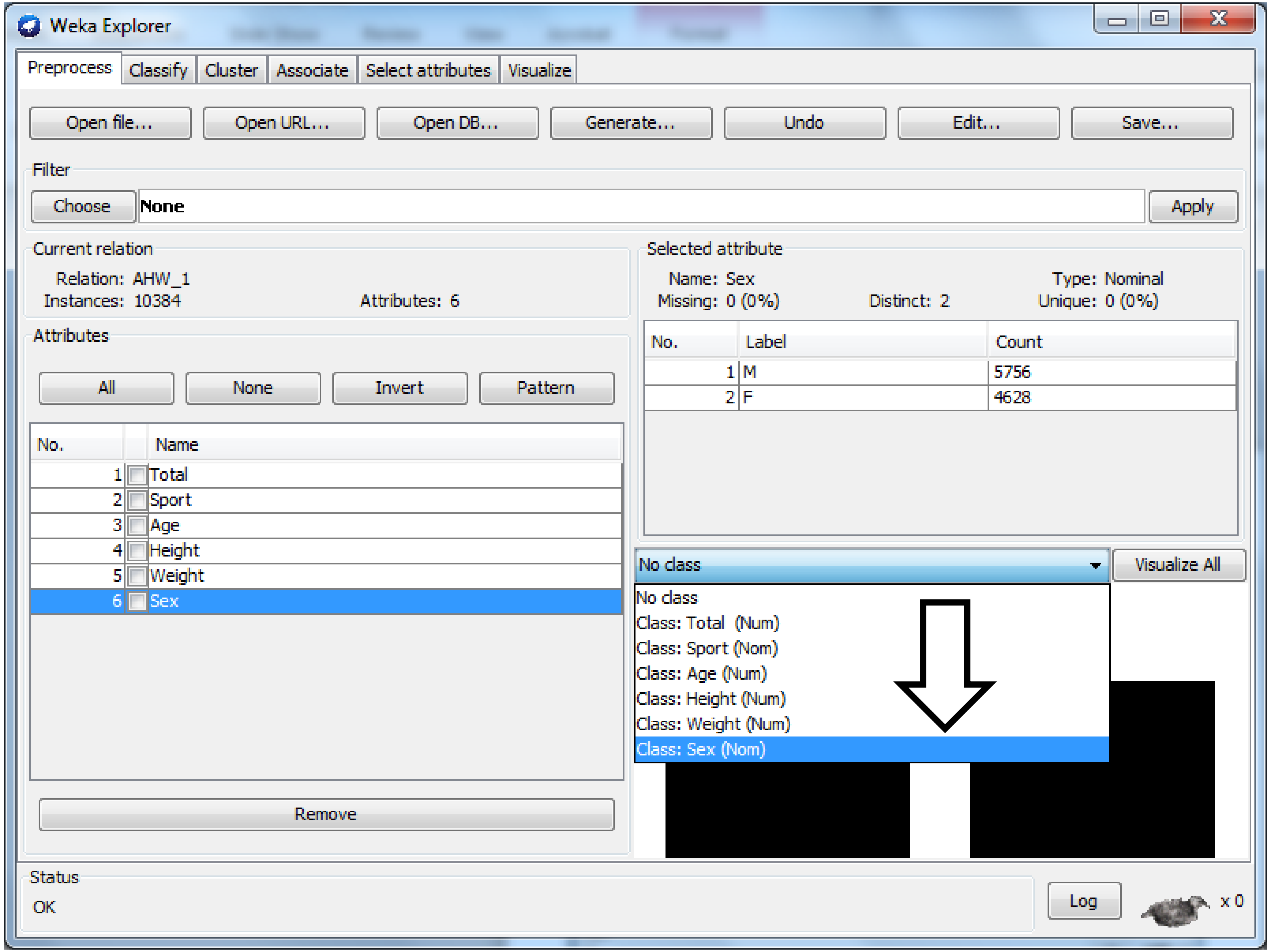

Open
Look in:
DataPrepExercise
AHW_1
Recent Items
Desktop
My Documents
Computer
Network
File name:
iguez\Documents\WorkPC\PACE\DBMCI\DataPrepExercise
Files of type:
CSV data files (*.csv)
Open
Cancel

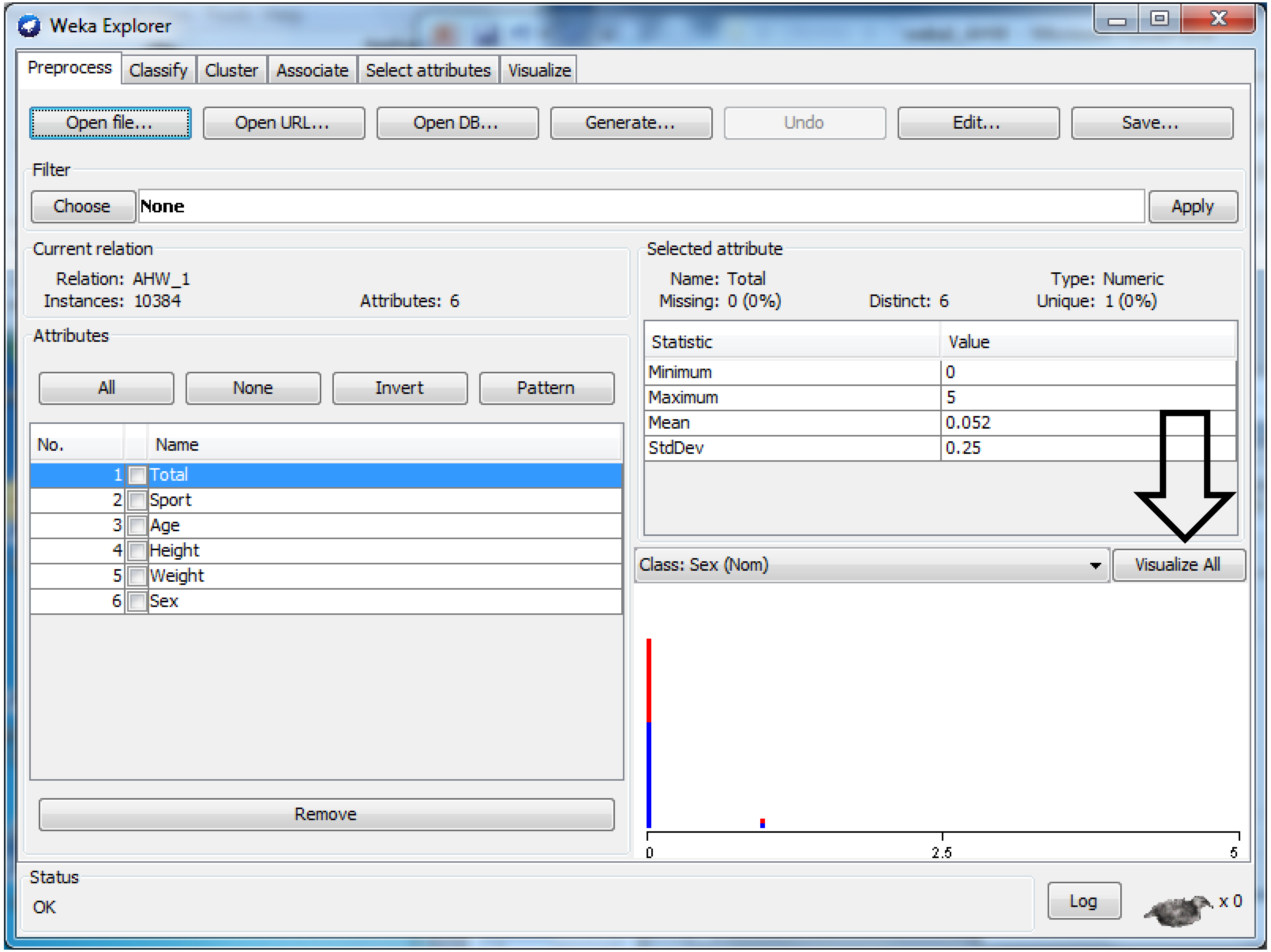

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: AHW_1
Instances: 10384
Attributes: 6
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Selected attribute
Name: Sex
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
No.
Label
Count
1 M 5756
2 F 4628
No class
Visualize All
No class
Class: Total (Num)
Class: Sport (Nom)
Class: Age (Num)
Class: Height (Num)
Class: Weight (Num)
Class: Sex (Nom)
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: AHW_1
Instances: 10384
Attributes: 6
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Selected attribute
Name: Total
Type: Numeric
Missing: 0 (0%)
Distinct: 6
Unique: 1 (0%)
Statistic
Value
Minimum 0
Maximum 5
Mean 0.052
StdDev 0.25
No class
Visualize All
0
[1.16, 1.2]
0
2.5
5
Status
OK
Log
x 0

