Summer School
for PhD students

Research Tasks, Methods & Data Structure

Kaloyan Haralampiev
Alexander Efremov

Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Agenda

Introduction

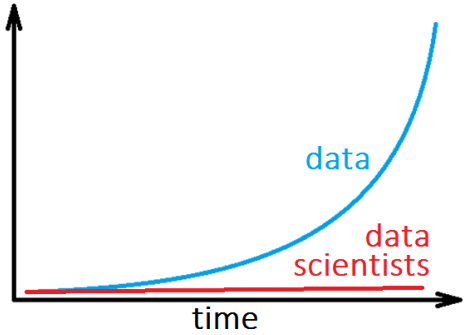
- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

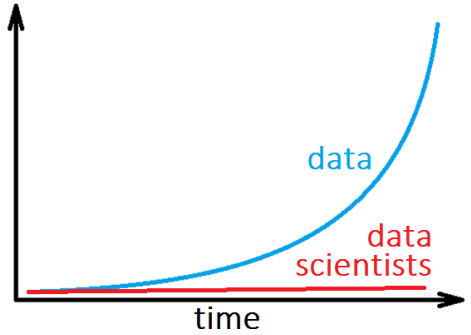
Why Data Analytics?

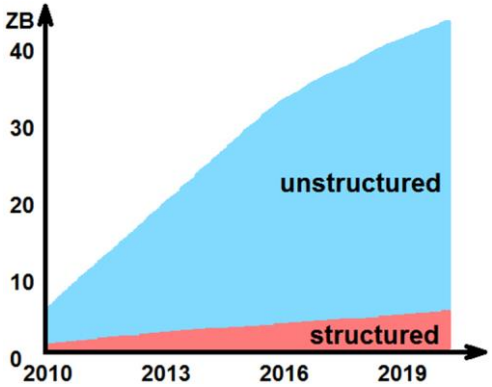
- Premises: Data, Computation Power, Clouds, etc.
- Reasons: Accurate, Fast, Cheep, Scalable, Measurable, etc.
- Data Types
- Structured Data
- Unstructured Data
  - Text
  - Images
  - Audio
  - Video



Why Data Analytics?

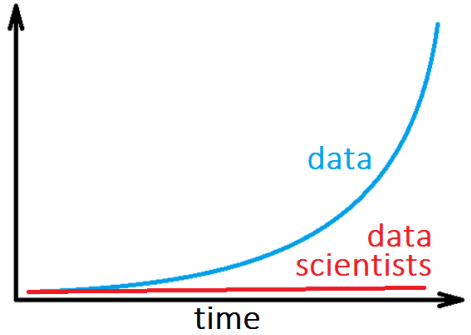
- Premises: Data, Computation Power, Clouds, etc.
- Reasons: Accurate, Fast, Cheep, Scalable, Measurable, etc.
- Data Types
- Structured Data
- Unstructured Data
  - Text
  - Images
  - Audio
  - Video

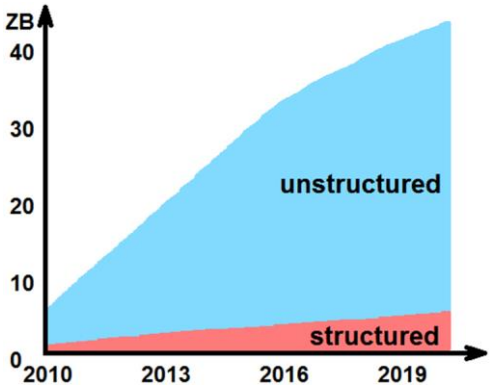




Why Data Analytics?

- Premises: Data, Computation Power, Clouds, etc.
- Reasons: Accurate, Fast, Cheep, Scalable, Measurable, etc.
- Challenges
  - **Data size**: big or small data
  - **Data types**: structured, unstructured data, mixed data
  - **Data sources**: CRM, SRM, Sales, PLM, SCM... /demogr., behav./
  - **Dirty data**: Missings and Outliers
  - **System**: multivariable, interconnections, dynamics, time-varying behavior...
  - **Tasks**: complex workflows





Agenda

Summer School

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Examples

- **Churn Reduction** for Service Providers, Mobile Operators
- **Target Market Offering** in Retail Industry
- **Recommender Systems** in E-commerce
- **Promotions Optimization** in Supermarkets
- **Order Strategy Optimization** in Storages
- **Optimal Accept/Reject Decision** in Credit Risk
- etc.

Agenda

Introduction

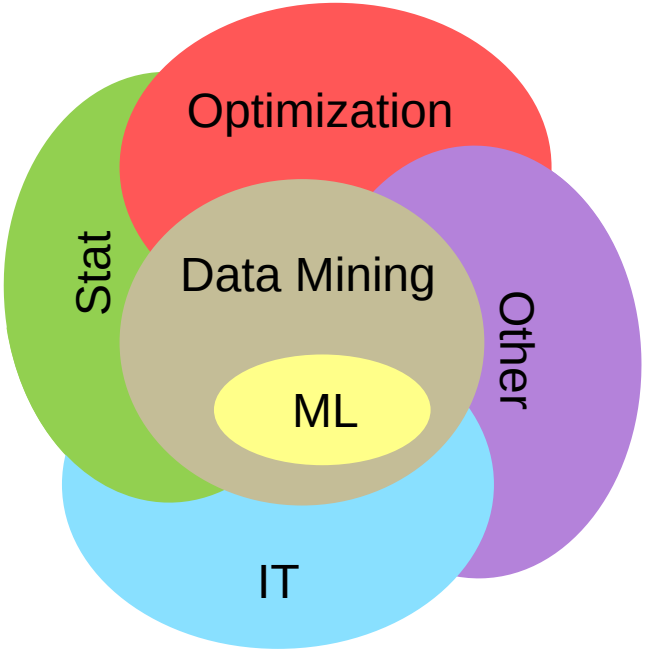
- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Terms

- Areas
  - Statistics
  - Machine Learning (ML)
  - Information Technology (IT) /data bases, parallel computations, .../
  - Optimization
  - Numerical methods
  - ...
  - DM
  - AI?
  - DS?
  - ...



Agenda

Introduction

- Why Data Analytics?
- Examples
- Terms
- DM Stages
- Automation

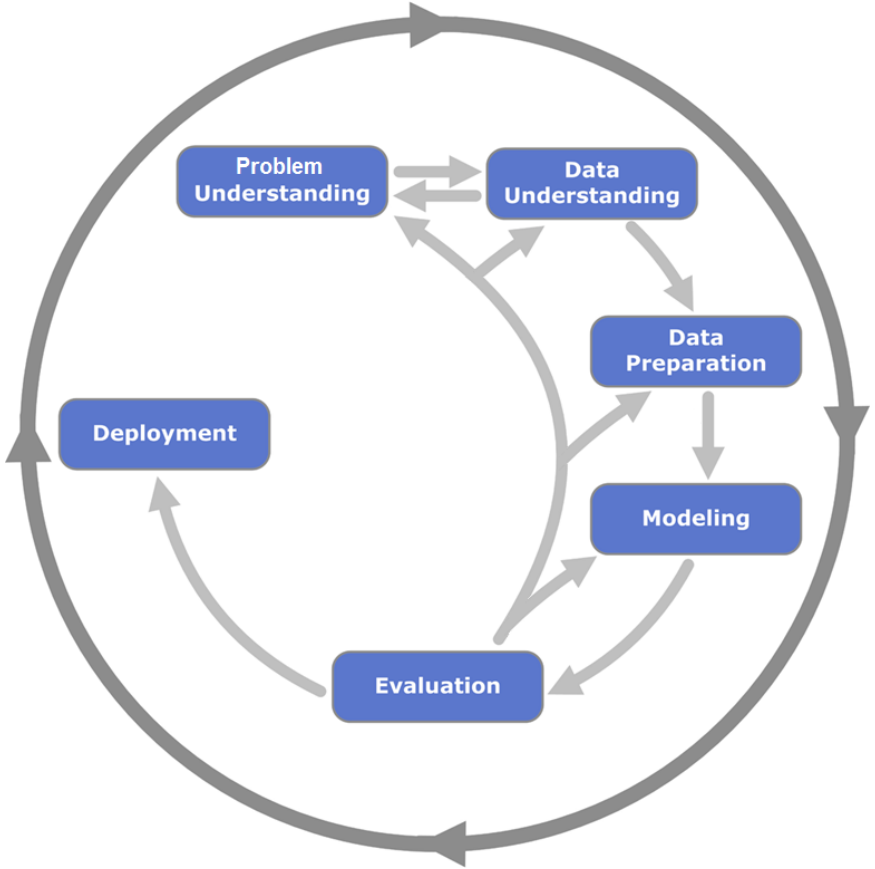
Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

DM Stages

CRISP – DM

Cross Industry Standard Process for Data Mining



Agenda

Introduction

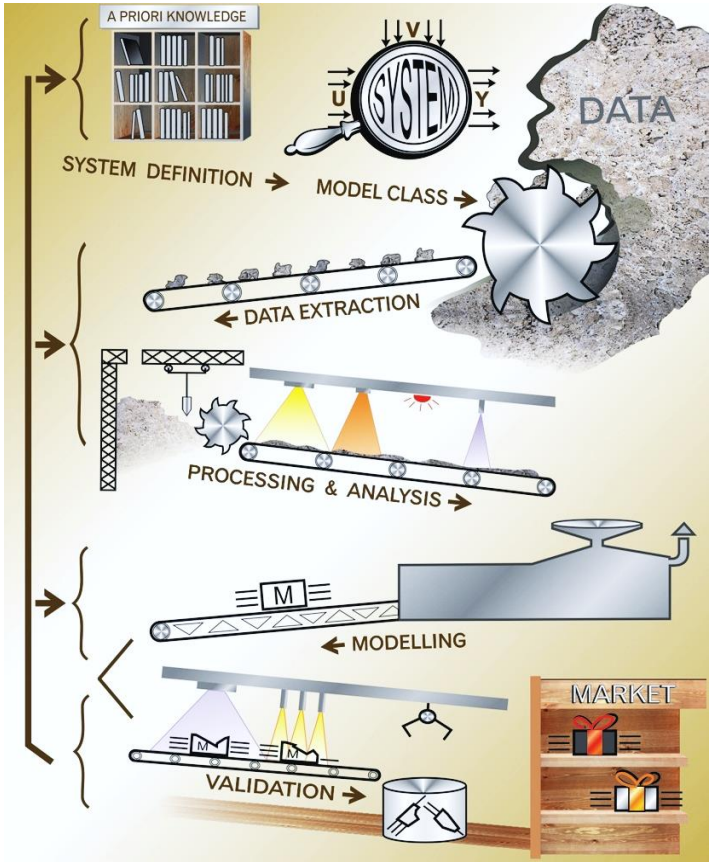
- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- Automation

Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Automation

- Automated DM Workflows
  - Many domain problems → One DM problem
- Excluding Human Role
  - Many variables, fast, cheep... solution
- Dangers
  - Wrong decisions and conclusions
- How to Reduce Risk?
  - A-priori & a-posteriori balance (Gray Box)



Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

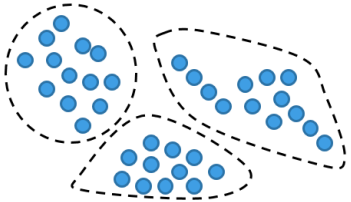
Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

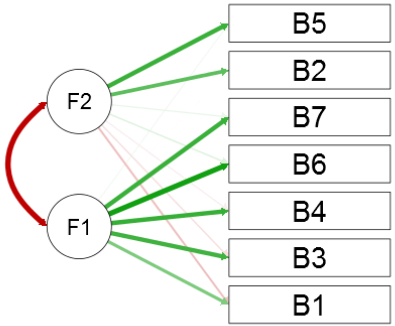
Relationship Between Research Tasks and Data structure

Research tasks	Data structure
Descriptive statistics	Variables only
Segmentation	
Dimension reduction	
Relations Analysis	Dependent variable(s) and independent variable(s)
Classification	
Time series analysis	Time as independent variable

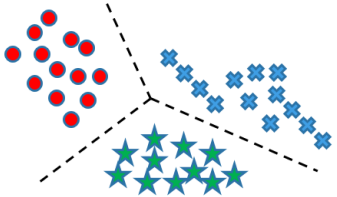
DM Tasks



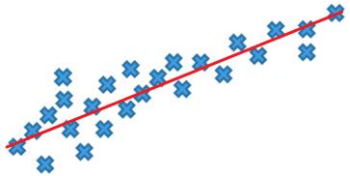
Segmentation



Dimension reduction



Classification



Relations Analysis

Time series analysis

Agenda

Introduction

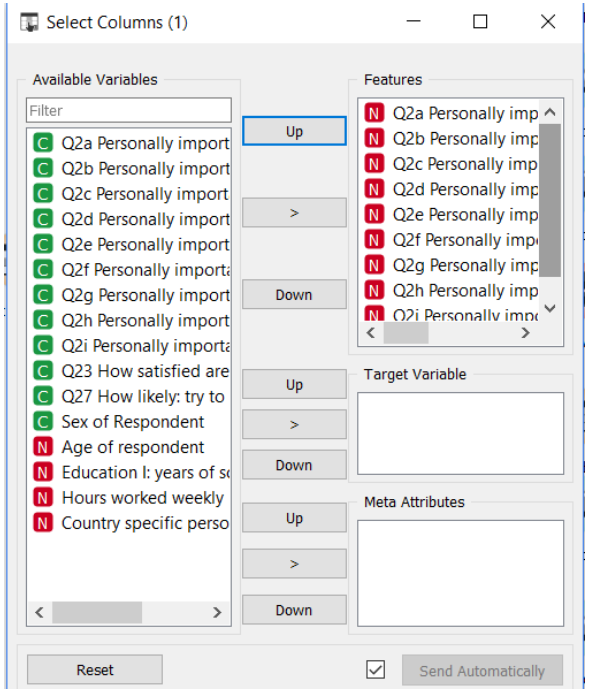
- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

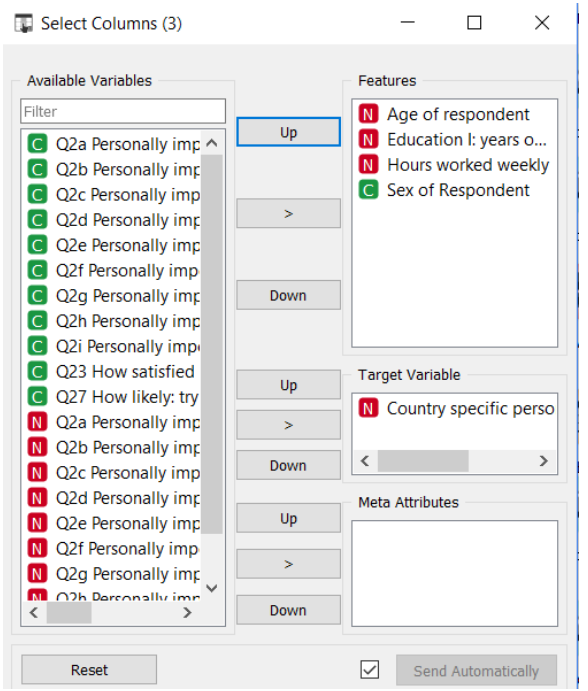
- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Data Structures / Examples

- Variables only



- Dependent variable(s) & independent variable(s)



Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

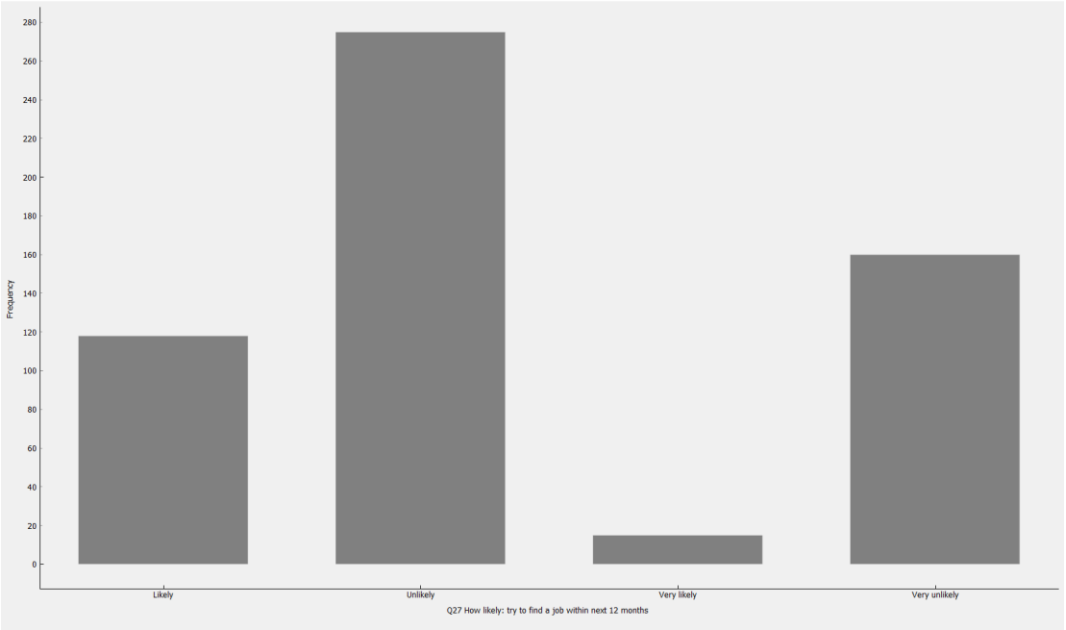
Research Task & Data Structure

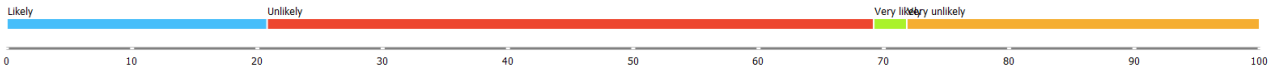
- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Descriptive Statistics

Scale	Statistical method
Nominal	Frequencies Percentages
Ordinal, rank or score	+ Cumulative frequencies + Cumulative percentages
Interval	+ Central tendency (mean, median, mode) + Dispersion (variance, standard deviation, range) + Skewness + Kurtosis + Quartiles, deciles, percentiles

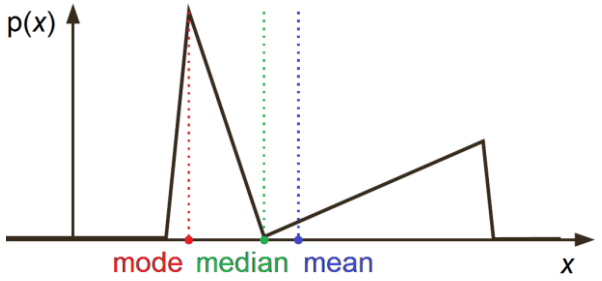
Descriptive Statistics / Example: Nominal & Ordinal Scales

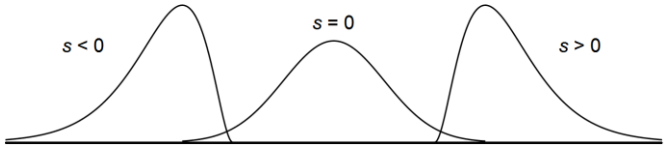




Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Descriptive Statistics





mean – arithmetic average $\bar{x} = 1/N \sum_k x_k$

mode – most frequent value(s) mo_x = argmax f(x)

median – middle value. Below & above me_x lies equal number of obs.