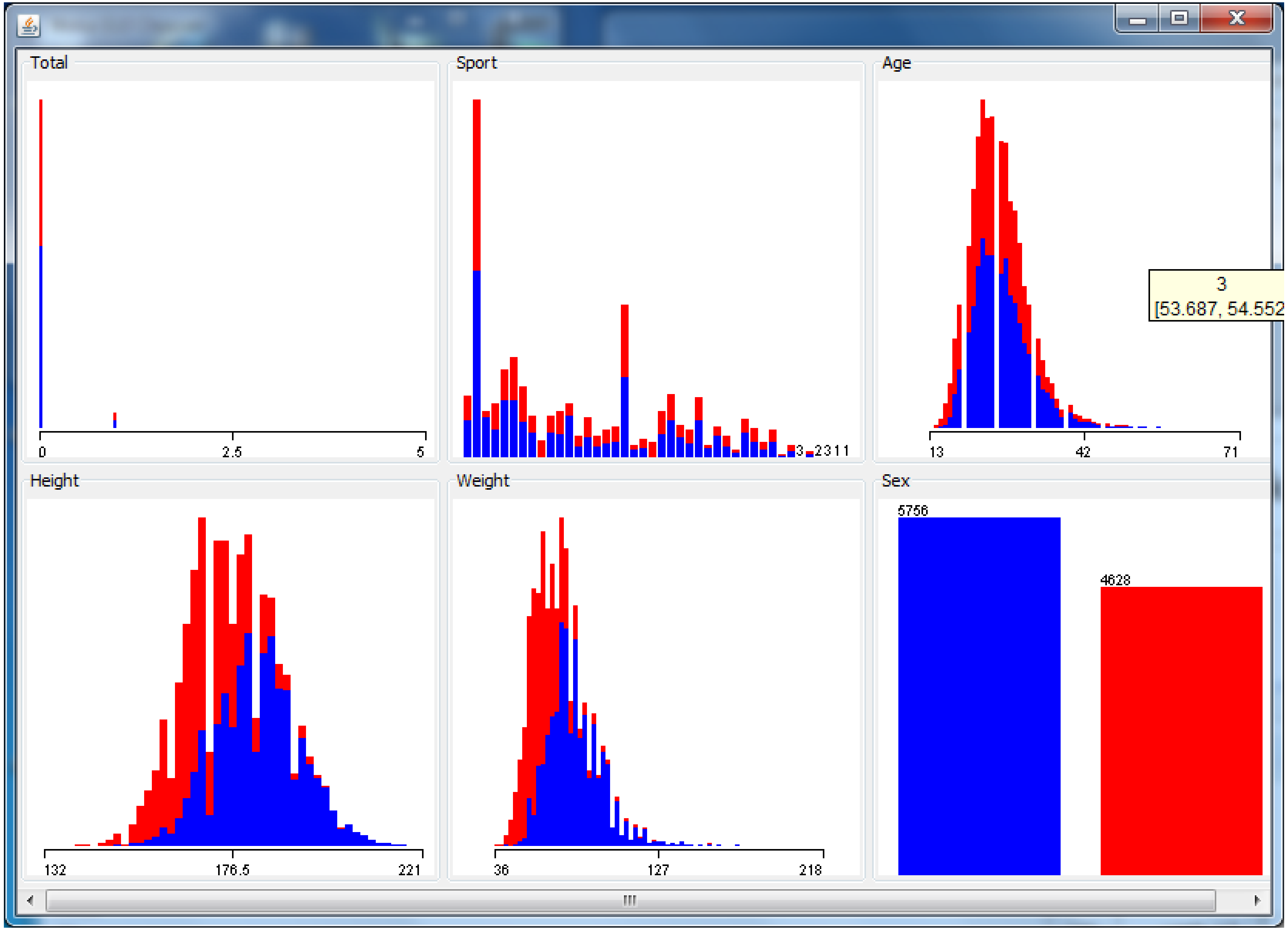


Total
0
2.5
5
Sport
3
2
3
1
1
Age
13
42
71
Height
132
176.5
221
Weight
36
127
218
Sex
5756
4628


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: AHW_1
Instances: 10384
Attributes: 6
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Selected attribute
Name: Sex
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
No.
Label
Count
1 M 5756
2 F 4628
No class
Visualize All
No class
Class: Total (Num)
Class: Sport (Nom)
Class: Age (Num)
Class: Height (Num)
Class: Weight (Num)
Class: Sex (Nom)
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: AHW_1
Instances: 10384
Attributes: 6
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Selected attribute
Name: Total
Type: Numeric
Missing: 0 (0%)
Distinct: 6
Unique: 1 (0%)
Statistic
Value
Minimum
0
Maximum
5
Mean
0.052
StdDev
0.25
Class: Sex (Nom)
Visualize All
0
2.5
5
Status
OK
Log
x 0

Are variables different for Male and Female?



Weka Exercise

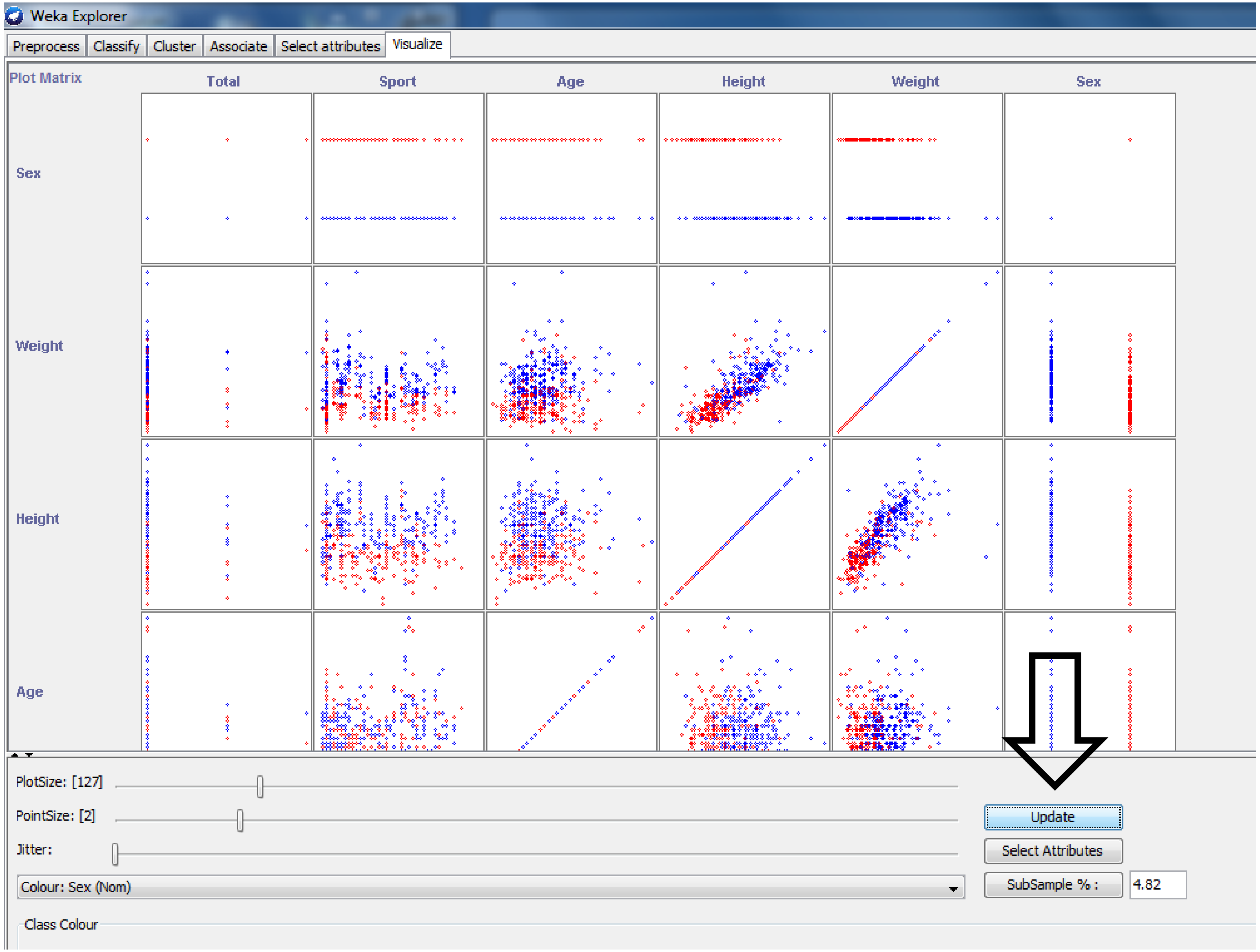
- Visualize scatter plots.

    (visualize tab)

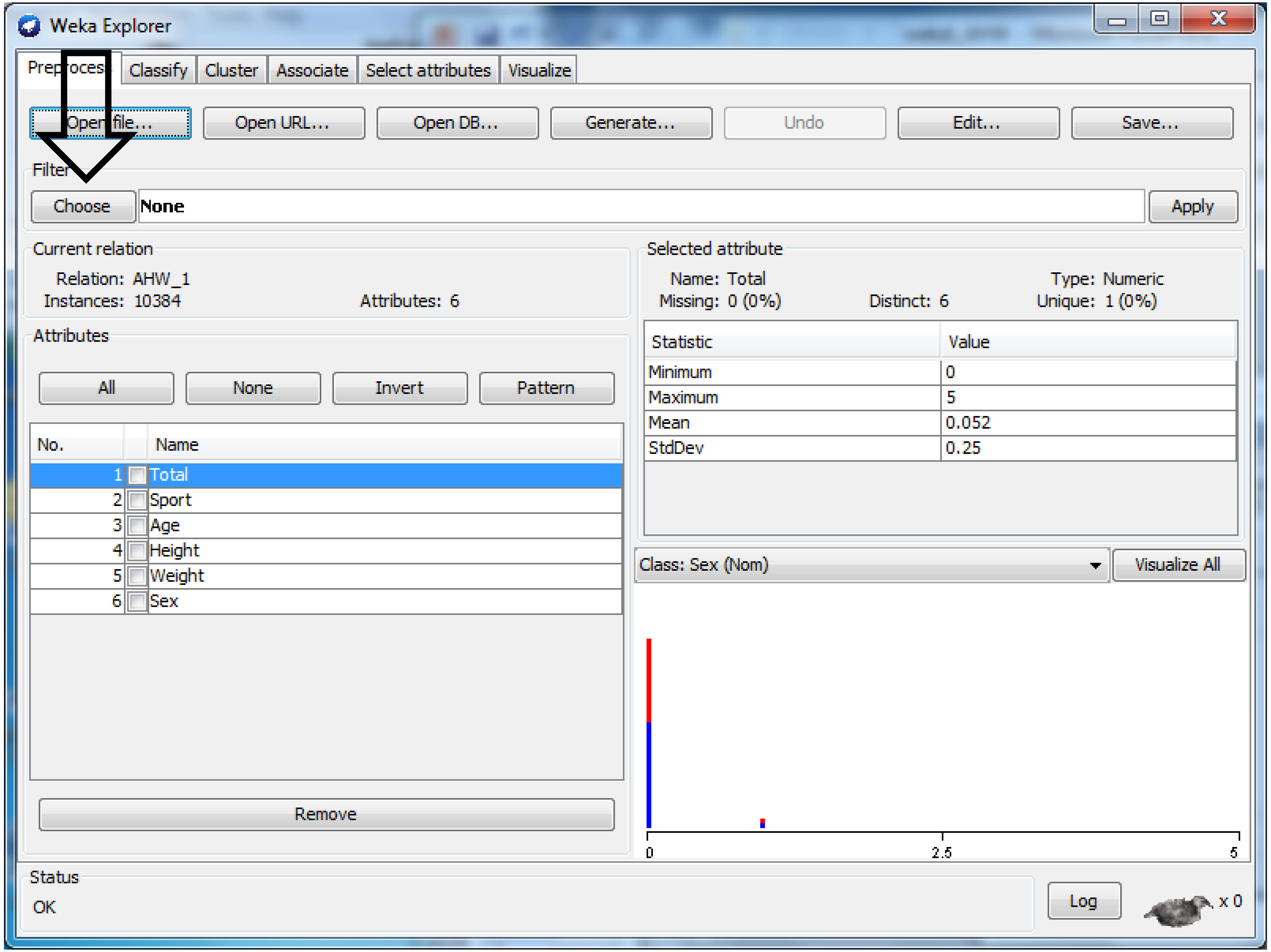
- Q:

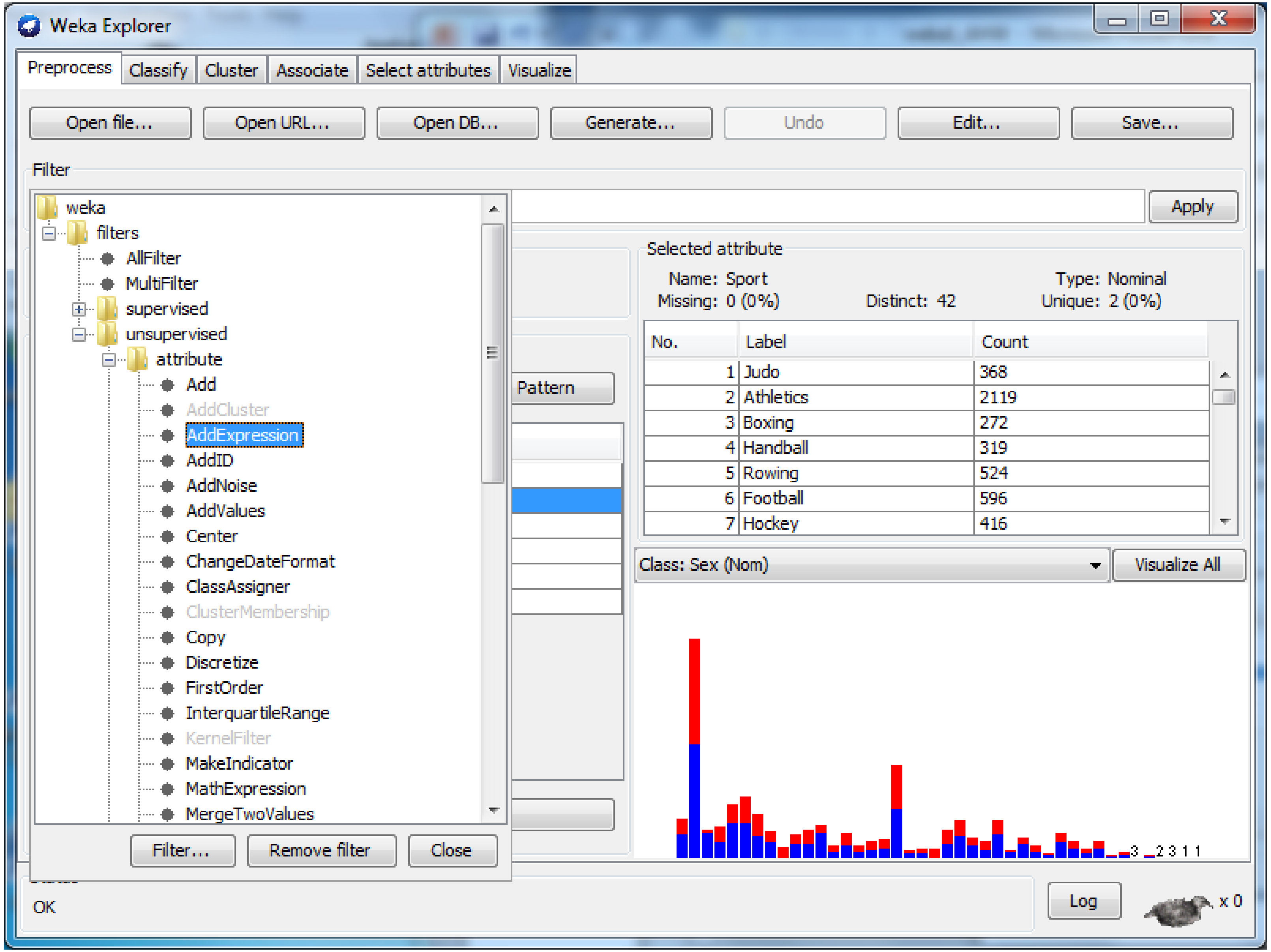
    Are there any ‘high’ correlations between variables?

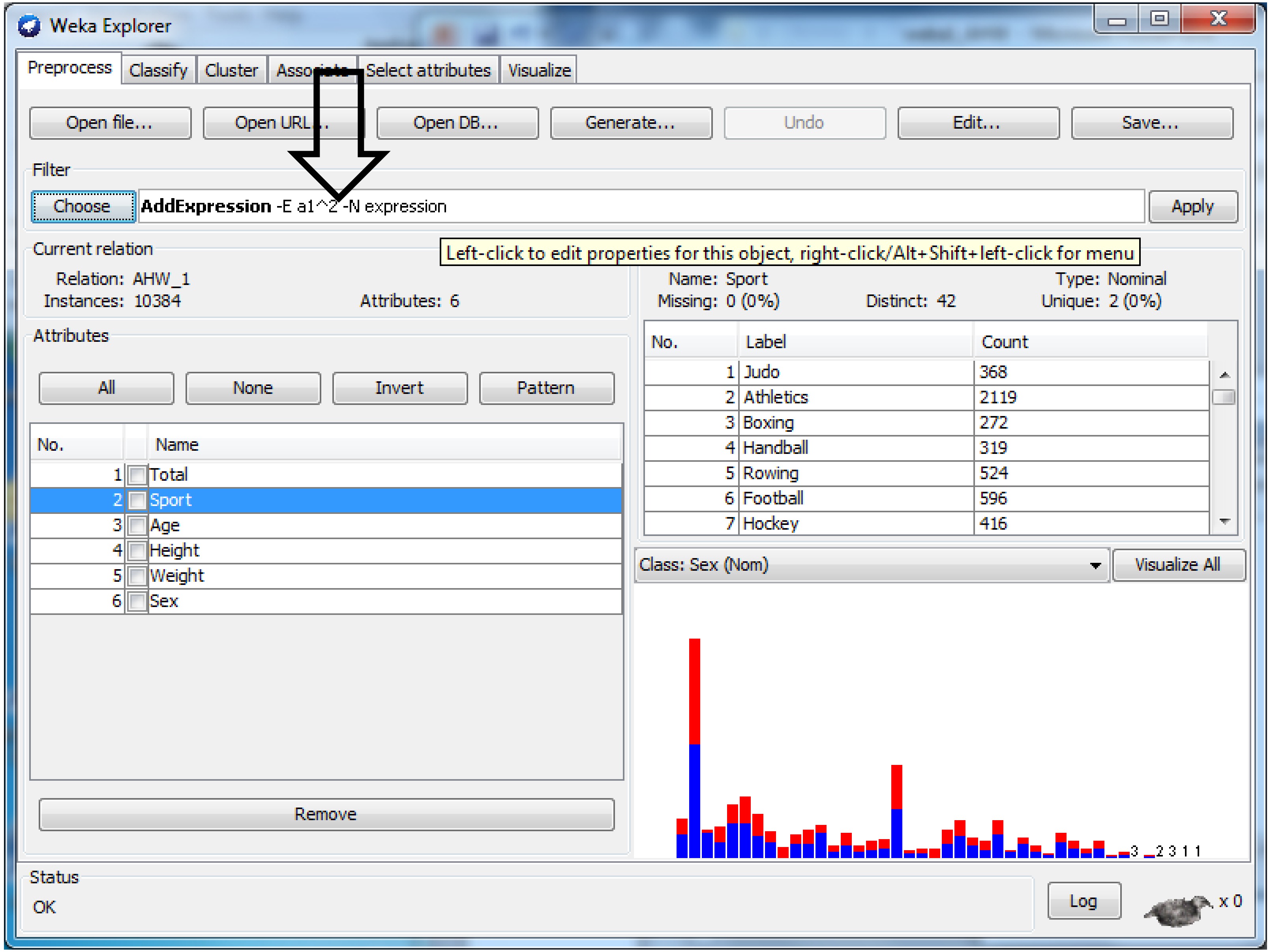
Visualize scatter plot, are there correlations?