quantiles – values dividing a set into equal-sized groups

quartiles – values dividing a set into 4 equal (25%) groups

quintiles – values dividing a set into 5 equal (20%) groups

percentiles – values dividing an ordered set into 100 equal (1%) groups /0.25 quantile = 1 quartile = 25 percentile/

variance – measures the spread of obs. around mean $D_x = m_2 N/(N-1)$

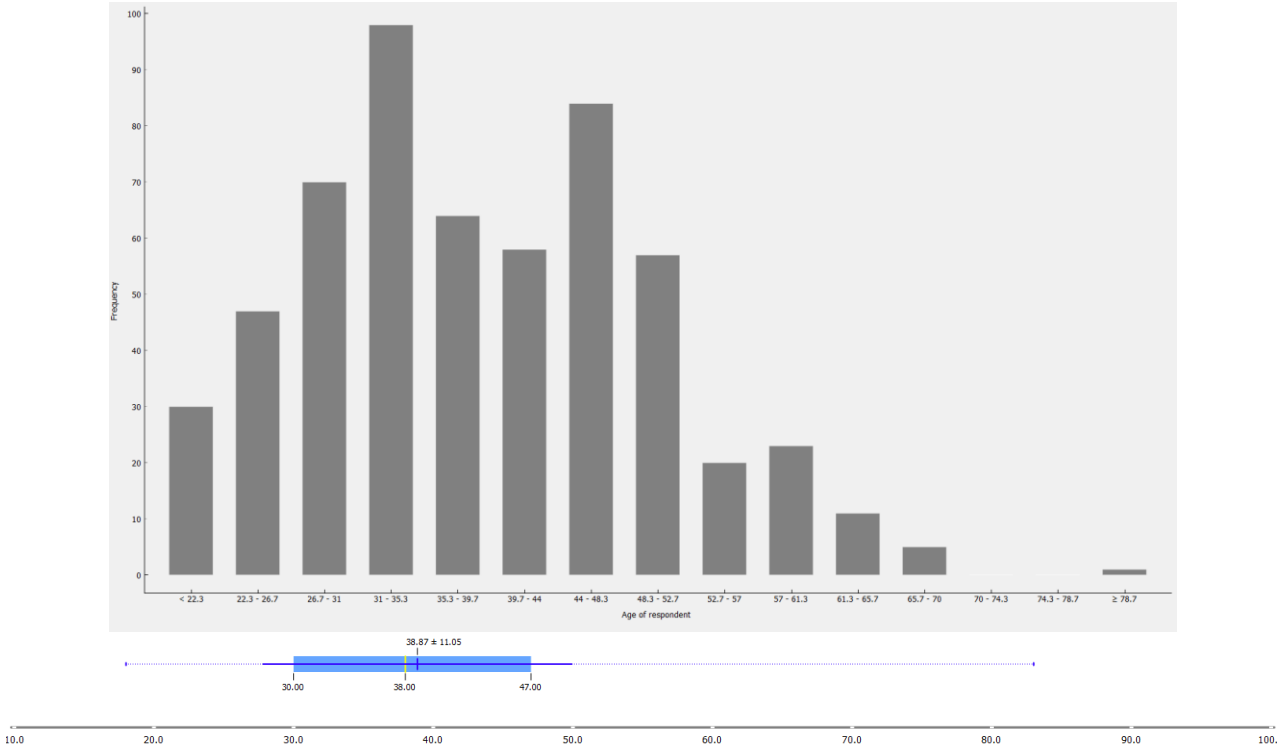
standard deviation – sqrt(variance) $\sigma_x = D_x^{1/2}$

skewness – distribution asymmetry (positive – right, negative – left oriented) $s = m_3 / m_2^{3/2}$

kurtosis – peak sharpness relative to a normal distribution (positive – above, negative – below Normal peak) $K = m_4 / m_2^2 - 3$

m_i – i-th central moment

Descriptive Statistics / Example: Interval Scales



Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- DM Stages
- DM Automation

Research Task & Data Structure

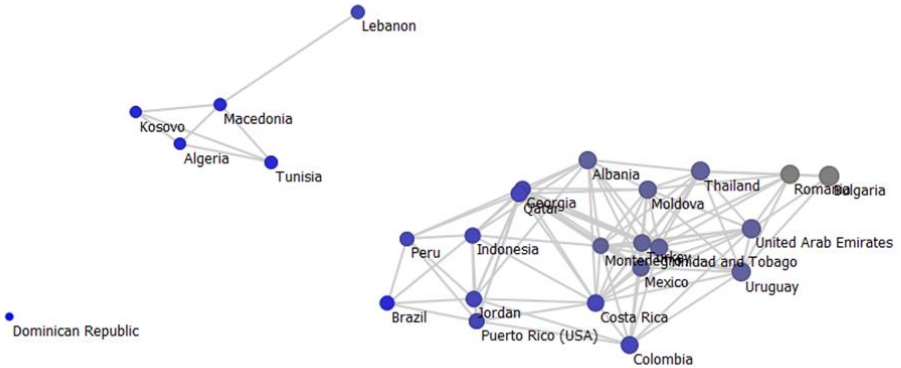
- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

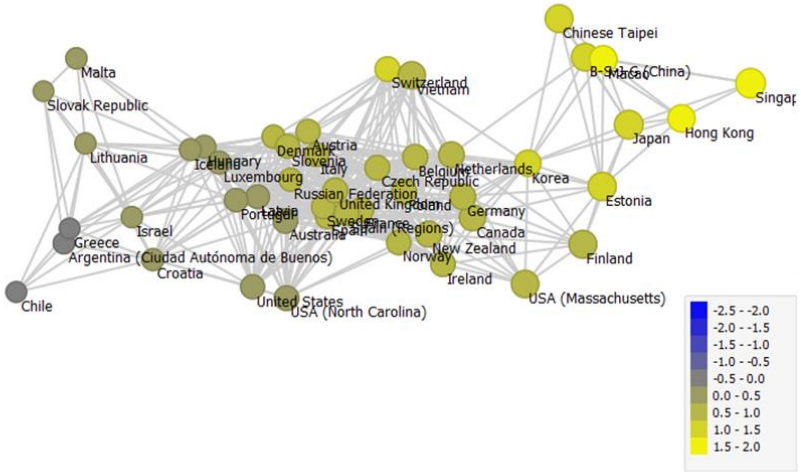
Segmentation

Scale	Statistical method
Nominal Ordinal or rank	TwoStep cluster
Binary Score Interval	+ Hierarchical cluster + Partitional cluster + Multidimensional scaling

Segmentation / Example: Multidimensional Scaling

Data: PISA (http://www.oecd.org/pisa/)
Software: Orange (https://orange.biolab.si/)

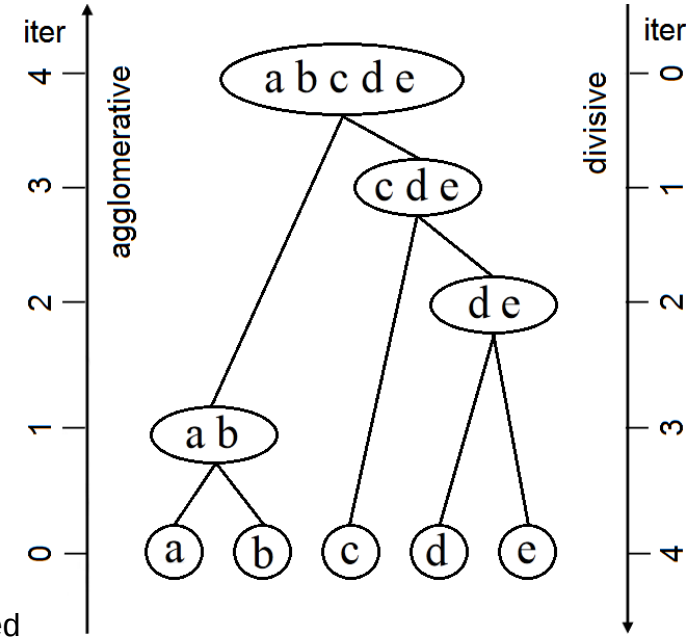




Clustering / Hierarchical

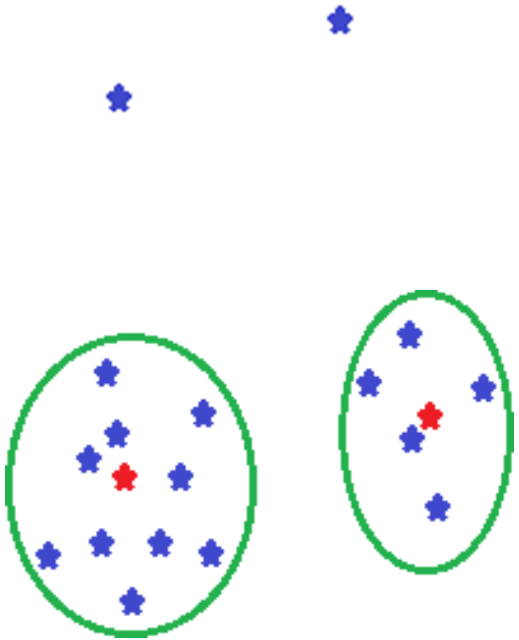
– Hierarchical Clustering Approaches

- Tree structure of clusters
- **Agglomerative**: bottom-up approach
  Start: each cluster is unique object
  Clusters are merged into larger clusters
  Until: all obj. are in one cluster / stopping rule is satisfied
- **Divisive**: top-down approach
  Start: one cluster with all objects
  Clusters are divided into smaller clusters
  Until: each cluster is unique object / stopping rule is satisfied