Make a new variable for Wt in pounds




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Apply
weka
filters
AllFilter
MultiFilter
supervised
unsupervised
attribute
Add
AddCluster
AddExpression
AddID
AddNoise
AddValues
Center
ChangeDateFormat
ClassAssigner
ClusterMembership
Copy
Discretize
FirstOrder
InterquartileRange
KernelFilter
MakeIndicator
MathExpression
MergeTwoValues
Filter...
Remove filter
Close
Pattern
Selected attribute
Name: Sport
Type: Nominal
Missing: 0 (0%)
Distinct: 42
Unique: 2 (0%)
No. Label Count
1 Judo 368
2 Athletics 2119
3 Boxing 272
4 Handball 319
5 Rowing 524
6 Football 596
7 Hockey 416
Class: Sex (Nom)
Visualize All
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
AddExpression -E a1^2 -N expression
Apply
Left-click to edit properties for this object, right-click/Alt+Shift+left-click for menu
Current relation
Relation: AHW_1
Instances: 10384
Attributes: 6
Name: Sport
Type: Nominal
Missing: 0 (0%)
Distinct: 42
Unique: 2 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
No.
Label
Count
1 Judo 368
2 Athletics 2119
3 Boxing 272
4 Handball 319
5 Rowing 524
6 Football 596
7 Hockey 416
Class: Sex (Nom)
Visualize All
Remove
Status
OK
Log
x 0

Weka Exercise

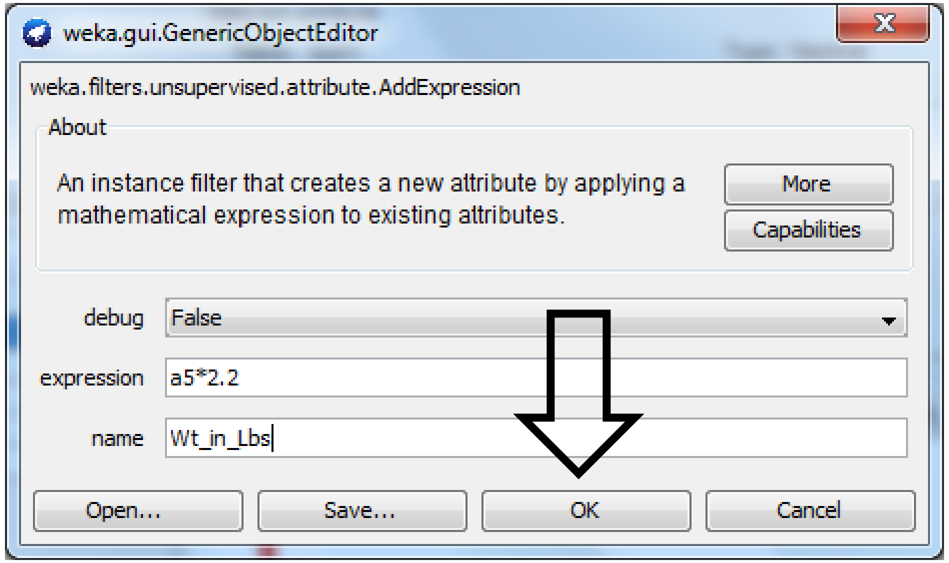
- The weight variable is kilogram units, but USA uses lbs. So make a new variable in weights (1 kilogram~2.2 lbs).

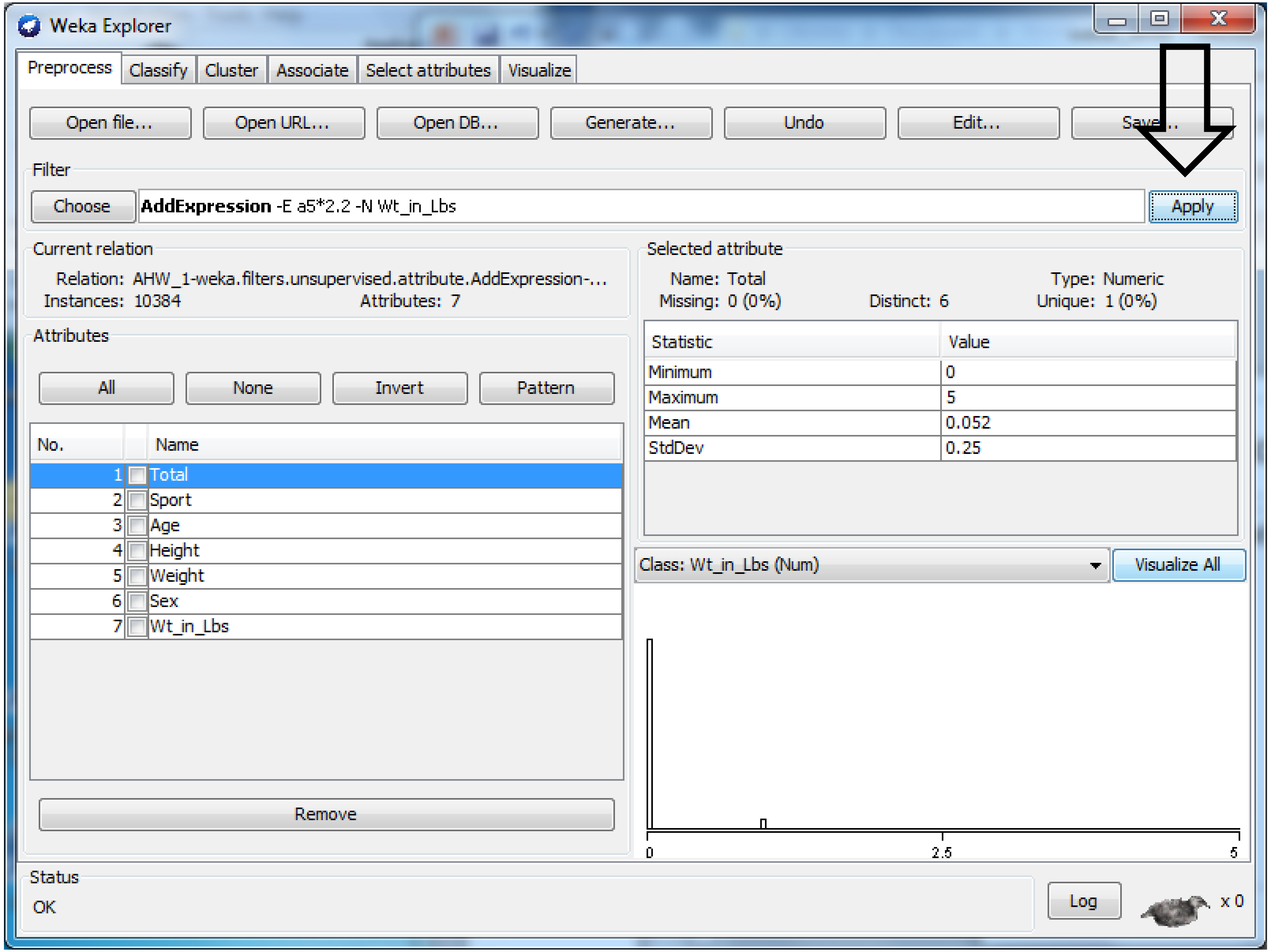
    (choose -> filters,unsupervised,attribute, addexpression

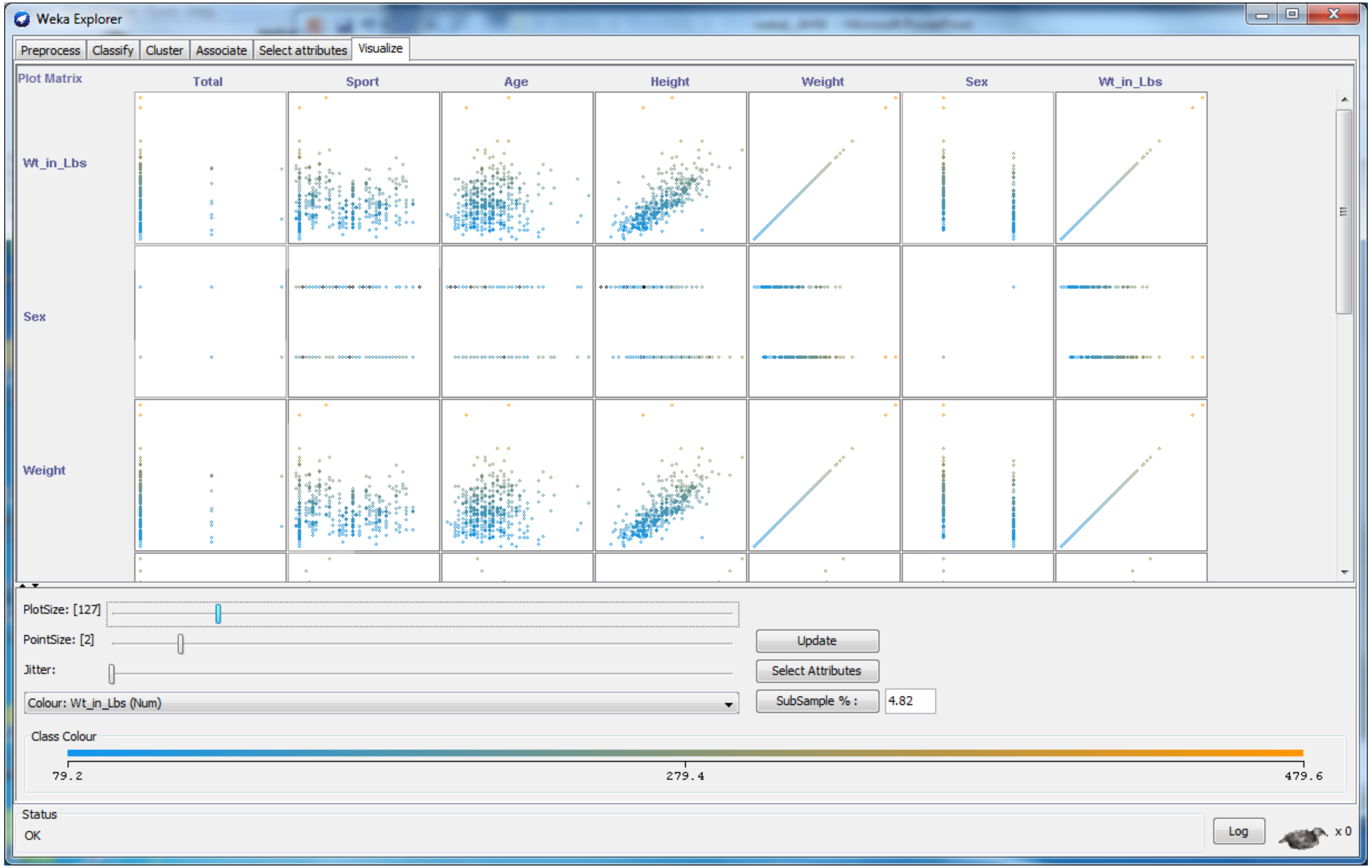
    type an expression, use a5 to indicate weight)

- Check out the correlations again. What do you see?

Make a new variable for Wt in pounds

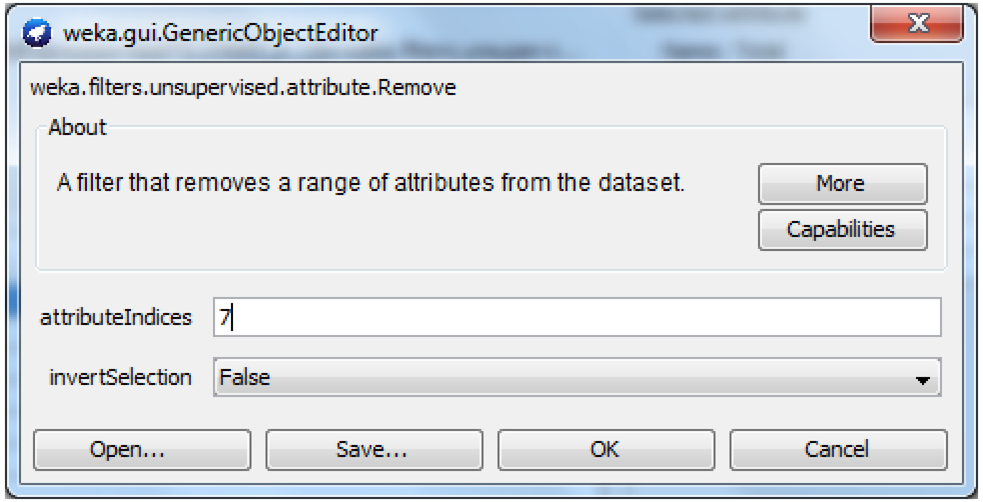



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
AddExpression -E a5*2.2 -N Wt_in_Lbs
Apply
Current relation
Relation: AHW_1-weka.filters.unsupervised.attribute.AddExpression-...
Instances: 10384
Attributes: 7
Selected attribute
Name: Total
Type: Numeric
Missing: 0 (0%)
Distinct: 6
Unique: 1 (0%)
Statistic
Value
Minimum
0
Maximum
5
Mean
0.052
StdDev
0.25
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
7 Wt_in_Lbs
Class: Wt_in_Lbs (Num)
Visualize All
0
2.5
5
Remove
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
Total
Sport
Age
Height
Weight
Sex
Wt_in_Lbs
Wt_in_Lbs
Sex
Weight
PlotSize: [127]
PointSize: [2]
Jitter:
Update
Select Attributes
SubSample % :
4.82
Colour: Wt_in_Lbs (Num)
Class Colour
79.2
279.4
479.6
Status
OK
Log
x 0

To Remove a Variable:

Choose -> weka -> filters -> unsupervised -> attributes -> Remove
And then apply



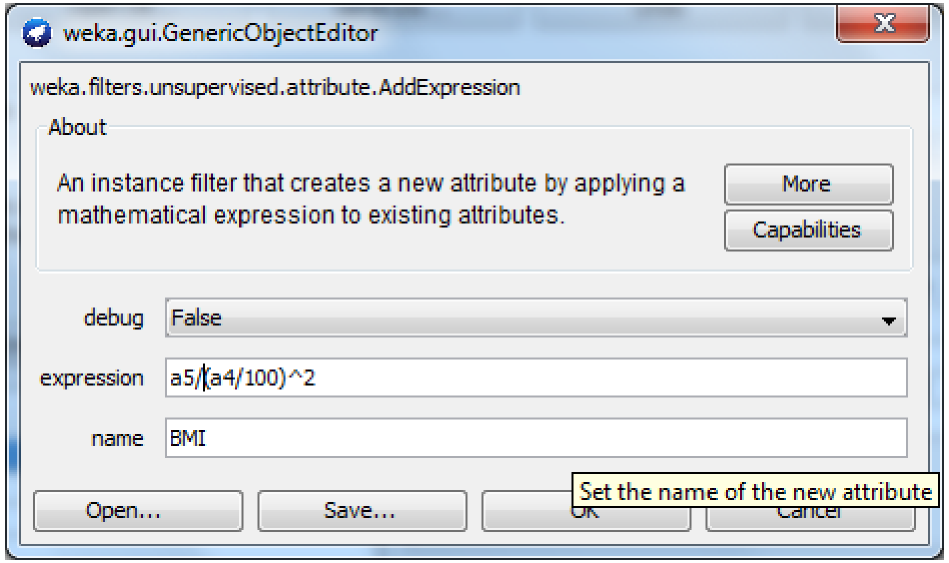
Weka cont.

- Add new variable *weight + height*
- Visualize scatter plot
  - question: Is this a useful variable?
- Repeat for

  Body Mass Index defined as Mass (kg)/Height(m) ^{2}
  - Note: Weight already in Kg. and Height is in cm. (so use a4/100)
  - question: Is this a useful variable?

Add new variable:
Body Mass Index = Mass (kg)/Height(m)^2

Weight already in Kg.
Height is in cm.



Is this a useful variable?

- linear combination *weight + height*

  depends on algorithm,

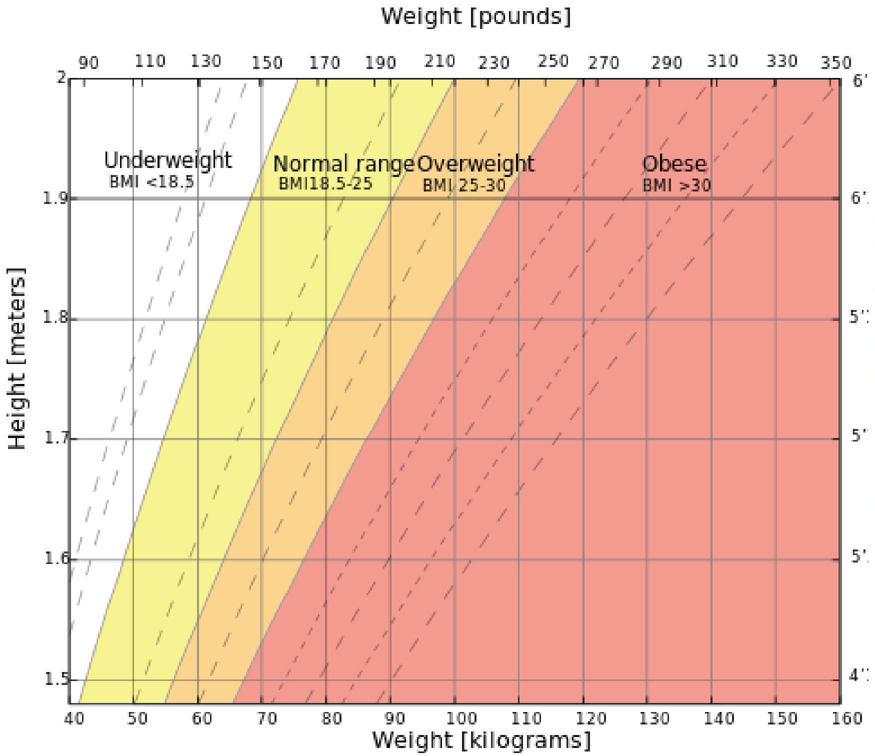
  some regression methods will find it (and it may obscure interpretation)

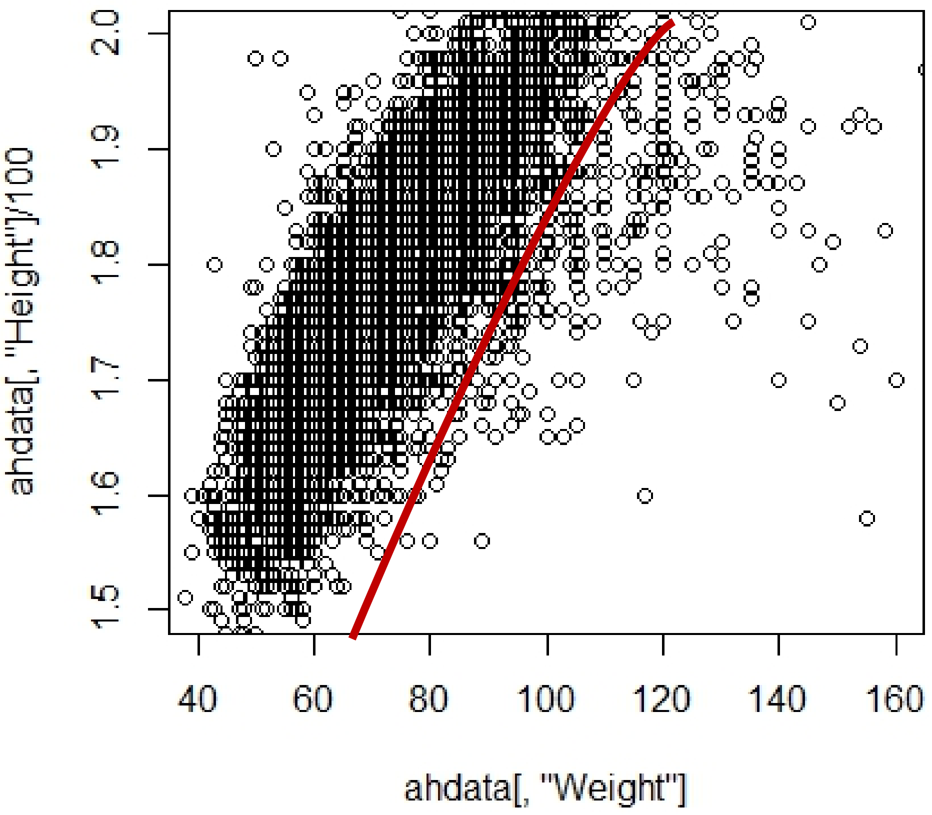
  other methods may not (ie decision tree)