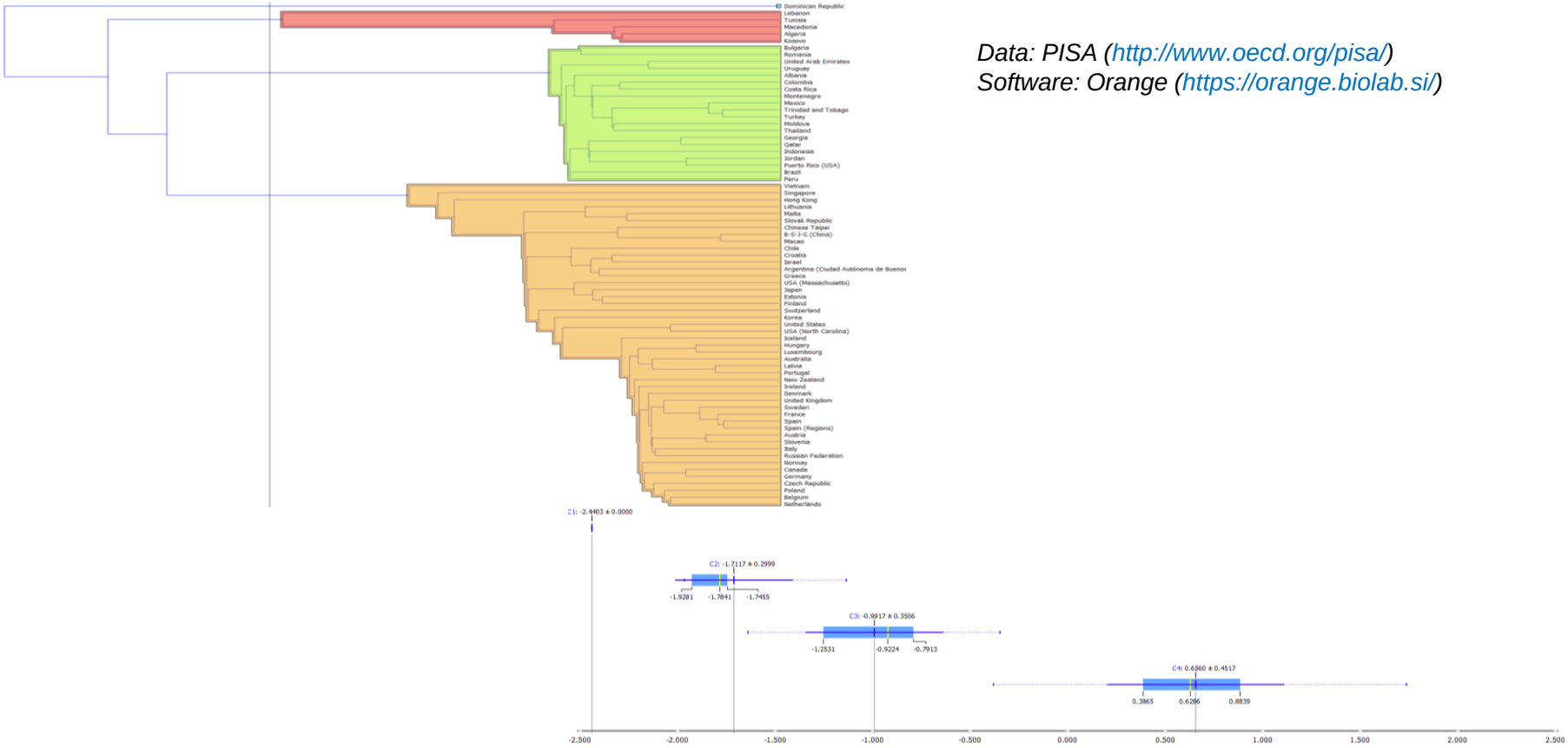


Clustering / Distances

- Distance Between Objects
  - Euclidian distance
  - Manhattan
  - Mahalanobis
  - Simple match coefficient
  - Mixed (Gower) distance...
- Distance Between Clusters
  - Single-link
  - Complete-link
  - Average-link...



Segmentation / Example: Hierarchical Cluster

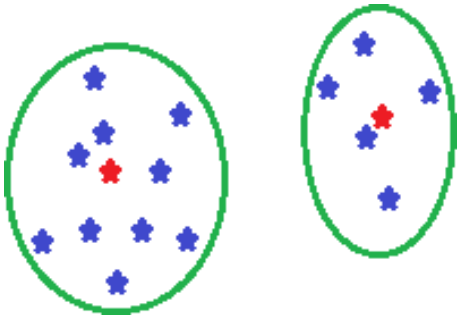


Data: PISA (http://www.oecd.org/pisa/)
Software: Orange (https://orange.biolab.si/)

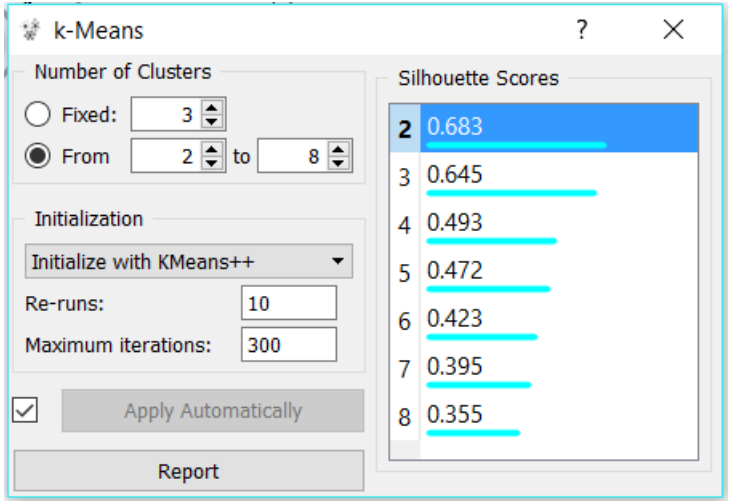
Clustering / Partitional

– Partitional Clustering

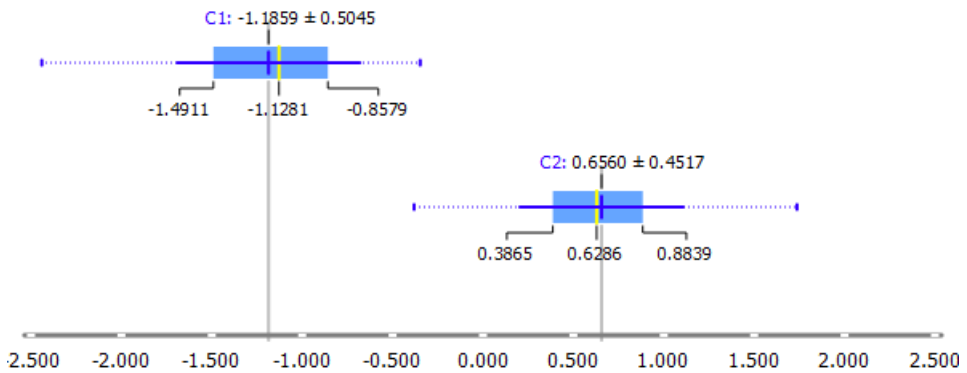
- K-means (for numeric attributes)
- K-modes (for ordinal and nominal attributes)
- K-medoids (for ordinal and nominal attributes)



Segmentation / Example: K-Means Cluster



Data: PISA (http://www.oecd.org/pisa/)
Software: Orange (https://orange.biolab.si/)



Clustering / Comparison: Hierarchical & Partitional Approaches

Summer School

- Hierarchical
  - Number of clusters is not required
  - No backtracking => no corrections
  - Slow approach: $O(N^2)$

    Distances: $(N^2 - N)/2$ first iteration (aglomer. clustering)

    (N = 1000 => D has 499 500 distances)

- Partitional
  - Number of clusters is required
  - Cluster members are rearranged iteratively
  - Fast (k-means)

    Distances: N x k per each iteration

    (N = 1000, k = 5 => 5000 distances)

Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- Automation
- DM Stages

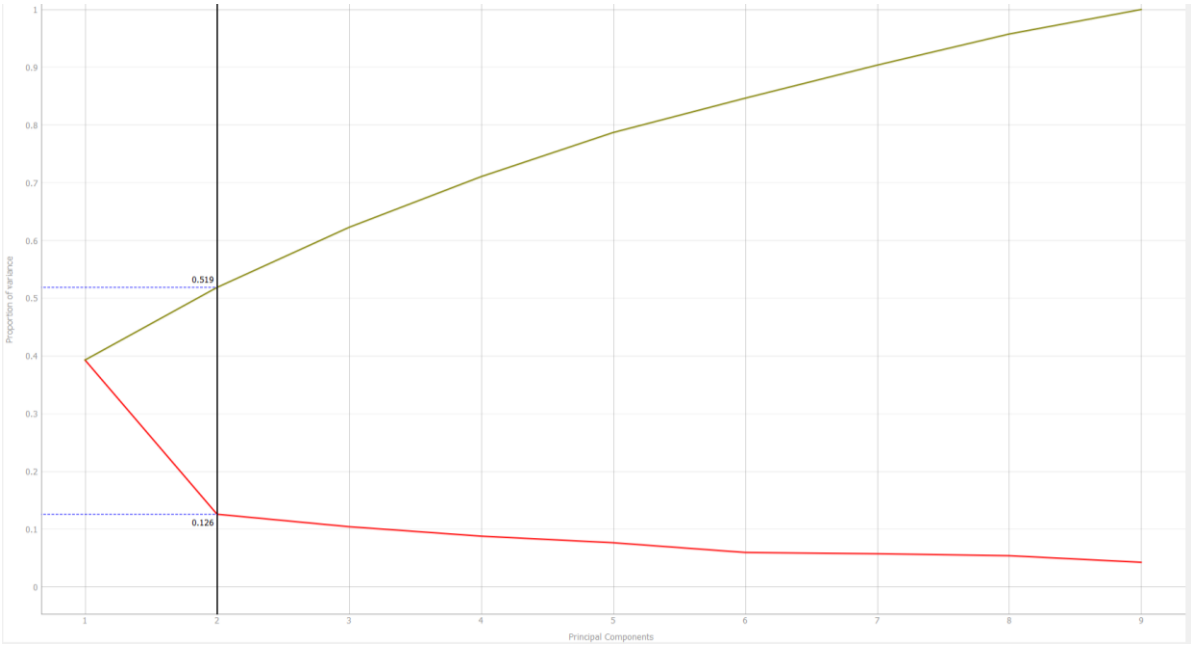
Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Dimension Reduction