- non-linear combination *height (kg)/weight(m)* ^{2}

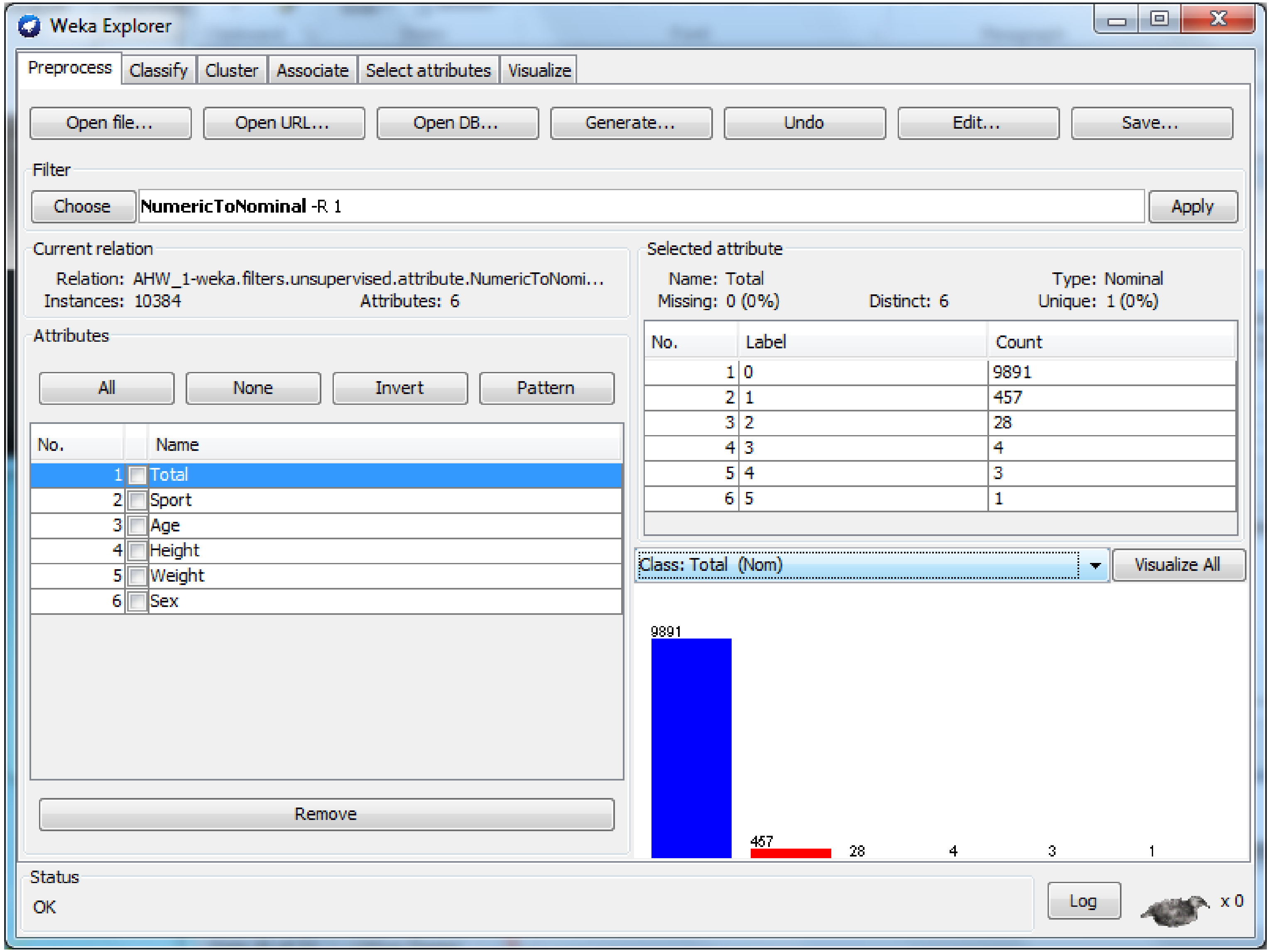
  could be very useful if BMI is apriori known to be important

Are athletes obese?

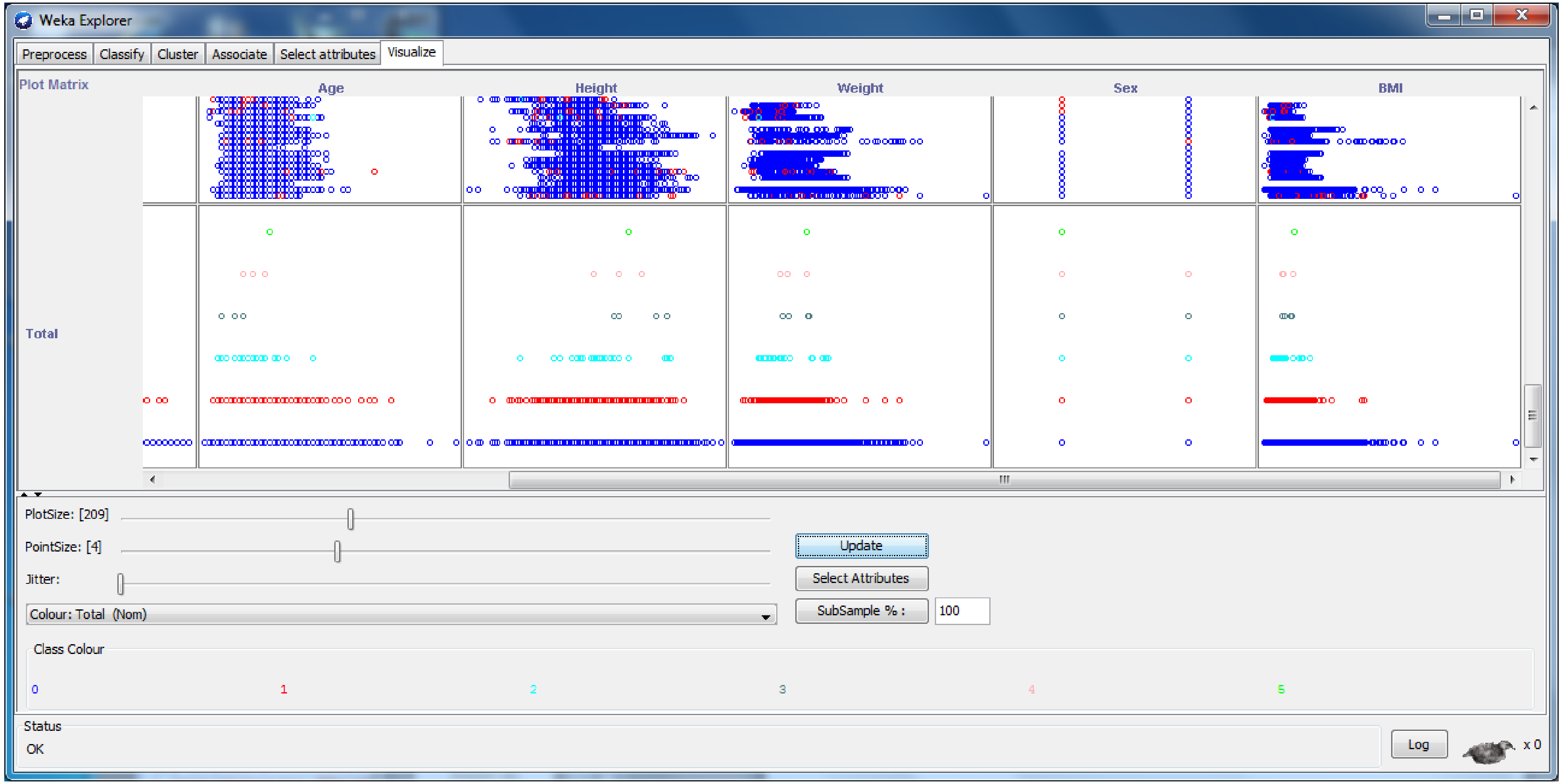




Choose -> weka -> filters -> unsupervised -> attributes -> NumerictoNominal,
Change Total to 0,1 nominal field, apply it, select it as the class



Visualize scatterplots of Total Class with Height, Weight, Sex, BMI

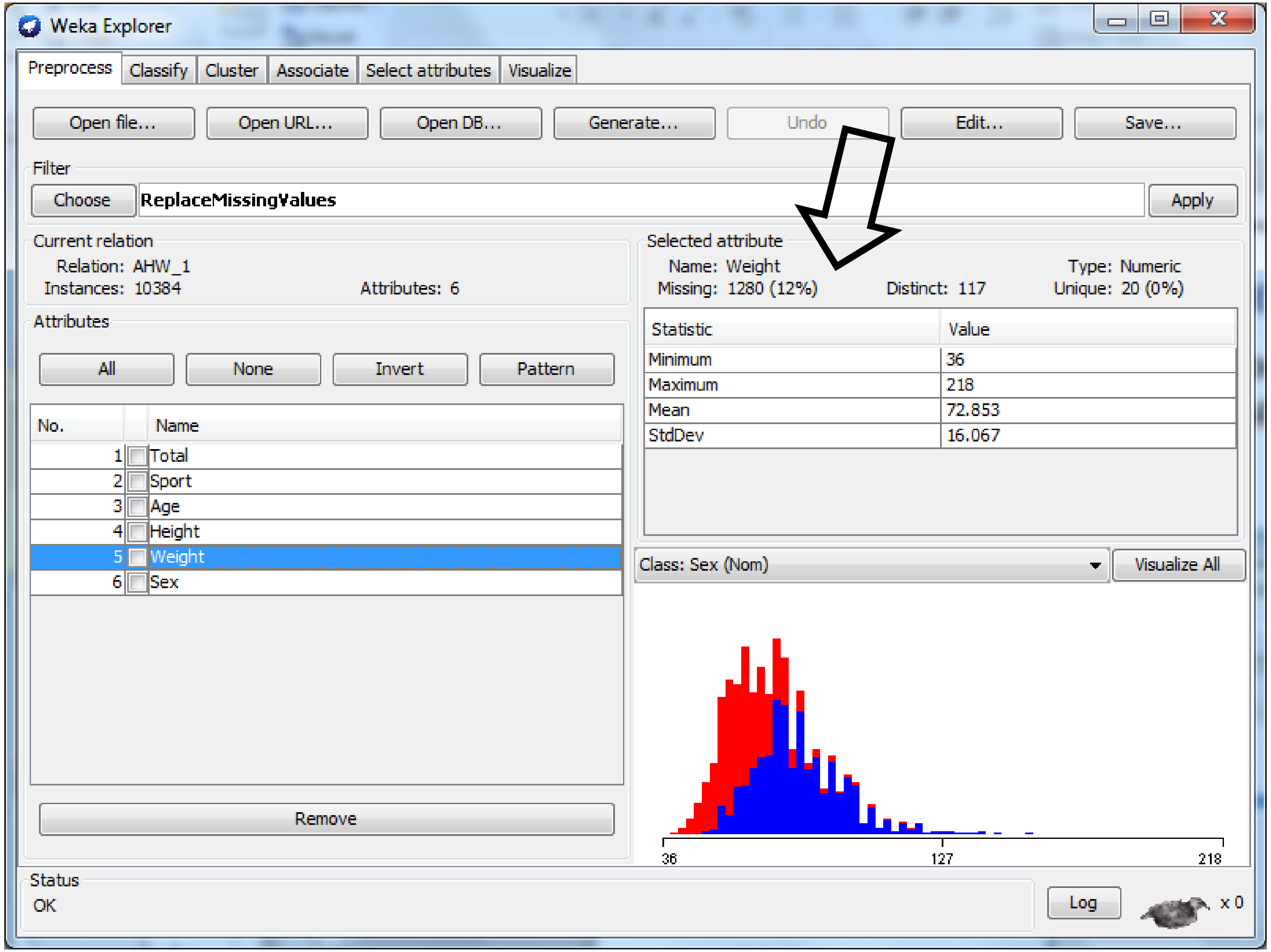


What else to do to predict Total medals won?

- Split data by sport (stratify)

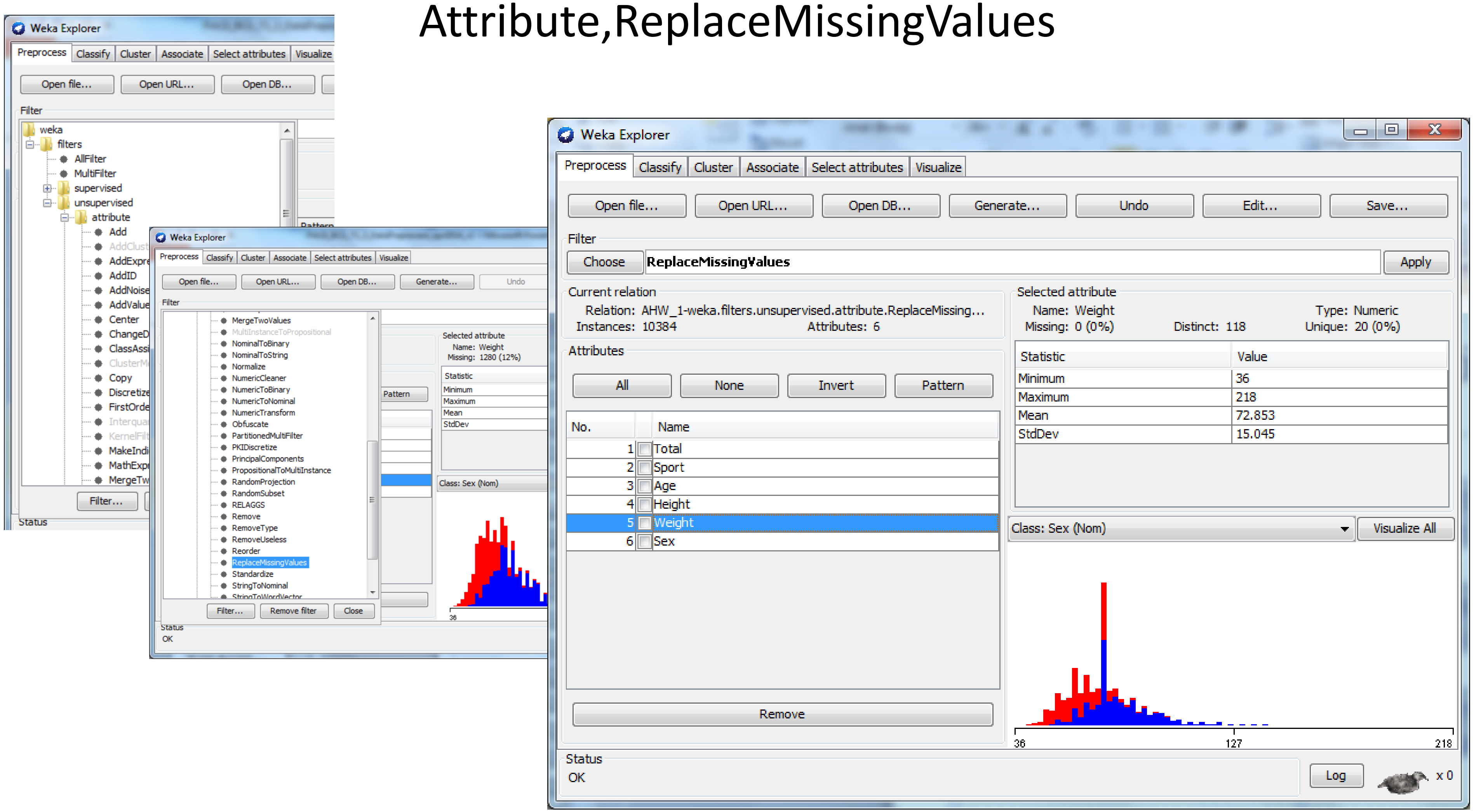
- Gather other data – previous winning, country winnings, etc..

Missing fields – is 12% too much?