Scale	Statistical method
Interval	Factor analysis
Binary	+ Likert scale + Cronbach's Alpha
Score	+ Categorical principal components

Dimension Reduction / Example: Factor Analysis



ZA6770_v2-1-0 components

	component	mportant: job se	mportant: high i	opportunities fo	ortant: an intere	ortant: work inde	ortant: help othe	rtant: a job useful	ortant: decide tim	int: contact with o
1	PC1	-0.242	-0.250	-0.292	-0.391	-0.380	-0.403	-0.358	-0.276	-0.360
2	PC2	-0.496	-0.672	-0.180	0.146	0.240	0.276	0.261	-0.186	0.109

ZA6770_v2-1-0 components

	PC1	PC2
1	1.730	2.715
2	1.482	-1.486
3	-4.197	-0.215
4	-2.780	-0.548
5	-1.798	-1.059
6	-0.226	-1.564
7	-0.186	0.778
8	-0.276	-0.119
9	-2.242	-0.556
10	-1.337	-0.517
11	-0.525	-0.175
12	-2.728	0.807
13	-0.162	0.570
14	-3.141	1.763
15	-1.993	1.073
16	0.236	0.295
17	-4.197	-0.215
18	-0.811	0.108
19	1.903	-2.033
20	0.893	0.032
21	1.456	-2.195
22	0.349	1.876
23	-0.687	0.766
24	1.056	0.820
25	0.193	0.731
26	-3.334	-0.125
27	0.393	0.052
28	-0.162	-0.217
29	-1.643	0.041
30	0.024	1.336
31	0.599	1.827
32	-3.885	-0.004
33	-2.371	-0.665
34	0.218	0.805
35	-4.197	-0.215
36	-1.579	0.117
37	1.773	0.148

Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Agenda

Introduction

- Why Data Analytics?
- Examples
- DM & Related Areas
- Automation
- DM Stages

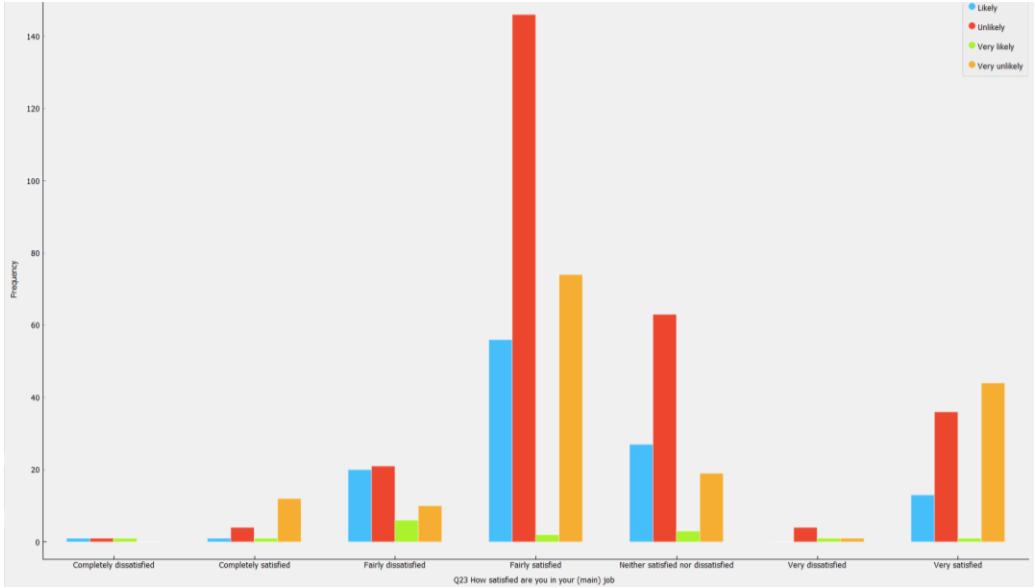
Research Task & Data Structure

- The Relationship
- Data Structures
- Descriptive Statistics
- Segmentation
- Dimension Reduction
- Relations Analysis
- Classification
- Time Series Analysis

Relations Analysis and/or Classification

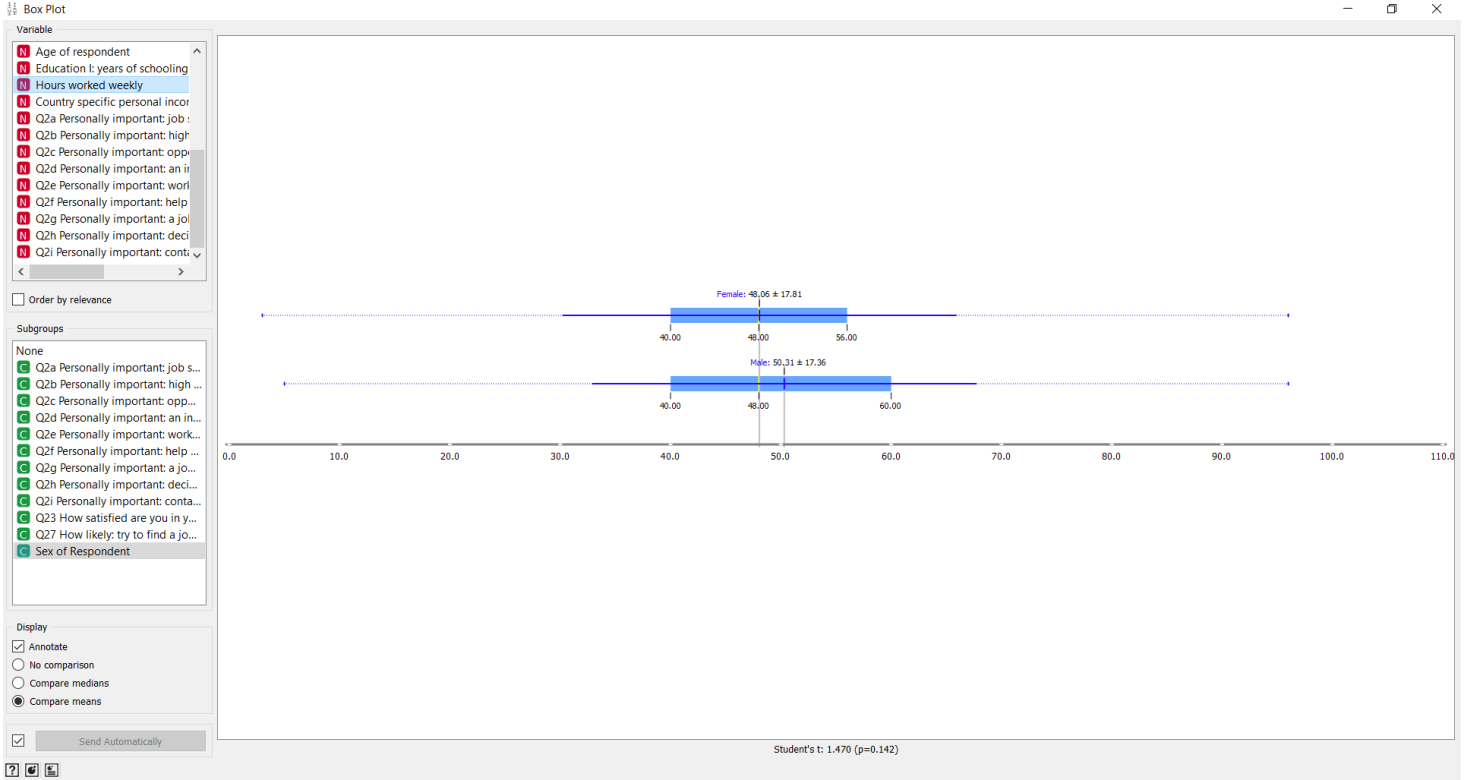
Independent variable(s)	Dependent variable(s)	
	Nominal Ordinal	Interval
Nominal Ordinal	Contingency tables Correspondence Logistic regression Rank correlation Classification trees Neural Networks	ANOVA GLM Rank correlation Classification trees Neural Networks
Interval	Discriminant Logistic regression Rank correlation Classification trees Neural Networks	GLM Regression Correlation Classification trees Neural Networks
Mix of nominal and/or ordinal and/or interval	Logistic regression Rank correlation Classification trees Neural Networks	GLM Rank correlation Classification trees Neural Networks

Relations Analysis, Classification / Ex: Contingency Tables



Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

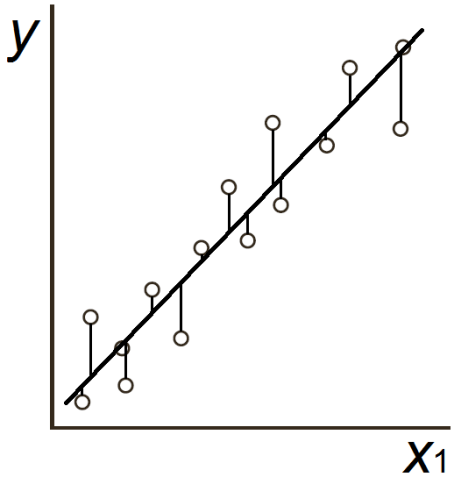
Relations Analysis, Classification / Ex: ANOVA



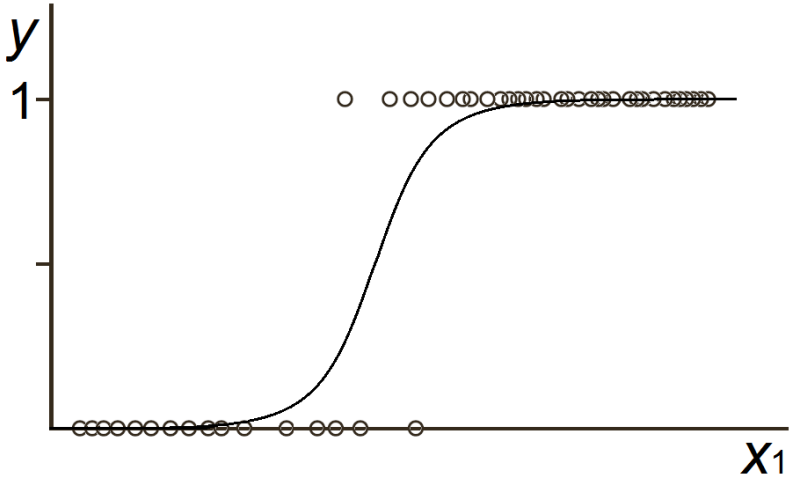
Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Regression / Linear & Logistic Models

- Linear Model

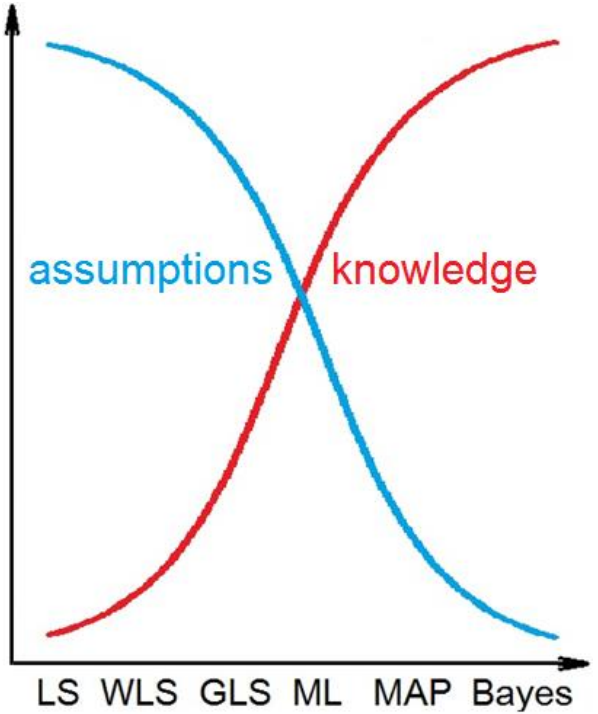


- Logistic Model

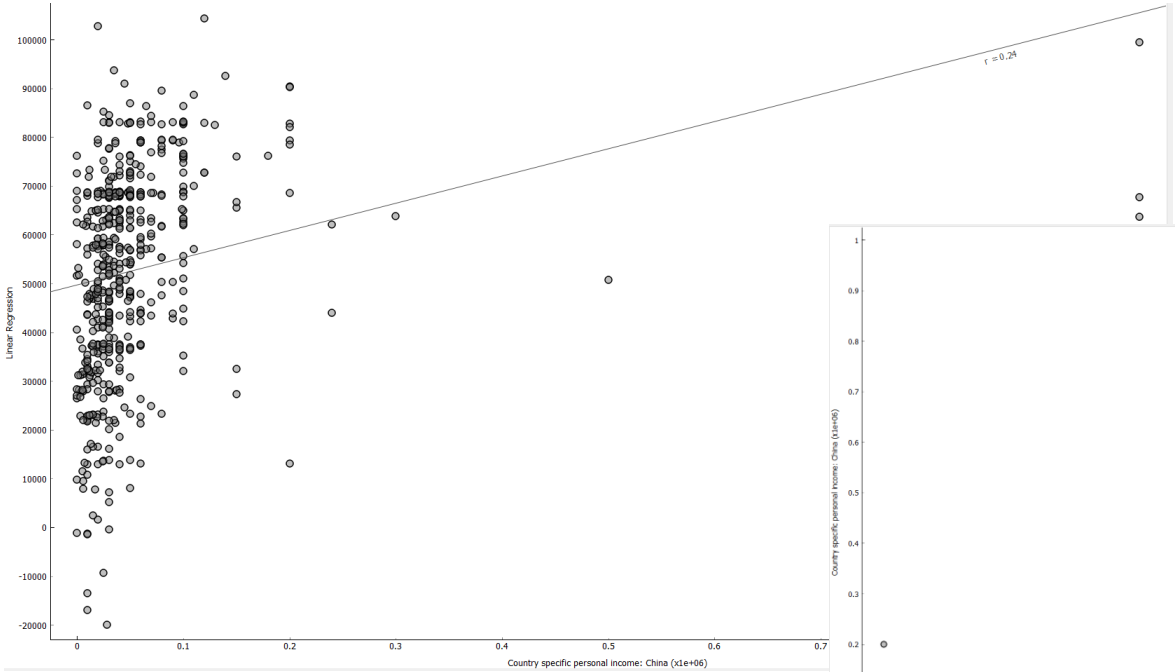


Regression / Estimators

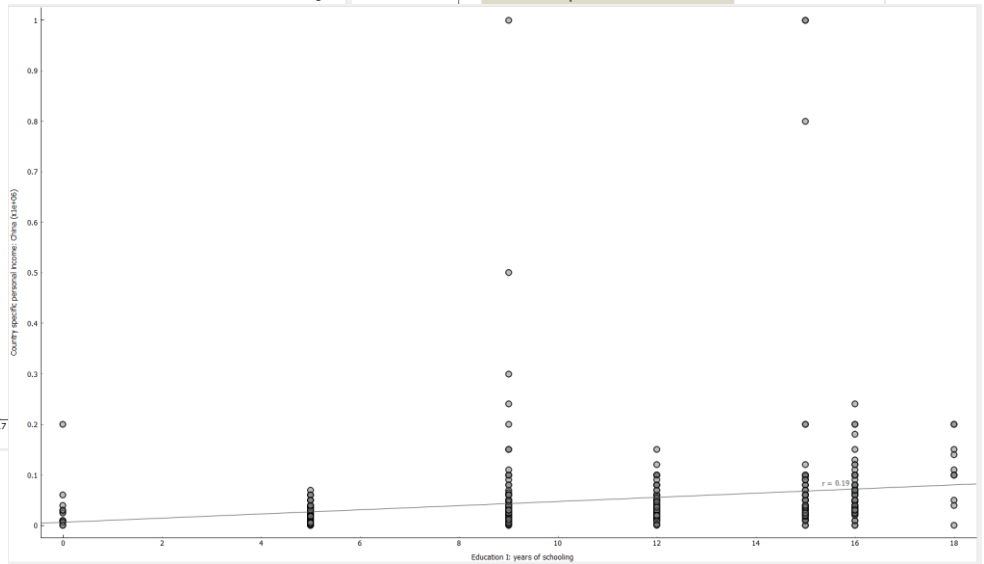
– System Knowledge & Assumptions
  - Bayes Method
    known: $p(\theta|y)$, $p(\theta)$, $los(\theta, \theta')$
  - Maximum A-Posteriori (MAP)
    known: $p(\theta|y)$, $p(\theta)$
    assumed: $los(\theta, \theta') = const$
  - Maximum Likelihood (ML)
    known: $p(\theta|y)$
    assumed: $p(\theta) = const$
  - Generalized Least Squares (GLS...)
    known: $\exists\ cov(e_{k1}, e_{k2})$
    assumed: $e \sim N(0, \Sigma_e)$
  - Weighted Least Squares (WLS)
    known: $\Sigma_e \rightarrow$ diag /W – weight matrix/
    assumed: $cov(e_{k1}, e_{k2}) = 0$
  - Least Squares (LS)
    assumed: $w_{kk} = const$



Relations Analysis, Classification / Ex: Regression & Correlation



	name	coef
1	intercept	44395.2142329
2	Age of respondent	-22.8504697
3	Education I: years of schooling	3675.8242980
4	Hours worked weekly	-663.8539557
5	Sex of Respondent=Female	-7141.3045064
6	Sex of Respondent=Male	7141.3045064



Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Relations Analysis, Classification / Ex: Logistic Regression