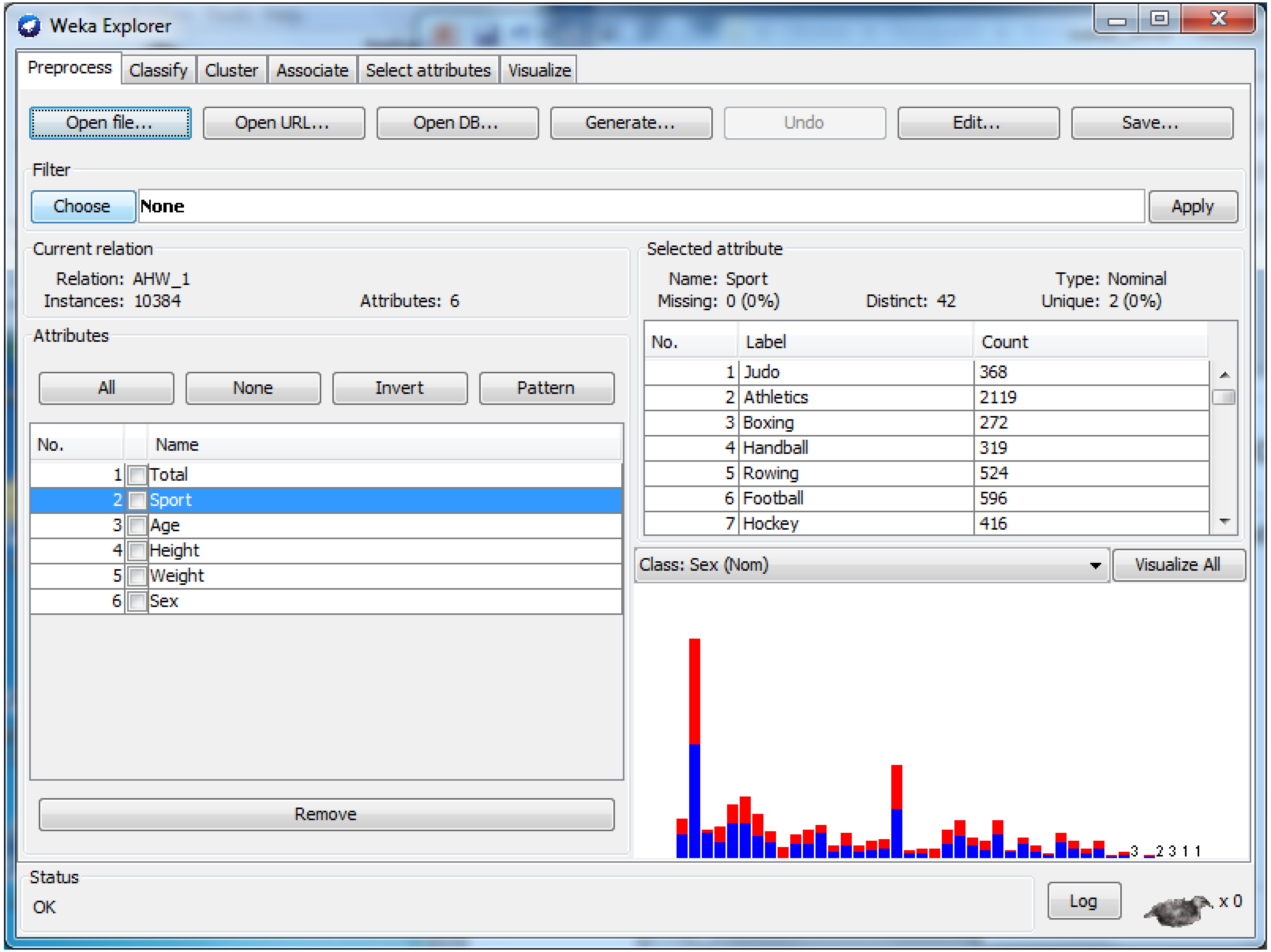


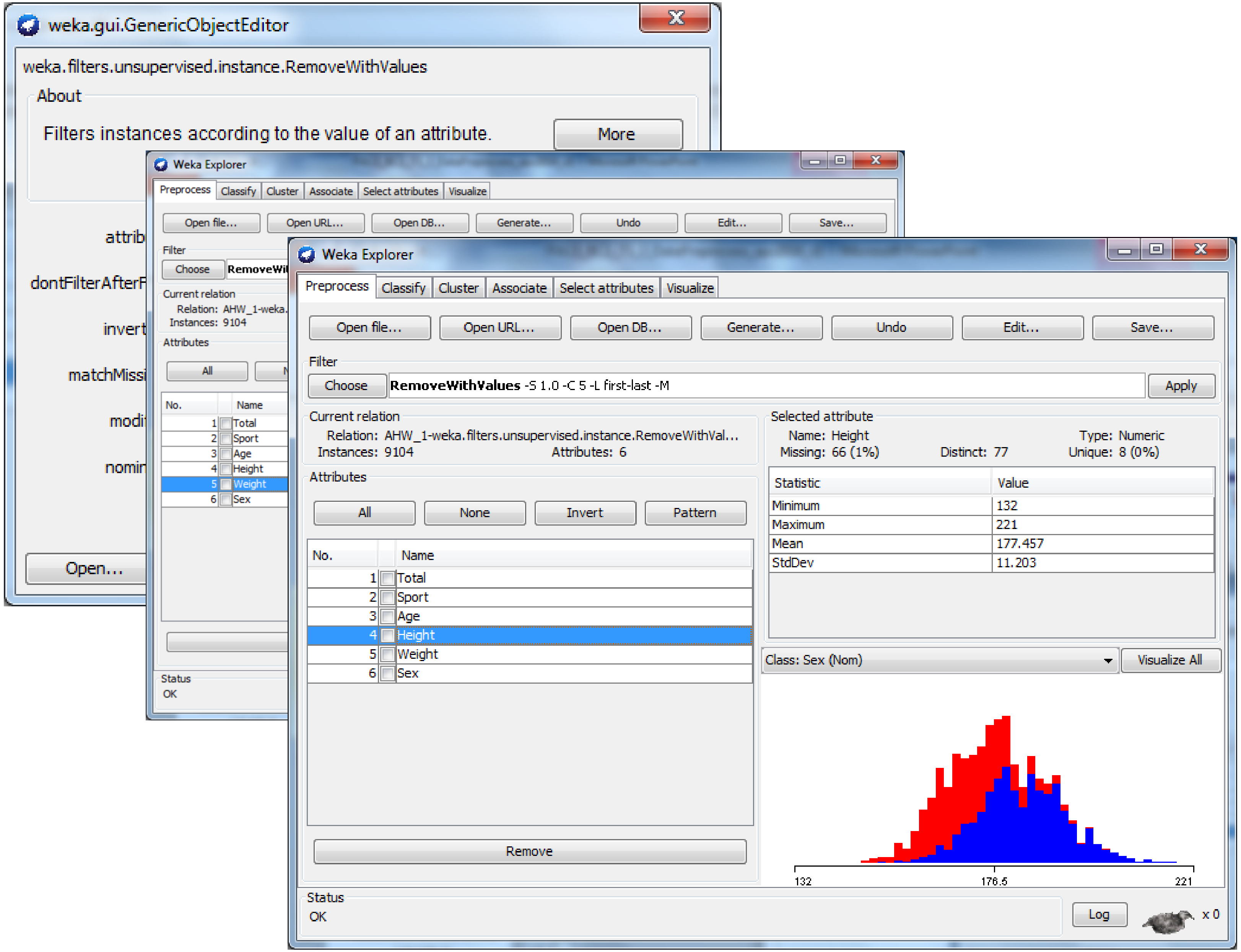
In Weka, for example

Preprocessing -> Choose -> Filter, Unsupervised, Attribute,ReplaceMissingValues

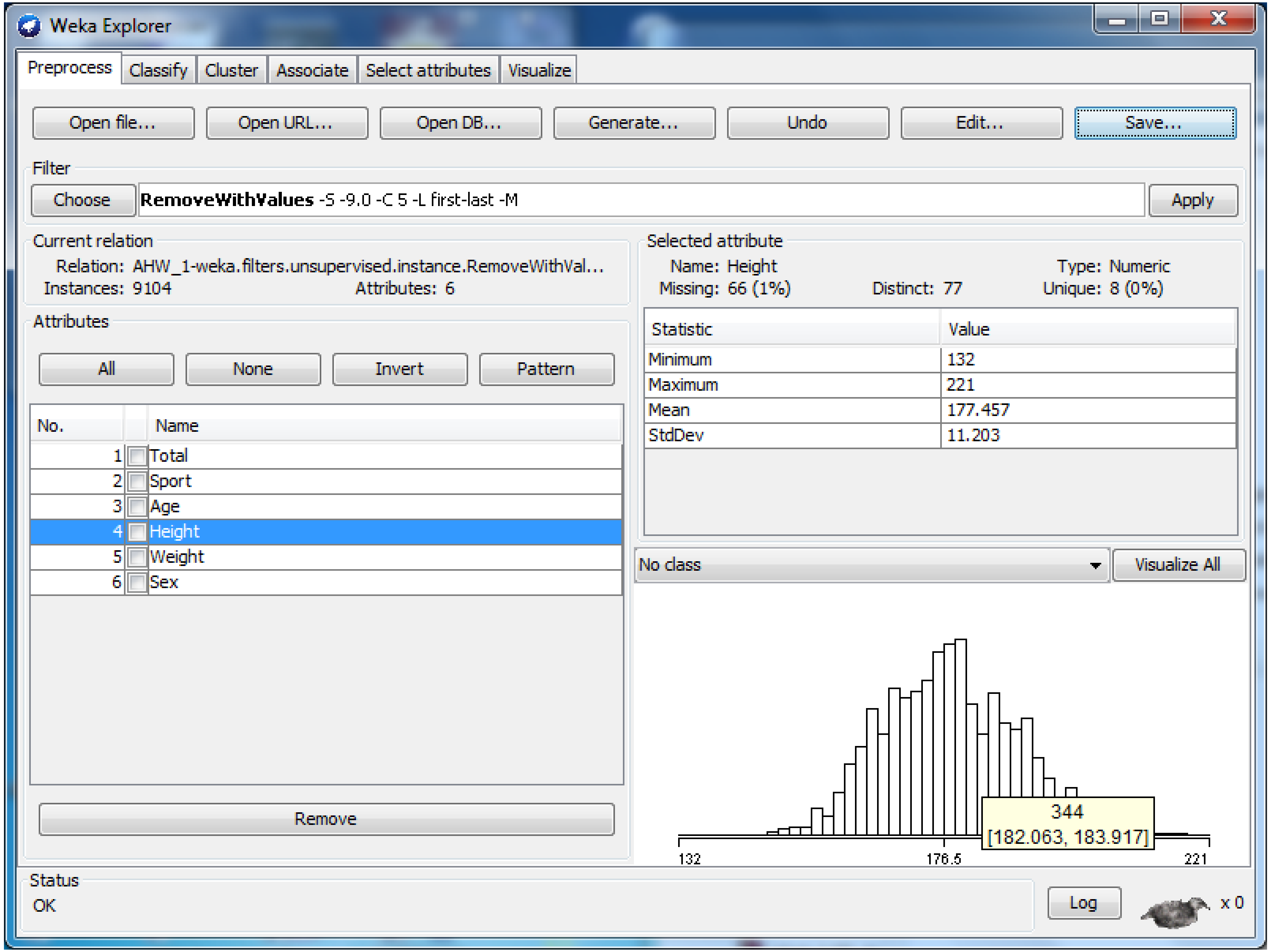


What about sport field – many nominals





weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveWithValues
About
Filters instances according to the value of an attribute.
More
Open...
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemoveWithValues -S 1.0 -C 5 -L first-last -M
Apply
Current relation
Relation: AHW_1-weka.filters.unsupervised.instance.RemoveWithVal...
Instances: 9104
Attributes: 6
Selected attribute
Name: Height
Type: Numeric
Missing: 66 (1%)
Distinct: 77
Unique: 8 (0%)
Statistic
Value
Minimum
132
Maximum
221
Mean
177.457
StdDev
11.203
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Class: Sex (Nom)
Visualize All
132
176.5
221
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemoveWithValues -S -9.0 -C 5 -L first-last -M
Apply
Current relation
Relation: AHW_1-weka.filters.unsupervised.instance.RemoveWithVal...
Instances: 9104
Attributes: 6
Attributes
All
None
Invert
Pattern
No.
Name
1 Total
2 Sport
3 Age
4 Height
5 Weight
6 Sex
Remove
Selected attribute
Name: Height
Type: Numeric
Missing: 66 (1%)
Distinct: 77
Unique: 8 (0%)
Statistic
Value
Minimum 132
Maximum 221
Mean 177.457
StdDev 11.203
No class
Visualize All
344
[182.063, 183.917]
132
176.5
221
Status
OK
Log
x 0