	name	Likely	Unlikely	Very likely	Very unlikely
1	intercept	-0.0004764	0.0005328	-0.0167649	-0.0138540
2	Q2a Personally important: job security=Important	-0.0003838	0.0027251	-0.0276124	-0.0299506
3	Q2a Personally important: job security=Neither important nor unimportant	0.0002190	-0.0001818	0.0038636	-0.0009423
4	Q2a Personally important: job security=Not important	0.0000628	-0.0000597	0.0114806	-0.0010184
5	Q2a Personally important: job security=Not important at all	-0.0000348	0.0000518	-0.0018987	-0.0004994
6	Q2a Personally important: job security=Very important	-0.0003396	-0.0020026	-0.0025980	0.0185567
7	Q2b Personally important: high income=Important	-0.0004100	0.0013847	-0.0750095	-0.0133372
8	Q2b Personally important: high income=Neither important nor unimportant	-0.0001370	0.0001077	0.0026382	-0.0002796
9	Q2b Personally important: high income=Not important	0.0001545	-0.0001883	0.0120869	-0.0003987
10	Q2b Personally important: high income=Not important at all	-0.0000348	0.0000518	-0.0018987	-0.0004994
11	Q2b Personally important: high income=Very important	-0.0000492	-0.0008231	0.0454182	0.0006609
12	Q2c Personally important: opportunities for advancement=Important	-0.0005200	0.0019866	-0.0562992	-0.0205469
13	Q2c Personally important: opportunities for advancement=Neither important nor unimportant	0.0006471	-0.0007970	0.0083093	0.0005045
14	Q2c Personally important: opportunities for advancement=Not important	0.0001803	-0.0002337	0.0056942	0.0000218
15	Q2c Personally important: opportunities for advancement=Not important at all	-0.0000507	-0.0000920	-0.0005985	0.0015844
16	Q2c Personally important: opportunities for advancement=Very important	-0.0007331	-0.0003310	0.0261293	0.0045821
17	Q2d Personally important: an interesting job=Important	-0.0003694	0.0016178	-0.0758658	-0.0166164
18	Q2d Personally important: an interesting job=Neither important nor unimportant	0.0002248	0.0000004	-0.0008385	-0.0046522
19	Q2d Personally important: an interesting job=Not important	0.0001164	-0.0002614	0.0037091	0.0013614
20	Q2d Personally important: an interesting job=Very important	-0.0004483	-0.0008240	0.0562303	0.0060531
21	Q2e Personally important: work independently=Important	-0.0006792	0.0012810	-0.0256730	-0.0143940
22	Q2e Personally important: work independently=Neither important nor unimportant	0.0005719	0.0003077	-0.0372891	-0.0090605
23	Q2e Personally important: work independently=Not important	0.0001687	-0.0004246	0.0025501	0.0025972
24	Q2e Personally important: work independently=Not important at all	-0.0000181	-0.0000379	-0.0000396	0.0005682
25	Q2e Personally important: work independently=Very important	-0.0005198	-0.0005934	0.0436866	0.0064351
26	Q2f Personally important: help other people=Important	-0.0002560	0.0015548	-0.0450567	-0.0181868
27	Q2f Personally important: help other people=Neither important nor unimportant	0.0009398	-0.0003257	0.0024383	-0.0088917
28	Q2f Personally important: help other people=Not important	-0.0002068	-0.0000276	0.0077005	0.0011156
29	Q2f Personally important: help other people=Very important	-0.0009534	-0.0006687	0.0181530	0.0121089
30	Q2g Personally important: a job useful to society=Important	-0.0005021	0.0020265	-0.0535275	-0.0224387
31	Q2g Personally important: a job useful to society=Neither important nor unimportant	0.0008483	0.0001600	0.0091207	-0.0128298
32	Q2g Personally important: a job useful to society=Not important	0.0004822	-0.0002377	0.0104211	-0.0023940
33	Q2g Personally important: a job useful to society=Not important at all	-0.0000607	-0.0001041	-0.0017362	0.0016310
34	Q2g Personally important: a job useful to society=Very important	-0.0012442	-0.0013119	0.0189569	0.0221775
35	Q2h Personally important: decide time of work=Important	0.0006245	0.0009762	-0.0644577	-0.0179720
36	Q2h Personally important: decide time of work=Neither important nor unimportant	-0.0011684	0.0008773	-0.0159336	-0.0031411
37	Q2h Personally important: decide time of work=Not important	-0.0005221	0.0001577	-0.0023412	0.0026153
38	Q2h Personally important: decide time of work=Not important at all	0.0000666	-0.0001818	-0.0028604	0.0015377

	Logistic Regression	r to find a job with
1	0.27 : 0.47 : 0.01 : 0.25 → Unlikely	Very unlikely
2	0.22 : 0.50 : 0.02 : 0.26 → Unlikely	Likely
3	0.14 : 0.50 : 0.02 : 0.33 → Unlikely	Very unlikely
4	0.23 : 0.45 : 0.02 : 0.29 → Unlikely	Unlikely
5	0.11 : 0.50 : 0.02 : 0.37 → Unlikely	Very unlikely
6	0.15 : 0.42 : 0.04 : 0.39 → Unlikely	Unlikely
7	0.19 : 0.47 : 0.01 : 0.33 → Unlikely	Very unlikely
8	0.19 : 0.43 : 0.00 : 0.38 → Unlikely	Very unlikely
9	0.32 : 0.44 : 0.09 : 0.15 → Unlikely	Unlikely
10	0.22 : 0.48 : 0.01 : 0.28 → Unlikely	Likely
11	0.19 : 0.52 : 0.01 : 0.28 → Unlikely	Very unlikely
12	0.24 : 0.44 : 0.04 : 0.29 → Unlikely	Unlikely
13	0.29 : 0.49 : 0.05 : 0.16 → Unlikely	Unlikely
14	0.23 : 0.49 : 0.03 : 0.25 → Unlikely	Unlikely
15	0.29 : 0.44 : 0.02 : 0.25 → Unlikely	Likely
16	0.09 : 0.49 : 0.02 : 0.39 → Unlikely	Very unlikely
17	0.22 : 0.48 : 0.01 : 0.28 → Unlikely	Likely
18	0.17 : 0.46 : 0.01 : 0.35 → Unlikely	Unlikely
19	0.17 : 0.27 : 0.00 : 0.57 → Very unlikely	Very unlikely
20	0.26 : 0.45 : 0.00 : 0.28 → Unlikely	Unlikely
21	0.22 : 0.45 : 0.04 : 0.29 → Unlikely	Unlikely
22	0.13 : 0.52 : 0.01 : 0.35 → Unlikely	Unlikely
23	0.29 : 0.46 : 0.01 : 0.23 → Unlikely	Likely
24	0.31 : 0.44 : 0.04 : 0.20 → Unlikely	Very unlikely
25	0.28 : 0.45 : 0.01 : 0.27 → Unlikely	Unlikely
26	0.15 : 0.50 : 0.01 : 0.34 → Unlikely	Very unlikely
27	0.20 : 0.50 : 0.01 : 0.29 → Unlikely	Likely
28	0.12 : 0.50 : 0.02 : 0.36 → Unlikely	Very unlikely
29	0.24 : 0.48 : 0.02 : 0.26 → Unlikely	Unlikely
30	0.27 : 0.40 : 0.00 : 0.32 → Unlikely	Likely
31	0.35 : 0.46 : 0.03 : 0.16 → Unlikely	Unlikely
32	0.31 : 0.43 : 0.07 : 0.20 → Unlikely	Very likely
33	0.24 : 0.48 : 0.02 : 0.25 → Unlikely	Unlikely
34	0.14 : 0.45 : 0.01 : 0.40 → Unlikely	Very unlikely
35	0.27 : 0.45 : 0.02 : 0.26 → Unlikely	Very unlikely
36	0.13 : 0.53 : 0.00 : 0.33 → Unlikely	Very unlikely
37	0.26 : 0.50 : 0.03 : 0.22 → Unlikely	Very unlikely
38	0.23 : 0.47 : 0.01 : 0.29 → Unlikely	Likely
39	0.12 : 0.50 : 0.00 : 0.38 → Unlikely	Very unlikely
40	0.12 : 0.53 : 0.01 : 0.34 → Unlikely	Very unlikely
41	0.18 : 0.45 : 0.01 : 0.36 → Unlikely	Likely

Actual

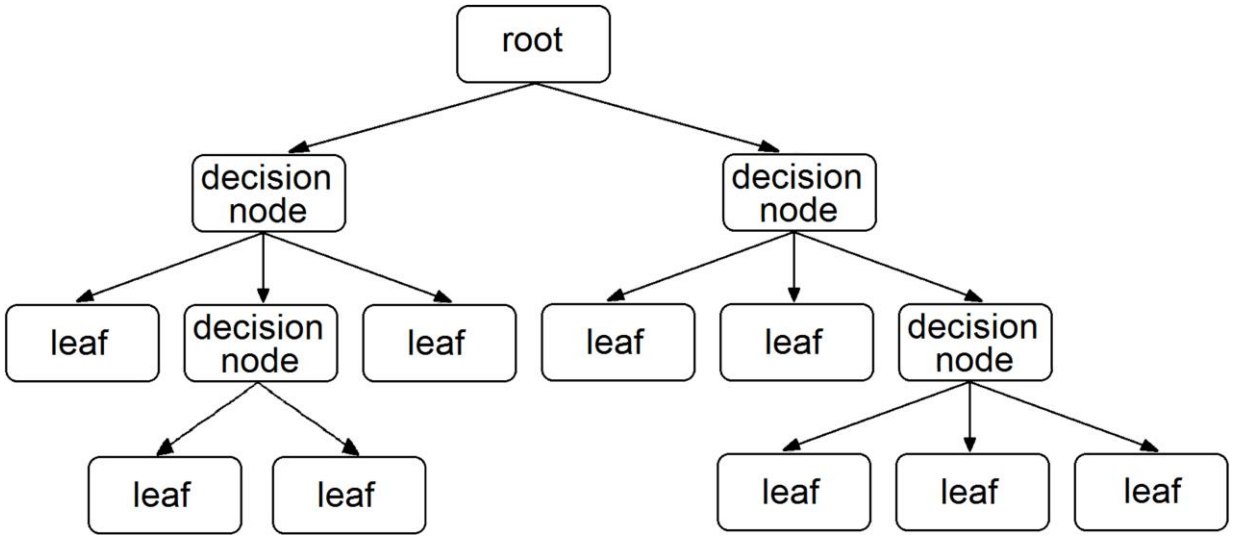
Predicted

	Likely	Unlikely	Very likely	Very unlikely	Σ
Likely	0.0 %	96.6 %	0.0 %	3.4 %	**118**
Unlikely	0.0 %	97.8 %	0.0 %	2.2 %	**275**
Very likely	0.0 %	100.0 %	0.0 %	0.0 %	**15**
Very unlikely	0.0 %	91.2 %	0.0 %	8.8 %	**160**
Σ	**0**	**544**	**0**	**24**	**568**

Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

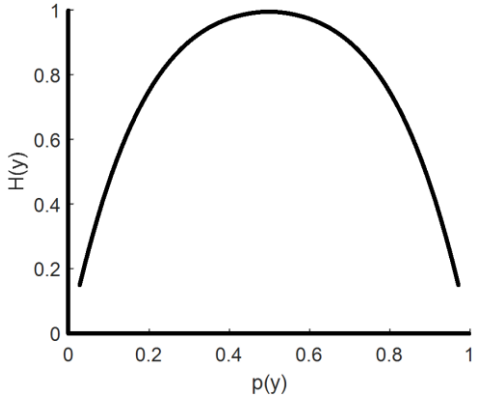
Decision Tree

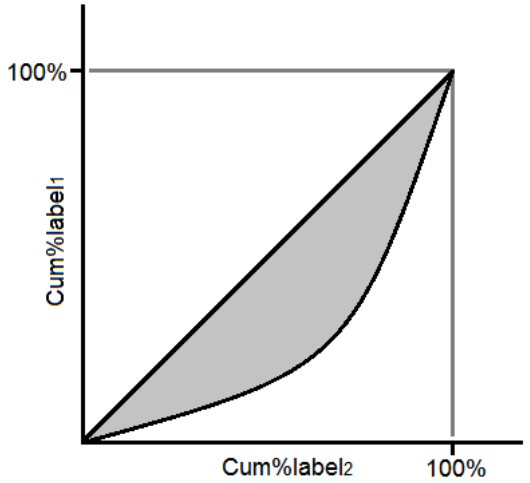
- DT Development
  - Initialization
  - Stage 1: Growing
  - Stage 2: Pruning



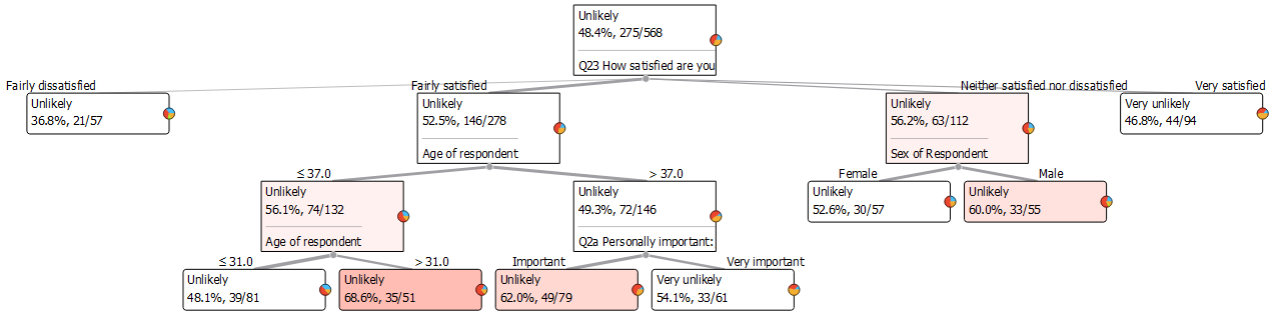
Decision Tree

- DT Criteria
  - Information Gain /H(y) – Entropy/
  - Gini
  - ChiSq
  - F-ratio
  - Twoing





Relations Analysis, Classification / Ex 1: Classification Trees



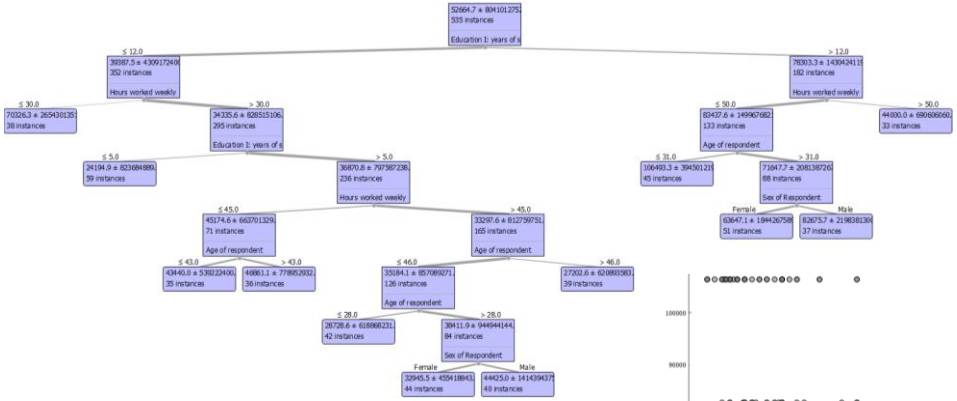
Predicted

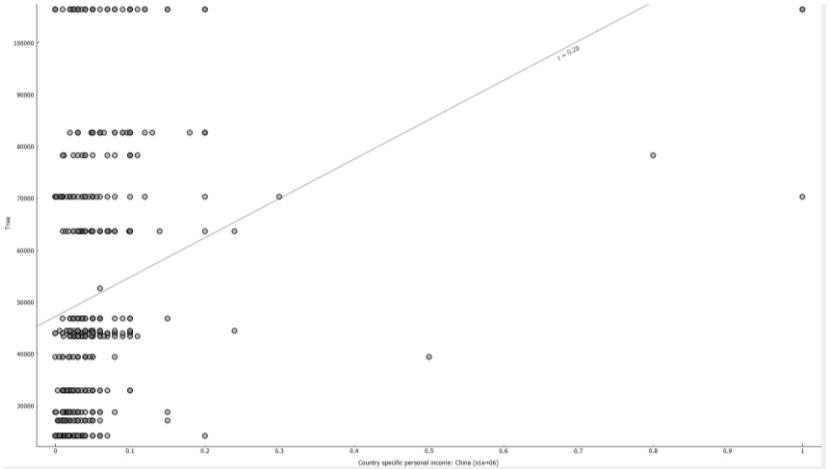
Actual	Likely	Unlikely	Very likely	Very unlikely	Σ
Likely	0.0 %	82.2 %	0.0 %	17.8 %	**118**
Unlikely	0.0 %	79.6 %	0.0 %	20.4 %	**275**
Very likely	0.0 %	93.3 %	0.0 %	6.7 %	**15**
Very unlikely	0.0 %	51.9 %	0.0 %	48.1 %	**160**
Σ	**0**	**413**	**0**	**155**	**568**

	Tree	' to find a job wit
1	0.35 : 0.37 : 0.11 : 0.18 → Unlikely	Very unlikely
2	0.18 : 0.69 : 0.02 : 0.12 → Unlikely	Likely
3	0.14 : 0.38 : 0.01 : 0.47 → Very unlikely	Very unlikely
4	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Unlikely
5	0.35 : 0.37 : 0.11 : 0.18 → Unlikely	Very unlikely
6	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Unlikely
7	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
8	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
9	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Unlikely
10	0.22 : 0.60 : 0.05 : 0.13 → Unlikely	Likely
11	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
12	0.18 : 0.69 : 0.02 : 0.12 → Unlikely	Unlikely
13	0.35 : 0.37 : 0.11 : 0.18 → Unlikely	Unlikely
14	0.18 : 0.69 : 0.02 : 0.12 → Unlikely	Unlikely
15	0.14 : 0.38 : 0.01 : 0.47 → Very unlikely	Likely
16	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
17	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Likely
18	0.10 : 0.62 : 0.00 : 0.28 → Unlikely	Unlikely
19	0.21 : 0.48 : 0.03 : 0.28 → Unlikely	Very unlikely
20	0.26 : 0.53 : 0.00 : 0.21 → Unlikely	Unlikely
21	0.22 : 0.60 : 0.05 : 0.13 → Unlikely	Unlikely
22	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Unlikely
23	0.26 : 0.53 : 0.00 : 0.21 → Unlikely	Likely
24	0.21 : 0.48 : 0.03 : 0.28 → Unlikely	Very unlikely
25	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Unlikely
26	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
27	0.14 : 0.38 : 0.01 : 0.47 → Very unlikely	Likely
28	0.26 : 0.53 : 0.00 : 0.21 → Unlikely	Very unlikely
29	0.22 : 0.60 : 0.05 : 0.13 → Unlikely	Unlikely
30	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Likely
31	0.22 : 0.60 : 0.05 : 0.13 → Unlikely	Unlikely
32	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Very likely
33	0.18 : 0.69 : 0.02 : 0.12 → Unlikely	Unlikely
34	0.10 : 0.62 : 0.00 : 0.28 → Unlikely	Very unlikely
35	0.21 : 0.48 : 0.03 : 0.28 → Unlikely	Very unlikely
36	0.26 : 0.53 : 0.00 : 0.21 → Unlikely	Very unlikely
37	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Very unlikely
38	0.35 : 0.37 : 0.11 : 0.18 → Unlikely	Likely
39	0.10 : 0.62 : 0.00 : 0.28 → Unlikely	Very unlikely
40	0.13 : 0.33 : 0.00 : 0.54 → Very unlikely	Very unlikely
41	0.38 : 0.48 : 0.01 : 0.12 → Unlikely	Likely

Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Relations Analysis, Classification / Ex 2: Classification Trees





	Tree	ecific personal inc
1	106493.333	100000.000
2	63647.059	?
3	63647.059	50000.000
4	32945.455	60000.000
5	24194.915	16800.000
6	24194.915	36000.000
7	63647.059	100000.000
8	82675.676	180000.000
9	44000.000	0.000
10	82675.676	100000.000
11	63647.059	80000.000
12	44425.000	50000.000
13	106493.333	0.000
14	82675.676	50000.000
15	106493.333	100000.000
16	24194.915	0.000
17	44000.000	100000.000
18	32945.455	100000.000
19	70326.316	300000.000
20	106493.333	150000.000
21	39387.500	40000.000
22	82675.676	80000.000
23	78303.297	100000.000
24	28728.571	30000.000
25	106493.333	120000.000
26	82675.676	100000.000
27	63647.059	100000.000
28	46861.111	24000.000
29	106493.333	70000.000
30	106493.333	200000.000
31	44425.000	48000.000
32	28728.571	20000.000
33	82675.676	60000.000
34	24194.915	?
35	78303.297	110000.000
36	63647.059	100000.000
37	106493.333	30000.000
38	63647.059	100000.000
39	82675.676	130000.000
40	82675.676	60000.000
41	106493.333	110000.000

Data: International social survey program (ISSP), 2015, China (http://zacat.gesis.org/webview/)
Software: Orange (https://orange.biolab.si/)

Research Tasks & Data Structure

THANKS

Contacts:

Kaloyan Haralampiev, Ph.D.

k_haralampiev@phls.uni-sofia.bg

http://kaloyan-haralampiev.info/

Assoc. Prof. of Sofia University “St. Kliment Ohridski”

https://www.uni-sofia.bg/

Alexander Efremov, Ph.D., Eng.

aefremov@gmail.com

https://bg.linkedin.com/in/aefremov

Assoc. Prof. at Technical University of Sofia

http://anp.tu-sofia.bg/aefremov/index.htm

Co-founder & Chief Scientist at A4E

https://www.a4everyone.com